De novo assembly of transcriptome and genome data from the Asia I *mtCOI* genetic clade of *Bemisia tabaci*

Mitulkumar V. Patel

A thesis submitted in partial fulfilment of the requirements of the University of Greenwich for the Degree of Doctor of Philosophy

Natural Resources Institute
UNIVERSITY OF GREENWICH

July 2016

DECLARATION

I certify that this work has not been accepted in substance for any degree, and is not concurrently being submitted for any degree other than that of Doctor of Philosophy being studied at the University of Greenwich. I also declare that this work is the result of my own investigations except where otherwise identified by references and that I have not plagiarised the work of others.

First supervisor

Signature : .....
Name : Professor Susan Seal
Date : 31 July 2016

Second supervisor

Signature : .....
Name : Dr. David Bailey
Date : 31 July 2016

Student

Signature : .....
Name : Mitulkumar Patel
Date : 31 July 2016

ACKNOWLEDGEMENTS

Pursuing a PhD is both an exciting and enjoyable experience. Through this challenging journey, it was my good fortune that I could work together with, and learn from so many outstanding scientists from the computational and molecular biology communities. The studies presented were carried out at the Natural Resources Institute, University of Greenwich. Funding for this work was provided from various projects within the University of Greenwich's research funding. Now standing at the terminal, I realize that I could never have succeeded without the help from many people. I would like to express my sincere gratitude to all these people.

First and foremost I would like to express my deepest gratitude to my mentors Professor Susan Seal and Dr. David Bailey for their excellent mentorship, encouragement, support, vision and believing in me. Special thanks to Professor Susan Seal for her enthusiasm and motivation that inspired me for research and guided me through all phases of my PhD. She always encouraged me to think as an independent scientist which is consistent with my long-term career aims. It was a great experience sharing office space with her and I could not have imagined a better mentor. Also, I would like to acknowledge students of the PhD student club at NRI who shared science, sports, lunches and dinners and coffee breaks with me.

I would like to acknowledge Roger Hoskins, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, and Madeline Crosby, Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, curators of the FlyBase database.

I am indebted to my parents for their unconditional love and unfailing support, and my companion in life Bena for accompanying me, her love, encouragement, and understanding.

ABSTRACT

Bemisia tabaci, the whitefly, is an economically important group of cryptic species that pose an ongoing and major threat to global food security. Molecular characterization from at least one of the species in the group is required urgently within whitefly community to define genetic differences across distinct populations and to facilitate the development of more effective insecticides to help farmers. In this thesis, the process of *de novo* assembly and characterizing transcriptome and genome data of an Asia I population is described along with steps towards the development of a genetic framework for insecticide discovery. A comprehensive transcriptome of adult females of an Asia I population was assembled from 0.864 million reads generated using the Roche 454 sequencing platform. A total of 29,418 contigs produced using CLC Genomics of which 8,563 were assigned putative functions. A draft genome 828 Mbp of the Asia I population was produced through *de novo* assembly of 206 million 250 bp paired-end reads generated on the Illumina MiSeq 2500 platform using Platanus. They were found to encode 41,981 PCGs and also contained 990 ncRNAs and repetitive elements (45.66%). A selection of 741 full-length genes were studied and found to contain larger intron sizes together with repetitive elements that contribute to the larger Asia I genome size and partly explain the difficulties encountered in genome assembly. Finally, in addition to the *B. tabaci* Asia I genome, one mitogenome and three endosymbiont genomes representing *Portiera*, *Wolbachia* and *Arsenophonus* were obtained from the same sequence library. The development of a genetic framework is described as a side-project which uses *Drosophila* essential genes as a reference to identify orthologs in other insects and validates with ChEMBL targets. The 'omics' data presented in this thesis provides a comprehensive sequence resource for Asia I populations and demonstrates the workflow of obtaining genetic information of host and its endosymbionts.

PUBLICATIONS

Articles

Articles in peer-reviewed journal

Collins C., **Patel MV**., Colvin J., Bailey D., and Seal S. (2013).
"Identification and evaluation of suitable reference genes for gene expression in the whitefly, *Bemisia tabaci*, by reverse transcription quantitative real-time RT-PCR." *Journal of Insect Science*, 14:63.

Seal S., **Patel MV**., Collins C., Colvin J., and Bailey D. (2012).
"Next Generation Transcriptome Sequencing and Quantitative Real-Time PCR Technologies for Characterisation of the *Bemisia tabaci* Asia I mtCOI Phylogenetic Clade." *Journal of Integrative Agriculture*, 11(2):281-292.

Articles in preparation

Patel MV., Colvin J., Bailey D., and Seal S. "Distinctive gene structure of the whitefly, *Bemisia tabaci*, identified using high-throughput genome and transcriptome sequencing."

Patel MV., Colvin J., Bailey D., and Seal S. "Draft genome sequence of the *Wolbachia*, an endosymbiont of *Bemisia tabaci* (Hemiptera: Aleyrodidae) species complex Asia I."

Conference abstracts

June, 2015 Ninth annual Arthropod Genomics Symposium, Kansas, USA
"*De novo* characterization of transcriptome and genome assemblies from the Asia I mtCOI genetic clade of *Bemisia tabaci*".
Patel MV., Colvin J., Bailey D., and Seal S.

May, 2013 First International Whitefly Symposium, Crete, Greece
"Characterization of next generation transcriptomic and genomic sequence data from the Asia I mtCOI genetic clade of *Bemisia tabaci*."
Patel MV., Cain S., Febrer M., Caccamo M., Colvin J., Bailey D., and Seal S.

June, 2011 Fifth annual Arthropod Genomics Symposium, Kansas City, United States of America
"Using insect pest genomics to identify new molecular targets for insecticide discovery and plant genetic engineering."
Bailey D., Colvin J., **Patel MV**., Seal S., Kersey P., Lawson D., Overington J., Caccamo M., Febrer M., Hogenhout S., Bass C., Denholm I., Gorman K., and Williamson M.

ABBREVIATIONS

AChE	Acetylcholinesterase
ANK	Ankyrin
ATP6	ATP synthase subunit 6
ATP8	ATP synthase subunit 8
BGMV	*Bean golden mosaic virus*
BGYMV	*Bean golden yellow mosaic virus*
BLAST	Basic Local Alignment Search Tool
bp	Base-pair
BUSCO	Benchmarking Universal Single-Copy Orthologs
BWA	Burrows-Wheeler Aligner
CDD	Conserved Domain Database
CDS	Coding sequence
CEG	Conserved eukaryotic genes
CLCV	*Cotton leaf curl virus*
CM	Covariance model
CMB	Cassava mosaic begomovirus
CMD	Cassava mosaic disease
COE	Carboxylesterases
COG	Cluster of orthologous group
DBG	de Bruijn Graph
DEG	*Drosophila* Essential Gene
DHU	Dihydrouridine
DNA	Deoxyribonucleic acid
E-value	Expect value
EBI	European Bioinformatics Institute
EC	Enzyme Commission
EcR	Ecdysone receptor
emPCR	Emulsion-based clonal amplification
ERR	Estrogen-related receptor
EST	Expressed sequence tag
GB	Giga byte
Gbp	Giga base-pairs
GMAP	Genomic Mapping and Alignment Program
GO	Gene ontology

GST	Glutathione-S-transferase
Gwl	Greatwall
HGSC	Honeybee Genome Sequencing Consortium
HMM	Hidden Markov Model
HSP	High-scoring segment pairs
IAGC	International Aphid Genomics Consortium
IGR	Insect growth regulator
IME	Intron-mediated enhancement
IPM	Integrated pest management
IUCN	International Union for the Conservation of Nature
Ka	Nonsynonymous
KAAS	KEGG Automatic Annotation Server
kbp	Kilo base-pairs
KEGG	Kyoto Encyclopedia of Genes and Genomes
Ks	Synonymous
LBD	Ligand-binding domain
LINEs	Long interspersed elements
LIYV	*Lettuce infectious yellows virus*
LTR	Long terminal repeat
Mbp	Mega base-pairs
MCL	Markov Cluster Algorithm
miRNA	MicroRNA
MISA	MIcroSAtellite
ML	Maximum Likelihood
MP	Mate-pair
mRNA	messenger ribonucleic acid
mtCOI	mitochondrial cytochrome oxidase I
MTD	Multi-target design
MYMV	*Mung bean yellow mosaic virus*
NCBI	National Center for Biotechnology Information
ncRNA	Non-coding RNA
ND4	NADH dehydrogenase subunit 4
ND5	NADH dehydrogenase subunit 5
NGS	Next generation sequencing
NL	Normalized
nr	non redundant
NRI	Natural Resources Institute
OLC	Overlap Layout Consensus
ORF	Open reading frame

P450	Cytochrome P450 monooxygenases
PacBio	Pacific Biosciences
PCG	Protein coding genes
PDB	Protein Data Bank
Perl	Practical extraction and report language
Pfam	Protein family database
PGAP	Prokaryotic Genome Annotation Pipeline
PIR	Protein Information Resource
QUAST	Quality assessment tool
RAST	Rapid Annotation using Subsystem Technology
RBD	RNA-binding domain
RNA	Ribonucleic acid
RNP	Ribonucleoprotein
RPKM	Reads Per Kilobase per Million mapped reads
rRNA	Ribosomal RNA
SCO	Single Copy Orthologs
SINEs	Short interspersed nuclear elements
SMRT	Single Molecule Real Time
SNP	Single nucleotide polymorphism
snRNA	Small nuclear RNA
snRNPs	Small nuclear ribonucleoproteins
SPCSV	*Sweetpotato chlorotic stunt virus*
SRA	Short Read Archive
SSRs	Simple sequence repeats
TE	Transposable element
TGSC	Tribolium Genome Sequencing Consortium
TMV	*Tomato mottle virus*
TRF	Tandem Repeat Finder
tRNA	Transfer RNA
TSA	Transcriptome Shotgun Assembly
TYLCV	*Tomato yellow leaf curl virus*
UL	Unnormalized
UTR	Untranslated region
WGS	Whole genome sequence

CONTENTS

FIGURES

TABLES

CHAPTER 1

Introduction

Insects are the largest and the most divergent group of species within the invertebrates, encompassing 29 orders and over 800,000 different insect species (Zdobnov and Bork, 2007; Vea and Grimaldi, 2016). Insects represent more than half of the total living organisms known (Chapman, 2009; Wilson, 2009). The order Hemiptera is the most diversified group of non-holometabolous insects (Kristensen, 1991; New, 2011) and is the fifth largest order of insects after Coleoptera, Diptera, Hymenoptera and Lepidoptera (Schuh and Slater, 1995; Grimaldi and EngeL, 2005; Cameron et al., 2006; New, 2011) within the class Insecta. Hemipteran insects such as whiteflies, aphids and scale insects have piercing or sucking mouthparts, and are the largest group of plant-feeding insects (Dolling, 1991; Chougule and Bonning, 2012). Whiteflies belonging to the species complex *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidae) have emerged as destructive pests of agricultural, ornamental and horticultural crops worldwide (Boykin et al., 2007; Wan et al., 2009; Barro et al., 2011; Li et al., 2011), infesting > 600 plant species in tropical and temperate regions (Martin et al., 2000; Cuthbertson et al., 2007; Naranjo and Ellsworth, 2009). *B. tabaci* whiteflies can cause extensive damage to a wide diversity of plant hosts ranging from major vegetables and salad crops to tropical and sub-tropical crops. Damage can be direct through feeding on plant phloem sap or indirect via acting as vectors for a large range of plant pathogenic viruses (Czosnek et al., 2002; Dalton, 2006; Jiu et al., 2006; Seal et al., 2006; Malka et al., 2016).

The complex is responsible for yield losses of billions of dollars annually (Dalton, 2006; Seal et al., 2006; Crowder et al., 2009; Naranjo et al., 2010). *B. tabaci* has recently been shown to represent a "cryptic species" complex including at least 37 morphologically indistinguishable, but genetically diverse populations (Dinsdale et al., 2010; Barro et al., 2011; Hu et al., 2011a; Alemandri et al., 2012; Liu et al., 2012; Parrella et al., 2012; Firdaus et al., 2013; Hu et al., 2014). To date, the method of choice for classification of members of this species complex has been pairwise genetic divergence (3.5%) within a fragment of the mitochondrial cytochrome c oxidase I (*mtCOI*) gene (Dinsdale et al., 2010; Hu et al., 2011a; Trautwein et al., 2012). Members of the *B. tabaci* species complex are on the list of "100 of the World's Worst Invasive Alien

Species" (http://www.issg.org/database/species/search.asp?st=100ss) by the International Union for the Conservation of Nature and Natural Resources (IUCN) (http://www.issg.org).

1.1 Why sequence the whitefly genome?

The relatively recent explosion of available whole genome sequences and expressed sequence tags (ESTs), and their use in insect transgenesis, has given a new direction to researchers within the field of entomology (IAGC, 2010; Wang et al., 2011, 2012; Fraser, 2012). A genome sequence reveals the gene structure, predicts the function of genes, and elucidates the genetic variations within the genome in comparison with other genomes (Stark et al., 2007; Behura et al., 2011). This combined with efforts made to quantify total mRNA, proteins and metabolites in any living organism, have led to the development of three new 'omics' fields: the transcriptomics (total mRNA) (Tariq et al., 2011), the proteomics (all proteins) (Wiśniewski et al., 2012) and the metabolomics (all metabolic pathways) (Chan et al., 2010). A genome sequencing project is a dynamic field that combines traditional biology with these bioinformatics areas, and many aspects are not possible without a whole genome sequence for comparison.

To date, there has been no genome sequence released for any *B. tabaci* populations. Efforts have, however, been made to characterize transcriptomes of several *B. tabaci* species (Leshkowitz et al., 2006; Wang et al., 2010a, 2011; Xie et al., 2012; Wang et al., 2012; Seal et al., 2012). Sequencing the genome of *B. tabaci* has become a major area of interest to characterize the total set of genes and their structures which may help to improve current understanding of diversity across *B. tabaci* species and relation between host range and bacterial endosymbionts. Whole genome sequencing of *B. tabaci* benefits whitefly biology by providing sequence information of the essential genes that are required at the different stages of life from egg to adult. The discovery of essential genes in *B. tabaci* could lead to the identification of a potential novel insecticide targets for this economically important pest.

Essential genes are the indispensable sets of genes required by an organism in order to survive under specific conditions. A better understanding of insect biology relies on identification and analysis of these essential genes, required for growth to a fertile adult (Koonin, 2000, 2003; Kobayashi et al., 2003; Zhang and Lin, 2009). The proteins encoded by the essential genes are generally found conserved across species (Zhang and Lin, 2009). Spradling and colleagues approached the identification of essential genes in *D. melanogaster* by knocking out these vital genes through mutations (Spradling et al., 1999).

Currently available chemical insecticides exert their effects on a very limited number of targets that are essential for electrical signaling in the insect nervous system. These include the voltage-gated sodium channel (He et al., 2011), GABA and the glutamate gated chloride channels (Buckingham

et al., 2005), and the nicotinic acetylcholine receptor (nAChr) (Matsuda et al., 2001) and AChE (Pang, 2006). Due to a variety of issues including the threat of emerging resistance to commonly used pesticides, and the undesired effects on beneficial insects, the task of identifying novel insecticidal targets has in the past decade emerged as one of the most significant challenges for the agrochemical industry (Hemingway et al., 2002; Oakeshott et al., 2003; He et al., 2007; Munhenga et al., 2008).

Using the power of informatics, genomes from different *B. tabaci* species can be compared to find the similarities and variations between their genomes. This comparison could help to develop multi-target insecticides for eliminating this agricultural pest and also provides a better understanding of how they may become resistant to insecticides, which is a major problem for pest control globally (Hemingway et al., 2002). Chemical control has been widely used to manage *B. tabaci* populations (Palumbo et al., 2001). The insecticides that have been used in the past to control this pest have included neurotoxic compounds, neonicotinoids and pyrethroids (Karunker et al., 2009; Luo et al., 2010; Yuan et al., 2012). However, *B. tabaci* has rapidly developed resistance to many insecticides upon extensive exposure to chemical insecticides (He et al., 2007; Roditakis et al., 2009; Luo et al., 2010) and due to endosymbiont activities (Mahadav et al., 2008). Insecticide resistance in Asian populations of the whiteflies has been widely studied where it is a major problem (Luo et al., 2010; Wang et al., 2010b; Xie et al., 2012; Yuan et al., 2012). Metabolic resistance to chemicals is the most common mechanism enhanced by overproduction of detoxification enzymes in herbivorous insects (Despres et al., 2007). In insects, detoxification mechanisms to plant toxins and chemical insecticides are mediated by three multigene families: glutathione S-transferases (GSTs), cytochrome P450 monooxygenases (cytochrome P450) and carboxylesterases (COEs) (Johnson, 1999; Claudianos et al., 2006; Schuler, 2011). *B. tabaci* species have developed defence strategies via changing structures and functions of such detoxification genes: the Middle East-Asia Minor 1 (MEAM1, previously referred to as the 'B biotype') and Mediterranean (MED, previously referred to as the 'Q biotype') species have inducible and constitutive defences respectively (Elbaz et al., 2012).

In previous years many insecticides have been discovered, and nowadays their mechanisms of action can be identified with the help of the target pest's genome sequence. Some studies have investigated the transcriptional response of whiteflies to different plant hosts using high-throughput RNA sequencing (Wang et al., 2010a, 2011, 2012).

Despite the availability of various transcriptome sequence information for the *B. tabaci* species complex, whole genome sequencing has become a major challenge mainly because of two reasons: genome complexity (size of the genome, size of the non-coding regions, high level of repetitive elements and heterozygosity) and innate complex biology (symbiosis) (Chu et al., 2013). The major aim of this study was to develop whitefly genome resources using next-generation sequencing technologies to assist in resolving the phylogeny of the whitefly species complex and to explore new strategies for controlling its most damaging members.

1.2 A genome is preferred over a transcriptome

During the last decade, sequencing and annotation of expressed sequence tags (ESTs) was the most cost-effective approach to understand the protein encoding capabilities of the whitefly species (Leshkowitz et al., 2006; Wang et al., 2010a, 2011, 2012; Xie et al., 2012; Seal et al., 2012; Upadhyay et al., 2015). However, recent innovations in Next Generation Sequencing (NGS) technologies have made sequencing faster and more cost-effective. Nowadays, sequencing machines are capable of sequencing the whole genome of any living organism within a few days at high speed with low sequence error rates (Loman et al., 2012).

Characterisation of a genome has become a major area of interest for the identification of genes, their organization and expression within the genome, identification of the total set of proteins and their biological functions, and homology and diversity with other species (IAGC, 2010; Ng and Kirkness, 2010). For any species, when there is an interest in a near-complete gene set, whole genome sequencing is a more robust approach than sequencing merely the expressed genes (transcriptome). While transcriptome sequencing can only provide information on the genes that were expressed at the time of sample collection, one may lose the lowly expressed genes (especially genes that are expressed at different life stages). In contrast, whole genome sequencing aims to yield a complete gene catalogue. Although the gene catalogue from a whole genome sequencing may still be incomplete due to the complexity of genome where certain regions are difficult to sequence and assemble, the gene catalogue will still be more complete from whole genome sequencing than even the deepest of transcriptome sequencing. Most importantly, many EST and transcriptome sequencing projects produce only an estimated set of protein coding genes, which will be insufficient as a large portion of the genome will be represented by non-coding regions (Gerstein et al., 2010).

A genome from at least one of the species in the whitefly complex has been required urgently for the past decade, both to begin to define functional differences within geographically and physiologically distinct populations, as well as to facilitate the development of novel pest-control technologies. In the present study, comprehensive transcriptome and genome sequencing of an Asia I population of the *B. tabaci* species complex were performed along with characterisation of endosymbiont genomes within this population.

1.3 Thesis structure

This thesis is organised as follows. Chapter 2 reviews literature relevant to the research questions addressed regarding the biology, bacterial symbiosis, host range and virus transmission, pest control and 'omics' data from research to date on members of the *B. tabaci* whitefly complex. Chapter 3 describes characterisation of *B. tabaci* Asia I transcriptome using next-generation sequencing and comparison with other whitefly populations where such data was available. After assessing the quality of raw sequencing reads, the assembly of the transcriptome data was performed using three different assemblers to try to achieve a high-quality assembly. The assembly was evaluated using different approaches and then functional annotations were assigned. The annotations were compared with existing transcriptomes from other whitefly clades and discussed for each of the analyses performed.

In Chapter 4, a new strategy is described that was developed to explore the quality of the Asia I genome assembly and at the same time observe structural relationships between ortholog genes across different insect species. The initial stage of the strategy used in Chapter 4 is based on mapping of full-length transcripts to the genome assembly using existing software and getting a glimpse of the gene structure of this whitefly prior to annotation of the genome. The next stage involves comparison of the whitefly gene structures with their corresponding orthologs from nine different insects. A detailed analysis was performed on the gene structure of Asia I species as there was no genomic information available for this pest. This large-scale comparative analysis revealed some interesting features which could provide a new insight in to the biology and evolution of these whitefly and also point towards the complexity of genome.

The very first draft genome of the *B. tabaci* Asia I population, is described in Chapter 5. The chapter outlines most of the analyses associated with any genome sequencing project. One of the main advantages of the strategy used in Chapter 4 is to train the gene model prediction algorithms which were used in the genome annotation pipeline. The effect of this was observed while evaluating the assembly as described in Chapter 4, where results were significantly improved while using gene models that were predicted by trained gene prediction algorithms. In Chapter 5, the genomic reads were filtered prior to assembly to remove any bacterial endosymbionts. The complete proteomes were compared within the hemipteran group of insects to reveal the gene expansion or contractions and results were discussed. This draft genome sequence of Asia I will provide a useful genomic resource to the community to better understand the genetic evolution of this insect and also provide the reference for other clades. The resource will also benefit via providing essential gene information for insecticide target discovery and play an important role in controlling this devastating pest.

In addition to the host genome, four additional genomes were obtained from the genome sequencing reads of Asia I population. Chapter 6 describes the methods and brief annotations of these four

genomes which were the mitochondrial genome and three bacterial endosymbiont genomes, namely *Portiera*, *Wolbachia* and *Arsenophonus*. The complete mitochondrial genome of *B. tabaci* was obtained and compared with the mitogenomes of other *B. tabaci* species. The genome sequence comparison between three *B. tabaci* species from the complex and one whitefly as an out group revealed similarity and diversity across them. As the *mtCOI* has been widely used to classify these *B. tabaci* species, a complete mitogenome will provide more genes to study the phylogeny across *B. tabaci* species complex. The genome assembly and annotation of *Portiera*, the primary endosymbiont of Asia I populations, is also described in Chapter 6. In addition, this *Portiera* genome was compared with the equivalent *Portiera* genomes from two *B. tabaci* species (MEAM1 and MED). Later in Chapter 6, the genome assembly and annotation of *Wolbachia*, a secondary endosymbiont of Asia I populations, is described. There is no genome available to date for this bacterial species from any whitefly host and therefore this draft assembly should help to understand the mechanisms involved in this symbiotic interaction. The PCGs of the *Wolbachia* genome from Asia I populations were compared to other *Wolbachia* PCGs from different insect hosts to study the genes gained or lost, and core genes shared across all. The genome of another secondary endosymbiont, *Arsenophonus*, was also obtained and annotated using the same approach as used for *Wolbachia*.

Chapter 7 describes the workflow to identify potential target genes to control pests and vectors via chemical or non-chemical approaches. The attempt was made at the beginning of this research study and established a genomic framework that utilizes the power of bioinformatics. The first phase of workflow shows how to obtain a set of corresponding orthologs from different species of interest (in this study, the aphid *Acyrthosiphon pisum* (Hemiptera: Aphididae), the malarial vector *Anopheles gambiae* (Diptera: Culicidae) and the honeybee *Apis mellifera* (Hymenoptera: Apidae)) using a highly curated set of essential genes from *Drosophila melanogaster* (Diptera: Drosophilidae). The workflow incorporate the beneficial pollinator *A. mellifera* to allow the best potential target genes for pests and vector species to be selected. The second phase was to confirm the biological importance of the gene orthologs using a protein family database (Pfam) and their structures using the Protein Data Bank (PDB) and the Conserved Domains Database (CDD). Further, it also validated the chemical tractability of the potential targets using the ChEMBL database. And the third phase focused on a comparison of active sites in the target genes across the different insect species included in the workflow to select the potential target gene for chemical insecticide development or non-chemical approaches like RNAi and CRISPR/Cas9. The complete genomic framework would be suitable to any species with genomic information available to identify potential target genes for effective control strategies.

Chapter 8 summarizes the findings of each chapter and highlights future directions for the methods and results presented in this thesis.

CHAPTER 2

Literature review

2.1 Whitefly - *B. tabaci*

B. tabaci was first described in 1889 by Panayiotis Gennadius, an Inspector of Agriculture, on tobacco crops in Greece and named the tobacco whitefly, *Aleyrodes tabaci* (Gennadius, 1889). Another whitefly was collected in 1897 on sweet potato in southeastern USA and described as sweet potato whitefly, *Aleyrodes inconspicua* (Quaintance, 1900). Later in 1914, this species were transferred to the genus *Bemisia* giving a new name to the type species of the genus, *Bemisia inconspicua* (Quaintance and Baker, 1914). By 1964, an additional 19 whitefly species were defined across 14 different countries on a variety of plant hosts (Perring, 2001). Synonmization of these species of whitefly into *B. tabaci* followed by the placement of *tabaci* into the genus *Bemisia* by Takahashi (1936).

The whitefly, *B. tabaci* (Figure 2.1), is genetically diverse and is distributed globally, being present across Asia, the Middle East, the Indian subcontinent, Africa, Australia, southern Europe, the Pacific and the Americas (Boykin et al., 2007; Dinsdale et al., 2010; Barro et al., 2011). Amongst over 1,200 known whitefly species worldwide (Mound and Halsey, 1978), *B. tabaci* represents a cryptic species complex, containing economically important pest and virus vectors for many agricultural, horticultural and ornamental crops worldwide (Boykin et al., 2007; Xie et al., 2012). Members of the whitefly species complex cause considerable crop damage globally (Pimentel et al., 2005) and are considered to be one of "world's top 100 worst invasive alien species" [http://www.issg.org/database/species/search.asp?st=100ss] by the International Union for the Conservation of Nature and Natural Resources (IUCN) list: http://www.issg.org.

FIGURE 2.1: Whitefly, *B. tabaci*. (http://pakagrifarming.blogspot.co.uk/2013/06/)

2.1.1 The whitefly species complex

B. tabaci consists a complex of cryptic species, originally considered as "biotypes", which are morphologically similar but genetically distinct (Brown et al., 1995; Liu et al., 2012; Pan et al., 2012). Although the genetic diversity and economic importance of the *B. tabaci* species complex have been recognized, there has been debate about its species status for decades with regard to whether *B. tabaci* is "a complex species or a species complex" (Brown, 2010; Dinsdale et al., 2010; Barro et al., 2011). A range of DNA-based techniques have been used for the identification of these whitefly species (Barro et al., 2000; Lisha et al., 2003; Brown and Idris, 2005; Rabello et al., 2008; Ahmed et al., 2010; Guo et al., 2012). As these species are morphologically indistinguishable, the study of the sequence diversity of a fragment of the *mtCOI* gene has been widely used to address the global relationships across *B. tabaci* species (Dinsdale et al., 2010; Hu et al., 2011a). The *mtCOI* was used by Frohlich et al. (1999) for the first time to delimit lineages of the *B. tabaci* complex using the 3' end of the *mtCOI* gene, a standard region adopted by subsequent researchers to understand the genetic diversity within the complex. The analysis by Dinsdale et al. (2010) provided a framework based on 3.5% pairwise genetic divergence in *mtCOI* suggesting that *B. tabaci* is a cryptic species complex including at least 24 morphologically indistinguishable species. Later, Hu et al. (2011a) added four more species based on phylogenetic analysis by Dinsdale et al. (2010) and increased the list of species to 28. Firdaus et al. (2013) and Hu et al. (2014) used the same approach and identified three and one more new species respectively which totals 32 putative species (Table 2.1). More species may be added to the list according to field surveys in India (Chowda-Reddy et al., 2012). Previously, *B. tabaci* was divided into more than 30 morphologically indistinguishable 'biotypes' (Banks and Markham, 2000; Perring, 2001; Liu et al., 2012) based on mating compatibility, virus transmission ability (Costa et al., 1993; Bedford et al., 1994; Maruthi et al., 2001, 2004; Shankarappa et al., 2007) and esterase profiles (Brown et al., 1995). It can be seen

from (Table 2.1) that there is no direct correspondence of biotypes with putative species identified by *mtCOI* sequence.

Species group	Species	Biotypes
Africa/Middle East/Asia Minor	Mediterranean	Q, J, L, and Sub-Saharan Africa Silverleafing
	Middle East-Asia Minor 1	B and B2
	Middle East-Asia Minor 2	
	Indian Ocean	MS
Asia I	Asia I	H, M, NA, and PCG-2
Asia II	Asia II 1	K, P, PCG-1, PK1, SY, and ZHJ2
	Asia II 2	
	Asia II 3	ZHJ1
	Asia II 4	
	Asia II 5	G
	Asia II 6	
	Asia II 7	Cv
	Asia II 9	
	Asia II 10	
	Asia II 11	
	Asia II 12	
Asia II India	Asia II 8	
Asia III	Asia III	
Asia IV	Asia IV	
Australia	Australia	AN
Australia/Indonesia	Australia/Indonesia	
China	China 1	ZHJ3
	China 2	
	China 3	
	China 4	
Italy	Italy1	T
	Italy2	
Japan1	Japan1	
Japan2	Japan2	
New World	New World	A, BR, C, D, F, Jatropha, N, R, and Sida
	New World 2	
Sub-Saharan Africa	Sub-Saharan Africa 1	
	Sub-Saharan Africa 2	S
	Sub-Saharan Africa 3	
	Sub-Saharan Africa 4	
	Sub-Saharan Africa 5	
Uganda	Uganda	

TABLE 2.1: The 37 putative species groups of *B. tabaci* species complex identified by Dinsdale et al. (2010), Hu et al. (2011a), Boykin et al. (2012), Firdaus et al. (2013), Hsieh et al. (2014) and Hu et al. (2014).

Phylogeny studies based on *mtCOI* sequence have revealed at least 37 distinct genetic groups within the *B. tabaci* complex based on sufficient evolutionary distance between each group indicating that they seem to be different species (Dinsdale et al., 2010; Barro et al., 2011; Hu et al., 2011a; Liu et al., 2012; Firdaus et al., 2013; Hu et al., 2014). Although the concept of *B. tabaci* as a complex of cryptic species has now been widely accepted, the nomenclature used to describe these species has been highly variable and confusing. Species of the *B. tabaci* complex have been referred to as

populations, haplotypes, biotypes, putative species, genetic groups, clades and species. Therefore, the need for a revised nomenclature of *B. tabaci* species has been approached recently (Boykin, 2014).

2.1.2 Whitefly economic impact

Members of the *B. tabaci* species complex have caused excessive crop damage worldwide and because of their host range across extensive areas, it has been difficult to estimate with confidence their economic impact. From its first identification in Greece in the late 19th century (Gennadius, 1889) until the late 1920s, *B. tabaci* was not considered as a serious pest. It was first described as a serious pest in the early 1930s on cotton in northern India (Hussain and Trehan, 1933). Later, severe infestations were reported on cotton in Sudan and Iran (1950), El Salvador (1961), Mexico (1962), Brazil (1968), Turkey (1974), Israel (1981), Thailand (1978), USA (1981) and Ethiopia (1984) (Basu, 1995). Other major crop damage was recorded due to the attack by the B-biotype in the 1990s (Baoli et al., 2007), and now *B. tabaci* is held responsible for crop losses and economic damage of over US $5 billion annually (Czosnek and Brown, 2010).

2.1.3 Whitefly host range and virus transmission

This hemipteran pest is capable of transmitting hundreds of different plant viruses from seven distinct groups: geminiviruses, luteoviruses, carlaviruses, potyviruses, closteroviruses, nepoviruses and a DNA-containing rod-shaped virus (Duffus, 1987, 1996). The geminiviruses (Family Geminiviridae: Genus *Begomovirus*) represent the most important and abundant group, containing at least 200 species, including *Tomato yellow leaf curl virus* (TYLCV), *Tomato mottle virus* (TMV), and *African cassava mosaic virus* (ACMV) (Fauquet et al., 2008). Viruses belonging to the genus *Begomovirus* are some of the most economically damaging plant viruses reducing crop yields from 20% to 100% (Brown and Bird, 1992). The major symptoms of this virus are leaf curling, yellow veining, yellow mosaics, stunting and vein thickening (Anon, 2001). Begomoviruses contain single-stranded circular DNA genomes unlike the majority of plant viruses, which possess RNA genomes. The only vectors known for begomoviruses are members of the *B. tabaci* species complex, which transmit the viruses in a persistent and circulative manner (Brown, 2007). There is no robust evidence available for the existence of other insect species able to transmit begomoviruses. Cassava mosaic disease (CMD) caused by cassava mosaic begomoviruses (CMBs) transmitted by *B. tabaci* was reported in Africa (Fauquet and Stanley, 2003). It has been reported that CMD causes 20 - 90% yield losses worldwide (Thresh et al., 1998; Dasgupta et al., 2003; Owor et al., 2004). *Tomato yellow leaf curl virus* (TYLCV) affecting tomato crops throughout the world has severely reduced production of tomatoes (Czosnek and Laterrot, 1997). There are several other whitefly-transmitted viruses that

can cause heavy yield losses including the virus species *Tomato mottle virus* (TMV), *Bean golden mosaic virus* (BGMV), *Bean golden yellow mosaic virus* (BGYMV), *Cotton leaf curl virus* (CLCV), *Mung bean yellow mosaic virus* (MYMV), *Lettuce infectious yellows virus* (LIYV) and *Sweetpotato chlorotic stunt virus* (SPCSV) (Oliveira et al., 2001). Begomovirus-induced diseases can be severe, due to poorly selective, and ineffective chemical control of the vector *B. tabaci*, as well as resistance mechanisms developing in whitefly populations to insecticides (He et al., 2007; Ma et al., 2007; Luo et al., 2010; Wang et al., 2010b; Xie et al., 2012; Yuan et al., 2012).

Populations of *B. tabaci* that are highly polyphagous have been recorded as serious pests on a large range of plant hosts and a major reason behind the worldwide distribution of this pest seems to be linked to international trade in plants and their products (Dalton, 2006; Baoli et al., 2007). According to previous studies, this pest can attack vegetables, ornamentals and a range of economically important crop plants: 52 host plants were reported in Israel by Avidov and Harpaz (1969), 172 plants were reported in Egypt by Azab et al. (1971) and Gameel (1972) reported 115 host plant species in Sudan. Mound and Halsey (1978) recorded 420 host plants from 74 different families worldwide and later, more than 500 host plants were reported by Greathead (1986). Amalgamating these reports has led to *B. tabaci* being known to feed on more than 600 host plant species worldwide in greenhouse as well as in the field (Gelman et al., 2005; Li et al., 2011) and crop hosts include sweet potatoes (*Ipomoea batatas*), cotton (*Gossypium hirsutum*), cassava (*Manihot esculenta*), tobacco (*Nicotiana tabacum*), and tomatoes (*Lycopersicon esculentum*). Recently, Li et al. (2011) reported that *B. tabaci* keeps expanding its plant host range, and recorded 361 different plant species belonging to 89 families in China alone, including host plants belonging to the Cucurbitaceae, Compositae, Cruciferae, Leguminosae and Solanaceae.

B. tabaci causes damage to plant hosts directly via feeding large amount of sap from phloem sap and indirectly by honey dew excretion on fruits and leaves. Honey dew can lead to sooty mould development, which interferes with photosynthesis and other processes in the leaf, which in turn leads to plant products of both poorer quality and yield (Byrne and Bellows, 1991).

2.1.4 Whitefly control strategies

For the past few decades, chemical insecticides have been the primary strategy to control *B. tabaci* species as a pest (Castle et al., 1996; Dennehy et al., 1996; Palumbo et al., 2001). In the past, 'imidacloprid', a chloronicotinyl systemic insecticide of the class neonicotinoids was the most effective insecticide to control whiteflies (Elbert et al., 1990; Mullins, 1993; Palumbo et al., 2001). However, there is currently no satisfactory insecticide to control these whitefly as it has become resistant to chemical insecticides in many parts of the world (Cuthbertson et al., 2011; Broughton et al., 2013).

Predictable chemical control of the whitefly, *B. tabaci* is difficult. A number of insecticides have effectively controlled this pest in the past but resistance has developed rapidly. These include neurotoxic compounds (bifenthrin, fenpropathrin, endosulfan, acephate, methamidophos, and amitraz), neonicotinoids, insect growth regulators (IGRs), and pyrethroids (Horowitz and Ishaaya, 1996; Prabhaker et al., 1998; Karunker et al., 2009; Luo et al., 2010; Yuan et al., 2012).

Insecticide resistance in *B. tabaci* is often associated with metabolic detoxification by oxidative and hydrolytic pathways (Byrne et al., 2000; Rauch and Nauen, 2003; Feng et al., 2010), although organophosphate and pyrethroid resistance in *B. tabaci* develops due to the modification in target site of acetylcholinesterase (AChE) and the voltage-gated sodium channel (Morin et al., 2002; Roditakis et al., 2009; Alon et al., 2008). Mutation in the *ace1* gene for AChE has been identified at position 331 (F to W) in *B. tabaci* biotype-B that triggers the resistance against organophosphate (Alon et al., 2008). Within the same species two mutations in the voltage-gated sodium channel have been identified at position 918 (M to V) and 925 (L to I) that makes the species resistant to pyrethroids (Morin et al., 2002). Cross-resistance between neonicotinoids and pymetrozine has also been identified in *B. tabaci* (Rauch and Nauen, 2003; Gorman et al., 2010). Insecticide resistance in *B. tabaci* species has been extensively studied in different populations across China (He et al., 2007; Ma et al., 2007; Luo et al., 2010; Wang et al., 2010b; Xie et al., 2012; Yuan et al., 2012).

The current control strategies are insecticide driven, but efforts are being made to integrate other methods including non-chemical, cultural and biological methods, and their combination has been found as providing the most effective control (Ellsworth and Martinez-Carrillo, 2001). Integrated pest management (IPM) systems have been developed for whitefly pest control which have the four cornerstones including chemical control, host plant resistance, cultural practices and biological control (Gilbertson et al., 2011). Cultural practices have been seen to be more effective because of their preventative nature in controlling whiteflies (Hilje et al., 2001). There are some other methods like natural enemy augmentation and conservation that could also be integrated for whitefly control (Gerling et al., 2001; Naranjo, 2001). The current goal to control whiteflies is to develop sustainable and ecologically-based management systems.

2.1.5 Whitefly life cycle and biology

The life cycle of *B. tabaci* has four main stages: egg, larvae (four stages), a pupa and adult, and ranges from 2.5 - 3 weeks depending on temperature, which optimally is 25 - 30°C (Gerling and Mayer, 1996; Johnson et al., 2005). Depending on the temperature, an adult female can lay 50 - 400 eggs on the undersides of leaves. The eggs are very small (up to 0.2 mm in length and 0.09 mm in diameter), whitish-yellow in colour, oval in shape and tapered at their distal end (Gerling and Mayer, 1996). Over a period of five to seven days, the eggs start hatching at 25°C and become brown. At the nymphal stage (first nymphal instar is 0.3 mm in length) there is only limited movement up to

few centimeters, and so this stage is known as a 'crawler'. The crawlers molt to a second nymphal instar in 2-3 days once they have begun feeding (Mohanty and Basu, 1987). The second, third and fourth nymphal instars are oval in shape, yellowish in colour and immobile. The nymphs excrete a waxy material outside their body that helps them to adhere to the leaf. The fourth nymphal instar is 0.6 mm long with atrophied legs, small eyes and antennae (Brar et al., 2005). The nymphs use their piercing-sucking mouth part to suck the plant sap. At the end of the nymphal stage, the pupal stage begins where the eyes become red, the body expands to 0.7 mm and becomes yellow in colour. Upon development of adult from pupa, the adult *B. tabaci* are 2 - 3 mm in length with white wings and yellow body (Gerling and Mayer, 1996; Johnson et al., 2005; Mann et al., 2009).

Whiteflies are arrhenotokous parthenogenetic insects, meaning an embryo that develops from an unfertilized egg results in a hemizygous male progeny (Gullan and Cranston, 2010; Thompson, 2011). Asexually produced males are haploid, containing half the number of chromosomes as females that are produced sexually and which are diploid (Byrne and Devonshire, 1996; Carriere, 2003; Horowitz et al., 2003). This haplodiploid reproductive system promotes insecticide resistance development in the hemizygous male insects (Horowitz et al., 2003; Crowder et al., 2009). This is because if the resistance genes result from mutations are open to selection from the outset in hemizygous males, even if a resistance gene is recessive, semi-dominant or dominant, resistance can still develop at the same rate (Denholm et al., 1998; Horowitz et al., 2003; Crowder et al., 2009). A comparative study between the greenhouse whitefly and the sweet potato whitefly, suggested that males are less tolerant to insecticides than females (Horowitz et al., 1988; Carriere, 2003). However, research on whiteflies reveals that *B. tabaci* resistant males are as tolerant as resistant females (Carriere, 2003; Horowitz et al., 2005).

2.1.6 Whitefly endosymbionts

Endosymbioses are commonly found in insects as up to 50% of all insects rely on intracellular bacteria for their proper development and growth (Kikuchi, 2009; Kuechler et al., 2011). One of the insect's defenses against parasitoids and pathogens can be beneficial symbiont(s) living within the insect host. *B. tabaci* hosts bacterial endosymbionts (Baumann, 2005) and there are two groups of endosymbionts in *B. tabaci*: primary endosymbionts and secondary endosymbionts. To date, there have been eight endosymbionts reported in *B. tabaci* including the primary essential endosymbiont "*Candidatus* Portiera aleyrodidarum" (*Oceanospirillales*; hereafter '*Portiera*') (Sloan and Moran, 2012a), as well as a range of secondary endosymbionts including *Rickettsia* spp. (*Rickettsiales*; hereafter '*Rickettsia*') (Gottlieb et al., 2006), "*Candidatus* Cardinium hertigii" (*Bacteroidales*; hereafter '*Cardinium*') (Weeks et al., 2003), "*Candidatus* Hamiltonella defensa" (*Enterobacteriales*; hereafter '*Hamiltonella*') (Zchori-Fein and Brown, 2002), *Arsenophonus* spp. (*Enterobacteriales*; hereafter '*Arsenophonus*') (Zchori-Fein and Brown, 2002), "*Candidatus*

Fritschea bemisiae" (*Chlamydiales*; hereafter '*Fritschea*') (Everett et al., 2005), *Candidatus* Hemipteriphilus asiaticus (hereafter '*Hemipteriphilus*') (Bing et al., 2013a) and *Wolbachia* spp. (*Rickettsiales*; hereafter '*Wolbachia*') (Nirgianaki et al., 2003). No correlation has been identified between *B. tabaci* species and their secondary endosymbionts on a global scale as the same species may harbour certain endosymbionts in one geographic region and not in the other (Gueguen et al., 2010; Skaljac et al., 2010; Marubayashi et al., 2014). These communities of secondary endosymbionts in *B. tabaci* can vary across different species and their geographical distributions. Although these secondary endosymbionts have different patterns of localisation inside *B. tabaci*, they all share bacteriocytes with *Portiera* (Gottlieb et al., 2008; Skaljac et al., 2010). These frequent infections of the primary endosymbiont along with different secondary endosymbionts make *B. tabaci* an interesting model species to study the metabolic complementation across endosymbionts.

Portiera is a primary endosymbiont residing in specialized cells called bacteriocytes in *B. tabaci* (Baumann et al., 2004) and provides carotenoids (Sloan and Moran, 2012a) as well as essential nutrients to supplement the restricted diet provided by host plants (Baumann, 2005). Other endosymbionts are secondary endosymbionts in *B. tabaci* and these appear to play significant (but not essential) roles in ecology and biology (Gottlieb et al., 2010; Brumin et al., 2011; Himler et al., 2011). For example, the detoxification of insecticide has been correlated to high bacterial densities in *B. tabaci* (Ghanim and Kontsedalov, 2009). The composition and density of endosymbionts in *B. tabaci* is very important as it has been reported to influence virulence of the endosymbiont and its transmission to the offspring (Kondo et al., 2005). Recent research by Pan et al. (2013) suggests that insecticide resistance and host plant adaptation changes the relative amount of endosymbionts *Portiera*, *Rickettsia*, *Hamiltonella* and *Cardinium* in *B. tabaci* MEAM1 population.

Portiera

Portiera, a gammaproteobacteria, is the primary endosymbiont of the whitefly species complex and therefore infects every individual *B. tabaci* species. This endosymbiont is confined in specialized host cells called the bacteriocytes and is vertically transmitted which has resulted in parallel evolution with their hosts for millions of years (Wilson et al., 2010; Husnik et al., 2013; Sloan et al., 2014). As a result of this endosymbiotic relationship, an extreme genome reduction and degradation has been observed in many insect hosts that harbour primary endosymbionts, predominantly in the Sternorrhyncha species, including *Portiera* in whiteflies (Sloan and Moran, 2012a, 2013; Santos-Garcia et al., 2012), *Buchnera aphidicola* in aphids (Shigenobu et al., 2000; Pérez-Brocal et al., 2006), *Carsonella ruddii* in psyllids (Sloan and Moran, 2012), as well as two mealybugs, *Tremblaya princeps* and *Moranella endobia* (McCutcheon and vonDohlen, 2011; Husnik et al., 2013; López-Madrigal et al., 2013). *Portiera* infected insect hosts use plant phloem sap as an unbalanced food source which is then complemented by the primary endosymbiont via providing essential amino acids to their hosts (Douglas, 1998). Similarly, *Portiera* has been known to provide essential nutrients to *B. tabaci* species (Baumann, 2005). Interestingly, Sloan and Moran (2012a)

have reported *Portiera* has the ability to synthesize carotenoids in whiteflies that insects generally obtain from their diet.

Portiera complete genome sequence have been obtained from two different *B. tabaci* species (MED and MEAM1) and analysed by Jiang et al. (2012). For the MED population, they reported a *Portiera* genome size of 350,928 base-pairs (bp) comprising 281 protein coding genes (PCGs) and 36 RNA coding genes. In contrast, for the MEAM1 population, the genome size reported was 351,658 bp comprising 277 PCGs and the same 36 RNA coding genes. Later, these *Portiera* genomes from MED and MEAM1 populations were found incomplete by Sloan and Moran (2012a) and Santos-Garcia et al. (2012). Sloan and Moran (2012a), sequenced and analyzed the *Portiera* genome from MEAM1 population and reported its size to be 358,242 bp (Genbank accession: CP003708) with GC content of 26.2% comprising 296 genes encoding 256 proteins, 33 transfer RNAs (tRNAs) and three ribosomal RNAs (rRNAs). Santos-Garcia et al. (2012), reported a slightly smaller 357,472 bp (Genbank accession: CP003835) long *Portiera* genome from the MED population with GC content of 26.1% and comprising 292 genes encoding 246 proteins, eight pseudogenes and 38 noncoding RNA genes (33 tRNAs, three rRNAs, rnpB, and tmRNA). These genome sizes were confirmed by Jiang et al. (2013) for *Portiera* genomes from both the MEAM1 and MED populations. They identified that a 6 kilo base-pairs (kbp) region containing three genes (yidC, membrane protein insertase; trmE, GTP-binding protein; gidA, tRNA uridine 5-carboxymethylaminomethyl modification enzyme) was absent in their previous assemblies (Jiang et al., 2013). Comparisons revealed 99.8% similarity between the *Portiera* genomes from the MEAM1 and MED which had similar gene synteny although the *Portiera* from MED population was found to be 770 bp shorter than that of MEAM1 population (Jiang et al., 2013).

Wolbachia

Wolbachia strains are the most widely distributed rickettsial endosymbionts across the major arthropod classes (Hilgenboecker et al., 2008) and are particularly prevalent in herbivorous insects of the Hemiptera suborder Sternorrhyncha (aphids, whiteflies, psyllids, scales and mealybugs) (Moran, 2001). They are transmitted maternally and enhance this process by manipulating the host's reproductive systems by, for instance, feminization of genetic males, parthenogenesis, cytoplasmic incompatibility, and killing male progeny (Stouthamer et al., 1999; Werren et al., 2008). In addition, *Wolbachia* benefits its insect hosts by provision of essential nutrition, (e.g. in bedbug (Hosokawa et al., 2010), increasing fitness in uzifly (Guruprasad et al., 2011), leaf-mining moth (Kaiser et al., 2010) and mosquito (Dobson et al., 2002)), increasing host stem cell proliferation (Fast et al., 2011), and also protecting its host from a range of RNA viruses (Hedges et al., 2008; Teixeira et al., 2008; Moreira et al., 2009; Glaser and Meola, 2010). Although the putative role of *Wolbachia* in *B. tabaci* species remains poorly defined, its infection has been identified in several *B. tabaci* species (Zchori-Fein and Brown, 2002; Nirgianaki et al., 2003; Bing et al., 2014) and has been found responsible for a cause of cytoplasmic incompatibility across its host species (Barro and Hart,

2000). A recent study by Xue et al. (2012) investigated the physiological roles of *Wolbachia* in development and reproduction of *B. tabaci* MED population by selectively eliminating it using the antibiotic 'rifampicin' and then comparing the biology of *Wolbachia*-infected and -uninfected MED population. For *Wolbachia*-infected MED population, they found that the *Wolbachia* is associated with the complete development of nymph, increased percentage of female progeny and adult life span, and decreased juvenile development time. The significant reduction in nymph body size was found in uninfected MED population which suggest *Wolbachia* may supply nutrient to its host (Xue et al., 2012).

Hamiltonella

Hamiltonella, a gammaproteobacteria, is a maternally transmitted endosymbiont harboured by some plant sap-sucking insects including whiteflies and aphids. In some aphids, this secondary endosymbiont provides defence against parasitoid wasps which results in high mortality of wasp larvae (Oliver et al., 2003) or an increase in the heat tolerance of the host by inducing the expression of host heat shock genes (Russell and Moran, 2006). In the whitefly species complex, *Hamiltonella* infection has been found in only two *B. tabaci* species including MEAM1 and MED (Gottlieb et al., 2008; Chu et al., 2011; Thierry et al., 2011; Bing et al., 2013b). Similar to the other secondary endosymbionts in *B. tabaci*, *Hamiltonella* shares bacteriocytes with the primary endosymbiont *Portiera* (Gottlieb et al., 2008; Skaljac et al., 2010). *Hamiltonella* has been identified as a mutualistic endosymbiont with a host-dependent metabolism, relying on its host primary endosymbionts for required nutrients (Degnan et al., 2009). The essential nutrients not provided by primary endosymbiont might be supplied by host secondary endosymbionts. The genome sequencing of *Hamiltonella* revealed gene annotations which indicate the ability of *Hamiltonella* to synthesize amino acids and cofactors (Rao et al., 2012a). Other studies suggest that *Hamiltonella* was associated with an increase in whitefly fitness (Su et al., 2013). *Hamiltonella* has also been found to facilitate the transmission of TYLCV (Gottlieb et al., 2010) through protein-protein interaction between *Hamiltonella* GroEL protein and TYLCV coat protein (Morin et al., 1999).

The draft genome sequence of *Hamiltonella* from a MED population was found to be 1.84 Mbp with a GC content of 40.3% and containing 372 scaffolds encoding 1,806 proteins (Rao et al., 2012a).

Rickettsia

Rickettsia are small (0.4 by 0.9 -m), gram-negative, α-proteobacteria harboured by both invertebrate and vertebrate hosts. In *B. tabaci*, a *Rickettsia* was first reported by Gottlieb et al. (2006). Later, *Rickettsia* was also identified in other *B. tabaci* species including Asia II 7, China 1, Asia II 3 and Indian Ocean populations (Bing et al., 2013b,a; Gueguen et al., 2010). Unlike other endosymbionts in *B. tabaci*, *Rickettsia* are not confined to the bacteriocyte but rather were found throughout the body of the whitefly. *Rickettsia* are maternally inherited via entering into oocytes (Gottlieb et al., 2006). Transmission of *Rickettsia* in *B. tabaci* has also been reported while feeding on host

plants, which implies rare interspecific horizontal transmission (Caspi-Fluger et al., 2012). This secondary endosymbiont has been associated with dramatically increasing the fitness of whitefly, resulting in sex ratio distortions with higher proportion of females, and faster developmental stages (Himler et al., 2011). In addition, *Rickettsia* has also been shown to confer resistance to heat shock (Brumin et al., 2011) and greater susceptibility to various insecticides (Kontsedalov et al., 2008). In fact, the highest susceptibility to acetamiprid, thiamethoxam, sporimesifen and pyripoxen was found in double-infected (*Rickettsia-Arsenophonus* or *Rickettsia-Wolbachia*) strains (Ghanim and Kontsedalov, 2009). This increased susceptibility has been demonstrated in both MEAM1 (Chiel et al., 2007) and MED populations of the whitefly complex (Ghanim and Kontsedalov, 2009).

The draft genome sequence of *Rickettsia* from a MEAM1 populations was determined to be 1.24 Mbp with a GC content of 32.1% and comprising 335 scaffolds and 1,247 proteins (Rao et al., 2012b).

Cardinium

Cardinium was first characterized in an *Encarsia* wasp, a parasitoid of *B. tabaci*, and it was proposed as the species type (Zchori-Fein et al., 2004). Subsequently, bacterial infections from the genus *Cardinium* were identified in arthropods including whiteflies, ticks, mites and spiders (Zchori-Fein and Perlman, 2004; Gruwell et al., 2009; Nakamura et al., 2009). The genus *Cardinium* has been classified into four subgroups (A, B, C and D) and strains, adopting a similar nomenclature to *Wolbachia* (Lo et al., 2002; Nakamura et al., 2009; Edlund et al., 2012). Similar to *Wolbachia* endosymbionts, *Cardinium* has been reported as a reproductive manipulator in several arthropods because of the diverse effects associated with it, including cytoplasmic incompatibility, feminization and parthenogenesis induction (White et al., 2011). However, these effects have not been reported on *B. tabaci* and hence it is possible that *Cardinium* might also be a mutualistic endosymbiont in this insect species complex. Previous studies on secondary endosymbiont co-infections in *B. tabaci* revealed that the species infected with *Cardinium* are, however, very unusual, particularly the C1 strain which is predominantly found in whiteflies (Gueguen et al., 2010).

The genome sequence of *Cardinium* cBtQ1 from a MED population is 1.065 Mbp long containing 11 contigs with a GC content of 35%, 709 PCGs, 35 tRNAs, three rRNAs and 156 pseudogenes (Santos-Garcia et al., 2014a). A comparison with the *Cardinium* cEper1 lineage revealed gene loss in *Cardinium* cBtQ1 which affected genes encoding cofactors and amino acid biosynthesis. Additionally, the *Cardinium* cBtQ1 genome also revealed a large proportion of transposable elements, which cause chromosomal rearrangements and have inactivated genes. Chromosomal duplication and a multicopy plasmid were also reported in the *Cardinium* cBtQ1 genome, which encodes certain proteins that play a role in gliding motility and potential insecticidal activity (Santos-Garcia et al., 2014a).

Arsenophonus

Bacterial endosymbionts falling into the *Arsenophonus* (Proteobacteria) genus are found in approximately 5% of total arthropods (Duron et al., 2008; Nováková et al., 2009). *Arsenophonus* has been found to be present in important insect pests including the whitefly, *B. tabaci* (Thao and Baumann, 2004), the lerp psyllid, *Aphis gossypii* (Hansen et al., 2007), and the soybean aphid, *Aphis glycines* (Wille and Hartman, 2009), but its functional role inside these pests remains poorly characterized. However, Rana et al. (2012) have found protein-protein interaction between the *Arsenophonus* GroEL protein and the *Cotton leaf curl virus* (CLCV) coat protein, and reported the active role of *Arsenophonus* in virus transmission in whitefly. Another active role of *Arsenophonus* has been documented in the parasitoid wasp, *Nasonia vitripennis*, the bacteria *Arsenophonus nasoniae* causing male egg mortality (Huger et al., 1985; Werren et al., 1986; Gherna et al., 1991; Duron et al., 2010). Previous studies have also suggested a defensive role of *Arsenophonus* inside its hosts. According to Hansen et al. (2007), during the geographical survey of the lerp psyllid, *Glycaspis brimblecombei*, they found a positive correlation between the frequency of *Arsenophonus* infection and parasitism, which suggested the potential of *Arsenophonus* to provide selective benefits to the psyllid populations under heavy parasitism pressure (Hansen et al., 2007).

Fritschea

Fritschea was first identified by Thao et al. (2003) in the *B. tabaci* species complex. It belongs to the order *Chlamydiales* and was classified into four families based on 16S and 23S rDNA sequence phylogeny (Thao et al., 2003). The genus *Fritschea* has unknown effects on its host and not much is known about its phenotypes. However, *Chlamydiales* endosymbionts in *B. tabaci* have only been found inside bacteriocytes in the gut, which are transmitted directly to oocytes from the parent (Everett et al., 2005).

Hemipteriphilus

A novel *Orientia*-like bacterial endosymbiont was detected in *B. tabaci* (Bing et al., 2013b) and later this bacterium was described as *Hemipteriphilus* in a China 1 population of the *B. tabaci* complex (Bing et al., 2013a). According to phylogenetic analysis of 16S rRNA and the *gltA* gene, this bacteria belongs to the *Alphaproteobacteria*, a subdivision of the *Proteobacteria* and is closely related to human pathogens of the genus *Orientia*. *Hemipteriphilus* was found in a China 1 population (tentative species) with the infection rate ranging from 46.2% to 76.8%. Similar to the other secondary endosymbionts, this endosymbiont was also found to be confined in the bacteriocytes of *B. tabaci* sharing this localization with the primary endosymbiont *Portiera* (Bing et al., 2013a).

2.2 Current status of whitefly 'omics' data

Using the power of NGS techniques, extensive genetic sequence information of any organism can be obtained from two principal forms namely DNA (whole genome sequencing) and mRNA (whole transcriptome sequencing). In the context of *B. tabaci*, whole transcriptomes of various population/species of the whitefly cryptic species complex have been sequenced and analyzed by research groups across the globe (Leshkowitz et al., 2006; Wang et al., 2010a, 2011, 2012; Xie et al., 2012; Seal et al., 2012; Upadhyay et al., 2015). Whole genome sequencing projects for MEAM1 and MED populations are ongoing and have not been published yet. However, various estimations of their genome sizes have been published as 640-682 Mbp (MEAM1 and MED) (Guo et al., 2015) and 680-690 Mbp (MEAM1) (Chen et al., 2015), which are lower than the previously estimated genome size 1,020 Mbp of MEAM1 population (Brown et al., 2005).

Thus to date, whitefly 'omics' data represents the genetic information of the whitefly species complex in the form of whole genome sequencing, whole transcriptome sequencing and mitochondrial genome sequencing. Genome sequence information of some of the aforementioned endosymbionts is also published (see earlier sections).

2.2.1 Transcriptome sequencing

Since the first publication of a *B. tabaci* transcriptome in 2006 (Leshkowitz et al., 2006), four other transcriptomes of different *B. tabaci* populations have been sequenced and characterized. The order of their publication has been as follows: MED (Wang et al., 2010a), MEAM1 (Wang et al., 2011; Xie et al., 2012), Asia II 3 (Wang et al., 2012) and Asia I populations (Seal et al., 2012; Upadhyay et al., 2015). Of these, Asia I population (Seal et al., 2012) was sequenced at an adult stage using Roche 454 and MEAM1 population was also sequenced using Roche 454 including egg, nymph and adult stages whereas the rest were sequenced at different developmental stages using shorter read length Illumina sequencing technology.

MED

A transcriptome of MED, a cryptic species of the *B. tabaci* complex, was sequenced and analyzed by Wang et al. (2010a). The cDNA libraries were prepared from total mRNA extracted from eggs, nymphs, pupae and adult whiteflies. Using the Illumina sequencing GAII platform, more than 43 million reads of 75 bp in length were generated. Initial assembly of these reads generated 4,274,766 contigs with length ranging from 22 bp to 2,189 bp using SOAPdenovo (Li et al., 2010). These contigs were further assembled into 170,115 scaffolds using paired-end (PE) joining and gap-filling. Using the TGICL program (Pertea et al., 2003), these scaffolds were clustered into 168,900 unique sequences including 1,206 clusters and 167,694 singletons. Sequence homology searches against

the non-redundant (nr) database at the National Center for Biotechnology Information (NCBI) returned 27,290 unique sequences with BLAST hits using an E-value cutoff of $1E^{-05}$. There were 20% of distinct sequences that had top-hits to *A. pisum*, followed by *Pediculus humanus corporis* (15%), *Tribolium castaneum* (12%), and *Apis mellifera* (10%). They also found 126 sequences scored highest homology with the limited existing *B. tabaci* sequences at the nr database and the majority of them were to heat shock protein and cytochrome P450s. The BLAST results were further annotated by assigning gene ontology (GO) terms to 7,330 sequences which were classified into 52 groups and 214 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were assigned to 11,104 sequences. The raw sequencing reads were deposited at the NCBI Short Read Archive (SRA) database under the accession number: SRX018661.

MEAM1

MEAM1, considered as one of the most invasive and destructive cryptic species of the *B. tabaci* complex, has been studied extensively. Transcriptome sequencing of MEAM1 has been carried out by three different groups: Leshkowitz et al. (2006), Wang et al. (2011) and Xie et al. (2012). The very first transcriptome of any *B. tabaci* species was sequenced by Leshkowitz et al. (2006), who generated independent cDNA libraries from eggs, immature instars (crawler to pupae) and adults of non-viruliferous whiteflies. There were 18,976 sequences generated in total from all the libraries of which only 9,110 sequences remained after quality, vector and adapter trimming, and mitochondrial sequences being removed. These 9,110 sequences then assembled into 3,843 singletons and 1,017 contigs using Staden gap4 (Bonfield et al., 1995). About 45% of the total singletons and contigs had a homology from nr, Nucleotide, Swissprot, EST databases with an E-value cut-off of $1E^{-06}$. The taxonomy distribution of the homolog sequences found that 58% of singletons and contigs had a sequence homology to insects including *A. gambiae*, *A. mellifera*, *B. tabaci* and *D. melanogaster*. All the raw sequences were deposited to the EST database at NCBI under the accession numbers EE595518-EE604534, EE674607-EE674699.

Wang et al. (2011) sequenced the transcriptome at developmental life stages including egg, nymph, pupa, adult female and male using the Illumina sequencing GAII platform. A total of 17 million reads of 75 bp in length were generated and assembled into 123,055 contigs using SOAPdenovo software (Li et al., 2010). These contigs were further assembled into 104,722 scaffolds using PE joining and gap filling, and these were clustered into 57,741 distinct sequences and 57,606 singletons. Of these 57,741 distinct sequences, only 15,922 (27.60%) were assigned functional annotation using BLASTX of which 4,711 sequence were assigned GO terms. In comparison with the MED population transcriptome, they found 3,585 pairs of high quality orthologs which inferred their sequence divergence with MEAM1 population. Analysing sequence divergence, they found average differences as 0.83%, 1.66% and 1.43% in coding, 5' untranslated and 3' untranslated regions respectively. The raw reads can be accessed from NCBI SRA database with the accession

SRX022878, and the assembled sequences can be searched at the Transcriptome Shotgun Assembly (TSA) database using the accession numbers HP643344 to HP701084 (Wang et al., 2011).

Xie et al. (2012) also sequenced the transcriptome of MEAM1 at developmental life stages including egg, nymph and adult, but using the Roche 454 GS FLX sequencing platform. After quality checking, trimming low quality reads and removing mitochondrial, rRNA and short reads (<100 bp), approximately 340 Mbp size of transcriptome was generated from 857,205 reads. The latter reads assembled into 178,669 unigenes including 23,694 isotigs and 154,975 singletons. Based on the sequence homology at nr database, of the total 178,669 unigenes, 30,980 had homology with insects, 17,881 from bacteria and 129,808 remained as non homologous at an E-value cut-off of $1E^{-05}$. Among the homolog sequences, 40,611 unigenes were assigned GO terms and 288 KEGG pathways were assigned to 6,917 unigenes. This de novo metatranscriptome analysis reported a wide diversity of bacterial endosymbionts in *B. tabaci* as well as genes representing the synthesis of amino acids in host endosymbiont relationships. Further in-depth analysis revealed putative molecular markers and potential insecticide resistance genes. They found over-expression of cytochrome P450 genes in the thiamethoxam-resistant *B. tabaci* compared to thiamethoxam-susceptible *B. tabaci* populations. The above Roche 454 transcriptome reads were deposited in the NCBI SRA database under accession number SRA036954.

Asia II 3

The transcriptome of Asia II 3, another cryptic species of the *B. tabaci* complex, was sequenced in 2012 (Wang et al., 2012). Using the GAII Illumina sequencing platform, 16 million reads of 74 bp in length were generated. These reads were assembled into 144,103 contigs with an average length of 201 bp using the SOAPdenovo assembly program (Li et al., 2010). These contigs were further assembled into 77,263 scaffolds with a mean size of 359 bp and clustered into 52,535 distinct sequences. Of these 52,535 sequences, only 16,596 (31.60%) were annotated based on BLASTX sequence homology against protein database at nr. Illumina sequencing reads were deposited at the NCBI SRA database under the accession number SRR062575. The assembled sequences were also deposited at the TSA database at DDBL/EMBL/GenBank under the accession numbers HP777244 to HP823074 (Wang et al., 2012).

Asia I

Asia I, one of the cryptic species from the *B. tabaci* complex, is the subject of this thesis. The transcriptome of Asia I population was sequenced and analyzed at the University of Greenwich (Seal et al., 2012). The transcriptome sequencing of adult females from an Asia I population from India was performed using the Roche 454 Titanium platform, and generated 310,094 single reads with a mean read length of 336 bp. The initial assembly of these single reads by CLC Genomics Workbench generated 14,217 contigs of which 3,821 were identified as core contigs based on their consensus sequence length. Of the total 3,821 core contigs, only 1,997 (52.26%) were assigned

functional annotations using the nr database and 6,714 GO terms including 2,618 terms associated with biological process, 2,439 associated with molecular function and 1,657 associated with cellular function. *A. pisum*, a hemipteran insect pest, was found to be the closest species to Asia I with the highest number of BLASTX top-hits. The assembly required manual curation as well as strategies to achieve accurate transcriptome assembly from this existing sequencing reads (Seal et al., 2012).

Recently, another transcriptome of Asia I population was sequenced from different developmental stages including egg, pupa and an adult using the Illumina HiSeq1000 platform (Upadhyay et al., 2015). A total of 1,324,517 scaffolds were assembled from 83 million PE reads using the ABySS assembler (Simpson et al., 2009). Using CAP3 program (Huang and Madan, 1999), these scaffolds were further assembled into 72,716 unitigs comprising 34,428 contigs and 38,288 singletons. Of the total 72,716 unitigs, only 21,129 (29.05%) were annotated with 52,847 GO terms and 131 KEGG pathways (Upadhyay et al., 2015). The assembly was deposited to the TSA database at DDBL/EMBL/GenBank under the accession numbers GAUC00000000 and GAUL00000000.

2.2.2 Mitogenome sequencing

The mitochondrion occupies a substantial portion of the cytoplasmic volume of eukaryotic cells and is responsible for various cellular functions including ATP production, apoptosis, energy transduction, detoxification and signal transduction (Torres et al., 2009; Reeve and Lightowlers, 2012). It has also been shown to be essential in the evolution of many complex species (Lang et al., 1999; van der Giezen, 2011). With few exceptions, most insect mitogenomes constitute 37 genes in total including 13 PCGs, 22 tRNAs and two rRNAs (rrnL and rrnS) (Clary and Wolstenholme, 1985; Boore, 1999). For the past two decades, the *mtCOI* gene has been used extensively as a molecular marker to distinguish different members of the *B. tabaci* complex based on their *mtCOI* phylogeny. Through these pivotal works on *mtCOI* phylogenies, *B. tabaci* has been indicated to be a complex of cryptic species with at least 37 distinct genetic groups (Dinsdale et al., 2010; Barro et al., 2011; Hu et al., 2011a; Alemandri et al., 2012; Liu et al., 2012; Parrella et al., 2012; Firdaus et al., 2013; Hu et al., 2014). The rise of *mtCOI* phylogeny as a molecular tool to demonstrate the delimitation of species in the *B. tabaci* has led to a complete mitochondrial genome comparison to better understand the global evolutionary genetic relationship across the species of this complex. The sequencing and analysis of complete mitogenomes has been facilitated by advances in NGS technologies (Behere et al., 2016). To date, the complete mitogenomes from only three species of the *B. tabaci* complex (New World I, MED and Asia I) have been published and hence comparisons on the phylogeny of complete mitogenomes versus the *mtCOI* have been limited (Thao et al., 2004; Wang et al., 2013; Tay et al., 2016).

New World I

New World I, one of the cryptic species of the *B. tabaci* complex, was the first species from the complex whose complete mitogenome was sequenced and characterized (Thao et al., 2004) (GenBank accession: AY521259). The complete mitogenome of New World I is 15,322 bp long consisting 13 PCGs, 22 tRNAs, two rRNAs and two repeat regions. Thao et al. (2004) also sequenced mitogenomes of other whiteflies (not belonging to the *B. tabaci* complex) including *Tetraleurodes acaciae* (GenBank accession: AY521626), *Neomaskellia andropogonis* (GenBank accession: AY572539), *Aleurochiton aceris* (GenBank accession: AY572538), *Trialeurodes vaporariorum* (GenBank accession: AY521265), and *Aleurodicus dugesii* (GenBank accession: AY521251) along with one psyllid, *Pachypsylla venusta* (GenBank accession: AY278317) and one aphid, *Schizaphis graminum* (GenBank accession: AY531391). They found very similar gene synteny in two whitely species (*T. vaporariorum* and *A. dugesii*), one aphid (*S. graminum*) and one psyllid (*P. venusta*) to that of the proposed ancestral gene order in insects. However, a different gene arrangement to that of the proposed insect ancestor were also seen in the remaining four whiteflies which included *B. tabaci*, *T. acaciae*, *N. andropogonis* and *A. aceris*. The rearrangement was caused by excision of a DNA fragment encoding for COI-tRNA^{gly}-NADH dehydrogenase subunit 3 (*ND3*)-tRNA^{ala}-tRNA^{arg}-tRNA^{asn} between ATP synthase subunit 6 (*ATP6*) and NADH dehydrogenase subunit 5 (*ND5*) genes.

MED

The mitogenome of a MED population was assembled by Wang et al. (2013) using the complete mitogenome of the New World I population as a reference. Initially, eight contigs were assembled from the MED population mitogenome sequencing by mapping reads to the New World I population reference mitogenome followed by contig joining and gap filling to build the complete mitogenome of MED population. The complete mitogenome of MED population is 15,632 bp long and encodes 37 genes in total including 13 PCGs, 22 tRNAs and two rRNAs (Wang et al., 2013), similar to the New World I population and most metazoan mitogenomes (Boore, 1999). The MED population mitogenome also contains 10 non-coding regions of at least 10 bp in size. They also found overlap across two pairs of genes (NADH dehydrogenase subunit (*ND4*) - *ND4l*) by 3 bp and (*ATP6* - ATP synthase subunit 8 (*ATP8*)) 10 bp on the same strand, which is common in insect mitogenomes (Beckenbach and Stewart, 2009; Negrisolo et al., 2011). Polyadenylated genes were also found such as cytochrome c oxidase subunit I *COXI*, cytochrome oxidase subunit II *COXII* and *ND5* with partial stop codons, which are completed by changing T to the TAA stop codon via polyadenylation. Similar gene arrangements were found by comparative analysis of the mitogenomes from MED and New World I populations. However, in-depth sequence comparison of 13 PCGs revealed 21.30% overall divergence between MED and New World I populations (Wang et al., 2013), which was higher than the divergence of marker gene *mtCOI* (14.9%) that has been used to delimit the species of the *B. tabaci* complex (Dinsdale et al., 2010; Barro et al., 2011; Hu et al., 2011a; Alemandri

et al., 2012; Liu et al., 2012; Parrella et al., 2012). The complete mitogenome of MED population was deposited at DDBL/EMBL/GenBank database under the accession number JQ906700.

Asia I

The complete sequence of an Asia I population mitogenome was sequenced and assembled by Tay et al. (2016). This represented the third complete mitogenome sequence published from the *B. tabaci* species complex. The complete mitogenome of this Asia I population is 15,210 bp in size and consists of 13 PCGs, 22 tRNAs, two rRNAs and a 467 bp putative control region with A+T repeats. Similar gene synteny was found between the Asia I population mitogenome and the above two mitogenomes of *B. tabaci* complex members New World I and MED. In addition, overall similarity was also found with the mitogenome of *Bemisia afer*, a whitefly of another *Bemisia* species complex (Chu et al., 2010), with only one $tRNA^{sera}$ being absent in *B. after*. Comparative analysis of the Asia I population mitogenome with the New World I and MED populations mitogenomes revealed no gene rearrangement but found one minor difference as $tRNA^{arg}$ found on "plus" strand in Asia I population whereas on "negative" strand in New World I and MED populations (Tay et al., 2016). The complete mitogenome of this Asia I population was deposited at DDBL/EMBL/GenBank database under the accession number KJ778614.

CHAPTER 3

De novo transcriptome assembly and characterization from *Bemisia tabaci* cryptic species Asia I

3.1 Introduction

The whitefly species complex, *B. tabaci*, exists as a complex of "cryptic species", originally considered as "biotypes" and containing at least 37 morphologically indistinguishable and genetically distinct cryptic species (Dinsdale et al., 2010; Barro et al., 2011; Hu et al., 2011a; Alemandri et al., 2012; Liu et al., 2012; Parrella et al., 2012; Firdaus et al., 2013; Hu et al., 2014). These species differ in host range, virus transmission, mating behaviour, isozymes, genetic composition and fecundity (Liu et al., 2007; Crowder et al., 2010). Among them, two genetic clades of the complex, MEAM1 and MED have been studied extensively and considered the most predominant and damaging species (Perring, 2001; Liu et al., 2007). In this particular study, the Asia I species of the *B. tabaci* complex from India was focused on due to the lack of genetic information of this important pest of cotton for which insecticide resistance development is a particular problem.

The genetic diversity and divergence across the *B. tabaci* species complex is very important to determine how species specific phenotypes have been formed. However, little is known of the molecular factors driving such differences across these species, and it is not clear how transcriptomes of these species have been affected by natural selection through divergence from a common ancestor. To date, most population studies involving *B. tabaci* genotyping have used DNA sequence based methods to differentiate *B. tabaci* species. These have included only a few genes such as *mtCOI*, 16S ribosomal DNA, *RNApyII*, *shaw*, *prp8* and the nuclear ribosomal intergenic transcribed spacer 1 (Boykin et al., 2007; Dinsdale et al., 2010; Barro et al., 2011; Hsieh et al., 2014). Partial *mtCOI* gene sequences (657 bp) have been the most widely used molecular marker for this species complex and can be carried out on a single insect (Dinsdale et al., 2010; Hu et al., 2011a; Liu et al., 2012). However, it is certain that the vast majority of important genetic traits will not be reflected in the *mtCOI* status of *B. tabaci*. Additionally, the identification of individual genes may not provide

an accurate description of genome-wide sequence divergence. It is therefore important to identify other genetic markers to obtain more robust genomic divergence across *B. tabaci* species and to complement the results of *mtCOI* genotyping. Transcriptome sequencing is a rapid way to obtain large numbers of nuclear as well as the mitochondrial gene markers and can represent a near complete catalog of expressed genes to study genome-wide sequence divergence.

Recent advancements in next generation sequencing technologies offer rapid and high-throughput sequence determination in genomic research for which less or no genomic resource is available (Gibbons et al., 2009). To date, there is no draft genome available for any *B. tabaci* species. Transcriptome sequencing using high-throughput, next generation sequencing has already provided a rich molecular resource for better understanding of molecular mechanisms of insecticide resistance and functional analysis of differentially expressed genes in the MED, MEAM1 and Asia II 3 populations (Wang et al., 2010a, 2011, 2012). Significant genetic divergence has also reported between these three species comparing their transcriptomes (Wang et al., 2010a, 2011, 2012). More genomic information from other *B. tabaci* species is needed to resolve further divergence between them. In this study, the transcriptome of adult female whiteflies of a *B. tabaci* Asia I inbred population, was sequenced and characterized. This transcriptome will provide a rich molecular resource for the identification of Asia I genes such as those involved in host adaptation, biological invasion and insecticide resistance. It will also provide a valuable resource to gain further insight into genome characteristics and can be used as evidence to generate more robust gene predictions from the Asia I genome assembly as it becomes available.

The transcriptomes from four *B. tabaci* species including Asia I, MEAM1, MED and Asia II 3 were compared to reveal the global genetic divergence across them and to identify gene orthologs which indicate signs of diversifying natural selection.

Specific full-length Asia I species genes were also compared with the corresponding ortholog genes in the fruit fly *D. melanogaster* (Adams et al., 2000), as this is a well-characterized model insect, as well as, the pea aphid *A. pisum* (IAGC, 2010) as a related hemipteran insect. In addition, corresponding gene orthologs from the termite *Zootermopsis nevadensis* (Dictyoptera: Termopsidae) and Asian citrus psyllid *Diaphorina citri* (Hemiptera: Psyllidae) (Holt et al., 2002), whose genomes have been recently sequenced, were also included in the comparison. Such phylogenetic approaches reveal the sequence divergence between different orders of Insecta class and also allow a detailed analysis of potential insecticide targets.

3.2 Methods

3.2.1 Establishment of the Asia I species colony

Originally, the Asia I population (adult females) used was collected from aubergine (egg-plant) in Coimbatore, South India (Rekha et al., 2005). At the Natural Resources Institute (NRI) quarantine insectary, a colony was reared from these insects on cotton, *Gossypium hirsutum* L. *cv.* Laxmi, in insect-proof cages ((26±1)°C, 14 h:10 h L:D, (70±10)%r.h). An inbred line was obtained via initialising ten sub colonies allowing single mating pairs of *B. tabaci*, to colonise a new *B. tabaci*-free cotton seedling. After 25 days, only one sub-colony was selected which had the most adults and single mating pairs, and these were placed on to ten new cotton seedlings and the remaining sub-colonies were discarded. This process was repeated for seven generations and the colony with most adults was selected and its population allowed to increase. The purity of the colony was confirmed, first by RAPD-PCR fingerprinting (Gawel and Bartlett, 1993) on 20 individuals, followed by *mtCOI* partial gene sequencing using Btab-UniR and Btab-UniL primers according to the protocol of Shatters et al. (2009). After amplification, a 745 bp PCR product was cloned into pGEM-T easy vector (Promega, USA) and sequenced to confirm the culture purity, before adults were collected for RNA extraction.

3.2.2 RNA isolation, library construction and 454 sequencing

Total RNA was isolated from two batches of 50 adult female whiteflies by placing the live whiteflies in extraction buffer (supplied in the RNAqueous4PCR kit; Ambion, Texas, USA), grinding them using a micropestle, and extracting total RNA according to the manufacturer's protocol. RNA integrity was confirmed using a 2100 Bioanalyzer (Agilent technologies). Purified mRNA from five μg of this total RNA was sent to Evrogen (Evrogen JSC, 16/10 Miklikho-Maklaya Street, Moscow, Russia) for oligo-dT primed mRNA amplification (Zhu et al., 2001) using three oligonucleotide primers: SMART Oligo II oligonucleotide (5' AAGCAGTGGTATCAACGCAGAGTACGCrGrGrG -3'), CDS-GSU primer (AAGCAGTGGTATCAACGCAGAGTACGCrGrGrG-3') and SMART PCR primer (AAGCAGTGGTATCAACGCAGAGTACGCrGrGrG-3').

Sequencing was performed on the GS FLX Titanium platform from 454 Life Sciences/Roche in early 2010. Libraries for sequencing were prepared using the GS FLX Titanium general library preparation kit (454 Life Sciences/Roche), with cDNA libraries purified and quantitated according to the manufacturer's general library preparation protocol from 454 Life Sciences/Roche before sequencing. Two cDNA libraries were generated, one normalized (here after referred to as 'NL') and the other one was kept as unnormalized (here-after referred to as 'UL') to characterise as many

genes as possible from the sequencing data. Fragmentation was performed to produce 400 - 1000 bp double stranded cDNA and a single stranded cDNA library isolated by melting each double stranded cDNA.

Emulsion-based clonal amplification (emPCR) of the single stranded cDNA library was carried out according to the emPCR method (454 Life Sciences/Roche). The single stranded cDNA library was immobilized onto magnetic streptavidin-coated beads, and the beads emulsified with the amplification reagents provided by 454 Life Sciences/Roche to make the final cDNA library, which was sequenced by GS FLX Titanium (454 Life Sciences/Roche).

3.2.3 Quality control and assembly

Raw reads produced by 454 GS FLX were processed further to remove adapter sequences, low quality reads, poly (A:T) tails, empty reads and reads below 30 bp in length using ESTclean (v1.0) (Tae et al., 2012). After preprocessing, all clean reads from NL and UL libraries were assembled separately using three different assemblers to get the best assembly: CLC (v7.0.4) (Brautigam et al., 2011), MIRA (v4.0) (Chevreux et al., 2004) and CAP3 (Huang and Madan, 1999). These assemblers were selected according to their algorithms and compatibility with the single long 454 reads. There are two types of assembly algorithms: de Bruijn Graph (DBG) approach and Overlap Layout Consensus (OLC) approach. In DBG approach, reads are decomposed into k-mers (a k-mer is a subsequence of a fixed-length, k) to build DBG. Each node of DBG corresponds to a k-mer and edges correspond to suffix-prefix matching between them. CLC is the most memory-efficient DBG assembler (Kumar and Blaxter, 2010). In the OCL approach, an overlap graph is built in which each node represents a read and edges correspond to overlap between reads which is computed using pair-wise sequence alignment. MIRA and CAP3 uses the OLC approach to build an assembly. The assembly was performed with the transcript length cut-off of 150bp using all three assemblers. To evaluate the assembly statistics, three assemblies were compared using the quality assessment tool (QUAST) (Gurevich et al., 2013) with transcript size >= 500 bp. QUAST can evaluate genome assemblies by comparing them using different metrics including GC (%), genome fraction (%), duplication ratio, mismatches per 100 kbp, total genes and indels per 100 kbp. The best assembly was selected with the following conditions: higher N50, lower number of ambiguous bases (N's). The harmonic mean of the total sequence sizes was calculated and "lookup" values were identified as 75% of the harmonic mean.

3.2.4 Homology searches and functional annotation

Homology searches and functional annotations of the assembled sequences were performed using Blast2GO (v.3.0) software (Götz et al., 2008). The analysis was carried out using a step-by-step

strategy guided by the software. Initially, to save time, all assembled contigs (here after 'transcript') and singletons were searched locally against the nr protein database using BLASTX (v2.2.29+) program with an E-value cut-off of $1E^{-05}$ and retrieving the maximum 20 hits. The transcripts and singletons that did not receive any BLASTX hits were searched against NCBI's TSA database using BLASTN (v2.2.29+) program with an E-value cut-off of $1E^{-10}$ and the search restricted to only *B. tabaci* sequences. For gene ontology annotation, transcripts with BLASTX hit were imported into Blast2GO and GO identifiers were mapped to the corresponding homolog gene identified by BLASTX. The annotation step in Blast2GO was performed at 0.05% probability level cut-off to reduce the false discovery rate. The annotation augmentation tool (ANNEX) of the Blast2GO was used to modulate the GO terms. The transcripts were then searched remotely against InterPro web server using InterProScan (v5.0) to predict the domains, motifs and classify proteins into families followed by merging the InterProScan GO IDs to existing annotation. The Enzyme Commission (EC) numbers were retrieved via direct mapping of GO terms to their corresponding enzyme codes followed by the metabolic pathways annotation, KEGG. In addition, all annotated transcripts were searched at KEGG Automatic Annotation Server (KAAS) using the single-directional best hit as recommended for ESTs to further enrich the pathway annotations. Finally, GOSlim was performed to extract the key ontological terms and mapping function to existing GO terms (Götz et al., 2008).

3.2.5 Identification of protein families

To reveal the shared and distinct proteins across *B. tabaci*, *A. pisum*, *D. melanogaster*, *Z. nevadensis* and *D. citri*, total proteins were analyzed. From the BLASTX search at nr database, protein sequences were predicted for UL and NL libraries, and all duplicate proteins were removed. Total proteins of *A. pisum* (v2.1b) (http://www.aphidbase.com), *D. melanogaster* (r6.04) (http://www.flybase.org), *Z. nevadensis* (v2.2) (NCBI) and *D. citri* (v1.1) (NCBI) were retrieved. All-vs-all BLASTP (v2.2.29+) was performed using proteins from these five insect species. The BLASTP results were further analysed using OrthoMCL (v2.0.8) (Li et al., 2003) with the alignment cut-off of 50% for both similarity and coverage. The ortholog pairs were grouped into clusters using Markov Cluster Algorithm (MCL) (v14-137) (Enright et al., 2002) with an inflation factor 1.5.

3.2.6 Estimation of transcriptome completeness

The total transcripts of *B. tabaci* were estimated by comparing assembled transcripts from two different libraries, NL and UL, and classifying into shared and unique transcripts. Initially, the transcripts of two libraries were merged into one and then clustered with 95% similarity threshold using CD-HIT-EST (v4.6) program (Fu et al., 2012). The cluster file was then processed to group transcripts into three categories: NL-specific, UL-specific and shared between both. Here, a 'cluster'

is a sequence composed of one or more transcripts. Furthermore, the completeness of the individual transcript was assessed from each library by translating into protein according to BLASTX results. The online tool 'TargetIdentifier' (v2.0) (Min et al., 2005) was used to translate transcripts into proteins using the BLASTX coordinates. This tool also compares the start codon and stop codon of the predicted protein with their homolog protein from nr database and output transcripts as a full-length, short full-length, possible full-length, ambiguous, partial and 3' sequenced partial. Transcripts were recognized as full-length if they had a 5' stop codon in the transcript followed by a start codon (ATG) or did not have a 5' stop codon but had an in-frame start codon (ATG) corresponding to a BLASTX hit.

3.2.7 Identification of repetitive elements and microsatellites

To identify repetitive elements in the transcriptome of Asia I population, all assembled transcripts were analyzed using 'RepeatMasker' (v4.0.5) using RMBLAST (v2.2.27+) alignment program and Repbase database for vertebrates (http://www.repeatmasker.org, Chen, 2004). The 'MIcroSAtellite' (MISA) (v1.0) tool was used to locate microsatellite repeats within assembled transcripts (Thiel et al., 2003). The assembled transcripts were searched for di-, tri-, tetra-, penta- and hexa-nucleotide repeats with the threshold 6, 5, 5, 5 and 5 respectively.

3.2.8 Estimation of gene expression in Asia I

To estimate the expression level of genes in Asia I population, all cleaned reads were mapped to the assembled transcripts from both libraries using CLC mapper (v7.0.4) with 50% similarity and 95% sequence length threshold. Each read should map to the reference over at least 50% of its length with 95% similarity to consider it as a mapped read. The read counts were extracted from both libraries and Reads Per Kilobase per Million mapped reads (RPKM) was calculated for each reference transcript as described by Mortazavi et al. (2008). These RPKM values were used to determine the gene expression level using the 'R' package, edgeR (Robinson et al., 2010).

3.2.9 Phylogenetic analysis with other *B. tabaci* species

The orthologs of full-length genes of Asia I population were identified in published transcriptomes including MEAM1 (Wang et al., 2011), MED (Wang et al., 2010a) and Asia II 3 species (Wang et al., 2012) according to their previous descriptions using BLASTP. Orthologs were also selected from other insects including *A. pisum*, *D. melanogaster*, *Z. nevadensis* and *D. citri*, and aligned using the MAFFT (v6.903b) program (Katoh and Standley, 2013). The phylogenetic tree was inferred using the Neighbor-Joining method (Saitou and Nei, 1987).

3.3 Results and discussion

3.3.1 454 sequencing summary

To enable a comprehensive view and profiling of the Asia I transcriptome, total mRNA was extracted from adult female and sequenced using Roche 454 GS FLX sequencing technology. A total of 301,094 single-end reads were generated with an average read length of 336 bp and GC content of 37.45% for the UL library. In contrast, the NL library produced 563,662 single-end reads with an average read length of 213 bp and GC content of 35.73%. After quality control, removal of adapter sequences, low quality reads, poly (A:T) tails, empty reads and reads below 30 bp in length, 276,861 (91.95%) cleaned reads were obtained for the UL library and 387,833 (68.80%) reads for the NL library (Figure 3.1). The reason for the high percentage of cleaned reads in NL library was the higher number of short reads (below 30 bp). This could be due to the random events occurred during the sequencing run or intrinsic properties of the library preparation as reported in previous studies (Hale et al., 2009; Yang et al., 2010; Ewen-Campen et al., 2011).

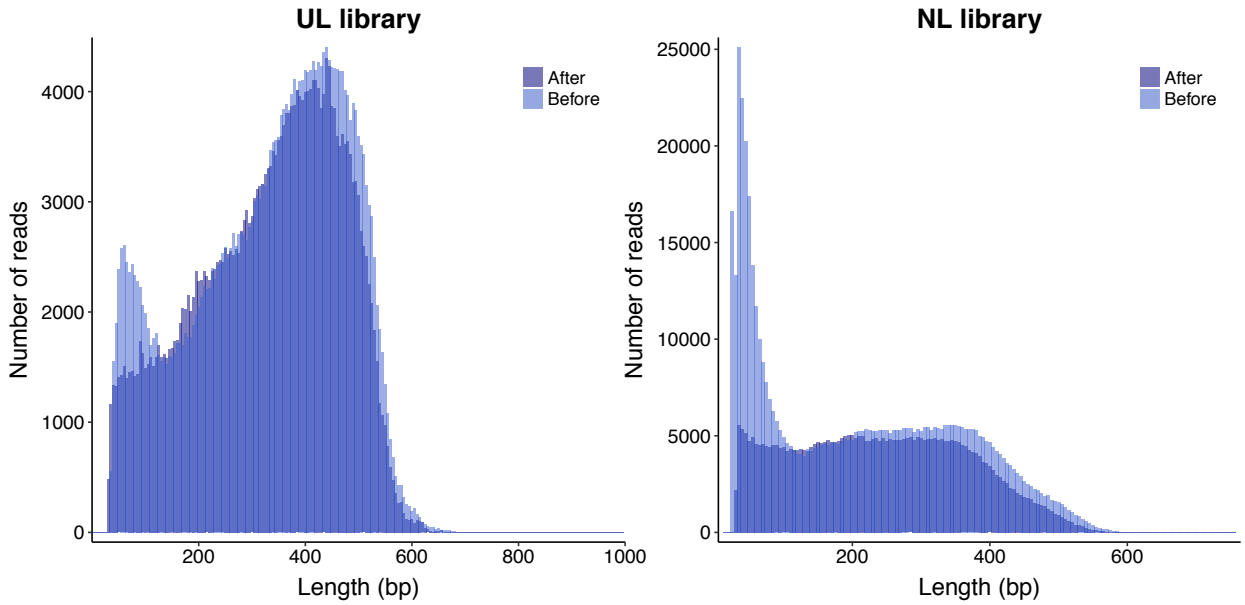


FIGURE 3.1: Read distribution of UL and NL libraries before (light blue colour) and after (dark blue colour) quality control.

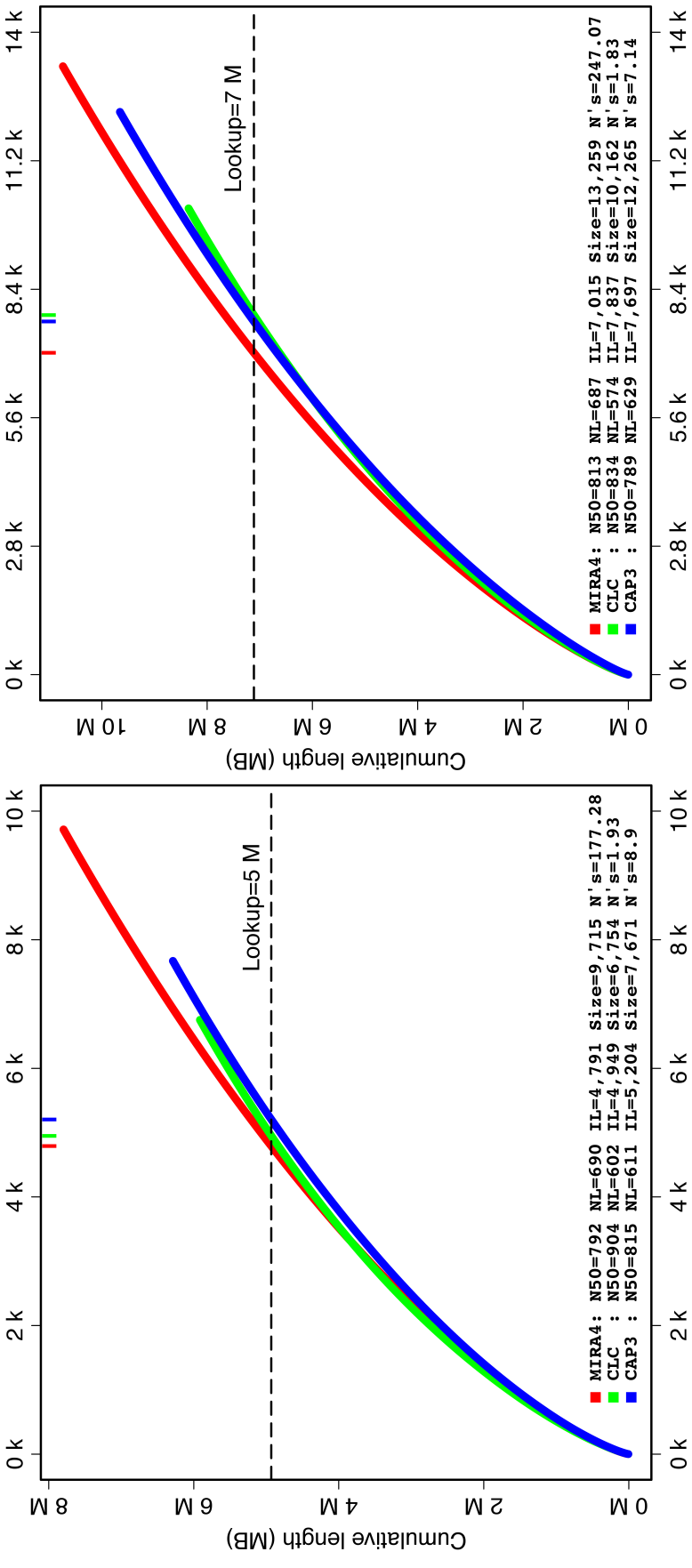
3.3.2 *De novo* assembly of transcriptome

To achieve a higher quality assembly, the cleaned reads from both libraries were assembled separately using different assemblers such as CLC, MIRA and CAP3 with the transcript length at least 150 bp. The comparison of assembly statistics helped to choose the best assembly based on N50 length (the

transcript size N in which half the assembly is represented in transcripts longer than N) of transcripts and low number of ambiguous bases (N's). For UL library, CLC had assembled 11,396 transcripts and 38,229 singletons, MIRA assembly had produced 15,937 transcripts and 121,735 singletons, and 12,883 transcripts and 29,862 singletons were produced by CAP3. Similarly for NL library, 24,341 transcripts and 68,784 singletons were produced by CLC, MIRA had assembled 28,699 transcripts and 87,515 singletons, and CAP3 had generated 27,871 transcripts and 78,469 singletons (Table 3.1). The CLC assembly was selected as the best assembly for both libraries because it had longer N50 transcripts, 904 in UL and 834 in NL, and the lowest number of ambiguous (N's) bases, 1.93 (per 100 kbp) in UL and 1.83 (per 100 kbp) in NL libraries (Figure 3.2). CLC assemblies of cleaned reads had a higher number of contigs from NL library compared to UL library (Table 3.1). The assembled contigs from NL library were also shorter with maximum contig length and an average contig length compared to UL library. These results were found similar to the assemblies generated from UL and NL cDNA in milkweed bug (Ewen-Campen et al., 2011).

		UL library	NL library
Sequencing			
	Number of reads	301,094	563,662
	Total bases (Mbp)	101.21	120.33
	Min. read length (bp)	40	18
	Max. read length (bp)	1004	760
	Mean read length (bp)	336.16	213.49
	GC content	37.45%	35.73%
	ESTclean	276,861	387,833
Assembly			
CLC	Number of transcripts	11,396	24,341
	Total bases (Mbp)	7.67	13.08
	Min. transcript length (bp)	150	150
	Max. transcript length (bp)	6,430	5,891
	GC content (>= 500 bp)	37.41%	36.06%
	N50 length (bp) (>= 500 bp)	904	834
	N50 transcripts (>= 500 bp)	2,239	3,530
	N's per 100 kbp (>= 500 bp)	1.93	1.83
MIRA	Number of transcripts	15,937	28,699
	Total bases (Mbp)	10.18	16.08
	Min. transcript length (bp)	150	150
	Max. transcript length (bp)	6,687	5,541
	GC content (>= 500 bp)	37.48%	36.08%
	N50 length (bp) (>= 500 bp)	792	813
	N50 transcripts (>= 500 bp)	3,391	4,675
	N's per 100 kbp (>= 500 bp)	177.28	247.07
CAP3	Number of transcripts	12,883	27,871
	Total bases (Mbp)	8.29	15.08
	Min. transcript length (bp)	150	150
	Max. transcript length (bp)	6,907	3,686
	GC content (>= 500 bp)	37.57%	35.99%
	N50 length (bp) (>= 500 bp)	815	789
	N50 transcripts (>= 500 bp)	2,657	4,443
	N's per 100 kbp (>= 500 bp)	8.90	7.14

TABLE 3.1: Sequencing and assembly summary for both libraries. The statistics are based on all transcripts (size >= 0 bp) unless stated size (>= 500 bp).



(A) UL library

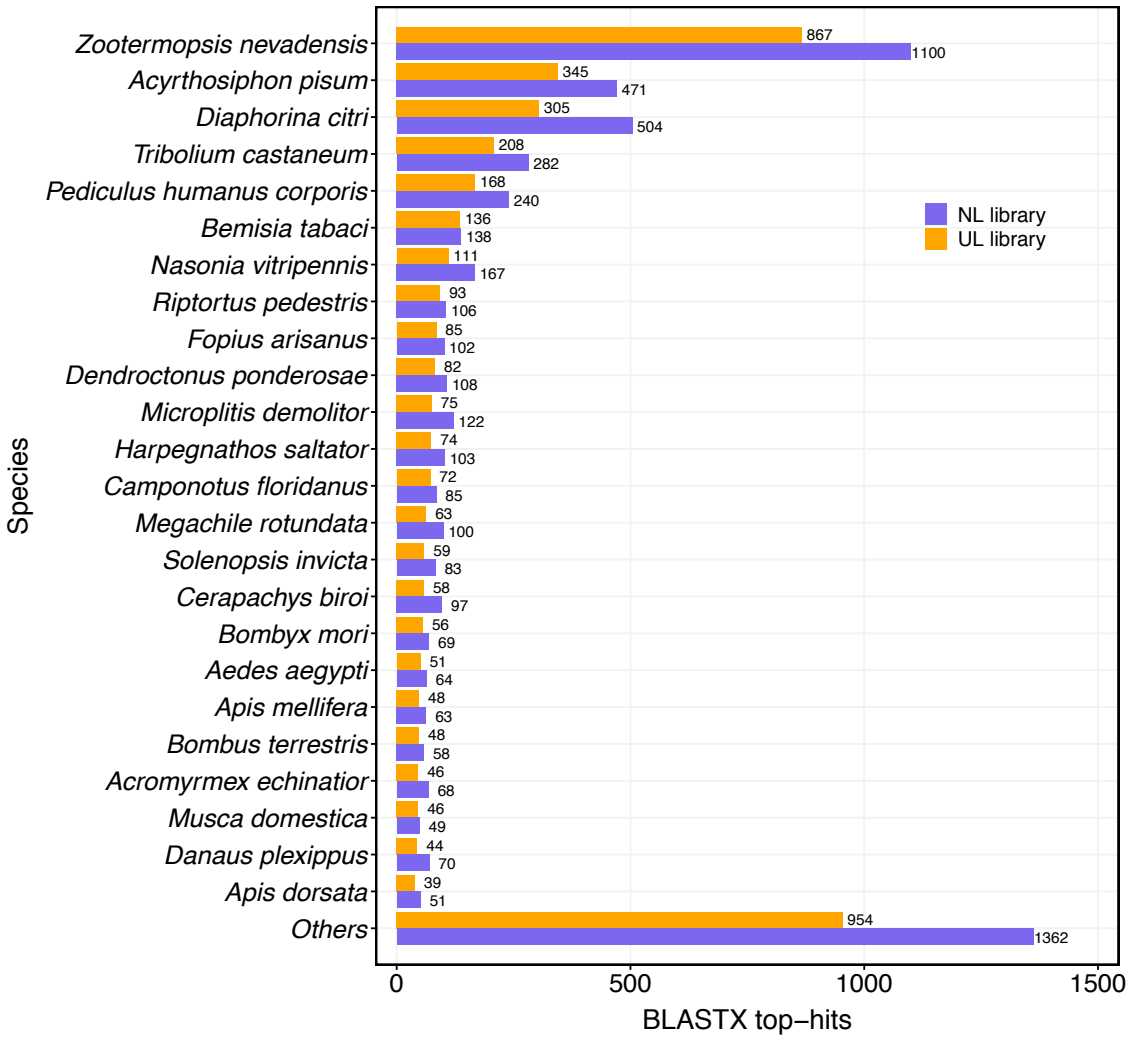
(B) NL library

FIGURE 3.2: Cumulative lengths of assembled transcripts from different assemblers. (A) Assembled transcripts from UL library, (B) Assembled transcripts from NL library. Transcripts with size >= 500 bp were selected from three assemblies and plotted according to length from lower to higher. The "lookup" value represents 75% of the harmonic mean of the total sequence sizes.

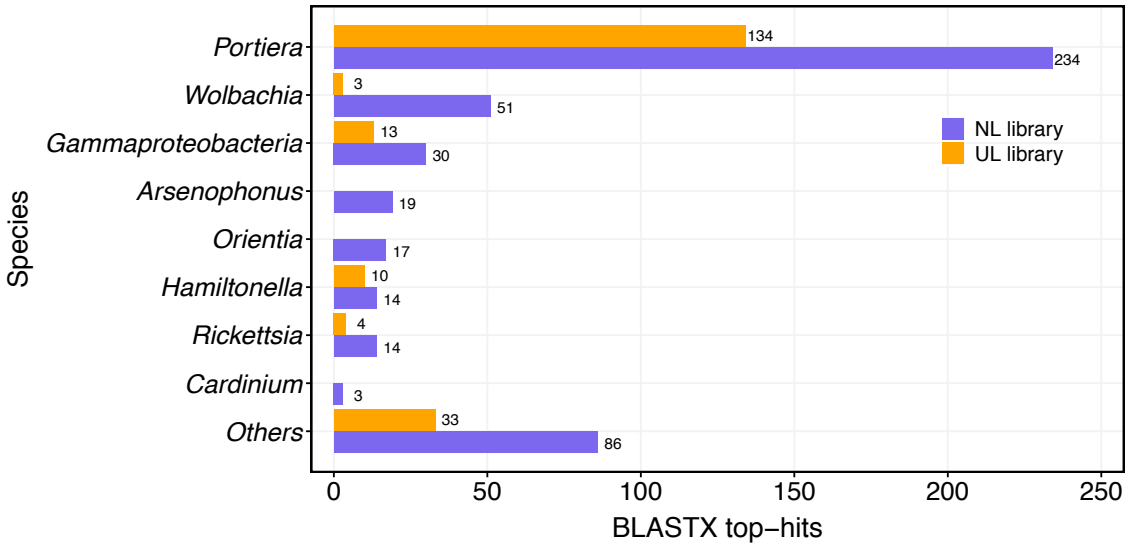
3.3.3 Functional annotation of Asia I transcripts

To determine the putative function, transcripts were subjected to the BLAST homology search at nr protein database using E-value cut-off of $1E^{-05}$. Of the total 11,396 transcripts from UL library, 4,330 (38.99%) gave a significant hit, corresponding to 4,020 unique proteins in the nr database (Appendix A, Table A3.1). Similarly, 6,129 transcripts (25.17%) from NL library showed a significant hit, corresponding to 5,565 unique proteins (Appendix A, Table A3.2). Because of lack of whitefly genome information, a relatively large proportion of assembled transcripts (61.01%: UL library, 74.83%: NL library) could not be aligned to known proteins at the nr database. The proportion of aligned transcripts (38.99%, UL library) to the annotated sequences at the nr database is higher than that of published whitefly transcriptomes including MEAM1 (16.95%) (Leshkowitz et al., 2006), MED (16.20%) (Wang et al., 2010a), MEAM1 (27.60%) (Wang et al., 2011), Asia II 3 (31.60%) (Wang et al., 2012) and MEAM1 species (28.50%) (Xie et al., 2012), and greenhouse whitefly (36.28%) (Karatolos et al., 2011). The taxonomic distribution of species with the top BLASTX hits from both libraries is shown in Figure 3.3. The number of top-hits in all cases are higher for NL than UL library library as would be expected with the NL library revealing more rarer transcripts.

The termite, *Z. nevadensis*, showed the highest number of top-hits (867 hits UL library, 1100 hits NL library) followed by other species of the kingdom Eukaryota Figure 3.3A. These top-hits were higher than two times the top-hits from *A. pisum* and *D. citri*. These results were found surprising as these two species (*Z. nevadensis* and *D. citri*) were not reported in previously published transcriptomes of *B. tabaci* with such a high number of top-hits (Wang et al., 2010a, 2011, 2012; Xie et al., 2012). This was because their genomes were only recently sequenced and annotated (Hunter et al., 2014; Terrapon et al., 2014). There were 136 and 138 top-hits found for *B. tabaci* from UL and NL libraries respectively, which were higher than the previously published transcriptomes MEAM1 (97 hits) (Wang et al., 2011), MED (124 hits) (Wang et al., 2010a), and Asia II 3 species (94 hits) (Wang et al., 2012). The results are surprising as *B. tabaci* and *Z. nevadensis* belong to different Orders (Hemiptera and Isoptera respectively) and do not share common characteristics like feeding behaviour and developmental life cycle. Of the total top-hits for *Z. nevadensis*, there were 656 hits and 805 hits found with percentage identity greater than 50% from UL and NL libraries respectively. The highly conserved genes shared between *B. tabaci* and *Z. nevadensis* (from 90% to 100% similarity) include actin 5C, serine/threonine-protein phosphatase, transitional endoplasmic reticulum ATPase, inosine-5'-monophosphate dehydrogenase 1, casein kinase II subunit alpha, eukaryotic translation initiation factor 2 and GTP-binding protein 128up (see Appendix A, Table A3.1 and Table A3.2 for more details). These results are considered to be erroneous hits and further analysis is required to reveal why they are revealed as positive top-hits based on their sequence identity and alignment coverage between *B. tabaci* and *Z. nevadensis* proteins.



(A) Eukaryota



(B) Bacteria

FIGURE 3.3: Taxonomic distribution of species based on BLASTX top-hits retrieved for each transcript (from both UL and NL libraries) with E-value above 10^{-05}. (A) Eukaryota and (B) Bacteria. The horizontal bars represent number of hits found in each species (on y-axis).

The BLASTX top-hits also contained homolog sequences from several bacteria from both libraries. Seven bacterial species were identified based on sequence homology. These included, from highest to lowest number of top-hits, *Portiera*, *Wolbachia*, *Arsenophonus*, *Orientia*, *Hamiltonella*, *Rickettsia* and *Cardinium* (Figure 3.3B). These results suggest the infection of particular bacterial endosymbiont in the Asia I population studied, but for low hit results such as *Cardinium* (n=3 in NL only), the hits may be false-positives.

To date, eight endosymbionts have been reported to infect *B. tabaci* (Baumann, 2005; Chiel et al., 2007; Bing et al., 2013a) of which seven are suggested to have high homology sequences present in the Asia I transcriptomes. There was no sequence homology found with *Fritschea* which suggests the absence of *Fritschea* inside our Asia I population. Absence of *Fritschea* was also reported by Bing et al. (2013b) in four native species (Asia II 3, Asia II 1, China 1, Asia II 7) and two invasive species (MEAM1, MED) of the *B. tabaci* cryptic complex.

Singletons from both libraries were also searched against the nr protein database using E-value cut-off of $1E^{-05}$. Of the total 38,229 singletons from UL library, 7,292 (19.07%) had a BLASTX hit and 12,297 (17.87%) hits of the total 68,784 singletons from NL library. Singletons had the lower number of BLAST hits compared to transcripts and the alignment length was also found very small, and therefore was not considered in further analyses. Transcripts (UL library: 7,066 transcripts, NL library: 18,212 transcripts) which did not have homology against nr database were further searched against TSA database with search restricted to *B. tabaci* sequences only. More than 94% (6,687 transcripts) and 87% (15,935 transcripts) of these transcripts were found to have a homolog sequence in TSA database. These results suggest that the majority of assembled transcripts were not assigned a biological function because of the very limited transcriptomic information available. However, it is most likely that the ongoing insect genome projects including those on the *B. tabaci* genomes will facilitate future annotation of such transcripts.

3.3.4 Gene Ontology classification and pathways

To classify the functions of annotated *B. tabaci* transcripts, Gene Ontology (GO) terms were assigned to each BLASTX result (Appendix A, Table A3.3 and Table A3.4). According to the BLASTX sequence homology and Blast2GO mapping strategy, only 3,193 transcripts (UL library) were assigned GO terms and grouped into three main categories: biological process (19 subcategories), molecular function (16 subcategories) and cellular component (11 subcategories). Similarly, 4,419 transcripts (NL library) were also categorized into biological process (18 subcategories), molecular function (15 subcategories) and cellular component (12 subcategories) (Figure 3.4).

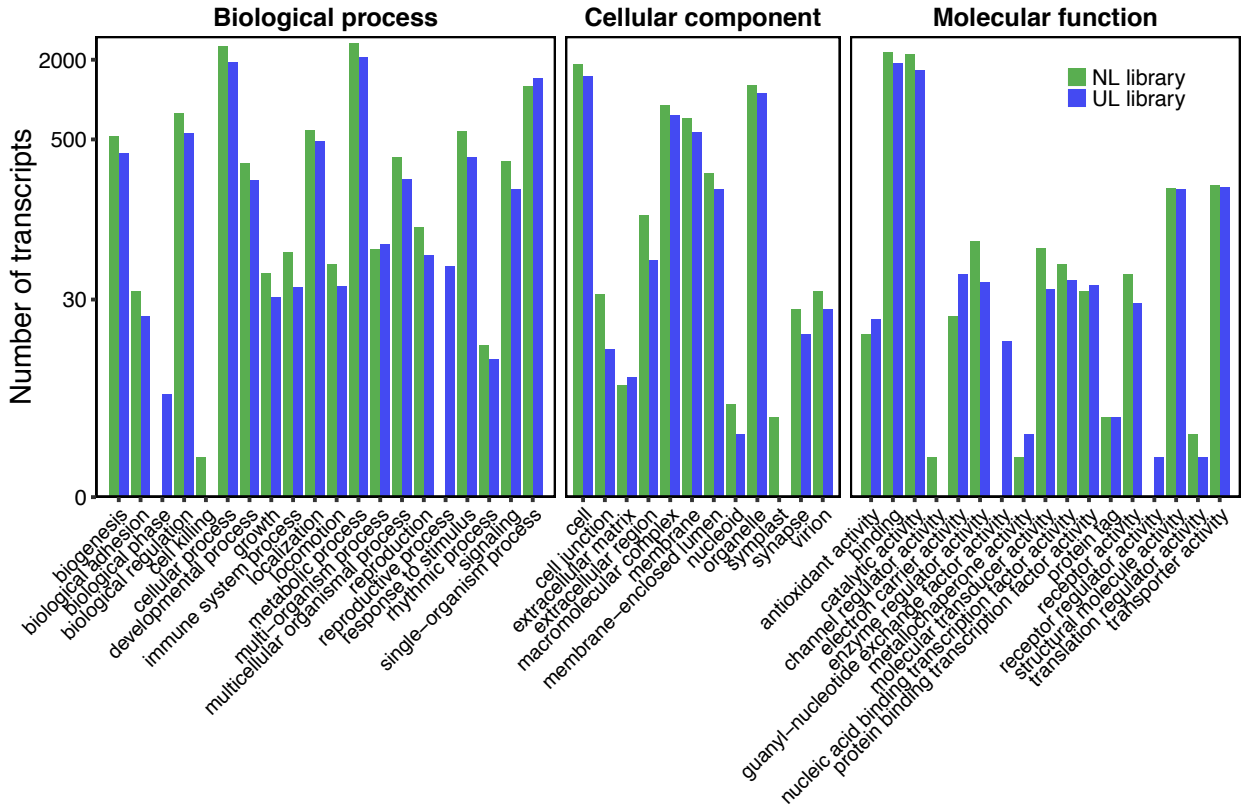


FIGURE 3.4: Gene Ontology classification of UL and NL libraries. The GO terms were classified into three main categories: biological process, cellular component and molecular function. The x-axis represents subcategory and y-axis shows number of transcripts belonging to a category.

As shown in Figure 3.4, the genes involved in biological process were highly represented (UL library: 8,296 transcripts, NL library: 10,360 transcripts) followed by cellular component (UL library: 4,255 transcripts, NL library: 5,261 transcripts) and molecular function (UL library: 4,237 transcripts, NL library: 5,249 transcripts) in both libraries. "Metabolic process" was found as the major subcategory that comprised 25.15% (UL library) and 25.65% (NL library) of the genes involved in biological process. "Cell" (35.01% UL library, 34.97% NL library) and "binding" (44.27% UL library, 43.15% NL library) were the most abundance subcategories associated with cellular component and molecular function respectively. These results are consistent with the transcriptomes of MED, MEAM1 and Asia II 3 species, where "metabolic process", "cell" and "binding" were the most abundant subcategories (Wang et al., 2010a, 2011, 2012). Genes associated with "cell killing" (biological process), "symplast" (cellular component), and "channel regulator activity" (molecular function) were not found in the UL library. Similarly, genes involved in "biological phase", "reproductive process" (biological process) and "guanyl-nucleotide exchange factor activity", "receptor regulator activity" (molecular function) were only found in the NL library Figure 3.4.

To assess the diversity of gene functions between *B. tabaci* species, GO annotations of transcriptome assemblies of different *B. tabaci* species including MEAM1, MED, Asia II 3 and Asia I were compared, resulting in similar distribution of gene functions over three main categories (Appendix B, Figure B3.1). This may suggest that the functions of genes are highly conserved across different *B. tabaci* species and not affected by bias in the construction of sequencing libraries. However, in most subcategories, genes were highly represented in MEAM1, MED and Asia II 3 species compared to Asia I species (Appendix B, Figure B3.1). This is probably because of the much larger amount of sequencing data produced for these three species: MEAM1 (1.27 GB) (Wang et al., 2011), MED (3.27 GB) (Wang et al., 2010a) and Asia II 3 (1.24 GB) (Wang et al., 2012). Interestingly, genes associated with biological process: "biological phase", "reproductive process", "single-organism process", cellular component: "nucleoid", "symplast" and molecular function: "electron carrier activity", "guanyl-nucleotide exchange factor activity", "protein tag" were not found in MEAM1, MED (except "electron carrier activity") and Asia II 3 species but they were present in Asia I species (Appendix B, Figure B3.1). However, another set of genes involved in biological process: "cell proliferation", "death", "pigmentation" and "viral reproduction" were missing in Asia I species but were present in MEAM1, MED and Asia II 3 (Appendix B, Figure B3.1). These results suggest two possibilities: (1) incomplete or stage specific transcriptome sequencing (2) incomplete transcriptome assembly in MEAM1, MED, Asia II 3 and Asia I species. In this study, transcriptome was sequenced and analyzed from adult Asia I females. In contrast, transcriptomes of MED, MEAM1 and Asia II 3 species were sequenced from four different developmental stages: egg and nymph, pupa and adult (Wang et al., 2010a, 2011, 2012).

KEGG pathway analysis was carried out on all assembled transcripts to identify the active pathways represented in the Asia I transcriptomes. From the UL library, a total of 11,396 distinct transcripts were mapped against the KEGG database. Of the total mapped transcripts, 540 transcripts were assigned to 108 unique KEGG pathways belonging to four distinct pathway maps including "metabolism", "genetic information processing", "environmental information processing" and "organismal systems" (Appendix A, Table A3.5). Similarly, 776 transcripts were assigned to 117 KEGG pathways from the NL library (Appendix A, Table A3.6). There was no transcript found for remaining pathway maps including "cellular processes", "human diseases" and "drug development" from either the UL or NL library. Among the KEGG pathways, 'purine metabolism' was highly represented by 200 transcripts followed by 'biosynthesis of antibiotics" (166 transcripts), 'pyrimidine metabolism' (113 transcripts) and 'oxidative phosphorylation' (71 transcripts) for the NL library. However, 'biosynthesis of antibiotics" were highly represented in the UL library (124 transcripts) followed by 'purine metabolism' (113 transcripts), 'oxidative phosphorylation' (71 transcripts) and 'pyrimidine metabolism' (64 transcripts). These results were different to annotated pathways in Asia II 3 as the highly represented pathway was 'starch and sucrose metabolism' (553 transcripts) followed by 'purine metabolism' (458 transcripts) and 'galactose metabolism' (183 transcripts) (Wang et al., 2012). The pathways including 'starch and sucrose metabolism' (17

transcripts, 16 transcripts) and 'galactose metabolism' (12 transcripts, 8 transcripts) were lowest represented in Asia I (both NL and UL libraries), which suggests either differences in the number of genes involved in metabolic pathways across *B. tabaci* species, or more likely is a reflection of the transcriptomes coming from different insect populations of different species that were under different conditions and so expressed different genes.

Enzyme classification of annotated transcripts shows that the transferases group of enzymes were highly represented enzymes associated with different pathways (n=74, n=87) of Asia I species followed by oxidoreductases (n=72, n=81), lyases (n=47, n=51), hydrolases (n=41, n=46), ligases (n=23, n=27) and isomerases (n=13, n=18) from the UL and NL libraries respectively (Table 3.2). The largest proportion of Asia I species enzymes were involved in the KEGG pathways "biosynthesis of antibiotics" and "aminoacyl-tRNA biosynthesis". It was interesting that the KEGG pathway "Biosynthesis of antibiotics" was never seen in previously published transcriptomes of MED, MEAM1 and Asia II 3 species (Wang et al., 2010a, 2011, 2012). Together the results of GO annotations and KEGG pathway annotations provide whitefly research a significant resource to understand the functions of essential genes and their involvement in metabolic pathways across different species of the *B. tabaci* complex.

	UL library		**NL library**		
Enzyme	Contigs	Pathways	Contigs	Pathways	Most represented pathway
Oxidoreductases	165	72	195	81	Biosynthesis of antibiotics
Transferases	158	74	252	87	Biosynthesis of antibiotics
Hydrolases	133	41	209	46	Biosynthesis of antibiotics
Lyases	45	47	61	51	Biosynthesis of antibiotics
Isomerases	10	13	19	18	Biosynthesis of antibiotics
Ligases	65	23	94	27	Aminoacyl t-RNA biosynthesis

TABLE 3.2: Classification of annotated enzymes into six main classes: oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. Total number of transcripts encoding enzymes and their associated pathways are also listed along with the KEGG pathway contain highest number of enzymes.

3.3.5 Prediction of transcriptome completeness and full-length cDNA

To estimate the completeness of the assembled Asia I transcriptomes, assembled transcripts from two libraries were clustered into three categories: NL-specific, UL-specific and shared clusters. Based on a 95% sequence similarity cut-off, 4,806 transcripts from UL and 5,415 from NL libraries were clustered into 4,442 unique clusters. There were 6,492 and 18,484 clusters found unique in UL and NL libraries respectively.

For the purpose of getting complete transcripts, "ortholog hit ratio" was calculated as described in O'Neil et al. (2010) by dividing the length of the putative coding region of the transcript by the total length of the corresponding ortholog. Each transcript and the corresponding BLASTX top-hits were considered orthologs and the aligned region of transcript was considered to be a conservative estimator of 'putative coding region'. The "ortholog hit ratio" gives an estimate of the number of transcripts represented by each transcript (Belleghem et al., 2012; Xu et al., 2013). Figure 3.5 shows that a higher number of transcripts fall below "ortholog hit ratio" 0.5 which suggest that many of the large number of assembled transcripts were not full-length or aligned completely to their corresponding ortholog. It is also true for both assemblies that the completeness of the transcripts decreases for longer transcripts. However, numerous transcripts with an "ortholog hit ratio" 1.0 suggest the completeness of some transcripts. An "ortholog hit ratio" >1 indicates insertions in transcripts and a few of these are visible (Figure 3.5). Of the total annotated 4,335 transcripts of the UL library, 533 had >0.9 ratio and 1,548 had >0.5. Similarly, 495 transcripts had >0.9 ratio and 1,695 had >0.5 from 6,171 annotated transcripts of the NL library.

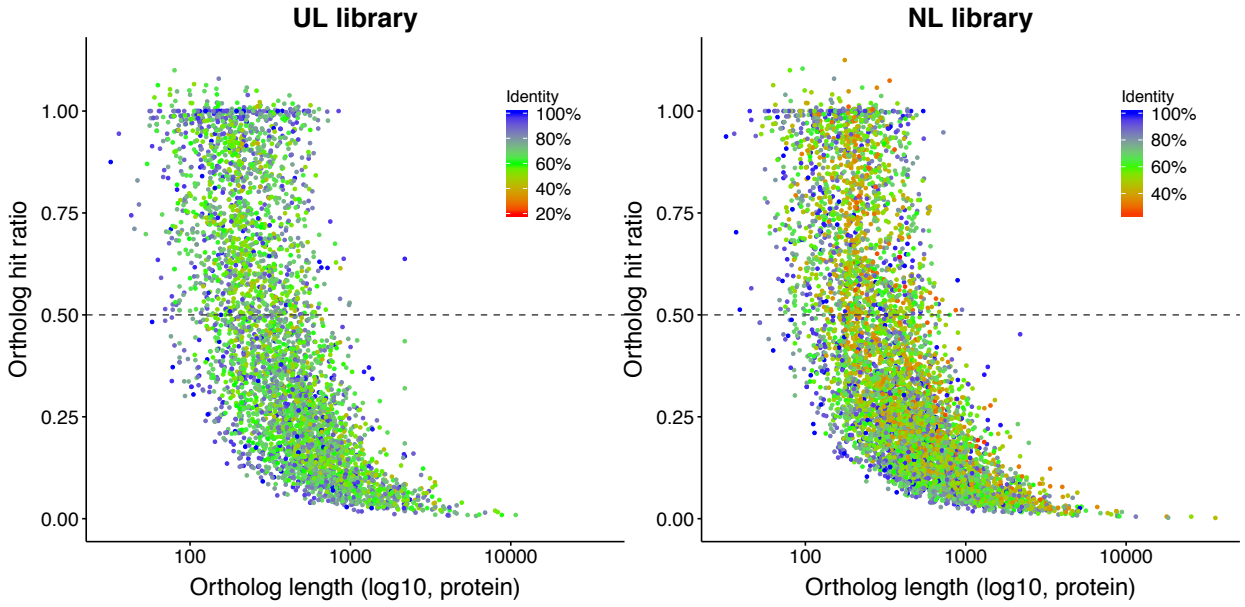


FIGURE 3.5: Distribution of "Ortholog hit ratio" for both libraries. The ratio was calculated for each transcript aligned at nr database with E-value above $1E^{-05}$. A ratio of 1.0 indicates the transcript is likely to be fully assembled whereas ratio >1.0 suggest insertions in transcripts.

Further, the completeness of transcripts was assessed based on the Open Reading Frame (ORF) of all annotated transcripts. The full-length transcripts are considered to be an important resource for many genetic and genomic studies by providing preliminary information of the putative coding genes. To identify the full-length transcripts encoding a complete ORF from the assembled transcriptome, all annotated transcripts were analyzed using TargetIdentifier. From the total 4,335 annotated transcripts with E-value cut-off of $1E^{-05}$ (UL library), 1,122 were identified as full-length with

complete ORF, length of ORF ranging from 90 bp to 2,532 bp and 161 as short full-length with size from 102 bp to 1,581 bp (Figure 3.6). Similarly, 1,251 were identified as full-length with ORF length from 45 bp to 1,920 bp from 6,171 annotated transcripts of NL library. Figure 3.6 also shows the highest number of identified full-length transcripts that were shorter than 750 bp, which suggest that the longer transcripts were not easy to assemble to full-length using current sequencing data.

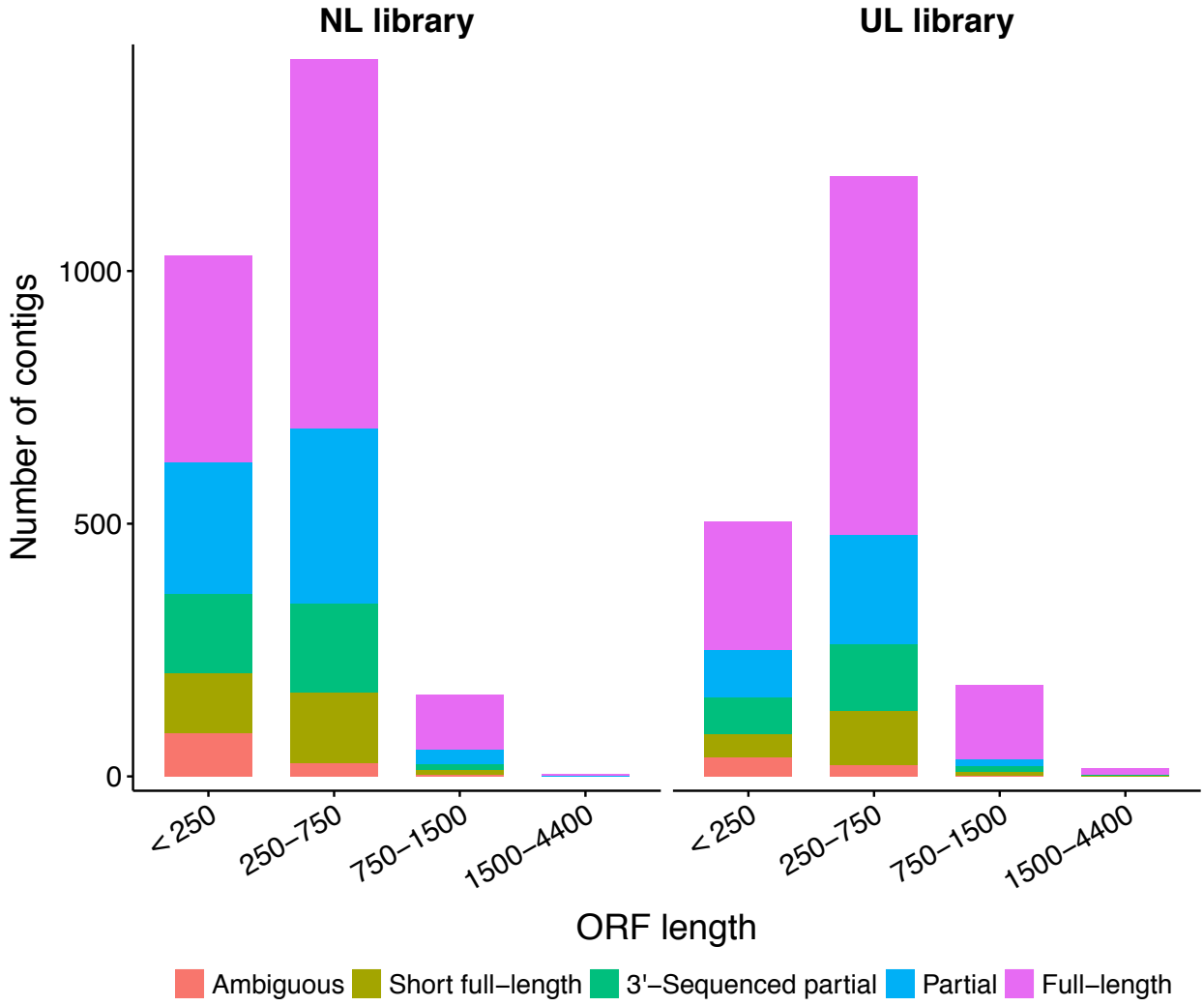
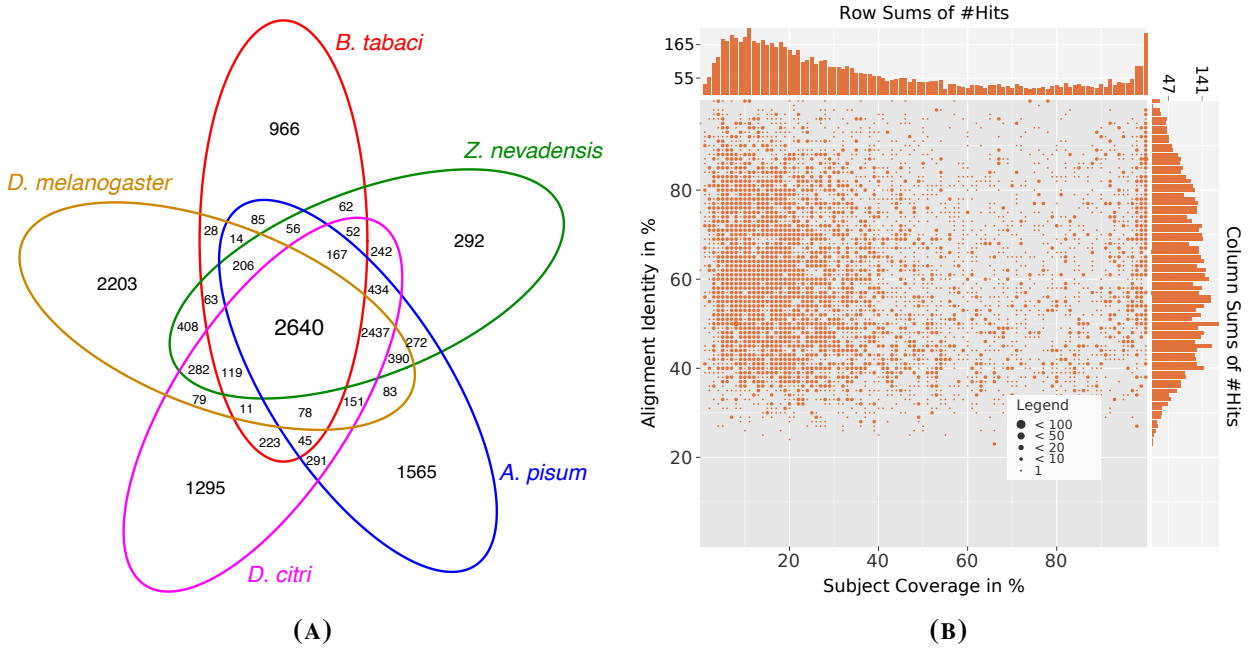


FIGURE 3.6: Length distribution of assembled transcripts ORF and grouped into five categories according to their completeness.

3.3.6 Protein orthologs analysis

Ortholog analysis of proteins from related species may identify protein families that are conserved within or between species. Best reciprocal BLAST hits between proteins gives paralogs if they are within species and orthologs if they are from different species. This approach is widely used to identify ortholog pairs between and within species (Moreno-Hagelsieb and Latimer, 2008). Here, protein orthologs were identified between *B. tabaci*, *A. pisum*, *D. melanogaster*, *Z. nevadensis*

and *D. citri*. There were 2,640 'Single Copy Orthologs' (SCO) found present in all five species referred as 'core orthologs' (Figure 3.7A). For *B. tabaci*, the highest number of SCO were found in *D. citri* (n=223) followed by *A. pisum* (n=85), *Z. nevadensis* (n=62) and *D. melanogaster* (n=28). These findings do not support the previous BLAST result where *Z. nevadensis* showed the highest proportion of top-hits compared to *A. pisum* and *D. citri*. These results indicate a possibility of false positive BLAST top-hits with partially aligned proteins and higher percentage identity between *B. tabaci* and *Z. nevadensis*.



(A) **(B)**

FIGURE 3.7: Identification of protein orthologs between *B. tabaci*, *A. pisum*, *D. melanogaster*, *Z. nevadensis* and *D. citri*. (A). The Venn diagram illustrates distinct and shared orthologs from the OrthoMCL analysis between *B. tabaci*, *A. pisum*, *D. melanogaster*, *Z. nevadensis* and *D. citri*. The numbers represent SCO from each species. (B). The alignment coverage and percentage identity of *B. tabaci* proteins (query) were plotted against their orthologs from *Z. nevadensis* (subject).

High conservation of SCO should have high coherence for protein lengths across species. 'Ortholog hit ratio' of each BLASTX hit from *Z. nevadensis* was also calculated to confirm the alignment coverage against *B. tabaci* proteins. BLASTX hits from *Z. nevadensis* were selected from both libraries (UL and NL) and corresponding proteins were predicted for *B. tabaci*. The alignment coverage and percentage identity of protein orthologs were compared between *B. tabaci* and *Z. nevadensis*. The majority of *B. tabaci* proteins were aligned at higher percentage identity ranges from 50%-100% but had lower alignment coverage (1%-40%) with *Z. nevadensis* proteins (Figure 3.7B). This finding indicates that the majority of *B. tabaci* proteins were incomplete or fragmented as they did not align to *Z. nevadensis* proteins in their complete length or at least 50% length and therefore should not be considered as true orthologs.

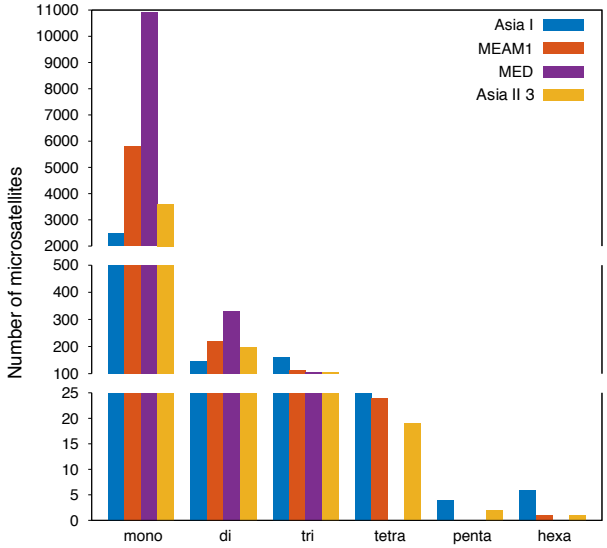
3.3.7 Identification of repetitive elements

The proportion of microsatellites and repetitive elements in the *B. tabaci* transcriptome was assessed by the MISA tool (Sonah et al., 2011) and RepeatMasker. Repeatmasking of all 11,396 (7,671,493 bp) assembled transcripts of the UL library resulted in identification of 90,149 bp (1.17%) of repeated sequences. Similarly, 169,414 bp (1.29%) of repeated sequences were identified from 13,081,869 bp of the NL library. The most frequently occurring repetitive elements were simple sequence repeats (SSRs) (0.73%, 0.68%) followed by low complexity (0.3%, 0.46%), retroelements (0.07%, 0.07%) including long terminal repeat (LTR) elements (n=27, n=45), short interspersed nuclear elements (SINEs) (n=1, n=3) and long interspersed elements (LINEs) (n=25, n=54), and DNA transposons (0.01%, 0.01%) from the UL and NL libraries respectively (Figure 3.8A). A total of 11,396 and 24,341 transcripts were examined by MISA resulting in 172 and 295 SSRs in UL and NL libraries respectively (Figure 3.8A). These results are not surprising as the NL library contains more transcripts than the UL library. From the UL library, the most abundant type of SSRs motif was mononucleotide (n=1,086), followed by trinucleotide (n=97), dinucleotide (n=58), tetranucleotide (n=16) and hexanucleotide (n=1). Similarly, mononucleotide (n=2,226) was the most abundant SSRs in the NL library followed by trinucleotide (n=132), dinucleotide (n=131), tetranucleotide (n=20), hexanucleotide (n=7) and pentanucleotide (n=5) (Figure 3.8B). This approach has been widely used to identify markers which have been found to be amplifiable and polymeric in validation studies (Kaur et al., 2011; Parchman et al., 2010). Microsatellites have been identified using this approach in the MEAM1 species and experimentally validated as a potential marker to distinguish species of the *B. tabaci* complex (Wang et al., 2014a), suggesting that this approach is very efficient for identifying molecular markers.

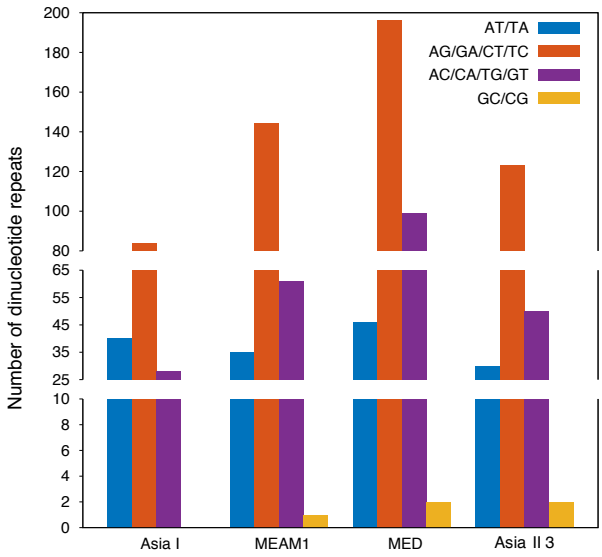
To achieve a complete set of microsatellites from the Asia I transcritpomes, a total of 29,418 transcripts (CD-HIT-EST clusters) were assessed by the MISA tool (Sonah et al., 2011). There were 3,033 microsatellites identified in Asia I species transcripts (Appendix A, Table A3.7), which were lower in numbers than previously reported microsatellites in MEAM1 (n=6,419), MED (n=11,711) and Asia II 3 species (n=4,115) (Wang et al., 2014a). Among six characterized microsatellites, mononucleotides were the most common, followed by tri-, di-, tetra-, hexa- and pentanucleotide repeats (Figure 3.8B). There were four microsatellites including tri-, tetra-, penta- and hexanucleotide the have more repeats in the Asia I transcritpome data than in equivalent MEAM1, MED and Asia II 3 data (Figure 3.8B). Based on six translation frames and complementary strand, (AT)n reads the same as (TA)n, (GC)n and (CG)n, (ATG)n and (TGA)n, (CAT)n and (ATC)n, and therefore mono-, di- and trinucleotide repeats can be grouped into 2, 4 and 10 unique classes respectively (Jurka and Pethiyagoda, 1995). Dinucleotide repeats in Asia I were found much lower than in MEAM1, MED and Asia II 3 (Figure 3.8C). However, the occurrence of trinucleotide was different in Asia I. The four classes including AAG, AAT, AAC and ATG were significantly higher in Asia I data than in MEAM1, MED and Asia II 3 data (Figure 3.8D).

	UL library	NL library
Retroelements	53 (0.07%)	102 (0.07%)
SINEs	1	3
LINEs	25	54
LTR elements	27	45
Retroviral	0	0
DNA transposons	12 (0.01%)	19 (0.01%)
Simple repeats	1,243 (0.73%)	2,018 (0.68%)
Low complexity	460 (0.3%)	876 (0.46%)
Microsatellites	172	295
Di-nucleotide	58	131
Tri-nucleotide	97	132
Tetra-nucleotide	16	20
Penta-nucleotide	0	5
Hexa-nucleotide	1	7

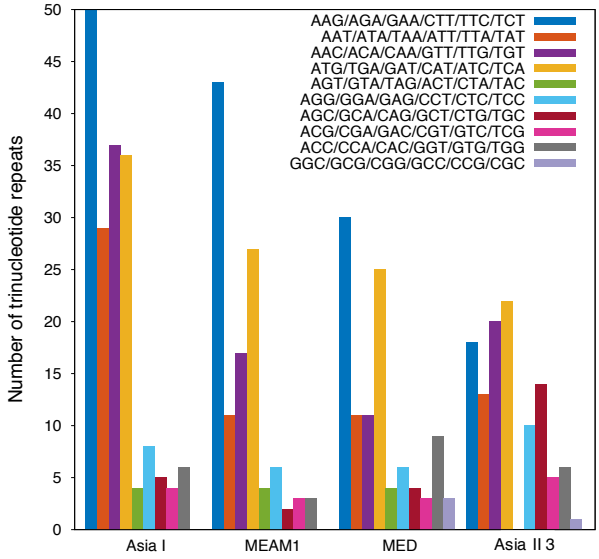
(A)



(B)



(C)



(D)

FIGURE 3.8: Distribution of microsatellite repeats in Asia I, MEAM1, MED and Asia II 3 species. (A) Summary of repetitive elements and microsatellites in the transcriptome of Asia I species. (B) Distribution of microsatellite repeats (C) Distribution of dinucleotide repeats. (D) Distribution of trinucleotide repeats.

3.3.8 Molecular variation across *B. tabaci* species

A number of full-length genes were selected to perform phylogenetic analysis across *B. tabaci* species. For this, orthologs of full-length genes such as CYP6CM1 and Hsp90 were identified in other *B. tabaci* species and insects. The complete coding sequence of CYP6CM1 gene was compared to reveal sequence diversity across *B. tabaci* species MEAM1, MED and Asia I. CYP6CM1, a typical cytochrome P450, associated with resistance to the imidacloprid in MEAM1 and MED

species (Karunker et al., 2008) has revealed number of amino acid substitutions in Asia I species compared to MEAM1 and MED species (Figure 3.9). However, the "heme-iron ligand" motif ("FGDGPRLCIA") was found conserved across three species (Figure 3.9), which confirms their involvement in oxidative metabolism of natural compounds as well as drugs (Guengerich, 2008; McLean et al., 2012).

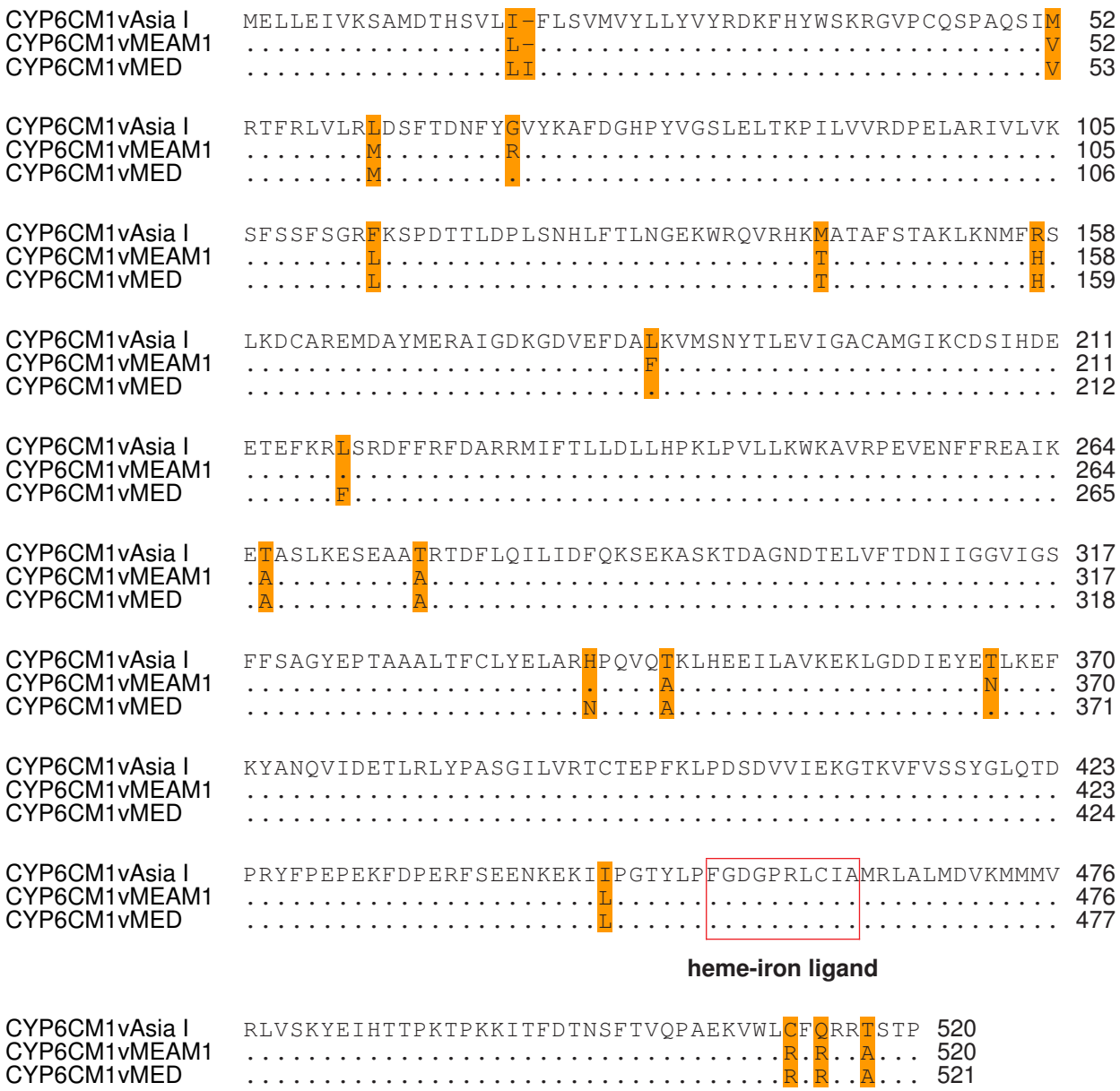
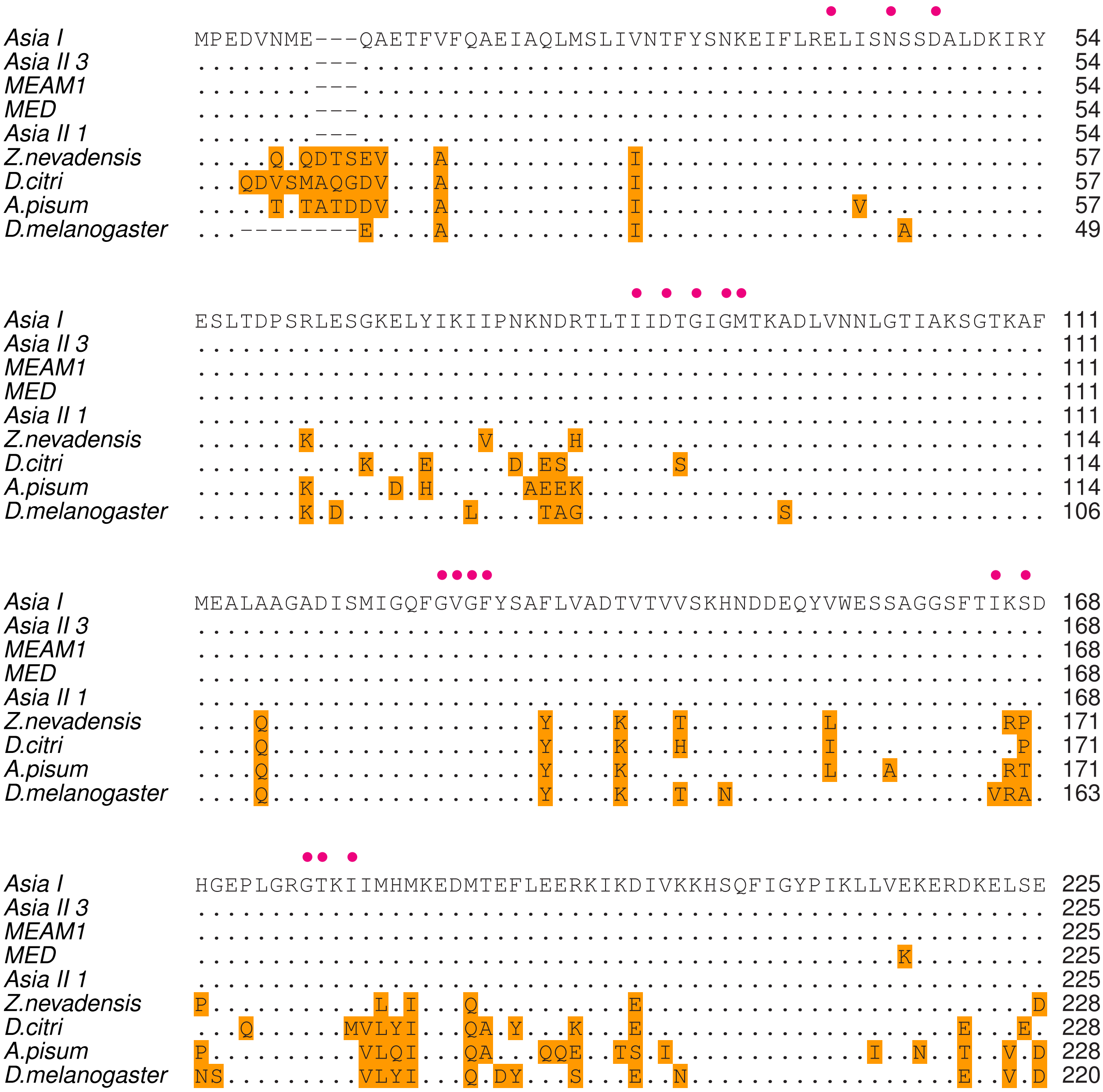
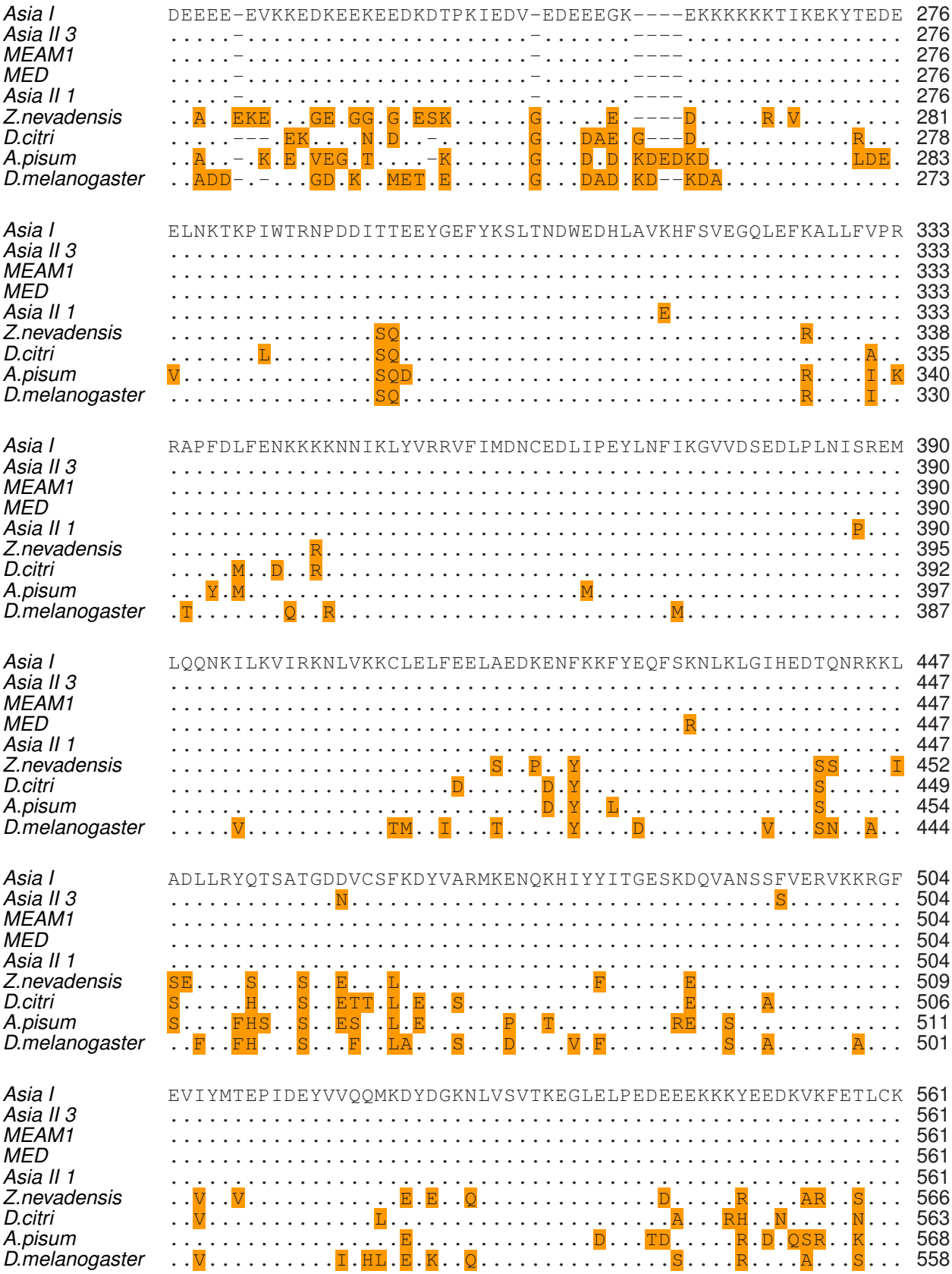


FIGURE 3.9: Multiple sequence alignment of CYP6CM1 orthologs across *B. tabaci* species MEAM1 (CYP6CM1vMEAM1), MED (CYP6CM1vMED) and Asia I (CYP6CM1vAsia I). The identical residues are presented as dots whereas non identical residues are highlighted as orange shaded blocks. The deletion mutation in Asia I and MEAM1 species is presented as dash (-) at 19 position. The "heme-iron ligand" motif ("FGDGPRLCIA") is highlighted in red box.

The comparison of heat shock family protein such as Hsp90 revealed a very high degree of conservation across *B. tabaci* species and also conserved domains across other insect families as represented by *Z. nevadensis*, *D. citri*, *A. pisum* and *D. melanogaster* (Figure 3.10). Hsp90 is a highly conserved essential protein from eukaryote to prokaryote except Archaea (Genest et al., 2011). The essential function of Hsp90 can be altered by inhibition of ATP binding site and therefore it is a well-known promising target for drugs (Donnelly and Blagg, 2008). The Hsp90 protein can be divided in to three domains, namely N-terminus, a middle domain and a C-terminus domain. The N-terminus of Hsp90 is the ATP-binding domain found highly conserved across orthologs of *B. tabaci* species but variations were found across orthologs in *Z. nevadensis*, *D. citri*, *A. pisum* and *D. melanogaster* (Figure 3.10). The ATP-binding sites were highlighted as red "bullets" (Figure 3.10). The conserved signature motif 'MEEVD' was shared by all orthologs at their C-terminus, which is a primary site for dimerization (Figure 3.10, highlighted in red box).




Asia I DEEEE-EVKKEDKEEKEEDKDTPKIEDV-EDEEEGK----EKKKKKKTIKEKYTEDE 276
Asia II 3 276
MEAM1 276
MED 276
Asia II 1 276
Z.nevadensis 281
D.citri 278
A.pisum 283
D.melanogaster 273
Asia I ELNKTKPIWTRNPDDITTEEYGEFYKSLTNDWEDHLAVKHFSVEGQLEFKALLFVPR 333
Asia II 3 333
MEAM1 333
MED 333
Asia II 1 333
Z.nevadensis 338
D.citri 335
A.pisum 340
D.melanogaster 330
Asia I RAPFDLFENKKKKNNIKLYVRRVFIMDNCEDLIPEYLNFIKGVVDSEDLPLNISREM 390
Asia II 3 390
MEAM1 390
MED 390
Asia II 1 390
Z.nevadensis 395
D.citri 392
A.pisum 397
D.melanogaster 387
Asia I LQQNKILKVIRKNLVKKCLELFEELAEDKENFKKFYEQFSKNLKLGIHEDTQNRKKL 447
Asia II 3 447
MEAM1 447
MED 447
Asia II 1 447
Z.nevadensis 452
D.citri 449
A.pisum 454
D.melanogaster 444
Asia I ADLLRYQTSATGDDVCSFKDYVARMKENQKHIYYITGESKDQVANSSFVERVKKRGF 504
Asia II 3 504
MEAM1 504
MED 504
Asia II 1 504
Z.nevadensis 509
D.citri 506
A.pisum 511
D.melanogaster 501
Asia I EVIYMTEPIDEYVVQQMKDYDGKNLVSVTKEGLELPEDEEEKKKYEEDKVKFETLCK 561
Asia II 3 561
MEAM1 561
MED 561
Asia II 1 561
Z.nevadensis 566
D.citri 563
A.pisum 568
D.melanogaster 558

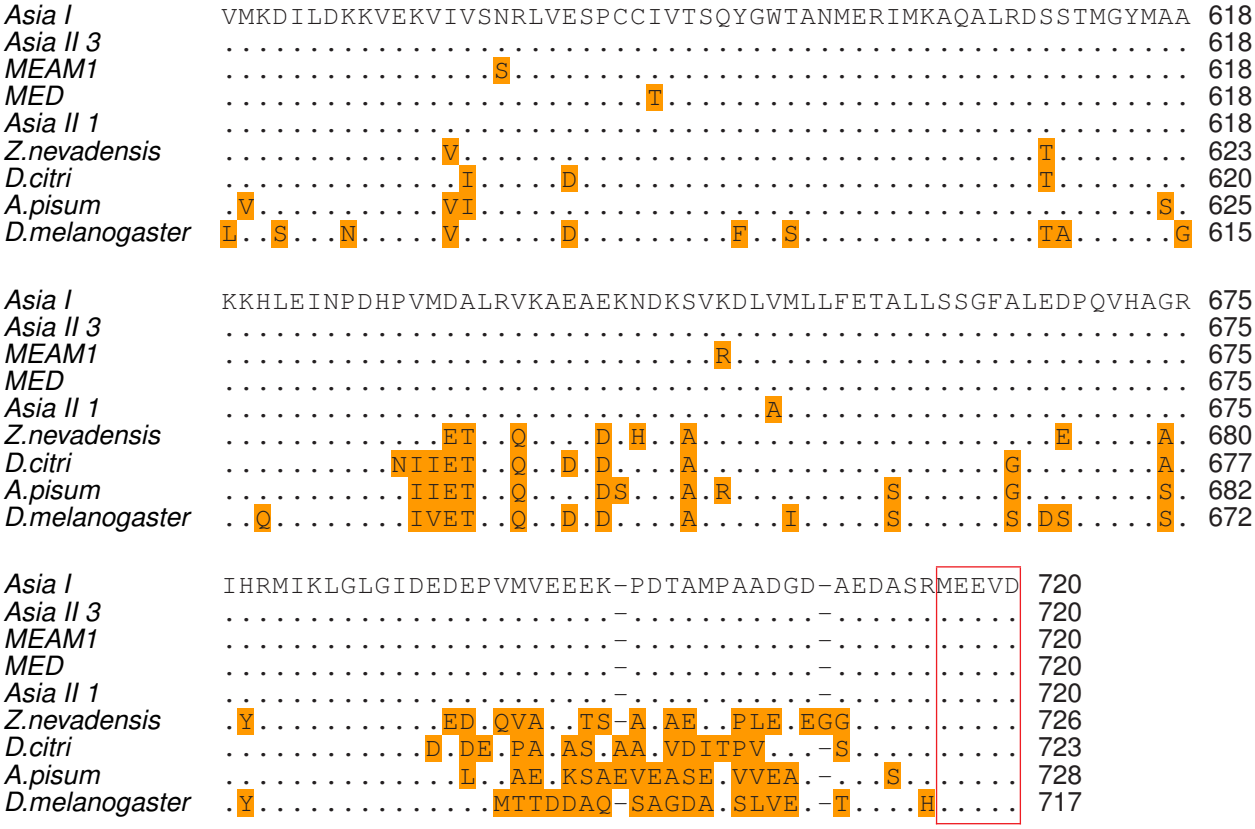


FIGURE 3.10: Multiple sequence alignment of Hsp90 orthologs across *B. tabaci* species MEAM1, MED, Asia II 3, Asia II 1 and Asia I, and *Z. nevadensis*, *D. citri*, *A. pisum* and *D. melanogaster*. The identical residues are presented as dots whereas non identical residues are highlighted as orange shaded blocks. The deletion mutation in Asia I and other species is presented as dash(-). The conserved motif 'MEEVD' at C-terminus is highlighted in red box. The red "bullets" (•) represents ATP-binding sites.

CHAPTER 4

Gene structure of *Bemisia tabaci* cryptic species Asia I

4.1 Introduction

Eukaryotic genome sequencing projects rely heavily upon ESTs and complete complementary DNA (cDNA) sequences for accurate and comprehensive gene discovery and to annotate gene structures using both automated and manual methods (Burge and Karlin, 1997; Zhang, 2002; Brent, 2005). ESTs are a valuable resource for gene discovery in genome projects and combined with the genome sequence, the expressed sequences delineate the gene structures by resolving exons and introns using splice alignment programs (Huang et al., 1997; Bailey et al., 1998). Producing accurate gene predictions is technically challenging for any genome project, as despite the availability of numerous tools there has been no single method that has been found to be able to elucidate it quickly and accurately (Zhang, 2002; Brent, 2005).

Generally, gene prediction programs including GENSCAN (Burge and Karlin, 1997), Fgenesh (Salamov and Solovyev, 2000), Glimmer-HMM (Majoros et al., 2004) and GeneMark.hmm (Lukashin and Borodovsky, 1998) employ probabilistic methods like Hidden Markov Models (HMMs). These HMMs are used to find the likelihood of partitioning a nucleotide into genomic features such as exons, introns and intergenic regions based on a prior set of probabilities. These tools can predict the PCGs in the genome but their prediction is far from a perfect prediction of complete gene structure (Burset and Guigo, 1996; Pavy et al., 1999; Guigo et al., 2000).

The correct gene structures or the individual features of the gene such as exons and introns can be obtained from the spliced alignments of proteins or transcribed sequences against the genome sequence of the same species or a closely related species. There are several tools available to perform splice alignments via aligning transcripts against genome sequence, including Genomic Mapping and Alignment Program (GMAP) (Wu and Watanabe, 2005), BLAT (Kent, 2002), spidey (Wheelan et al., 2001), sim4 (Florea et al., 1998) and gap2 (Huang et al., 1997). Only a few tools

are available to perform spliced alignment of proteins to genome such as 'exonerate' (Slater and Birney, 2005) and Spaln (Gotoh, 2008). Additionally, there are a number of programs available to integrate spliced alignments into gene prediction methods (Birney and Durbin, 2000; Salamov and Solovyev, 2000; Yeh et al., 2001).

In the pilot phase of the genome sequencing project of Asia I, the splice alignment approach was used in order to obtain more robust gene predictions for the Asia I genome. This approach was to provide accurate gene structures for several hundreds of complete genes which then could be used as an 'evidence' to train the *ab initio* gene predictors in a later genome annotation phase of the project. It would also reveal the complexity of gene structures in Asia I species prior to the annotation and this would help to optimise corresponding parameters in the genome annotation pipeline to obtain more accurate gene predictions. This splice alignment approach had already proved very effective in resolving the gene structures and improving the genome annotation in *Arabidopsis* (Haas et al., 2002; Seki et al., 2002). The overall strategy used in this chapter was as follows: selection of full-length transcripts from the transcriptome sequencing (as described in Chapter 3) as a prerequisite for automated, high-precision gene structure annotation via mapping onto the pilot phase genome assembly. Full-length transcripts facilitate the resolution of complex gene structures in Asia I species and also ensures the gene boundaries with the start and stop codons within the genome.

Here in this chapter, the use of a set of 119 transcripts to explore the quality of a draft genome assembly is reported, at the same time observing the structural relationships between these genes and their orthologs from three more closely related hemipterans: *A. pisum*, *Nilaparvata lugens* (Hemiptera: Delphacidae) and *Diaphorina citri* (Hemiptera: Psyllidae), two dipterans: *A. gambiae* and *D. melanogaster*, two hymenopterans: *A. mellifera* and *Nasonia vitripennis* (Hymenoptera: Pteromalidae), one phthirapteran: *Pediculus humanus* (Phthiraptera: Pediculidae) and one coleopteran: *Tribolium castaneum* (Coleoptera: Tenebrionidae) species for which genome sequence already exist. Annotation has permitted us to identify many genome segments of particular biological interest, including those encoding key metabolic, developmental and insecticide target genes. Access to accurate gene models will accelerate population studies and assist in the development of management strategies for this devastating disease vector.

4.2 Methods

4.2.1 Establishment of the Asia I species colony

The Asia I species colony used in this study was collected originally from aubergine (egg-plant) in Coimbatore, South India (Rekha et al., 2005), and reared in the NRI quarantine insectary on cotton, *G. hirsutum* L. cv Laxmi, growing in insect-proof cages ((26+1)oC, 14h:10h L:D, (70+10)% r.h.). The identity of the colony was confirmed as described in Chapter 3 (section methods) by partial *mtCOI* gene sequencing. In order to obtain an inbred line, ten sub-colonies were initiated by allowing single mating pairs of *B. tabaci* (Asia I species) each to colonise a new, *B. tabaci* (Asia I species)-free cotton seedling.

4.2.2 Sequencing and assembly of cDNA

The methodology for sequencing and assembly of representative transcriptome from the whitefly, *B. tabaci* Asia I species, has been described previously and discussed in Chapter 3 (Seal et al., 2012). Full-length transcripts were searched against complete proteomes of *A. pisum*, *N. lugens*, *D. citri*, *D. melanogaster*, *A. gambiae*, *A. mellifera*, *N. vitripennis*, *T. castaneum* and *P. humanus* to identify orthologous genes using Inparanoid v7.0 (Östlund et al., 2010). *Daphnia magna* (Cladocera: Daphniidae) was used as an outgroup species for each Inparanoid run to remove any false ortholog pairs.

4.2.3 Pilot phase: genome sequencing and assembly

DNA was extracted from Asia I species females and males, collected from India and sequenced using an Illumina GAII platform. Additionally, three sequencing lanes were also sequenced using Illumina HiSeq 2000 platform, one for short insert PE library (400 bp) and two for long insert mate-pair (MP) libraries (3 kbp). The MP libraries were sequenced to 100 bp and trimmed 36 bp in upstream direction. CLC genomics workbench (v7.0.4) was used to perform *de novo* assembly of raw reads with minimum contig length restricted to 200 bp. Long insert (3 kbp) MP libraries were used by SSPACE (v3.0) (Boetzer et al., 2011) to construct scaffolds. It was later found that the 98.00% of the total reads from both MP libraries had 200-400 bp insert size instead of 3 kbp insert size. These MP reads can not be used to assemble the scaffolds as they may misassemble or miss-join the contigs. These reads were filtered before scaffolding to only select MP reads with 3 kbp insert size.

4.2.4 Genomic mapping and alignment

Full-length transcripts were mapped onto assembled genome scaffolds using GMAP (v2014-12-18) (Wu and Watanabe, 2005) to identify intron-exon boundaries and possible SNPs. A GMAP database was built from the genome assembly FASTA sequence file using the gmap_build program. Full-length transcripts were then mapped in batches to these genome scaffolds with an alignment option to produce full alignments in multithread mode at 12 cpu to run the GMAP program faster. Results were obtained in GFF3 file format, identifying the gene structure for each mapped full-length transcript.

4.3 Results

4.3.1 Full-length cDNA transcripts

The methodology for producing a representative transcriptome from the whitefly, *B. tabaci* Asia I species, has been described previously (Seal et al., 2012) and was also discussed in Chapter 3. The complete open reading frames of cDNA transcripts were first extracted using 'OrfPredictor' (Min et al., 2005) based on their sequence homology with closest ortholog and alignment coordinates. Among 8,563 annotated transcripts, 741 appeared with completeness of their open reading frames and therefore termed as full-length transcripts. These full-length transcripts were selected as initial data set for further analysis, four of which are listed in Table 4.1A. The corresponding Pfam domains are also listed in Table 4.1B for the selected four transcripts. An analysis of a further 721 transcripts, with annotations, is presented in Appendix A, Table A4.1.

Transcript	Length[a]	Protein	Closest ortholog	Genbank	Pident[b]
Transcript14	375	NADH dehydrogenase B14 subunit	*D. busckii*	gi|924555259	86
Transcript72	729	eukaryotic translation initiation factor 6	*D. citri*	gi|662209025	82
Transcript77	1329	clathrin coat associated protein ap-50	*R. pedestris*	gi|501295227	92
Transcript103	354	Vacuolar ATPase G subunit	*G. atropunctata*	gi|90820012	93

[a]Length : Amino acids
[b]Pident : BLASTP %identity

(A)

Transcript	Pfam id	Domain	Description
Transcript14	PF05347.12	Complex1_LYR	NADH dehydrogenase complex protein
Transcript72	PF01912.15	eIF-6	eukaryotic translation initiation factor 6 family
Transcript77	PF01217.17	Clat_adaptor_s	Clathrin adaptor complex small chain
Transcript77	PF00928.18	Adap_comp_sub	Adaptor complexes medium subunit family
Transcript103	PF03179.12	V-ATPase_G	Vacuolar (H+)-ATPase G subunit

(B)

TABLE 4.1: *B. tabaci* (Asia I species) transcripts used for initial intron-exon structural analysis. Four transcripts were selected from the Asia I species transcript profile for detailed intron-exon structural analysis. The table lists orthologs (A) and corresponding Pfam-based domains (B).

4.3.2 Genome: sequencing and draft genome assembly

Sequencing of seven PE lanes using Illumina GAII technology generated 3,124,129 sequences with maximum read length 80 bp and a total residue count of 1,036,713,302. In addition, one short insert (400 bp) and two long insert (3 kbp) libraries were also sequenced using Illumina HiSeq 2000 and generated 386,403,442, 357,614,632 and 381,489,430 reads respectively with a maximum read length of 100 bp. *De novo* assembly of these reads with CLC Genomics produced 343,393 contigs ranging in size between 200 bp and 657,201 bp, with an average length of 2,040 bp and 14,238 contigs with N50 length of 12,472 bp. The largest contig size was 658,076 bp and total contig size was 700,602,328 bp (701 Mbp). SSPACE constructed 290,838 scaffolds from 343,393 contigs using long insert mate pair reads, with an average size of 2,596 bp, a total residue counts of 755,155,629 bp (755 Mbp) and 6,969 scaffolds with an N50 length of 26,653 bp (Appendix A, Table A4.2). This draft genome assembly was found very fragmented and incomplete, as 200 Mbp (26.55%) of 755 Mbp were gaps (N's) leaving only 554 Mbp without gaps. Nevertheless, this 554 Mbp assembly was sufficient to predict gene structures in the Asia I species genome and could provide a valuable resource for accurate gene predictions as and when a more complete genome assembly becomes available.

The estimated size of the *B. tabaci* (MEAM1 and MED species) genome is 640-680 Mbp based on flow cytometry (Chen et al., 2015; Guo et al., 2015). The pilot phase draft assembly covers 554 Mbp of the genome in 290,838 scaffolds. In future, more deep sequencing will be required to obtain a more complete genome assembly for Asia I.

4.3.3 Genomic mapping and alignment

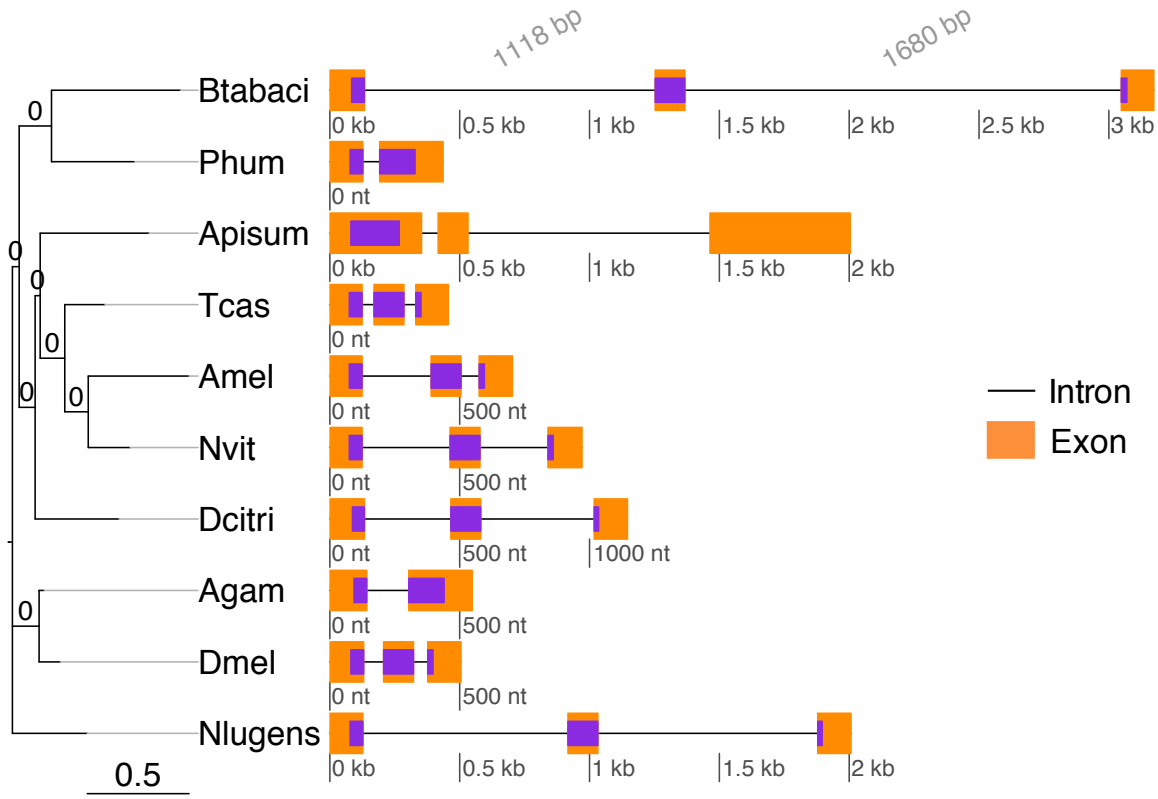
To confirm the mutual accuracy of transcriptome and genome sequence sets, all of the full-length cDNA transcripts were cross-mapped to genome scaffolds using GMAP (Wu and Watanabe, 2005), both to establish the length of each gene and confirm its assembly quality based on percentage identity between the transcriptome and genome.

Mapping 741 full-length transcripts onto the genome assembly resulted in 441 transcripts showing 100% sequence identity and showing 100% coverage to the preliminary genome assembly. The remaining 126 transcripts had 98-99.9% sequence identity and 174 transcripts had less than 99% sequence coverage. The 174 transcripts were mapped with 100% identity to discontinuous scaffolds and thus their intron-exon structure could not be determined. They were therefore not placed in the initial dataset. Of the total 741 transcripts, 126 were mapped on to consensus scaffolds with 100% coverage but mismatches occurred within the codon triplet, which could produce alternative isoforms. Of the total 741, an initial dataset of 567 (441+126) transcripts were chosen for further analysis based on sequence homology to previously characterized orthologs, both to

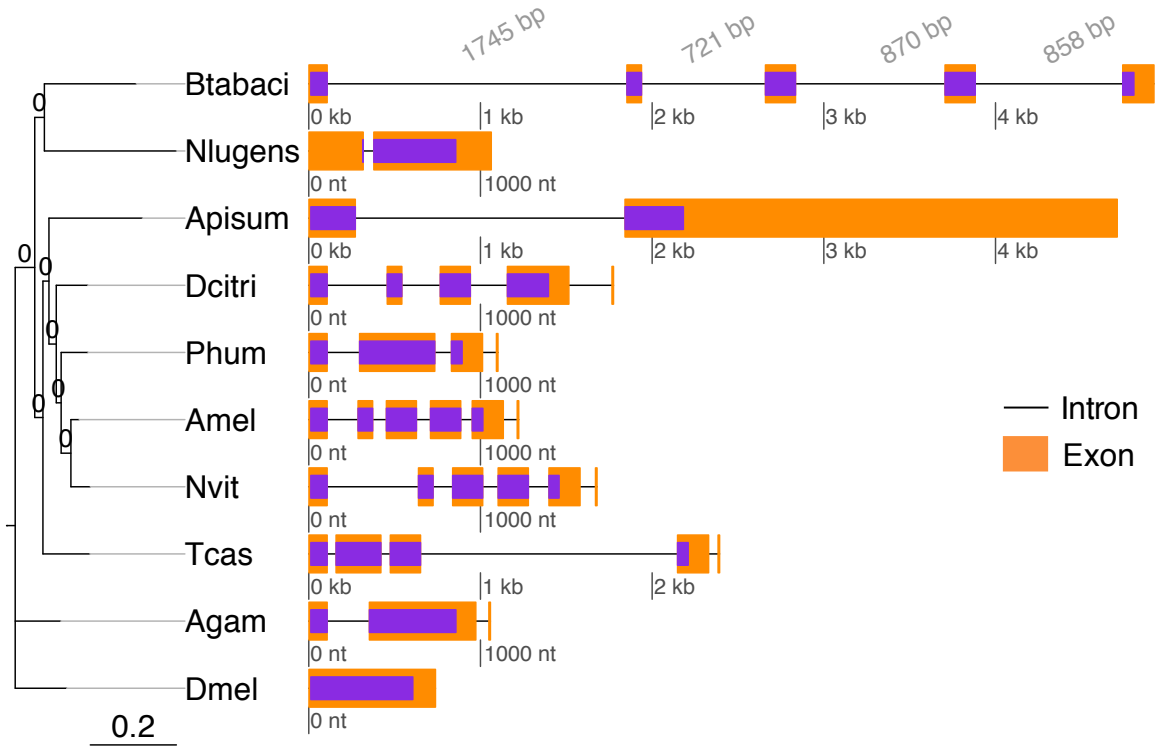
explore sequence divergence between Asia I species and other *B. tabaci* species, and to provide a set of orthologous genes present in whiteflies and other insects, which could be used with confidence for comparative gene structure analysis.

Using this genome assembly, gene models were produced for a subset of 119 of the 567 genes, adjudged to cover the entire open reading frame of their cognate proteins (by comparison with orthologs from other insects, see Appendix A, Table A4.3) and their introns not containing any ambiguous bases (N's). A representative selection of these gene models is shown in Figure 4.1, together with comparisons of their orthologs from *A. pisum*, *N. lugens*, *D. citri*, *D. melanogaster*, *A. gambiae*, *A. mellifera*, *N. vitripennis*, *T. castaneum* and *P. humanus*. In each case, the *B. tabaci* (Asia I species) open reading frame has also been compared with those of the other insects by Pfam analysis to confirm the identity of conserved functional domains, as indicated in Figure 4.1.

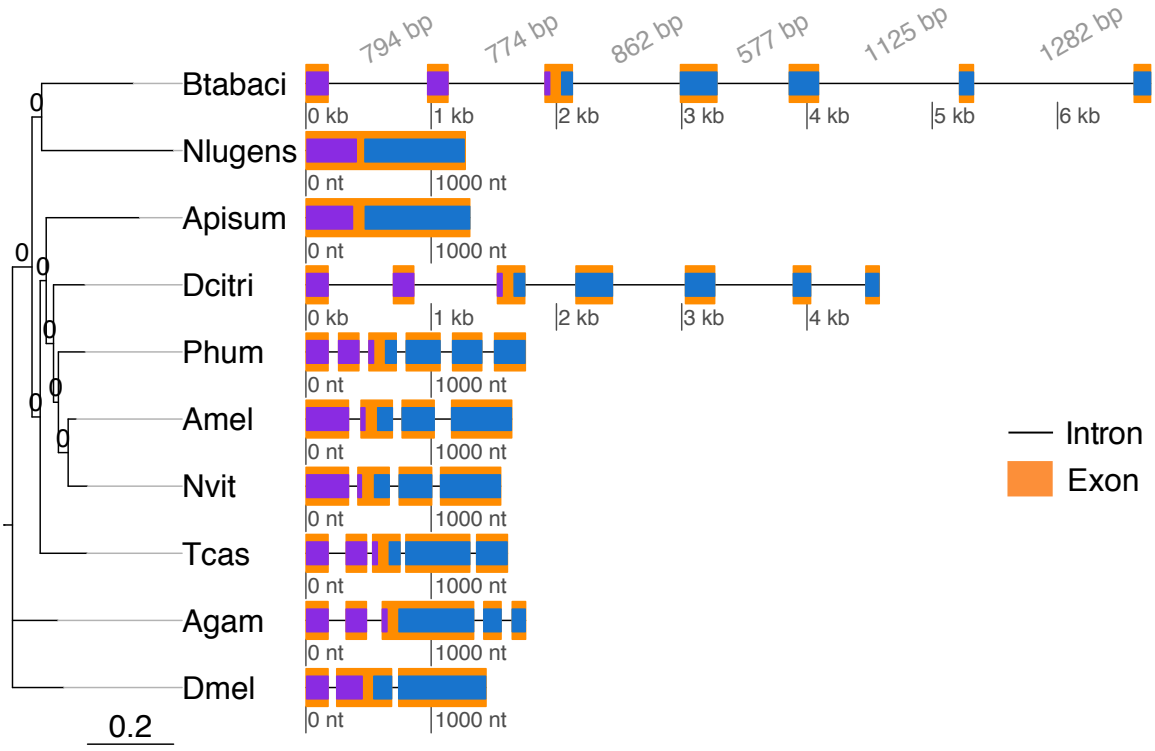
The gene structure comparison of orthologs across 10 insects revealed a huge variation in gene length, because of one single parameter, namely intron length. For each gene ortholog, it was observed that the gene lengths (exon+intron) were 3-5 times longer in *B. tabaci* (Asia I species) than in *A. pisum*, *N. lugens*, *D. citri*, *D. melanogaster*, *A. gambiae*, *A. mellifera*, *N. vitripennis*, *T. castaneum* and *P. humanus*. The selected gene models also show variability in intron sizes and their frequencies across *B. tabaci* (Asia I species) and the remaining nine insect genes. Although the introns of *B. tabaci* (Asia I) genes were much longer in comparison to introns of other nine insects compared in this study (Figure 4.1). The domains were found conserved (colour coded in Table 4.1B) across all nine insects regardless of their intron size and frequency.



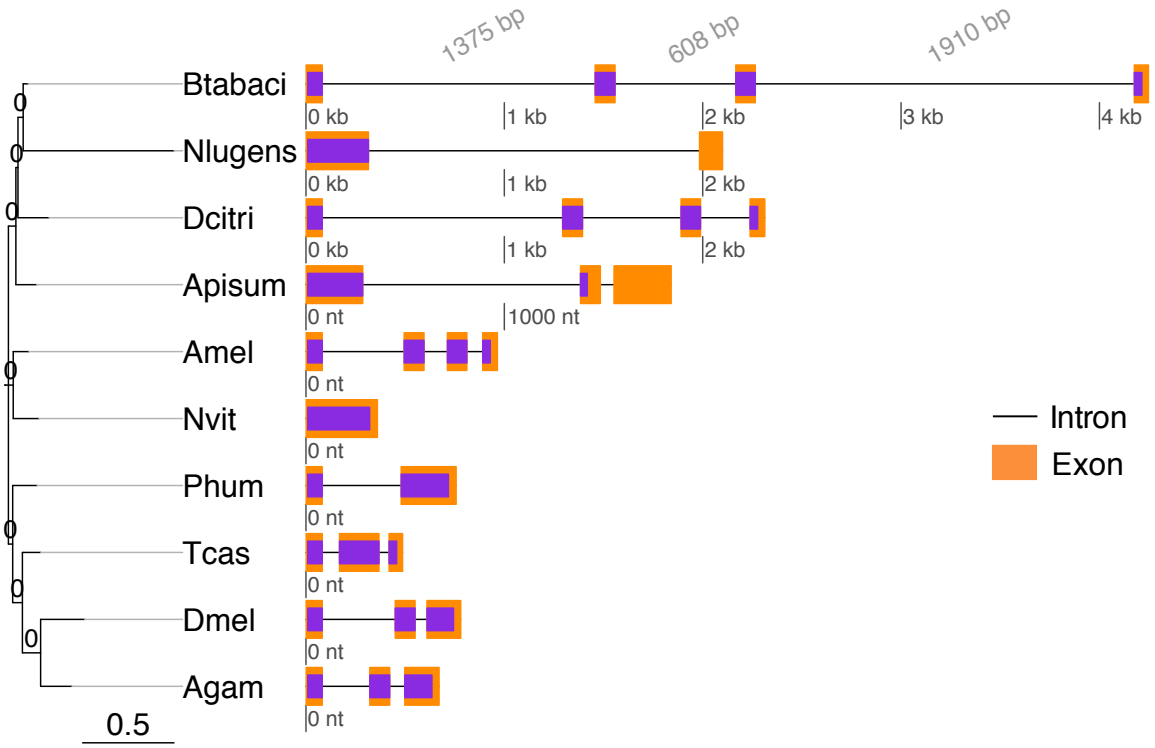
(A) Transcript14- NADH dehydrogenase B14 subunit



(B) Transcript72 - eukaryotic translation initiation factor 6



(C) Transcript77 - clathrin coat associated protein ap-50



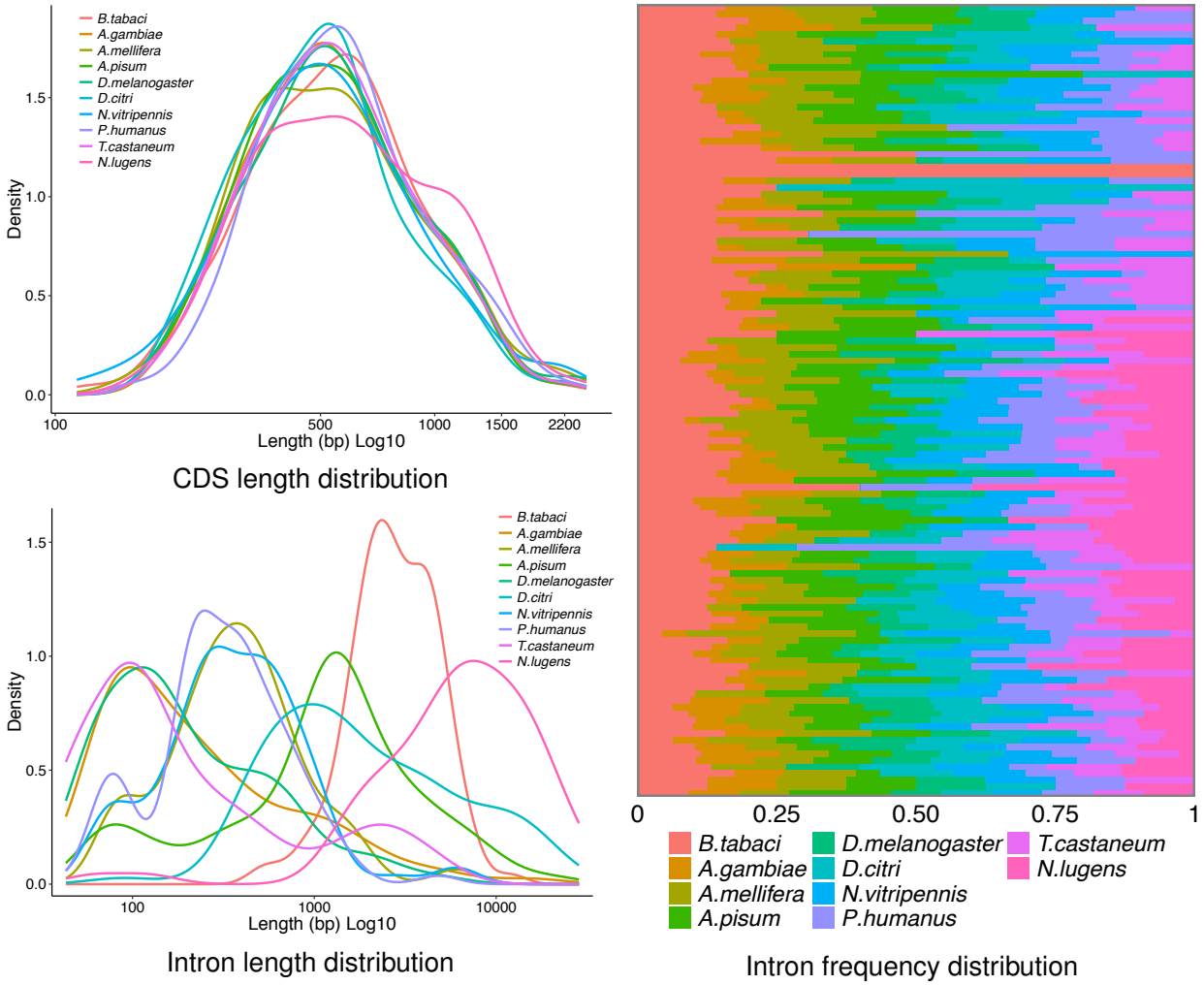
(D) Transcript103 - vacuolar ATPase G subunit

FIGURE 4.1: Intron-exon structure analyses for selected genes across 10 insects. Four genes, selected from the initial set of 119 transcripts from *B. tabaci* (Asia I species, Btabaci), were compared to orthologs present in other insect genomes (*A. pisum* (Apisum), *N. lugens* (Nlugens), *D. citri* (Dcitri), *A. mellifera* (Amel), *N. vitripennis* (Nvit), *A. gambiae* (Agam), *D. melanogaster* (Dmel), *T. castaneum* (Tcas) and *P. humanus* (Phum). Genes are arranged based on phylogenetic analysis using Muscle on protein sequences to show intron-exon boundaries within their coding sequences, with their 5' start codons aligned at the left. Introns are black horizontal line connecting two exons and their sizes are placed at the top for each *B. tabaci* (Asia I species) transcript. The exons of each gene are colour coded, with violet, blue and brown indicating the location of individual Pfam domains within each gene, and orange indicating the remaining coding sequence. Identifiers for the Pfam domains in each gene are given in Table 4.1B.

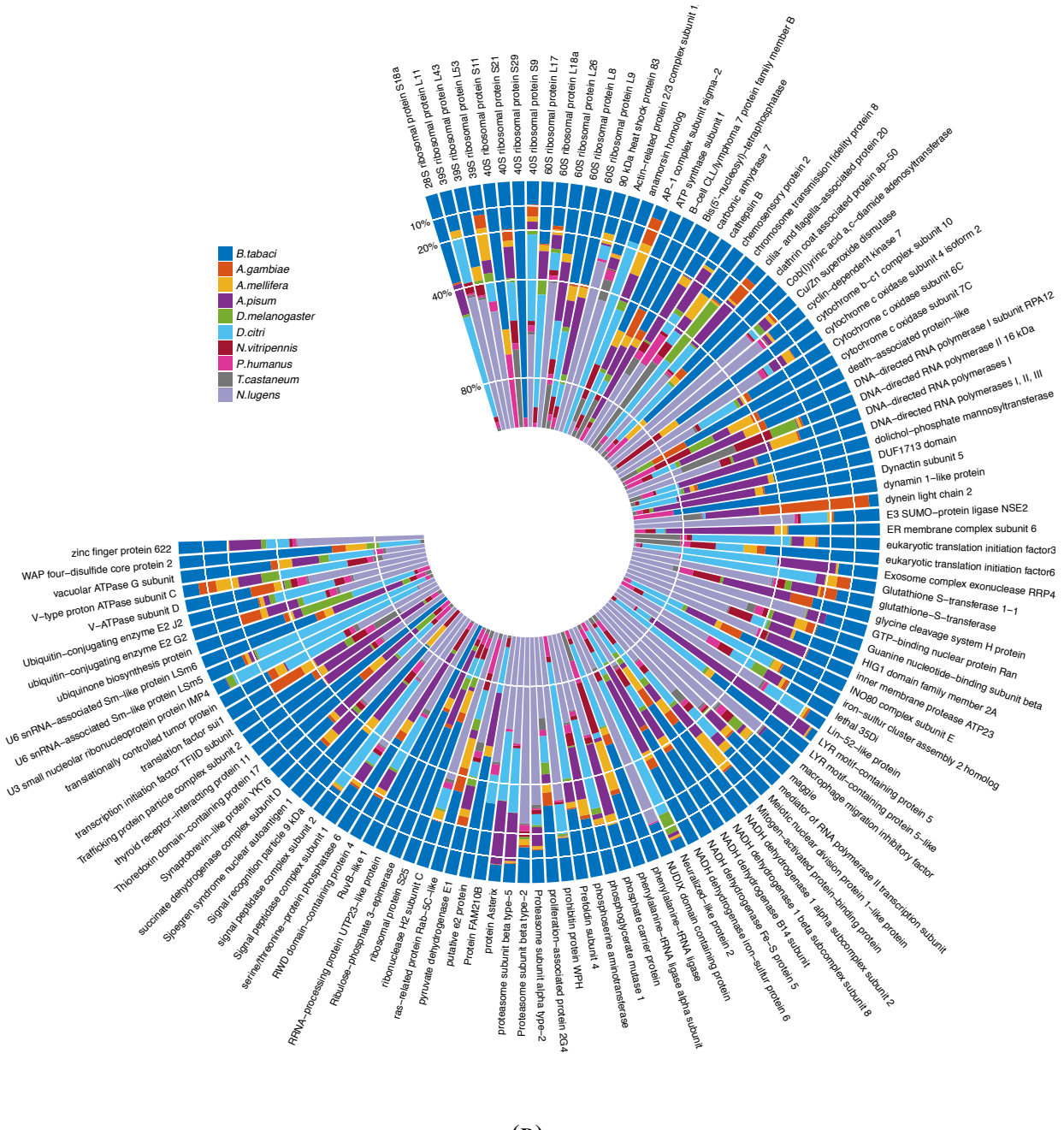
4.3.4 Gene complexity analyses

An overall assessment of the size and complexity of the subset of 119 *B. tabaci* (Asia I population) genes in the initial dataset shows that the gene span in *B. tabaci* (Asia I) is much longer than in *A. pisum*, *N. lugens*, *D. citri*, *D. melanogaster*, *A. gambiae*, *A. mellifera*, *N. vitripennis*, *T. castaneum* and *P. humanus*. The closest orthologs (106 genes) of 119 *B. tabaci* (Asia I population) genes were identified in *A. pisum* followed by *D. melanogaster* (105 orthologs), *A. mellifera* (104 orthologs), *P. humanus* (102 orthologs), *A. gambiae* (100 orthologs), *N. vitripennis* (99 orthologs), *T. castaneum* (99 orthologs), *D. citri* (88 orthologs) and *N. lugens* (72 orthologs). There were only 37 core orthologs found which were present in all 10 insects species (Appendix A, Table A4.4). Further to add more significance to the above gene structure comparison results, the length distribution of the coding sequence (CDS) and non-coding sequences (introns) of 119 ortholog genes were compared across all 10 insect species. Not surprisingly, the most conserved CDS length distribution was observed across all 10 insects (Figure 4.2A). All 10 insects showed a very similar pattern with the highest density peak at 500 bp; the CDS length distribution ranges from 114 bp to 2,511 bp. The opposite effect was observed in intron length distribution across these 10 insects (Figure 4.2A). The intron length distribution peaks at 2,205 bp in *B. tabaci* (Asia I) which was significantly higher than that of 1,100 bp, 836 bp, 268 bp, 230 bp, 218 bp, 102 bp, 91bp and 75 bp in *A. pisum*, *D. citri*, *A. mellifera*, *P. humanus*, *N. vitripennis*, *A. gambiae*, *D. melanogaster* and *T. castaneum* respectively (Figure 4.2A). However, the intron length distribution peak in *N. lugens* was found at 3,988 bp which was higher than that of *B. tabaci* (Asia I) and the remaining eight species. The variations in total gene size between these insects are clearly due to the size of their introns. These results demonstrate that the introns are much longer in *B. tabaci* (Asia I) genes than their orthologs in *A. pisum*, *D. citri*, *A. mellifera*, *P. humanus*, *N. vitripennis*, *A. gambiae*, *D. melanogaster* and *T. castaneum* regardless of their sequence homology in their coding regions.

The complexity of gene structure in all 10 insects was also analysed by comparing order and frequency of intron-exon within a gene. Of the total 119 genes, 90 and 83 were found with a greater number of introns in *B. tabaci* (Asia I) compared to two dipteran insects *A. gambiae* and *D. melanogaster* respectively, whereas only 67, 62, 41 and 67 were found in *B. tabaci* (Asia I) compared to hemipteran insects *N. lugens*, *D. citri* and *A. pisum* respectively. Similarly, 79, 67, 68 and 54 genes were found in *B. tabaci* (Asia I) with more number of introns than in *T. castaneum*, *N. vitripennis*, *P. humanus* and *A. mellifera* respectively (Figure 4.2A).



(A)



(B)

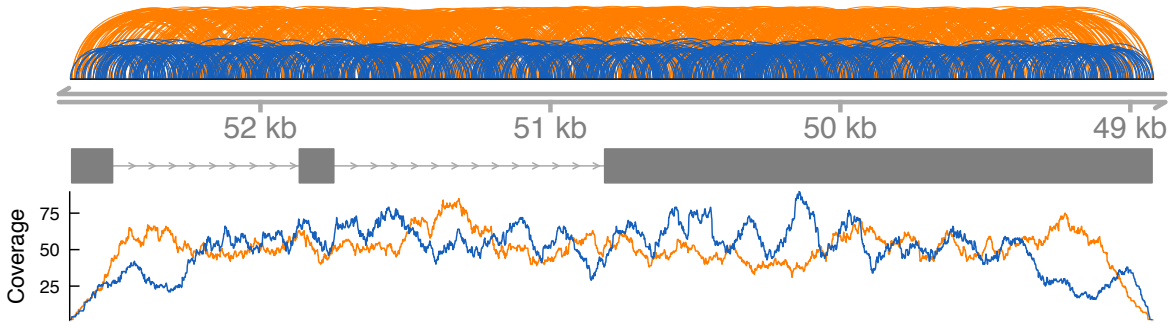
FIGURE 4.2: Length distribution of CDS and introns for selected 119 genes in *B. tabaci* (Asia I species) and their corresponding orthologs in other nine insects including *A. pisum*, *N. lugens*, *D. citri*, *D. melanogaster*, *A. gambiae*, *A. mellifera*, *N. vitripennis*, *T. castaneum* and *P. humanus*. (A) CDS density, intron density and intron count distribution across 10 insects, and (B) Intron size distribution.
*Figure 4.2A and Figure 4.2B has the same colour coding and order of genes from bottom to top.

The variation of total intron length in 119 gene orthologs from all 10 insects was shown in a "polar bar plot" (Figure 4.2B). In comparison with the presence of introns in *B. tabaci* (Asia I) genes, there were 13 genes found with no intron in *A. pisum* followed by 12 genes in *A. gambiae*, 11 genes in *D.*

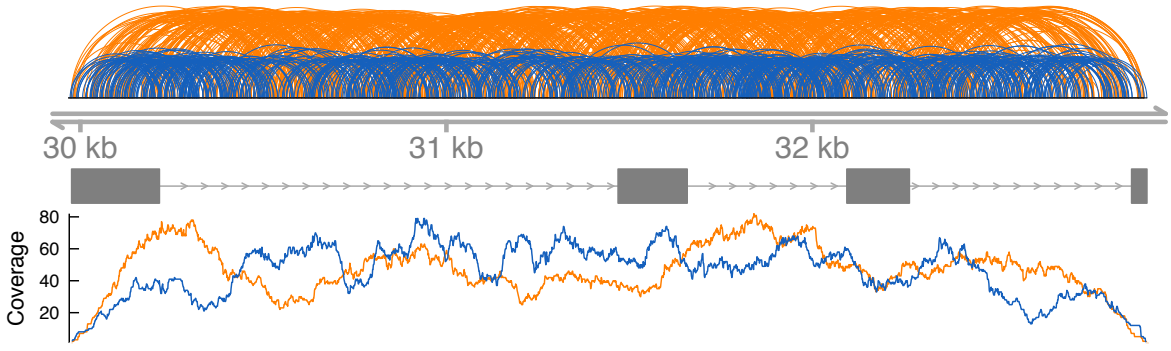
melanogaster and *T. castaneum*, nine genes in *D. citri* and *N. vitripennis*, six genes in *N. lugens*, and three genes in *A. mellifera* and *P. humanus* (Appendix A, Table A4.4).

4.3.5 Validation of genome assembly

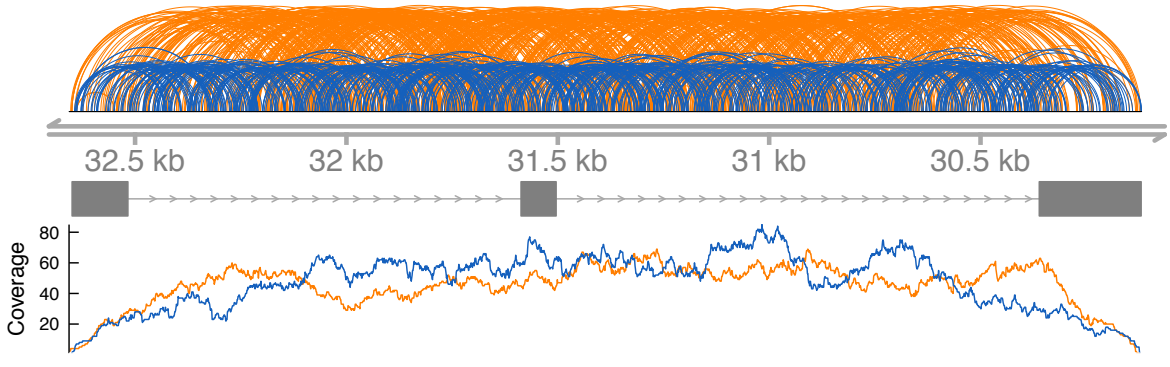
To validate the genome assembly and the intron length, two PE read libraries with different insert sizes were aligned to four randomly selected genome scaffolds encompassing transcripts listed in Appendix A, Table A4.1. These four transcripts were "Transcript82", "Transcript80", "Transcript85" and "Transcript79" which encodes "90 kDa heat shock protein", "GTP-binding nuclear protein Ran", "glutathione S-transferase" and "Glyceraldehyde 3-phosphate dehydrogenase" respectively. The mapping support of two PE libraries with insert size 200 bp and 450 bp over introns and exons is shown in Figure 4.3A-D. The PE reads mapped to assembled scaffolds with an average per-base coverage of 50x. These PE mappings confirm the exon-intron-exon structure along with their lengths and the accuracy of the assembled Asia I species genome scaffolds. The results also suggest that the PE reads not only support the coding regions (exons) of the gene but also support the longer non-coding regions (introns).



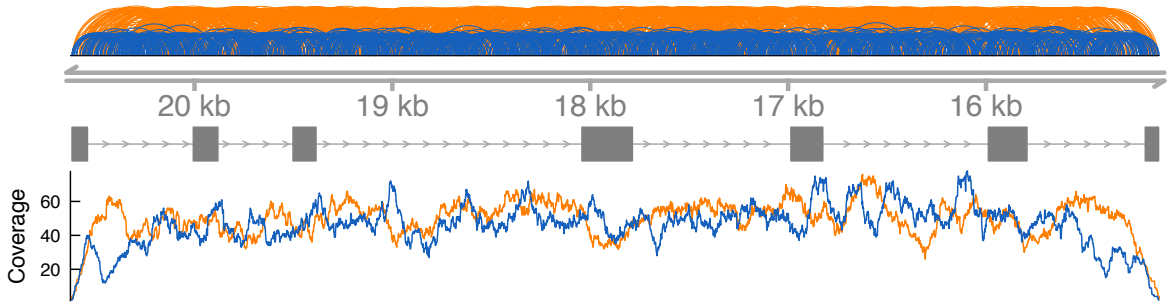
(A) Transcript82



(B) Transcript80



(C) Transcript85



(D) Transcript79

FIGURE 4.3: Genome assembly validation via PE read mapping. (A) Transcript82, (B) Transcript80, (C) Transcript85 and (D) Transcript79. The top arcs represent two different PE read libraries with insert size 200 bp (blue) and 450 bp (orange). The orientation of the genomic scaffold is 5'->3'. The genomic scaffold coordinates are shown in the middle along with the exons (dark grey block) and introns (grey line with arrows). The PE reads mapping coverage for each base is also shown as a line graph at the bottom.

4.3.6 Intron splice sites

In eukaryotes, the most protein-coding genes contain coding exons alternated with non-coding introns. The non-coding introns are removed and the coding exons are joined into a specific order to form a mature mRNA before they get exported to the cytoplasm. This process is known as mRNA splicing (Wang and Burge, 2008). The synthesis of a correct protein product requires accurate splicing of introns at specific sites. Every intron contains two splice signals that are essential for splicing and exon definition, the 5' splice site also known as donor splice site and the 3' splice site also known as the acceptor splice site (Breathnach and Chambon, 1981; Burset et al., 2000). The 5' donor splice site of each intron begins with the "canonical" dinucleotide GT followed by few varying nucleotides. Similarly, the 3' acceptor splice site of each intron ends with another "canonical" dinucleotide AG and followed by other varying nucleotides (Figure 4.4). Mutations occurring at these di-nucleotides lead to mis-splicing which results in exon skipping or intron retention. Splice site mutations also lead to human diseases including "Frasier syndrome", "atypical

cystic fibrosis", "neurofibromatosis type 1" and "familial dysautonomia" (Faustino and Cooper, 2003; Wang and Cooper, 2007; Tazi et al., 2009).

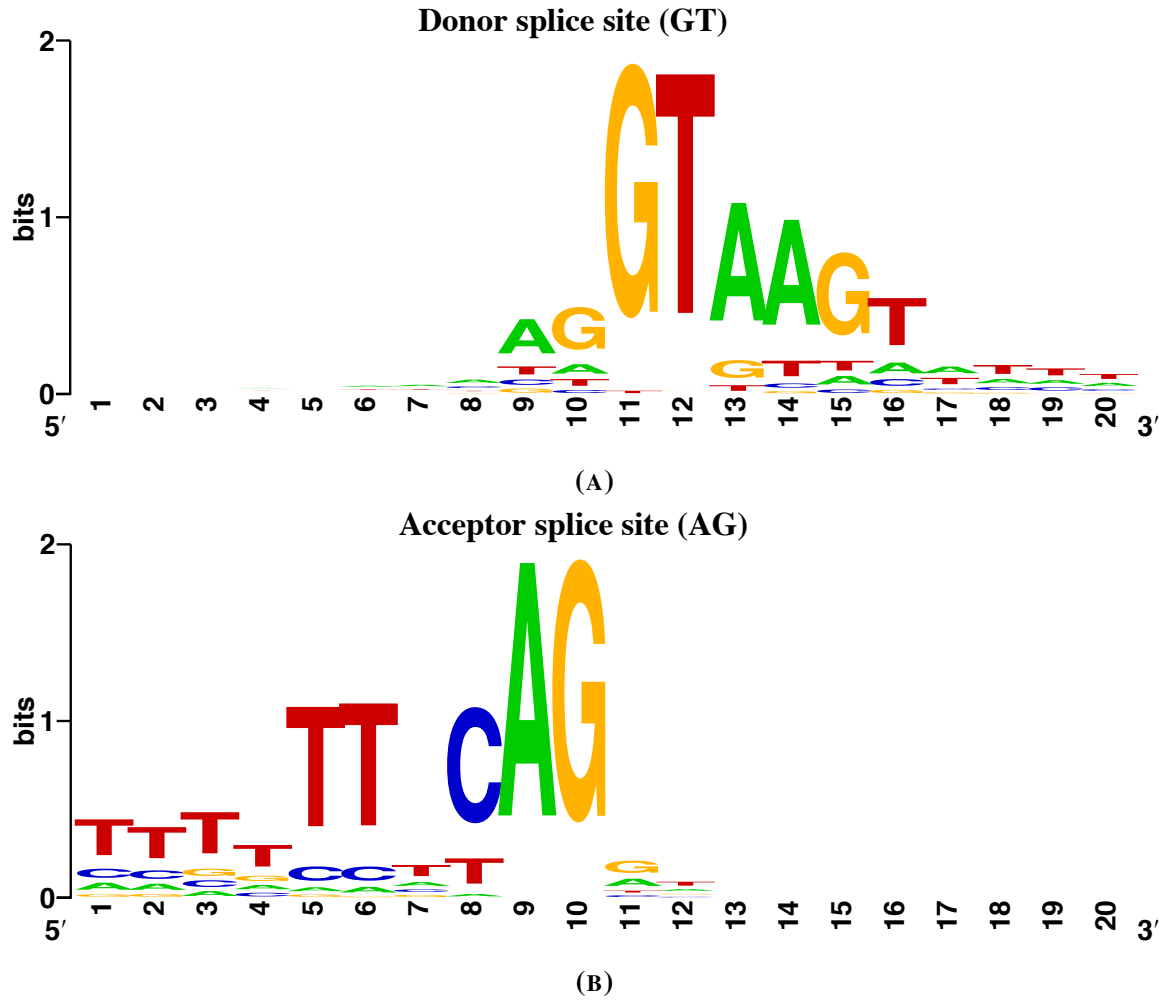


FIGURE 4.4: Intron splice site motifs in *B. tabaci* (Asia I species) genome. (A) 5' donor splice site and (B) 3' acceptor splice site.

4.3.7 Usage of nucleotides and di-nucleotides in introns

Genes containing two or more introns were selected for intron position based nucleotide composition analysis. The composition of nucleotides and di-nucleotides in introns up to position eight are shown in Figure 4.5. Occurrence of A and T nucleotides were predominantly higher than that of G and C regardless of their intron position. Di-nucleotides composed of A and T (AA, AT, TA, TT) were also higher than those of G and C (CC, CG, GC, GG) (Figure 4.5). These results suggest that the A+T nucleotides and di-nucleotides are most common in introns. T and TT were the most highly represented nucleotides and di-nucleotides where-as G and CG were the lowest represented. A high composition of G+C and CG dinucleotide leads to the development of CpG island, which was identified in 40-50% of the human genes (Suzuki et al., 2001). In *B. tabaci* (Asia I), G+C and CG

composition of the introns were found much lower and therefore a high frequency of CpG islands were not identified. Of the total 539 introns, only 17 (3.15%) introns containing CpG islands were identified.

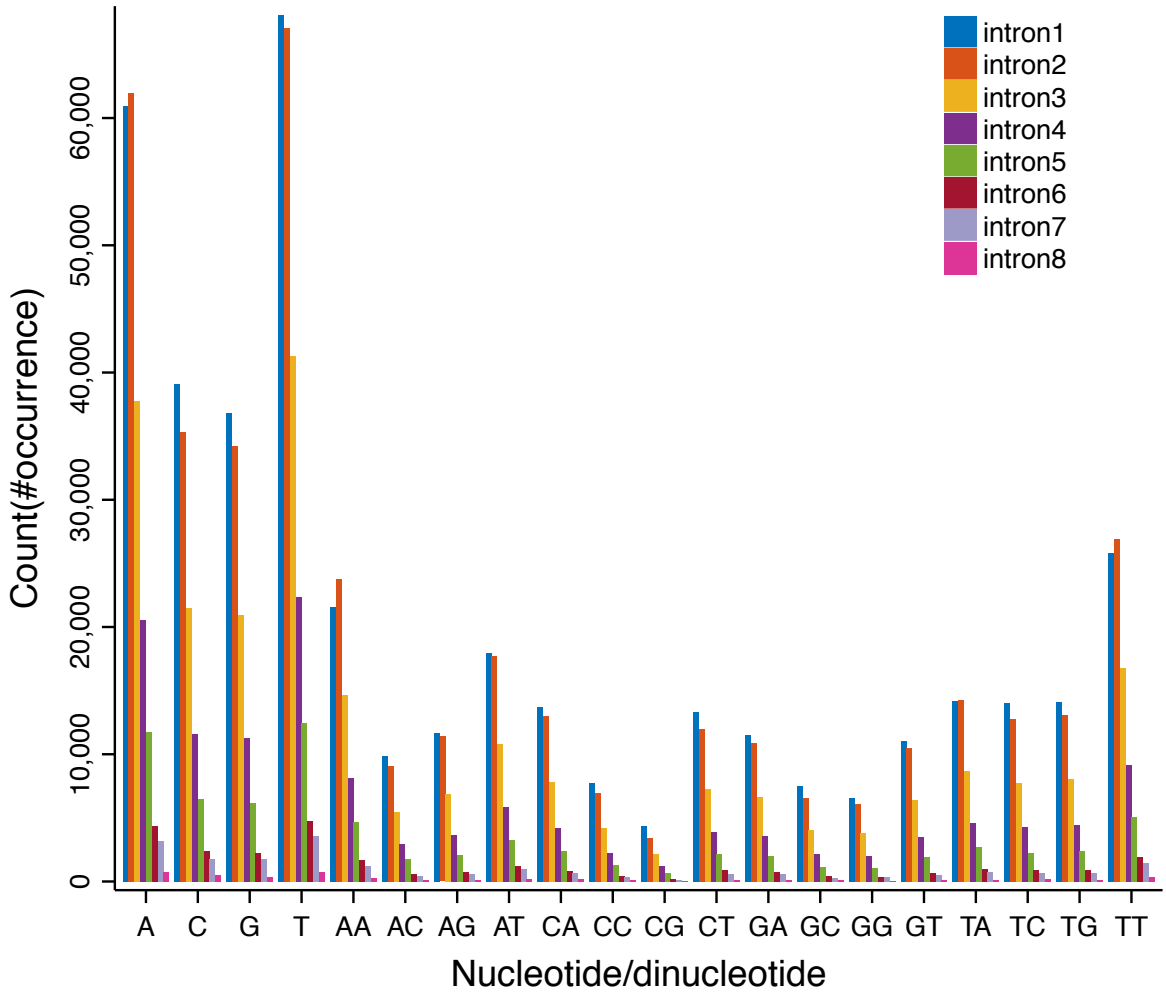


FIGURE 4.5: Frequencies of nucleotides and di-nucleotides in introns of *B. tabaci* (Asia I species) genome. Each colour bar represents occurrence of nucleotides and di-nucleotides in introns at various positions.

4.3.8 U12-type spliceosomal introns in *B. tabaci*

Introns are classified into four major groups namely group I, II, III and spliceosomal introns (Saldanha et al., 1993; Sharp and Burge, 1997). The first three groups of introns undergo self-splicing where-as the fourth group of introns require complex machinery called spliceosomes. A spliceosome comprises four small nuclear ribonucleoproteins (snRNPs) and more than a hundred non-snRNP proteins which associate with snRNPs during the event of splicing (Lamond, 1993; Wahl et al., 2009; Will and Lührmann, 2011). There are two distinct types of spliceosomal introns; U12-type and U2-type introns that are spliced by the minor and major spliceosomes respectively. U12-type introns are highly conserved and comprise less than half a percent of all spliceosomal introns (Sharp

and Burge, 1997; Levine and Durbin, 2001). The insect U12-type intron containing genes were retrieved from the U12DB (v1.0) intron database (http://genome.crg.es/cgi-bin/u12db/u12db.cgi) (Alioto, 2007). The translated protein sequences of these 71 genes were used to identify potential orthologs in *B. tabaci* (Asia I) using TBLASTN (v2.2.29+) (Altschul et al., 1990) and exonerate (Slater and Birney, 2005). The ortholog genes containing introns were identified along with their splice sites. This result was compared with the previous findings of U12-type introns in 20 insect genomes (Janice et al., 2012). Of the total 71 U12-type intron containing genes, only 36 genes were identified in *B. tabaci* (Asia I) (Table 4.2). This includes two of AT-AC U12-subtype and 34 of GT-AG U12-subtype intron. U12-type intron containing genes in *B. tabaci* (Asia I) were found higher than equivalent ones in 15 Dipteran species including 12 *Drosophila* and three mosquito species regardless of their genome size. However, they were found lower in *B. tabaci* (Asia I) than in *A. mellifera*, *N. vitripennis* and *P. humanus* (Table 4.2). The insect genomes harbour such a low number of U12-type introns indicating the deletion of this type of introns which leads to their conversion to U2-type intron (Janice et al., 2012). Since the draft genome assembly was used in this study, the U12-type intron containing genes in *B. tabaci* (Asia I) were still an incomplete set and therefore further assembly and analysis is required once an improved genome assembly is available.

Species	Introns available	U12-type introns identified	Genome size in Mbp
Aedes aegypti	71	17	1376
Anopheles gambiae	70	17	278
Apis mellifera	71	63	264
Bemisia tabaci	71	35	~650
Bombyx mori	68	32	530
Culex quinquefasciatus	71	15	579
Drosophila ananassae	71	17	185.82
Drosophila erecta	71	19	146.7
Drosophila grimshawi	71	19	234.72
Drosophila melanogaster	71	19	176.04
Drosophila mojavensis	71	19	166.26
Drosophila persimilis	70	17	176.04
Drosophila pseudoobscura	71	19	156.48
Drosophila sechellia	71	19	166.26
Drosophila simulans	71	19	146.7
Drosophila virilis	71	19	332.52
Drosophila willistoni	71	17	205.38
Drosophila yakuba	71	19	166.26
Nasonia vitripennis	71	46	335
Pediculus humanus	71	39	107.58
Tribolium castaneum	71	34	160

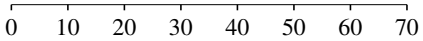


TABLE 4.2: U12-type intron containing genes in *B. tabaci* (Asia I species) and other insect genomes. The introns available are the ortholog genes found for that particular insect and red bar represents U12-type intron containing genes for corresponding insect.

4.4 Discussion

The production of a reference genome sequence for the whitefly, *B. tabaci* (Asia I population) represents a major objective in pest management and biology, since this insect is a pest by direct feeding damage as well as by being capable of transmitting hundreds of different viruses that are major threats to food security. However, at a size of approximately 640-680 Mbp (Chen et al., 2015; Guo et al., 2015), the *B. tabaci* (Asia I) genome is larger than most insect genomes including its closest hemipterans *A. pisum* and *D. citri*. This poses several challenges for genome assembly. In this chapter, the assembly and characterisation of a set of 119 transcribed genes is described, which enable us to quality control the accuracy of the templates to be used for subsequent genome assembly studies.

Validated gene models for the 119 genes were produced and described, enabling the side by side comparison of the genes from this hemipteran insect with others which have been sequenced from the same group (*A. pisum*, *N. lugens* and *D. citri*), and with more distant relatives (*D. melanogaster*, *A. gambiae*, *A. mellifera*, *N. vitripennis*, *T. castaneum* and *P. humanus*). Using Pfam to analyse the coding sequences, good concordance in overall domain content between *B. tabaci* (Asia I) and these nine other insect species was found, but a striking difference in the size and often the organisation of their introns can be seen, with the introns present in *B. tabaci* (Asia I) in general being larger and more numerous than those present in the other insect genes examined. This genetic expansion may contribute to the increased size of the *B. tabaci* (Asia I) genome, but perhaps more interestingly may also reflect a more dynamic genome. Studies are ongoing to determine whether the intron organisation observed in Asia I is replicated precisely in other *B. tabaci* species. Nevertheless, the gene structures examined in detail in this chapter via aligning PE reads to the assembled scaffolds to confirmed the accuracy of the assembly and their intron lengths. All four examined scaffolds showed good PE read support for each base with an average coverage of 50x. Consistent with this result, intron splice motifs including donor splice site and acceptor splice site were also found as conserved motifs in *B. tabaci* (Asia I) introns, which confirmed their start and end sites within gene.

The length of introns vary across a wide range of eukaryotic genomes, which suggest selective pressure of an unknown origin (Deutsch and Long, 1999). The intron-exon structural analysis of 119 orthologs from 10 insects revealed some interesting results. An average intron length of intron-containing gene orthologs in *B. tabaci* (Asia I) was 3,126 bp, which was longer than 351 bp, 427 bp, 517 bp, 521 bp, 565 bp, 572 bp and 1,813 bp in the corresponding gene orthologs from *D. melanogaster*, *P. humanus*, *T. castaneum*, *N. vitripennis*, *A. mellifera*, *A. gambiae* and *A. pisum* respectively. However, the average intron length was longer in *D. citri* (3,234 bp) and *N. lugens* (7,744 bp) than that in *B. tabaci* (Asia I). The genomes of seven of these eight insects are smaller than that of *B. tabaci* (Asia I) with *N. lugens* only containing a larger genome (1,140 Mbp). This therefore suggest that the larger introns in *B. tabaci* (Asia I) are highly likely to be one main

contributor to it's relatively large genome size. This is supported by the similar contribution of introns to a larger genome size in *N. lugens*. With an average intron length of 3,126 bp, introns of the "proliferation-associated protein 2G4" were 12,061 bp, "eukaryotic translation initiation factor 3 subunit 6" were 6,907 bp, "V-type proton ATPase subunit C" were 4,495 bp, and those of other genes were about 495 to 7,651 bp.

It is not only the intron length, but also the number of introns per gene that varies across eukaryotes (Deutsch and Long, 1999). The average number of introns per intron-containing gene was higher in *B. tabaci* (Asia I species) (2.92) followed by 2.83, 2.63, 2,58, 2.50, 2.38, 2.33, 1.90, 1.65 and 1.53 in *N. lugens*, *A. mellifera*, *P. humanus*, *A. pisum*, *D. citri*, *N. vitripennis*, *T. castaneum*, *D. melanogaster* and *A. gambiae* respectively. A positive correlation between the intron density of a gene and its expression level was reported in the human genome; genes with higher intron densities were more highly expressed than those with lower intron densities (Comeron, 2004). Similarly, introns have stimulatory effects on gene expression and this has been termed 'intron-mediated enhancement' (IME) (Mascarenhas et al., 1990). IME has been observed in many diverse eukaryotes (Duncker et al., 1997; Ho et al., 2001; Xu and Gong, 2003; Juneau et al., 2006), and the same trend was also supported in plants (Clancy and Hannah, 2002; Fiume et al., 2004; Rose, 2008). The higher intron density in *B. tabaci* (Asia I) together with their longer sizes indicates that the introns play a potentially highly significant role in gene regulation and potentially genome expansion in *B. tabaci* (Asia I).

Intron position within individual domains (i.e. where introns are present) is largely conserved between the species examined here. The positioning of these intron-exon boundaries and their functional significance remains to be determined, although the potential correlation between intron position and the nucleotide/dinucleotide composition of that intron was also examined. In this study, A+T nucleotides and di-nucleotides were found most common in introns. Where-as G+C and CG composition of the introns were found lowest represented. These results suggest that the *B. tabaci* (Asia I) introns are more AT enriched than GC which is to be expected from their overall GC content of 35.56%. Only 17 (3.15%) introns were found containing CpG islands which is much lower than the estimated percentage in entire human genome (40-50% genes) (Larsen et al., 1992; Suzuki et al., 2001).

The mRNA transcripts have permitted us to identify the size and location of introns and exons in the *B. tabaci* (Asia I) genome, and access to this accurate gene models will accelerate population studies and assist in producing high quality genome assembly and more robust gene annotation of this devastating disease vector. Further analysis examining more number of introns and genes to identify the potential role of introns in *B. tabaci* (Asia I) is recommended.

CHAPTER 5

Draft genome sequence of *Bemisia tabaci* cryptic species Asia I

5.1 Introduction

The whitefly, *B. tabaci* genome is one of the i5K target genomes (Robinson et al., 2011). The assembly and annotation of the *B. tabaci* Asia I species genome, the major aim of this research, is important for studying whitefly population diversity and its temporal evolution. Access to a complete *B. tabaci* genome sequence will permit detailed genome-wide association studies, providing a basis for studying other aspects of whitefly behaviour and response, including traits such as insecticide resistance and their relationships to various endosymbionts. Despite an ever-increasing global importance and economic impact, a draft whole genome of any *B. tabaci* species has not yet been reported. Major challenges have been proposed due to its innate complex biology (bacterial symbiosis) and genome complexity (size of genome, high level of heterozygosity and repetitive elements) (Chu et al., 2013). At present, as a genome sequence is not available for any *B. tabaci* species, comparative and pan-genomic approaches cannot be used to study global genetic differences and molecular mechanisms underlying the biological invasions of Asia I species.

Despite the availability of transcriptomes from different *B. tabaci* species including Asia I (as described in Chapter 3), MEAM1 (Leshkowitz et al., 2006; Wang et al., 2011; Xie et al., 2012), MED (Wang et al., 2010a) and Asia II 3 (Wang et al., 2012) , these do not provide complete genetic information for these species. Transcriptome sequencing has provided only a subset of genes that were expressed at different developmental stages of life and thus represents incomplete genetic information for those species. To obtain the complete set of genes for any species requires whole genome sequencing. Availability of genome sequences for species that fall in the *B. tabaci* species complex, will allow comparative genomic studies. They will enable further insights of the genetic diversity across this species complex and also the identification of conserved and highly diverged genes, recognition of gene families which have contracted or expanded and the evolution of *B. tabaci* members. The genome sequence of the Asia I species will also open the door for future

investigations on host-endosymbiont relationships and lateral gene transfer mechanisms between insect hosts and their endosymbionts.

A pilot phase of the Asia I genome sequencing project was conducted and the draft assembly (v1.0) was produced as described in Chapter 4. The genome assembly (v1.0) from the pilot phase was very fragmented and contained only 50% of the genome according to the estimated genome size of 640-680 Mbp from the MEAM1 and MED species (Chen et al., 2015; Guo et al., 2015). Therefore another deep sequencing of whole genome was required in order to capture more of the Asia I species genome. To achieve the high quality draft genome, the Illumina sequencing platform was employed again, but this time using longer reads (250 bp PE) and a construction suited to the 'DISCOVAR' assembly software.

In this chapter, a higher quality draft genome sequence of Asia I species is reported along with the prediction of gene models, identification of repetitive elements, non-coding RNAs, presence of endosymbionts, and gene families with their essential roles in detoxification and sex determination in Asia I species. The genomes of the closest hemipteran insects were also compared to identify orthologous genes and their structural complexity, and reveal key genetic variations which may underlie species-specific biological function across this Order. To date, this represents the first draft genome sequence from the *B. tabaci* cryptic species complex, although release of genomes for both the MED and MEAM1 species is expected to be imminent.

5.2 Methods

5.2.1 DNA extraction and sequencing

High quality genomic DNA was extracted from adult female and male whiteflies using a standard protocol (Blood and Tissue DNA extraction kit, Qiagen, Germany). The single Illumina (250 bp PE) sequencing library was constructed using a PCR-free protocol recommended for the DISCOVAR *de novo* assembler (v52488). The Illumina PCR-free library was prepared using the 'with-bead pond library' protocol as described in Fisher et al. (2011). DISCOVAR requires specific Illumina PE library with approximately 450 bp of fragment, from which 250 bp reads with inward facing are sequenced on Illumina HiSeq or MiSeq genome sequencers. Whole-genome shotgun sequencing of Asia I species was performed using Illumina MiSeq 2500 in a Rapid Run Mode at The Genome Analysis Centre (TGAC), Norwich, UK.

5.2.2 *De novo* genome assembly

Genomic sequences of endosymbionts were expected to be present in the Asia I species genome reads and were filtered to achieve symbiont-free Asia I genome sequences prior to assembly. The 250 bp reads generated by Illumina HiSeq 2000 were first aligned to reference genomes of endosymbionts using Burrows-Wheeler Aligner (BWA) (v0.7.12) (Li and Durbin, 2009) and filtered using SAMtools (v1.1) (Li et al., 2009). The raw reads were aligned to the genome sequence of seven endosymbionts including *Ca.* Portiera aleyrodidarum (BT-B, NCBI taxonomy id: 1206109, accession: CP003708, CP003868), *Ca.* Portiera aleyrodidarum (BT-Q, NCBI taxonomy id: 1239881, accession: CP003835, CP003867), *Wolbachia* pipientis (wBol1-b, NCBI taxonomy id: 1238452, accession: CAOH01000000), *Wolbachia* endosymbiont of *Culex quinquefasciatus* (wPel, NCBI taxonomy id: 570417, accession: NC_010981), *Wolbachia* endosymbiont of *Drosophila simulans* (wNo, NCBI taxonomy id: 1236908, accession: CP003883), *Wolbachia* endosymbiont of *D. melanogaster* (wMel, NCBI taxonomy id: 163164, accession: NC_002978), *Wolbachia* endosymbiont of *Drosophila simulans* (wRi, NCBI taxon id: 66084, accession: NC_012416), *Wolbachia* endosymbiont of *Drosophila simulans* (wHa, NCBI taxon id: 1236909, accession: CP003884), *Wolbachia* endosymbiont strain TRS of *Brugia malayi* (wBm, NCBI taxon id: 292805, accession: NC_006833), *Arsenophonus* endosymbiont str. Hangzhou of *Nilaparvata lugens* (NCBI taxon id: 1247024, accession: JRLH01000000), *Arsenophonus* nasoniae DSM 15247 (NCBI taxon id: 1121018, accession: AUCC01000000), *Ca.* Hamiltonella defensa MED (NCBI taxon id: 1163751, accession: AJLH02000000), *Rickettsia* sp. MEAM1 (NCBI taxon id: 1182263, accession: AJWD01000000), *Orientia* tsutsugamushi str. Boryong (NCBI taxon id: 357244, accession: NC_009488) and *Cardinium* cBtQ1 (NCBI taxon id: 1354314, accession: NZ_CBQZ010000000).

Prior to assembly, genome size and read coverage from raw and endosymbiont filtered reads were estimated by Jellyfish (v2.2.0) (Marçais and Kingsford, 2011) using k-mer approach. DISCOVAR *de novo* (http://www.broadinstitute.org/software/discovar) genome assembler was used to perform an assembly on the filtered set of reads containing Asia I species genome reads without endosymbiont reads. Further genome assembler such as Platanus (v1.2.1) (Kajitani et al., 2014) was also used to build *de novo* assembly on the same set of reads for the evaluation and comparison purposes. Platanus ran assembly in three steps: contig assembly from a de Bruijn graph with the k-mer range 32-167 (k-mer step size 10), scaffold assembly from the contigs using PE reads and finally, gap-filling on scaffolds using the PE reads.

5.2.3 Core eukaryotic genes: CEGMA, BUSCO

Core Eukaryotic Genes Mapping Approach (CEGMA) (v2.32) (Parra et al., 2007, 2009) and Benchmarking Universal Single-Copy Orthologs (BUSCO) (v1.1b1) (Simao et al., 2015) pipelines were used to assess the gene complement of the Asia I species genome assembly. CEGMA was run on the genome assembly using '–vrt' parameter to allow for vertebrate-sized longer introns to be identified. CEGMA reports a subset of the 248 most highly conserved eukaryotic genes (CEGs), and least paralogous CEGs. The detailed information of these CEGs is also given in the report if they are found in a complete or partial form. In the CEGMA pipeline, the predicted proteins were scored based on their alignment to a HMMER profile built for each CEGs. The alignment fraction of each predicted protein can range from 20-100%. The protein is grouped as 'full-length' if the fraction exceeds 70%, otherwise it is grouped as 'partial'.

BUSCO assesses the completeness of the genome assembly using orthologs from OrthoDB (Kriventseva et al., 2015) (www.orthodb.org). The BUSCO pipeline was run on the assembly in a 'genome' mode using the arthropod data set. The BUSCO pipeline has three phases: 1) use of BUSCO consensus sequences to identify the candidate regions of the genome to be assessed using TBLASTN; 2) Gene structure prediction using Augustus; and 3) Assessment of predicted genes using HMMER and lineage-specific profiles to classify them into four categories: complete (C), duplicated (D), fragmented (F) and missing (M).

5.2.4 Repeat annotation

The repetitive elements of the Asia I species genome were annotated in two ways: 1) RepeatMasker (v4.0.5) (http://www.repeatmasker.org/) was used to identify homologous repeats against RepBase TE library (Jurka et al., 2005) (update: 20150807) and 2) RepeatModeler (v1.0.8) (Smit A, Hubley R. RepeatModeler-1.0.8, http://www.repeatmasker.org/RepeatModeler.html) including RECON (v1.08) (Bao and Eddy, 2002), RepeatScout (v1.0.3) (Price et al., 2005) and Tandem Repeat Finder (TRF)

(v4.07b) (Benson, 1999) was used to construct a *de novo* repeat library from Asia I species genome scaffolds. This Asia I species-specific repeat library was then used by RepeatMasker to identify and classify the additional high and medium copy repeats in the genome. The identified repeats from both steps were merged and reprocessed using RepeatMasker. Tandem repeats including satellites, simple repeats and low complexity repeats were also predicted using TRF using Match=2, Mismatch=7, Delta=7, PM=80, PI=10, Minscore=50 and MaxPeriod=12.

5.2.5 Analysis of non-coding RNAs

Non-coding RNAs (ncRNAs) were predicted using INFERNAL (v1.1) (Nawrocki et al., 2009) and tRNAscan-SE (v1.3.1) (Lowe and Eddy, 1997) software packages. There were four types of ncRNAs predicted in Asia I species genome includes microRNA (miRNA), ribosomal RNA (rRNA), transfer RNA (tRNA) and small nuclear RNA (snRNA). MicroRNA (miRNA), rRNA and snRNA were predicted by INFERNAL using the Rfam database (v11) (Griffiths-Jones et al., 2005). Transfer RNAs (tRNAs) were predicted by tRNAscan-SE using eukaryote parameters and a covariance model (CM) which scores based on their secondary structures and sequence.

5.2.6 Asia I genome annotation

Gene structure predictions for the Asia I species genome was performed using the MAKER2 (v2.31) (Holt and Yandell, 2011). MAKER2 is a portable and easily configurable automated genome annotation pipeline. The MAKER2 pipeline was configured to run in four steps: 1) repeat masking the genome 2) *ab initio* gene prediction 3) Expressed Sequence Tags (ESTs) and/or protein evidence alignment and finally, 4) revising the final gene models. Repetitive elements of the Asia I species genome were first identified and masked using RepeatMasker (v4.0.5) (http://www.repeatmasker.org/) with the *de novo* repeat library from the Asia I species genome generated using RepeatModeler v1.0.8 (http://www.repeatmasker.org/RepeatModeler.html). SNAP (v2010-07-28) (Korf, 2004) and Augustus (v3.1.0) (Stanke et al., 2006) were employed to produce *ab initio* gene models. ESTs and proteins were derived from the RNA sequencing experiments of the Asia I species and aligned using BLASTX and TBLASTN integrated into MAKER2. Core eukaryotic genes identified by CEGMA (Parra et al., 2007, 2009) and BUSCO (Simao et al., 2015) were also supplied as a protein evidence for TBLASTN search into MAKER2. The MAKER2 predicted gene models from the Platanus assembly were further evaluated using CEGMA (Parra et al., 2007, 2009) and BUSCO (Simao et al., 2015). Finally, evidence derived from EST and protein homology alignments, and *ab initio* gene predictions were then integrated and revised in MAKER2 to generate a consensus gene set (Official Gene Set: OGS v1.1).

Functional annotation of the MAKER2 predicted proteins was performed using BLASTP top-hits to the nr database with the E-value cut-off of $1E^{-05}$. Gene motifs and domains were annotated by InterProScan (v5.15) (Jones et al., 2014) against several publicly available databases including Pfam, ProDom, TIGRFAMs, PIRSF, HAMAP, PRINTS, PANTHER, Gene3D, SUPERFAMILY, PROSITE, COILS and SMART. The KEGG annotation was also performed at KAAS server using SBH method (Moriya et al., 2007). The KEGG pathway for each gene was derived from the best KO hit.

5.2.7 Identification of gene orthologs

To identify orthologous groups and potential gene family expansion and contraction within the order Hemiptera, insect lineages with sequenced genomes were compared. Complete proteomes of Asia I species (OGS v1.1), *A. pisum*, *N. lugens* and *D. citri* were retrieved. All-vs-all BLASTP was performed using the OrthoMCL (v2.0.8) (Li et al., 2003) and the best reciprocal hits were determined. Orthologs, co-orthologs and inparalogs were classified using Markov clustering algorithm (MCL) (v14-137) (Enright et al., 2002).

5.2.8 Gene family analysis

Proteins beloning to the highly diverse gene family, GSTs, were identified from the BLASTP top-hits of MAKER2 gene set (OGS v1.1) against the nr database. The non-redundant set of Asia I species GSTs were aligned using MUSCLE (Edgar, 2004) to the corresponding protein orthologs from 10 other insects whose genomes were available. A maximum-likelihood phylogenetic tree was generated using FastTree 2 (Price et al., 2010) and illustrated with iTOL (Letunic and Bork, 2011).

5.3 Results and discussion

5.3.1 Genome sequencing and assembly

Genomic DNA extracted from Asia I species (adult female and males) was sequenced using the Illumina MiSeq 2500 approach with a coverage depth of 70x. A total of 210,053,096 reads were generated from a PE library with mean length 251 bp and insert size ranges from 450-750 bp. After removal of reads corresponding to endosymbionts, a total of 206,195,230 (98.16%) reads were obtained (Table 5.1, Appendix B, Figure B5.1). The initial draft assembly of Asia I genome by Platanus resulted in 1,571,703 scaffolds that spanned 828 Mbp with a N50 (the scaffold size N at which 50% of the genome assembly is in scaffolds longer than N) scaffold size of 2,110 bp, 1.18% gaps, and GC content of 39.57% (Table 5.1). DISCOVAR *de novo* assembler produced 1,974,766 scaffolds spanned 1.43 Gbp with N50 scaffold size of 1,386, 0.01% gaps and GC content of 39.94% (Appendix A, Table A5.1). The Platanus assembly therefore had a higher N50 than DISCOVAR *de novo* and the assembly size was closer to the estimated genome size of 640-690 Mbp (Chen et al., 2015; Guo et al., 2015) in comparison to DISCOVAR *de novo* assembly which was more than double the size. The above scaffold N50 values were significantly smaller than that of other sequenced insects such as *D.citri* (38 kbp) (Hunter et al., 2014), *A. pisum* (86.9 kbp) (IAGC, 2010) and *N. lugens* (356.6 kbp) (Xue et al., 2014). This is thought to be due to these other higher N50 values having been generated from MP and fosmid libraries. The scaffold N50 metric alone does not tell the sequence contiguity and therefore can not be used to select the best assembly. Together with the cumulative plot (Gurevich et al., 2013), the Platanus assembly (v1.1) was selected the best assembly with the longest scaffolds (steepest cumulative scaffold length curve, Appendix A, Figure B5.2) and used in the subsequent analysis in this chapter. The assembled genome of Asia I is A + T rich (60.44%), exhibiting only 39.57% GC content, which is higher than that of three published hemipteran genomes including *A. pisum* (29.6%) (IAGC, 2010), *N. lugens* (34.6%) (Xue et al., 2014) and *Diaphorina citri* (38.06%) (Hunter et al., 2014), and lower than *D. melanogaster* (42%) (Adams et al., 2000). In addition, the genome of Asia I is also larger than that of *A. pisum* (464 Mbp) (IAGC, 2010) and *D. citri* (485 Mbp) (Hunter et al., 2014) but smaller than *N. lugens* (1,141 Mbp) (Xue et al., 2014).

5.3.2 Assessment of genome assembly

The completeness of the Asia I genome assemblies from both assemblers was assessed using two different approaches, CEGMA and BUSCO. Using CEGMA, a set of 248 core eukaryotic genes were searched for in both genome assemblies which resulted in 81.45% and 67.74% of them being aligned to universal single copy orthologs as complete where-as 95.56% and 91.94% as partial in

Sequencing	Insert size	Length	Coverage	Total reads	Filtered
Illumina PE	450-750 bp	251 bp	70x	210,053,096	206,195,230
Assembly	Number	N50	GC%	Max. len.[a]	Total length
Contigs	1,747,531	809 bp	39.56 %	116.496 kbp	885.842 Mbp
Scaffolds	1,571,703	2.1 kbp	39.57 %	343.103 kbp	828.220 Mbp
Annotation	Number	Max. len.[a]	Avg. len.[b]	Total length	% of genome
Predicted genes	41,981	40.821 kbp	662 bp	27.77 Mbp	3.35 %
CDS	123,183	30,687 bp	215 bp	26.58 Mbp	3.19 %
Exons	124,106	30,687 bp	223 bp	27.72 Mbp	3.33 %
Introns	82,373	56,293 bp	2,200 bp	181.24 Mbp	21.87 %

[a]Maximum length, [b]Average length.

TABLE 5.1: Asia I species genome statistics: sequencing, assembly (v1.1) and annotation features (OGS v1.1).

Platanus and DISCOVAR *de novo* assembly respectively (Appendix B, Figure B5.3). Similarly, another set of 2,675 core arthropod genes were searched using BUSCO which resulted in 22% complete, 1% duplicate, 22% fragmented or partial and 55% missing in the DISCOVAR *de novo* assembly (Appendix B, Figure B5.3). In contrast, in the Platanus assembly, they were found as 35% complete, 1% duplicate, 32% fragmented or partial and only 32% missing (Figure 5.1). According to the CEGMA and BUSCO results of both assemblies, these supported the Platanus assembly having achieved a higher number of complete protein-coding genes in the genome compared to the DISCOVAR *de novo* assembly. The Platanus assembly also captured a higher number of fragmented and partial genes from CEGMA and BUSCO than the DISCOVAR *de novo* assembly (Appendix B, Figure B5.3). Based on these findings and previous assembly statistics comparisons of Platanus and DISCOVAR *de novo* assemblies, the Platanus assembly showed substantially better contiguity statistics compared to the DISCOVAR *de novo* assembly and was therefore selected for further analysis in this chapter.

The Platanus assembly was assessed further using the same set of datasets and programs. In addition to the scaffolds, the MAKER2 predicted gene models were also evaluated. Using both CEGMA and BUSCO for scaffolds and MAKER2 predictions, the complete set of genes were increased by 18% (BUSCO) and 18.94% (CEGMA) for MAKER2 gene set than that in scaffolds (Figure 5.1). In Chapter 4, a highly distinctive intron size and frequency distribution partly explained the difficulties encountered in genome assembly and suggested that the validated genes could produce more robust gene annotation. This is supported by the CEGMA and BUSCO results where MAKER2 gene set had a higher number of complete genes which were predicted using validated genes as evidence. The programs CEGMA and BUSCO failed to predict more complete genes from scaffolds because of the effect of size and frequency of introns on these programs.

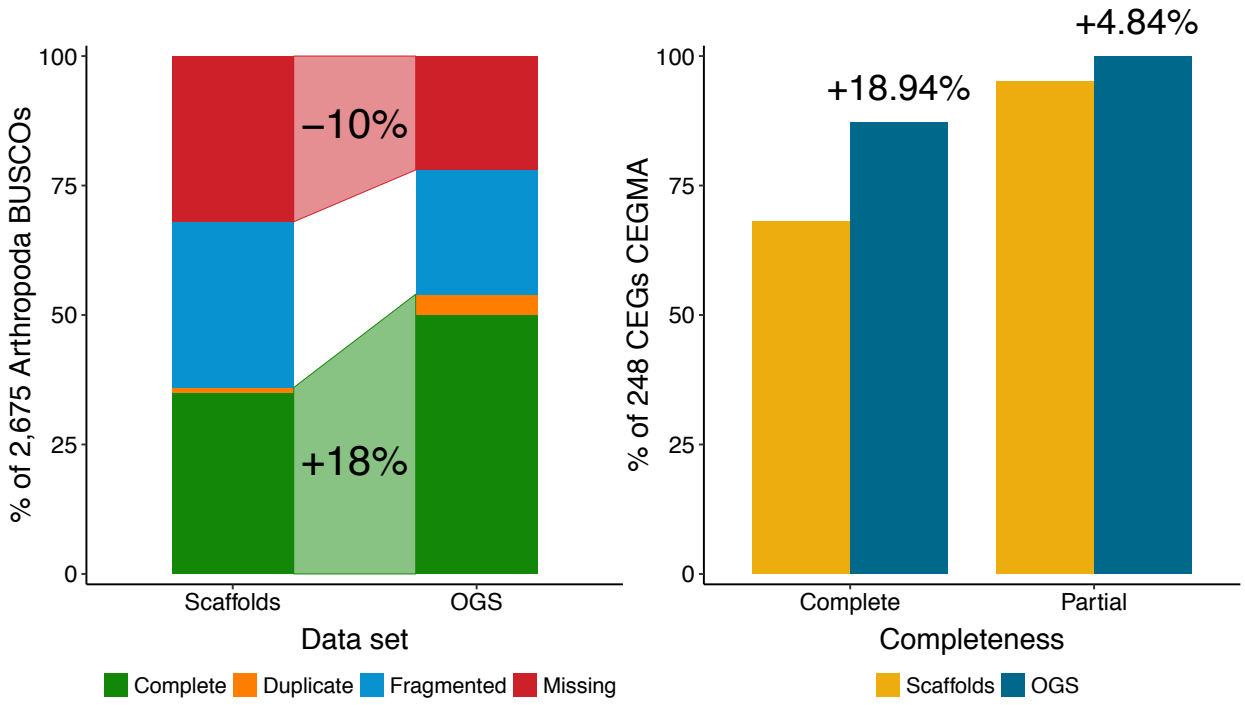


FIGURE 5.1: Assessment of Asia I genome Platanus assembly using BUSCO and CEGMA before (scaffolds) and after (OGS v1.1) MAKER2 gene model prediction.

5.3.3 Endosymbionts in Asia I species

To date, eight bacterial endosymbionts have been reported in *B. tabaci* and only four have had their genome sequence reported including *Portiera*, *Hamiltonella*, *Rickettsia* and *Cardinium*. No genome sequence is reported for the remaining four endosymbionts of *B. tabaci* including *Arsenophonus*, *Wolbachia*, *Fritschea* and *Hemipteriphilus*. To achieve the endosymbiont-free Asia I species genome library for assembly, all the genome sequencing reads from Asia I species were aligned to genome sequences of *Portiera*, *Hamiltonella*, *Rickettisia* and *Cardinium*, and three endosymbionts from other insect hosts including *Arsenophonus*, *Wolbachia* and *Hemipteriphilus*. There was no reference genome sequence available for *Fritschea*. Of the total Asia I species genome reads, only 1.83% reads showed matches with the above endosymbionts and were therefore filtered prior to the assembly step. Of the total aligned reads, the highest proportional of reads were corresponded to those of *Portiera* and its genomes reported (BT-B (21.40%), BT-Q (19.71%), B (20.97%) and Q (20.10%)) (Table 5.2). This result was expected as *Portiera* is obligate endosymbiont of all *B. tabaci* species. The second highest number of reads were aligned to *Wolbachia* endosymbionts including wPel (3.28%), wMel (0.08%), wHa (0.14%), wNo (4.90%), wRi (0.10%), wBm (0.0%) and wBol1-b (3.83%) followed by *Arsenophonus* endosymbionts (5.22%) and *Rickettisia* (0.22%). Only small proportional of reads were aligned to the *Hemipteriphilus* (0.007%) and *Hamiltonella* (0.0%) and it seems probable these are false hits. There was no corresponding read found to *Cardinium* sequences (Table 5.2).

Endosymbiont	Genome size	Mapped reads		Accession
		in pairs	broken pairs	
Portiera				
Portiera-BT-B	358,242	736,958	96,093	CP003708
Portiera-BT-Q	357,472	678,738	82,879	CP003835
Portiera-B	351,658	722,250	93,210	CP003868
Portiera-Q	350,928	692,262	83,163	CP003867
Wolbachia				
wPel	1,482,455	112,998	19,290	NC_010981
wMel	1,267,782	2,793	453	NC_002978
wHa	1,295,804	4,852	639	CP003884
wNo	1,301,823	168,887	6,995	CP003883
wRi	1,445,873	3,634	455	NC_012416
wBm	1,080,084	1	1	NC_006833
wBol1-b	1,377,933	131,932	19,611	CAOH01000000
Arsenophonus				
ArsenophonusNIL	2,953,863	178,658	4,129	JRLH01000000
ArsenophonusNAS	3,670,548	1,115	101	AUCC01000000
Rickettsia				
Ricketsia	914,329	7,674	3,793	AJWD01000000
Hamiltonella				
Hamiltonella	1,800,792	6	0	AJLH02000000
Hemipteriphilus				
Orientia	2,127,051	275	5	NC_009488
Cardinium				
Cardinium	1,012,588	0	0	NZ_CBQZ010000000

TABLE 5.2: Total endosymbiont genomic reads filtered from Asia I species genome reads via reference mapping approach.

5.3.4 Repetitive elements

Identification of repetitive elements is an essential step in any genome sequencing project, because unidentified repeats can affect the quality of predicted gene models and subsequent annotation analyses (Lorenzi et al., 2008). A significant proportion of the Asia I species genome is occupied

by repetitive elements (45.66%) including transposable elements (TEs) and tandem repeats (Table 5.3), which is a larger fraction than that has been found in *A. pisum* (33.3%) (IAGC, 2010) and *Tribolium castaneum* (42%) (Wang et al., 2008), but a lower fraction than that measured in published hemipteran insect, the brown planthopper (BPH), *N. lugens* (48.6%) (Xue et al., 2014). Most of these repeats were interspersed repeats, which account for 371 Mbp (44.88%). Of the remaining repeats, unclassified were the most abundant and accounting for 336 Mbp (40.58%) followed by DNA elements (2.31%), long interspersed elements (LINEs) (1.79%), simple repeats (0.63%), low complexity (0.18%), short interspersed nuclear elements (SINEs) (0.13%) and long terminal repeat (LTR) elements (0.08%) (Table 5.3).

sequences:	1,571,703		
total length:	828,219,547 bp		
GC level:	39.57%		
bases masked:	378,204,637 bp		
Repeat type	**Elements**	**Length occupied**	**% of sequence**
SINEs:	7,913	1,060,582 bp	0.13 %
ALUs	0	0 bp	0.00 %
MIRs	0	0 bp	0.00 %
LINEs:	94,440	14,799,207 bp	1.79 %
LINE1	0	0 bp	0.00 %
LINE2	0	0 bp	0.00 %
L3/CR1	2214	231928 bp	0.03 %
LTR elements:	3,811	629,568 bp	0.08 %
ERVL	0	0 bp	0.00 %
ERVL-MaLRs	0	0 bp	0.00 %
ERV_classI	0	0 bp	0.00 %
ERV_classII	789	119,898 bp	0.01 %
DNA elements:	114,170	19,158,058 bp	2.31 %
hAT-Charlie	9,651	1,896,023 bp	0.23 %
TcMar-Tigger	0	0 bp	0.00 %
Unclassified:	2,317,841	336,081,535 bp	40.58 %
Total interspersed repeats:		371,728,950 bp	44.88 %
Small RNA:	4,740	465,778 bp	0.06 %
Satellites:	0	0 bp	0 %
Simple repeats:	117,641	5,253,744 bp	0.63 %
Low complexity:	27,631	1,513,412 bp	0.18 %

TABLE 5.3: General statistics of repetitive elements identified in Asia I species genome.

The activity of these TEs and their impact on evolution of eukaryotic genomes is substantial (Biémont and Vieira, 2006; Feschotte and Pritham, 2007). Previous studies suggest that the TE content could play a significant role influencing genome size in many insect species (Comeron, 2001; Fernández-Medina et al., 2011). The impact of TE content was seen in *Aedes aegypti*, where the genome was double in size due to presence of TEs (Nene et al., 2007).

5.3.5 Non-coding RNAs

In addition to the protein-coding genes, ncRNAs were also identified from the draft genome assembly of Asia I. A total of 990 ncRNAs, which accounted for approximately 0.011% of the Asia I genome, were identified including 52 miRNA, 430 tRNA, 416 rRNA and 92 snRNA (Table 5.4). The miRNA and snRNA genes play significant roles in the regulation of eukaryotic gene expression. Predicted ncRNAs in the Asia I genome were lower than that of other hemipterans such as in *A. pisum* (163 miRNA) (IAGC, 2010) and *N. lugens* (372 miRNA, 1,982 tRNA and 198 snRNA) (Xue et al., 2014). The Asia I ncRNAs might be underestimated as they were predicted from an imperfect match between a draft genome sequence and known ncRNAs. Another possible factor that may contribute to lower ncRNAs in Asia I is that Asia I may possibly possess unidentified ncRNAs.

Type	Count	Average[a]	Total[b]	% of genome
miRNA	52	2,471.33	4,804	0.00058
tRNA	430	73.95	31,799	0.00384
rRNA	416	3,210.85	45,640	0.00551
5S	326	94.98	30,965	0.00374
5_8S	88	132.86	11,692	0.00141
SSU	2	2,983	2,983	0.00036
snRNA	92	1,466.19	12,069	0.00146
CD-box	6	398.33	1,195	0.00014
HACA-box	1	106	106	1.3E^{-05}
splicing	85	961.85	10,768	0.00130

[a]Average length, [b]Total length.

TABLE 5.4: Non-coding RNAs in Asia I species genome.

5.3.6 Asia I species gene models

Gene models in Asia I species were predicted using MAKER2 annotation pipeline (Holt and Yandell, 2011), combining both evidence-based and *ab initio* methods. The two-pass workflow was used with the MAKER2 pipeline as described by an online group (Holt, 2015) and it seems to be the best

approach to incorporate multiple *ab initio* predictions along with the EST and protein alignments as an evidence from the same species or closely related species. While running second phase of MAKER2, it is very important to make a decision on whether to keep the *ab initio* gene predictions where no ESTs or protein evidence is available (with AED score > 1). Keeping only gene models with evidence (AED score < 1) will remove all novel proteins that were predicted by *ab initio* programs. MAKER2 can produce more than the expected number of genes when it allow all the predicted gene models without filtering based on AED score, and some of them may be genuine. Moreover, the appropriate AED score cut-off might be different for different species and would need to be considered as discussed by Holt and Yandell (2011).

The main aim of this chapter was to assemble and annotate Asia I species genome and produce a comprehensive gene set by combining evidence-based and *ab initio* gene models. A total of 41,981 proteins-coding genes (PCGs) (OGS v1.1) were predicted with an average gene size of 662 bp, a coding sequence size of 214 bp and 2.95 exons per gene. This predicted protein count is higher than that of in other sequenced hemipterans including *A. pisum* (33,267 PCGs) (IAGC, 2010) and *N. lugens* (27,571 PCGs) (Xue et al., 2014). However, coding density in Asia I species (3.19%) was higher than that in *N. lugens* (2.74%) and lower than *A. pisum* (6.45%) due to the variation in their genome sizes. These results suggest that the contiguity of the *de novo* assembly and *ab initio* gene prediction did not limit the ability to identify PCGs in the Asia I data. Variations in the PCGs counts across these three species may be a consequence of species-specific gene expansions or contractions that have arisen as a result of gene duplication or gene loss and subsequent divergence, which has been one of the major reason for phenotypic differentiation across species (Robertson et al., 2003; Hahn et al., 2007). However, another possibility that PCGs were overestimated in Asia I species genome could be because of false-positive gene predictions. These can arise because of interference of putative TEs, bacterial symbionts and heterozygous gene copies. Further analysis will be required for manual examination of predicted gene models to assure their accuracy. The likelihood of a higher number of gene predictions can be investigated further when additional genomes are sequenced for hemipteran species.

For all PCGs, 49.85% were assigned preliminary functions based on BLASTP homology at nr database, 51.32% had similarity to domains in InterPro database, 40.32% were assigned Gene Ontology (GO) terms, encompassing molecular functions (32.81%), biological processes (29.04%), and cellular components (22.76%) (Appendix A, Table A5.2; Appendix B, Figure B5.4). KEGG pathways were mapped to 12.88% PCGs and EC was obtained for 7.08% PCGs (Appendix A, Table A5.3). Over half (50.15%) of PCGs remained unidentified when searched against the nr database and they are expected to be correctly assembled, and annotated based on ESTs evidence. Only a small proportion (6.17%) of unidentified PCGs were indeed expressed, as identified from the transcriptome assembly at a sequence identity threshold of 95% (Appendix A, Table A5.4). However, EST evidence was not found for a large proportion (93.83%) of unidentified PCGs at present and the likelihood of over-prediction remained after the three rounds of MAKER2 annotation pipeline

using *ab initio* programs, SNAP and AUGUSTUS. These unidentified PCGs can be assessed and evaluated using RT-PCR and sequencing analysis.

5.3.7 Asia I species genome complexity

Several challenges have been faced during the assembly and annotation of the Asia I species genome due certainly in part to two genomic features, namely large introns as well as high degree of repetitive sequences. Introns have been identified as a major component of any eukaryote genome including even smaller genomes from the taxonomically diverse lineages (Venter et al., 2001; Stein et al., 2003). Intron frequencies and their substantial proportion of the genome were compared across seven insects including *B. tabaci*, *A. pisum*, *N. lugens*, *D. citri*, *T. castaneum*, *P. humunus* and *Z. nevadensis* to assess their genome complexity (Figure 5.2). The intron frequency distribution shows there is no positive correlation between their frequencies and genome sizes as the higher the intron frequency does not always mean a longer genome (Figure 5.2, scatter plot). For example, *A. pisum* had the highest intron frequency of 146,873 but the genome size is lower than that of *B. tabaci*, *N. lugens*, *D. citri* and *Z. nevadensis*. However, the proportion of genome they account for was consistent with their genome sizes as the longer the genome, the longer the proportion (Figure 5.2, barplot). Despite their different genome sizes, two hemipteran insects, *B. tabaci* (Asia I, 828 Mbp draft genome assembly) and *N. lugens* (1,140 Mbp) showed similar distribution of introns in their genomes, 26.27% for *B. tabaci* and 24.38% for *N. lugens*. This is highly similar with the human genome (24%) though it is three times longer than these two insects (Venter et al., 2001). Interestingly, *A. pisum* was found with the highest number of introns and the highest percentage they occupied in the genome. The intron frequency and total intron sizes were shown to vary considerably across all seven insects which has also been commonly found across animal taxa through comparison of intron composition of whole genomes between and within taxa (Deutsch and Long, 1999; Lynch and Conery, 2003; Yandell et al., 2006; Zhu et al., 2009). A detailed study on large-scale genomic comparison will be interesting to understand the evolution and diversity of introns across an ever-increasing number of published genomes.

Despite this, genome may also contain repetitive sequences such as TEs, present in all genomes from bacteria to eukaryotes. Most commonly, these repetitive sequences represent a substantial fraction of the genome, although they vary within and between taxa (Bennetzen, 2005; Biémont and Vieira, 2006; Schnable et al., 2009; de Koning et al., 2011). These TEs have been found responsible for larger genome size in animals and plants. The proportion of TEs representing up to 77% are reported in the frog, *Pelophylax esculentus* genome which is relatively big 5.5-7.8 Gbp (Biémont and Vieira, 2006). In plants such as maize and rice, the proportion of TEs is variable and can account for up to 85% and 65% of the genome respectively (Bennetzen, 2005; Zuccolo et al., 2007; Lisch and Bennetzen, 2011).

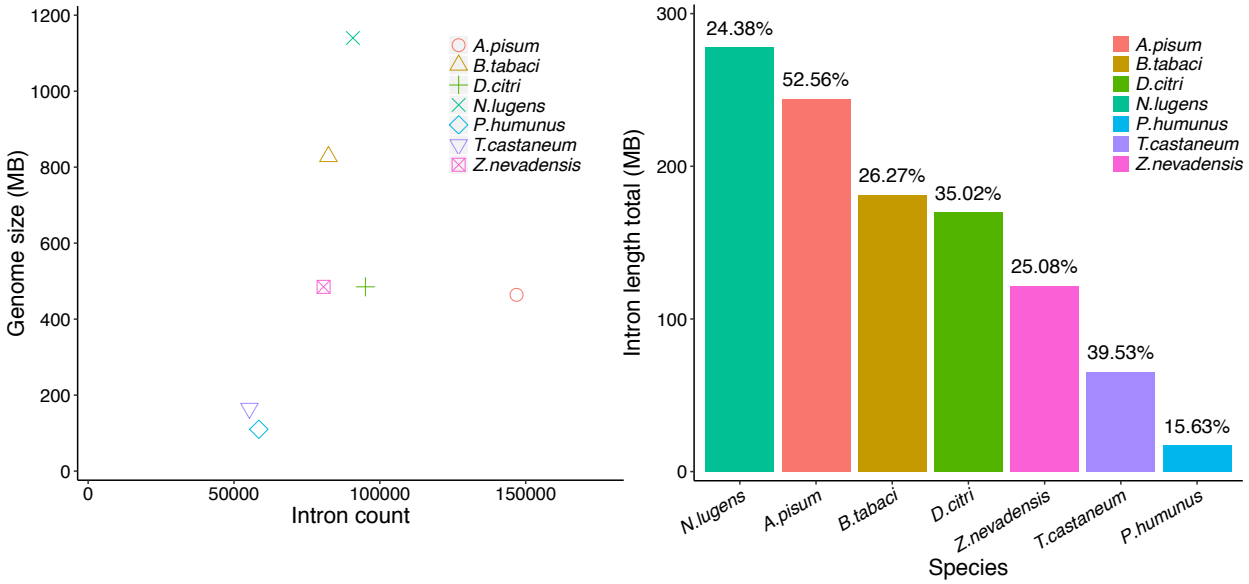


FIGURE 5.2: Distribution of intron counts and their total length in seven insect genomes. The scatterplot on left represents correlation between genome size and total number of introns for each species. The barplot on the right side illustrates that how much portion of genome occupied by introns in each species and the percentages are placed on top of each bar.

5.3.8 Gene orthology across hemipterans

Orthology analysis was performed among proteomes of four hemipteran species including *B. tabaci*, *A. pisum*, *N. lugens* and *D. citri* by clustering them into orthologous groups to determine hemipteran-specific single copy orthologs (SCOs) and whether *B. tabaci* genome has expanded or contracted gene families relative to the other hemipteran species. A total of 4,528 ortholog groups, representing 2,006 SCOs, were found conserved across all four hemipteran genomes (Figure 5.3). *B. tabaci* and *D. citri* shared the highest number of orthologs (583 orthologs) followed by 544 with *N. lugens* and 351 with *A. pisum*. These results agree with BLASTX top-hit species distribution where the highest number of BLASTX top-hits from hemipteran species was retrieved from *D. citri* (Appendix B, Figure B5.5). Among these four hemipteran species, the highest number of orthologs were found in *B. tabaci* (9,441) followed by *A. pisum* (9,164), *D. citri* (8,876) and *N. lugens* (8,498) (Figure 5.3). There were 1,575 paralog groups found unique to *B. tabaci*, representing 6,194 proteins (14.75%) of its predicted proteome (OGS v1.1).

These *B. tabaci* specific proteins were then analysed to determine GO terms (Ashburner et al., 2000) which are enriched in these proteins. Of the total 6,194 proteins, 3,052 (49.27%) proteins were found with a homolog, but not deemed sufficient of a homolog as to be considered an ortholog to, known proteins at NCBI, and 74.41% of these were annotated with 1,221 unique GO terms at level 2 (general function categories). These GO classification results were found consistent with the transcriptome of Asia I and previously sequenced transcriptomes of other

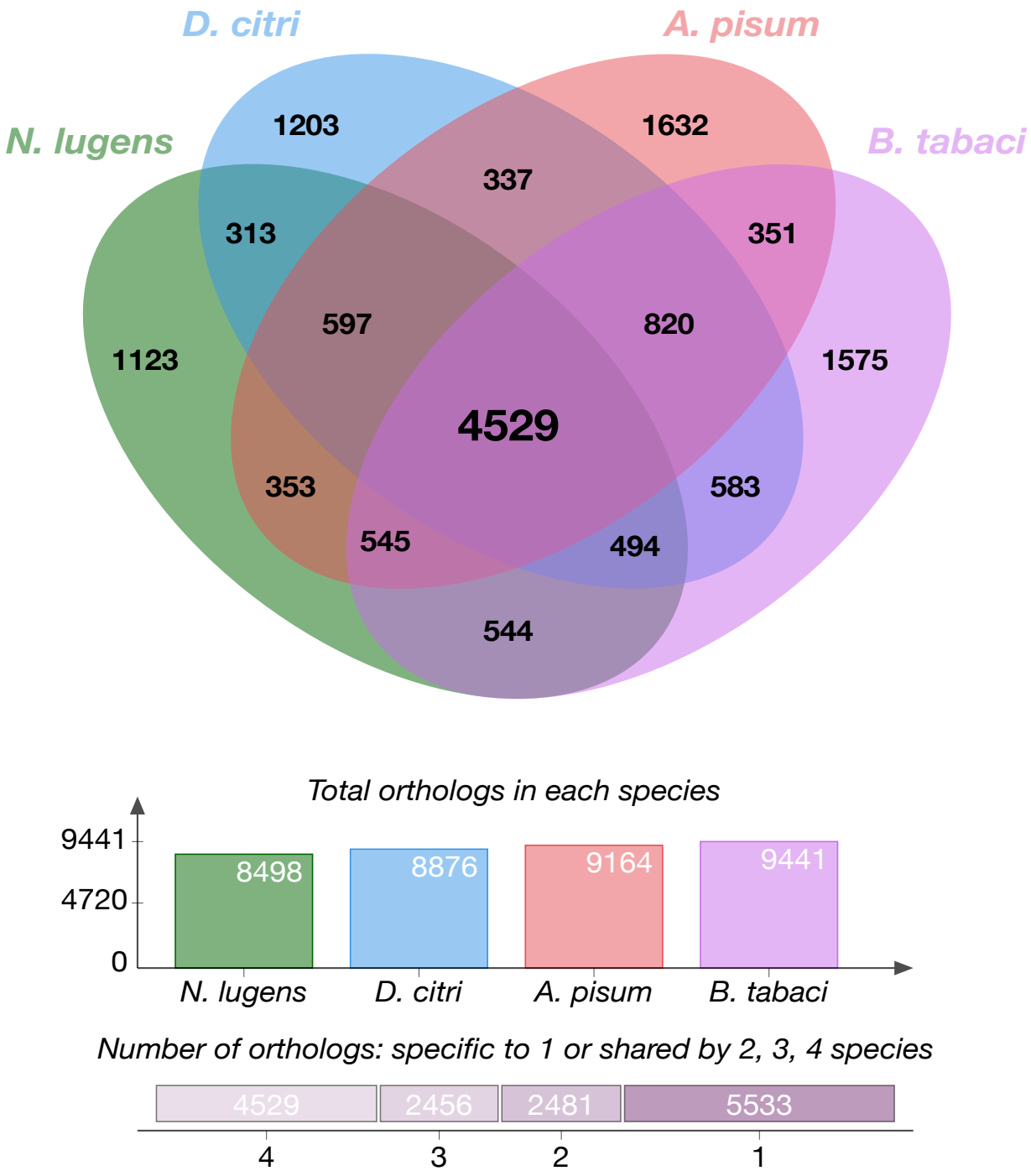


FIGURE 5.3: Orthology analysis of *B. tabaci* (purple) predicted proteins against the proteomes of *A. pisum* (red), *D. citri* (blue) and *N. lugens* (green). The barplot represents total number of orthologs identified in each species along with their counts. Orthologs count that were unique to species and shared between two, three or four species is also represented as horizontal bars at the bottom.

B. tabaci species where metabolic process, catalytic activity, binding and cell were the most abundant subcategories (Wang et al., 2010a, 2011, 2012). To identify the active pathways associated with these *B. tabaci* specific proteins, all 6,194 proteins were mapped with EC number to the

reference canonical pathways in KEGG. The EC classification revealed that the hydrolases (22%, 1,579 enzymes) group of enzymes were mostly represented in *B. tabaci* specific proteins followed by transferases (520 enzymes), oxidoreductases (376 enzymes), ligases (162 enzymes), lyases (144 enzymes) and isomerases (25 enzymes). A total of 129 KEGG pathways were assigned to 1,990 (32.12%) unique proteins. The highly represented pathways were thiamine metabolism (656 proteins), purine metabolism (326 proteins) and aminobenzoate degradation (146 proteins). Interestingly, thiamine metabolism pathway was not found in previous transcriptomes of MEAM1, MED and Asia II 3 species (Wang et al., 2010a, 2011, 2012) where-as it was found activated in the Asia I transcriptome. Thiamine metabolism is the sub-pathway that lies under the pathway class 'metabolism of cofactors and vitamins'. Other sub-pathways from the 'metabolism of cofactors and vitamins pathway' were found enriched in MEAM1 and MED guts (Ye et al., 2014). In addition, several KEGG pathways were also found activated in Asia I species which includes drug metabolism - cytochrome P450, nicotinate and nicotinamide metabolism, ascorbate and aldarate metabolism, glycosylphosphatidylinositol(GPI)-anchor biosynthesis, metabolism of xenobiotics by cytochrome P450, tyrosine metabolism, other types of O-glycan biosynthesis, drug metabolism - other enzymes, porphyrin and chlorophyll metabolism, starch and sucrose metabolism, pentose and glucuronate interconversions, and retinol metabolism. These pathways have been identified as gut-specific pathways and their functions appeared very similar in MEAM1 and MED species (Ye et al., 2014), and it is very highly likely that these pathways also perform similar function in populations of the Asia I species. Among these gut-specific pathways in Asia I, metabolism of xenobiotics by cytochrome P450, drug metabolism - cytochrome P450 and drug metabolism - other enzymes were reported with their association in insecticide resistance in MED species (Nauen et al., 2002; Horowitz et al., 2005; Ghanim and Kontsedalov, 2007; Yuan et al., 2012).

5.3.9 Glutathione S-transferases (GSTs)

GSTs are a gene family of multifunctional enzymes involved in detoxifying cells of natural and artificial molecules in both eukaryotes and prokaryotes (Rogers et al., 1999; Ranson et al., 2000; Lumjuan et al., 2005). In general, GST-mediated reactions involve the conjugation of the thiol group from glutathione (GSH; y-glutamyl-cysteinyl-glycine) to hydrophobic and electrophilic toxicants including many insecticides, drugs and herbicides. By this mechanism, they can increase the solubility of the substrate and eliminate them from a cell, and specifically target GSH multidrug transporters. In insects, GSTs represents a highly diverse gene family and can be classified into two major classes: microsomal and cytosolic GSTs. The membrane-bound microsomal GSTs are structurally and evolutionarily distinct from the cytosolic GSTs. The cytosolic GSTs class is larger than the microsomal GSTs class and can be further divided into six major subclasses: delta, epsilon, omega, sigma, theta and zeta. Among these six subclasses, delta and epsilon are specific to insects

(Friedman, 2011), and thought to be involved in detoxification mechanisms (Enayati et al., 2005; Friedman, 2011; Nardini et al., 2012).

A total of 23 GSTs were identified in the *B. tabaci* Asia I genome, representing each of the six subclasses of cytosolic GSTs: 11 delta, two epsilon, one omega, four sigma, one theta and two zeta (Figure 5.4), and remaining two were microsomal GSTs. This number is the same as that found in *B. tabaci* MED species, which also has 23 GSTs (Yang et al., 2016). *B. tabaci* has more GSTs than found in any other currently available hemipteran genomes; *A. pisum* contains 22 GSTs (20 cytosolic GSTs and two microsomal GSTs) (IAGC, 2010; Ramsey et al., 2010) and *N. lugens* has only 11 GSTs (9 cytosolic GSTs and two microsomal GSTs) (Xue et al., 2014). Moreover, even the hymenopteran *A. mellifera* and *N. vitripennis* have less with 11 (Claudianos et al., 2006; HGSC, 2006) and 19 GSTs (Oakeshott et al., 2010) respectively, and the exopterogote parasite *P. humanus* contains 11 GSTs (Kirkness et al., 2010). However, this pattern is not followed as the dipterans have consistently expanded GST gene families including *D. melanogaster* and *A. gambiae*, which have 37 and 28 GSTs respectively (Ding et al., 2003). The detailed comparative study across available insect genomes may shed more light on GST expansion and contraction patterns and also differentiate their evolution either by ecological adaptation or by shared ancestry.

In hemipterans, the microsomal GSTs were equally distributed as two GSTs found in *B. tabaci*, *A. pisum* and *N. lugens*. These microsomal GSTs are structurally different in comparison with the cytosolic GSTs, although both major classes are involved in detoxification of xenobiotics and in protection against oxidative stress (Hayes et al., 2005). Of the total 21 cytosolic GSTs, delta GSTs were the largest class of GSTs in *B. tabaci* (11 GSTs), which is consistent with findings in the other insects (Ding et al., 2003; IAGC, 2010; Oakeshott et al., 2010; Xue et al., 2014). Interestingly, *B. tabaci* has the same number of delta GSTs that were identified in *A. gambiae* but one more than in *A. pisum*. Further, two epsilon GSTs were present in *B. tabaci* genome but, as in *A. pisum* genome, no epsilon GSTs were identified. These delta and epsilon GSTs are unique to insects and occupy the majority of the GSTs consistent with their role in detoxification of insecticides (Tene et al., 2013; Djègbè et al., 2014). In *B. tabaci* MED species, a high level of resistance to thiamethoxam, a second generation neonicotinoid insecticide, has been reported due to the over-expression of several GSTs (Yang et al., 2016). Activities of multiple GSTs were also found associated with cross-resistance to other insecticides including imdacloprid, nitenpyram and acetamiprid in *B. tabaci* (MED species) (Yang et al., 2016). Similarly, in *A. gambiae*, GSTD1, the delta GSTs was found associated with insecticide metabolism via directly detoxifying DDT and pyrethroid (Huang et al., 2012; Tene et al., 2013). Sigma class of GSTs were the second largest GSTs in *B. tabaci* (four GSTs). Structural and catalytic roles of sigma GSTs have been defined in previous studies (Clayton et al., 1998; Singh et al., 2001). Additionally, some members of this class were found actively associated with a 4-hydroxynonenal, a by-product of lipid peroxidation, and therefore they thought to be involved in protecting from oxidative stress (Singh et al., 2001). The global distribution of the *B. tabaci* is partially because of its high fitness parameters such as high fecundity, wide range of plant hosts and

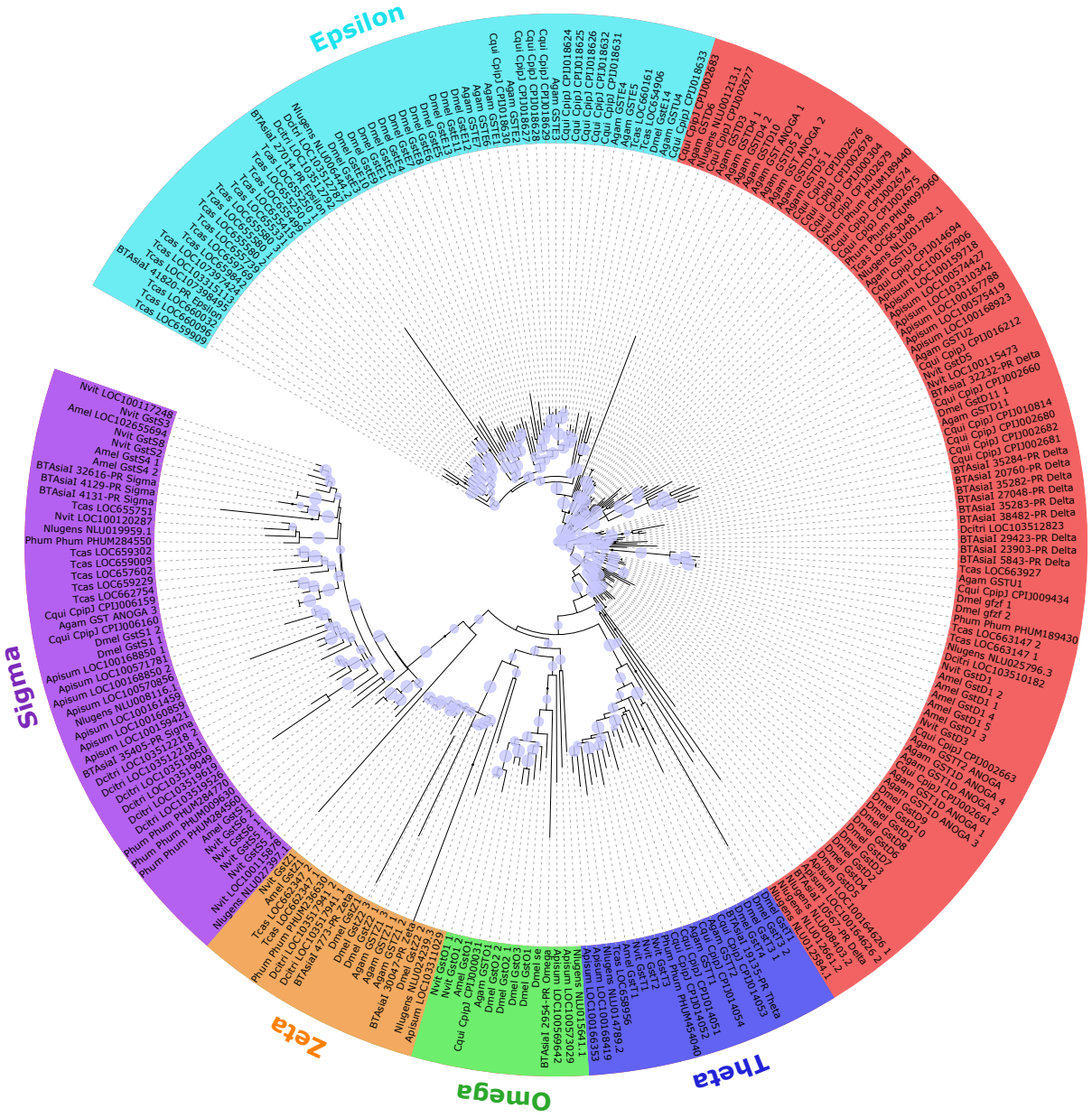


FIGURE 5.4: Phylogeny of GSTs across 11 insects including four hemipterans: BTAsiaI, *B. tabaci*; Apisum, *A. pisum*; Dcitri, *D. citri*; Nlugens, *N. lugens*; three dipterans: Dmel, *D. melanogaster*; Agam, *A. gambiae*; Cqui, *Culex quinquefasciatus*; two hymenopterans: Amel, *A. mellifera*; Nvit, *N. vitripennis*; one coleopteran: Tcas, *T. castaneum*; one phthirapteran: Phum, *P. humanus*; The tree was rooted and supported by the bootstrap values represented as circles at each node where the value is >=50.

ability to survive at a high rate under different environmental stresses (Lu and Wan, 2008). The remaining GSTs represented in *B. tabaci*, theta, zeta and omega, are ubiquitously distributed in nature which indicate their roles in endogenous metabolic processes. An expansion in zeta class of GSTs was observed in *B. tabaci* in comparison with *A. pisum*. The zeta GSTs are distributed as one or two in most insects and play an important role in the degradation of phenylalanine and catalysis of tyrosine (Board et al., 1997). One omega class of GSTs was identified in *B tabaci*, similar to *N.*

lugens and *A. gambiae* but less than two omega GSTs found in *A. pisum* and *A. mellifera*. In *Apis cerana*, the induced expression of omega GSTs by various abiotic stresses suggested their protective role against oxidative stresses (Brock et al., 2013).

5.3.10 Sex determination gene in Asia I species

The complete coding sequence of sex determination gene, Transformer-2 sex-determining gene *tra2* in Asia I species, was predicted based on its sequence homology with the published *tra2* gene in MEAM1 species (Xie et al., 2014). The coding sequence was further confirmed from genomic mapping and resulted in two splice variants of *tra2* in *B. tabaci* Asia I clade (*tra2-266, tra2-143*) (Figure 5.5A). The largest transcript (*tra2-266*) was 801 bp and contains eight exons and seven introns whereas *tra2-143* was 432 bp and contains four exons and three introns. The total size of introns was 12,884 bp and 4,131 bp for *tra2-266* and *tra2-143* respectively (Figure 5.5A).

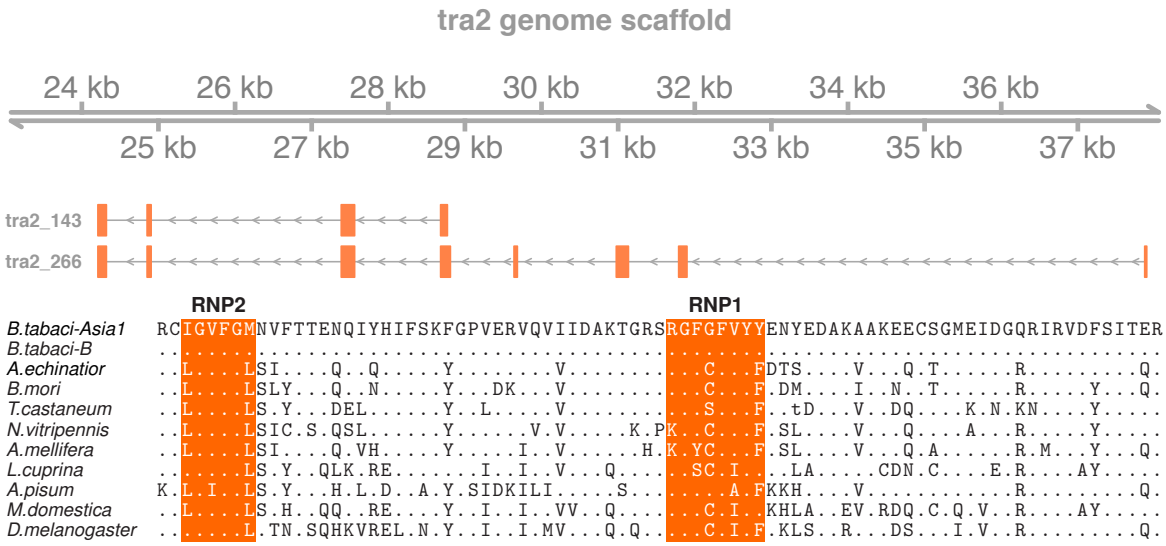


FIGURE 5.5: Transformer 2 (*tra2_143* and *tra2_266*) - putative sex determination gene in *B. tabaci*. Sequence alignment of RNA-binding Domain (RBD) of *tra2* gene across hemipteran (*B. tabaci* Asia I and MEAM1, *A. pisum*), hymenopteran (*A. mellifera*, *N. vitripennis*, and *A. echinatior*), lepidopteran (*B. mori*), coleopteran (*T. castaneum*) and dipteran (*D. melanogaster*, *L. cuprina* and *M. domestica*). The identical residues are represented as dots and RNP elements are highlighted in orange colour.

Sequence comparison of *B. tabaci tra2-266* gene with other insects revealed conservation only in structural features within RNA-binding domain (RBD), which indicate RBD is evolutionary constrained (Figure 5.5B). Due to the evolutionary constrain, only RBD domain of *tra2* gene was compared across hemipteran (*B. tabaci* Asia I and MEAM1, *A. pisum*), hymenopteran (*A. mellifera*, *N. vitripennis*, and *Acromyrmex echinatior*), lepidopteran (*Bombyx mori*), coleopteran (*T. castaneum*) and dipteran (*D. melanogaster*, *Lucilia cuprina* and *Musca domestica*) insects (Figure 5.5B). The pairwise comparison shows 100% sequence identity between two *B. tabaci* species Asia

I and MEAM1. The closest homolog to Asia I was *A. echinatior* (78%) followed by *B. mori* (74%), *T. castaneum* (73%), *N. vitripennis* (70%), *A. mellifera* (70%), *L. cuprina* (69%), *A. pisum* (66%), *M. domestica* (62%) and *D. melanogaster* (60%). Sequence comparison of RBD domains of *tra2* gene across insects revealed various base mutations within ribonucleoprotein (RNP) elements, which are the designated binding nucleotides of RNA. There were two putative RNP elements, RNP1 (eight amino acid) and RNP2 (six amino acid) found within RBD domain (Figure 5.5B), which were structured like a barrel ($\beta\alpha\beta\beta\alpha\beta$). These RNP elements bind to the *dsx* repeat element in *D. melanogaster* and other dipteran insects (Nagai et al., 1990; Amrein et al., 1994) but *dsx* repeat motif was found absent in *A. mellifera* ((Saccone et al., 2008; Concha et al., 2010; Permpoon et al., 2011; Nissen et al., 2012). RNP1 element of *B. tabaci* Asia I is composed of eight amino acids of which four amino acids were found different in *A. mellifera* and three in *D. melanogaster*. The mutation of an essential amino acid arginine (R) to lysine (K) was only found in two hymenopteran insects, *A. mellifera* and *N. vitripennis*, which is an essential residue for female *dsx* splicing in *D. melanogaster* (Amrein et al., 1994; Nissen et al., 2012). RNP2 element is six amino acids long and differs from other insects at first and six amino acids (I to L and M to L) and third amino acid in *A. pisum* (V to I) (Figure 5.5B).

The genetic structure of this *tra2* gene was found with its largest intron in *B. tabaci* (12,884 bp) compared to that of in other insects including *A. gambiae* (5,444 bp), *D. melanogaster* (3,256 bp) and *A. pisum* (977 bp), but smaller than that in *A. mellifera* (18,967 bp). This *tra2* gene was identified as an essential regulator gene associated with female splice regulation at two levels of the sex determination system of honeybee (Nissen et al., 2012). The similar results were also seen in beetles, where female specific transcript contains longer intron than male-specific transcript of *tra2* gene (Shukla and Palli, 2012). The recent study on determining sexual differences in whiteflies at transcriptome level revealed *tra2* gene as a more highly expressed gene in females than males (Xie et al., 2014). In this study, the genomic DNA was isolated from both adult females and males *B. tabaci*, and therefore it is still unclear whether the *tra2* gene of *B. tabaci* is male-specific or female-specific. It will be interesting to see the intron-exon structure in male-specific *tra2* gene of *B. tabaci* to determine its regulatory role in sex determination system of whiteflies.

CHAPTER 6

Mitochondrial and endosymbiont genomes assembled from *Bemisia tabaci* Asia I genome data

6.1 Introduction

The major aim of this study was to obtain the genome of Asia I species, although genome sequences of mitochondria and three bacterial symbionts including one primary endosymbiont (*Portiera*) and two secondary endosymbionts (*Wolbachia* and *Arsenophonus*) were also obtained from the same genomic sequence libraries.

6.1.1 Mitogenome of Asia I species

To date a partial fragment of the *mtCOI* gene, has been used extensively to distinguish *B. tabaci* species within the complex of cryptic species comprising at least 37 morphologically similar species (Dinsdale et al., 2010; Barro et al., 2011; Hu et al., 2011a; Alemandri et al., 2012; Liu et al., 2012; Parrella et al., 2012; Firdaus et al., 2013; Hu et al., 2014). However, a better understanding of genetic relationships within the *B. tabaci* complex can be achieved via comparing species through more extensive sequences, ideally nuclear as well as mitochondrial sequences. From the genomic sequence data generated, it was straightforward to assemble a complete mitogenomes. This was done as it is not only interesting for phylogenetic analysis, but could also benefit the design of improved primers for PCR amplification. To date, only four whitefly species have their mitogenome sequence reported, which includes New World I (Thao et al., 2004), MED (Wang et al., 2013), Asia I (Tay et al., 2016) and *Bemisia afer* (Wang et al., 2016b). The comparison of these four mitogenomes revealed similar gene arrangement in all four species with two differences as tRNA-Ser2 was absent in *B. afer* and tRNA-Arg was on the "negative" strand in the MED species (Tay et al., 2016).

Here in this chapter, the complete mitochondrial genome of Asia I (hereafter 'mtAsia I') species of *B. tabaci* species complex is reported. In addition, four published mitogenomes of the same complex were compared with mtAsia I to reveal the similarity and diversity across mitochondrial genes and their arrangements.

6.1.2 *Portiera* genome from Asia I species

Portiera is the primary endosymbiont harboured by whiteflies, including members of the *B. tabaci* cryptic species complex (Baumann, 2005). Previous studies suggested that the *Portiera* endosymbiont is vertically transmitted and provides with essential nutrients to their hosts (Baumann, 2005). The genome sequence of this primary endosymbiont revealed genes associated with biosynthesis of carotenoids in *B. tabaci* host (Sloan and Moran, 2012a). To date, only two *B. tabaci* species including MEAM1 and MED have their primary endosymbiont genome reported (Jiang et al., 2012; Sloan and Moran, 2012a; Santos-Garcia et al., 2012) and compared with each other to reveal their genetic diversity across different species (Jiang et al., 2012, 2013). The comparative genomics of these primary endosymbionts will greatly benefit the better understanding of their genetic relationships within the *B. tabaci* species complex. The *Portiera* genome of Asia I species will provide an opportunity for an exhaustive identification of bacterial origin genes in the host genome (when the Asia I species genome becomes available) showing persistent associations with the endosymbionts. Here in this chapter, the genome sequence of *Portiera* from Asia I species is also reported along with its sequence comparison with genomes of two primary endosymbionts from MEAM1 (GenBank: CP003868 (hereafter 'B'), CP003708 (hereafter 'BT-B')) and MED species (GenBank: CP003867 (hereafter 'Q'), CP003835 (hereafter 'BT-Q')).

6.1.3 *Wolbachia* genome from Asia I species

Wolbachia strains are the most widely distributed rickettsial endosymbionts across the major arthropod classes (Hilgenboecker et al., 2008) and are particularly prevalent in herbivorous insects of the Hemiptera suborder Sternorrhyncha (aphids, whiteflies, psyllids, scales and mealybugs) (Moran, 2001). They are transmitted maternally and enhance this process by manipulating host reproductive systems by, for instance, feminization of genetic males, parthenogenesis, cytoplasmic incompatibility, and killing male progeny (Stouthamer et al., 1999; Werren et al., 2008). Recent studies show that *Wolbachia* strains can be either parasitic or mutualistic (Glaser and Meola, 2010; Hosokawa et al., 2010). *Wolbachia* strains have been reported to provide mutualistic nutritional benefits to insect-host species such as the bedbug (Hosokawa et al., 2010), increased fitness in the uzifly (Guruprasad et al., 2011), leaf-mining moth (Kaiser et al., 2010) and mosquito (Dobson et al., 2002), and by increasing host stem-cell proliferation (Fast et al., 2011).

Wolbachia strains have been used successfully as environment-friendly biotechnologies to control disease vectors. For example, the mosquito *A. aegypti* infected with wMel from *D. melanogaster*, demonstrates a reduced ability to become infected with the dengue virus with a subsequent reduction in dengue virus transmission by these mosquito populations (Walker et al., 2011). They are as a result also targets for genetic engineering strategies for vector control (Kean et al., 2015).

The function of the secondary endosymbiont, *Wolbachia* in *B. tabaci* remains poorly defined as there is no genomic sequence available for this endosymbiont from any *B. tabaci* species. To date, the complete genome sequences of 10 *Wolbachia* strains have been sequenced: wMel, wAu, wBm, wPip-Pel, wRi, wOo, wOv, wCle, wHa and wNo, and the genomes of another 10 *Wolbachia* strains have been assembled/annotated as drafts: wDa, wMelPop, wRec, wDs-VAL, wGm, wAlbB, wCp Mol, wPip JHB, wDi and wBol1-b (http://www.ncbi.nlm.nih.gov/genome/?term=Wolbachia). To better define and understand the physiological role of *Wolbachia* in *B. tabaci* species, the genome sequence of the *Wolbachia* endosymbiont from the Asia I species was assembled and annotated in this chapter. The draft genome will assist research on elucidating the endosymbiotic relationship between *B. tabaci* and *Wolbachia*.

6.1.4 *Arsenophonus* genome from Asia I species

Members of the genus *Arsenophonus* (Enterobacteriaceae) are intracellular endosymbionts known to infect approximately 5% of the arthropods (Nováková et al., 2009). The genus *Arsenophonus* possess four different phenotypes including phytopathogenicity (Zreik et al., 1998), male-killing (Gherna et al., 1991), non-specific horizontal transmission (Thao and Baumann, 2004) and obligate endosymbiont (Trowbridge et al., 2006). These phenotypes indicate that the members of the genus *Arsenophonus* represent a diverse and widespread endosymbiotic lineage that serves as a model to study mechanisms by which molecular evolution influences symbiotic relationships with their hosts.

B. tabaci harbours *Arsenophonus* as a facultative endosymbiont and its multiple infections and evidence for horizontal transmission in *B. tabaci* has been reported (Thao and Baumann, 2004). The association of this endosymbiont in manipulation of host reproduction of *B. tabaci* has also been reported (Thierry et al., 2011). To date, there is no genomic information available for this endosymbiont from any species of the *B. tabaci* complex. Only two genomes of *Arsenophonus* strains have been sequenced from any insect: *Arsenophonus nasoniae* (Wilkes et al., 2010) from a parasitic wasp and *Arsenophonus* endosymbiont of the brown planthopper *N. lugens* (Xue et al., 2014).

To provide an opportunity to explore the evolution and the ecological spread of this pervasive endosymbiont across species of the cryptic complex, the genome sequence of the *Arsenophonus* endosymbiont of Asia I species was assembled and annotated in this chapter.

6.2 Methods

6.2.1 Mitogenome of Asia I species

Using NGS methodology, the genome sequence of Asia I species was sequenced and assembled as described in Chapter 5. The mitogenome of mtAsia I was obtained as a single scaffold from the draft genome assembly of Asia I species based on BLAST homology with published mitogenomes. The genome sequencing reads were mapped to the scaffold using CLC genomics workbench (v7.0.4) (CLC bio, Aarhus, Denmark) to identify single nucleotide polymorphisms (SNPs) and the assembly coverage for that scaffold. The mtAsia I mitogenome was annotated using MITOS (Bernt et al., 2013) with the codon Table 5 invertebrate. The annotated gene tracks were visualized using BRIG (Alikhan et al., 2011) along with their GC contents and read mapping coverage. The mitogenomes of Asia I, New World I and MED species and, *B. afer* were compared with mtAsia I using BLASTN (Altschul et al., 1990) and visualized using BRIG. Two major phylogenetic methods, Bayesian inference (BI) and Maximum Likelihood (ML) were used to compare topologies between two datasets: partial *mtCOI* with 657 bp and 13 PCGs with 10,961 bp, from of mtAsia I, New World I and MED species and, *B. afer*. The multiple alignment of *mtCOI* and 13 PCGs from these four species were performed separately by MAFFT (v7.299) (Katoh and Standley, 2013). BI was conducted using MrBayes (v3.1.2) (Ronquist and Huelsenbeck, 2003). Bayesian analysis was performed in combination with an exact model of molecular evolution as well as a rapid approximation of posterior probability tree using Markov Chain Monte Carlo (MCMC) (Huelsenbeck et al., 2004). MrBayes was run using four incrementally heated chains and run for 30 million generations with tree sampling every 1,000 generations and a burn-in of 7,500 trees. All runs reached a plateau in likelihood score, which was indicated by the standard deviation of split frequencies. In addition, Tracer (1.5) (Drummond and Rambaut, 2007) was used to test the convergence of the run, all effective sample size (ESS) values were larger than 200 for each of runs.

6.2.2 *Portiera* genome from Asia I species

The genomic DNA reads from the *B. tabaci* Asia I species were aligned using Burrows-Wheeler Aligner (BWA) (v0.7.12) (Li and Durbin, 2009) to the complete genome sequences of two *Portiera* from MEAM1 and MED species to extract *Portiera*-specific reads as described in Siozios et al. (2013). These raw reads were assembled into contigs using CLC genomics workbench (v7.0.4) (CLC bio, Aarhus, Denmark) and joined into scaffolds using SSPACE (v3.0) (Boetzer et al., 2011). The scaffolds were aligned to reference genomes of MEAM1 and MED species using the genome finishing module in CLC genomics workbench to link scaffolds by overlap edges. The paired end reads were mapped back to the resulted complete scaffold to close the gaps and correct base calling.

The complete *Portiera* genome was annotated at Rapid Annotation using Subsystem Technology (RAST) (v2.0) (Overbeek et al., 2014) to predict protein coding genes (PCGs), and tRNAs and rRNAs were identified by tRNAScan-SE (v1.21) (Lowe and Eddy, 1997) and RNAmmer (v1.2) (Lagesen et al., 2007) respectively. Tandem Repeats Finder (Benson, 1999) was used to identify tandem repeats in the *Portiera* genome of *B. tabaci* species including Asia I, MEAM1 and MED. The complete proteomes of *Portiera* from Asia I, MEAM1 and MED species were compared using nucmer (Kurtz et al., 2004) and protein orthologs were clustered using Proteinortho (Lechner et al., 2011). Phylogenetic analysis reported in this study was carried out using MAFFT (v7.182) (Katoh and Standley, 2013) and tree was constructed based on RAxML BS and ML (Stamatakis, 2006) algorithm using GTRGAMMA substitution model with 1,000 bootstrap replicates.

6.2.3 *Wolbachia* genome from Asia I species

The genomic DNA reads from Asia I species were aligned to the genome sequences of 20 *Wolbachia* strains to extract *Wolbachia*-specific reads using BWA (v0.7.12) (Li and Durbin, 2009) as the similar approach is described in Siozios et al. (2013). DISCOVAR *de novo* (Weisenfeld et al., 2014) was used to perform *de novo* assembly of these reads into contigs. The raw reads were mapped back to the assembled contigs using the BWA (Li and Durbin, 2009) to correct potential single nucleotide polymorphisms (SNPs). RAST (v2.0) (Overbeek et al., 2014) and NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (v2.10) (Angiuoli et al., 2008) were used for the genome annotation. tRNAs and rRNAs were predicted using tRNAScan-SE (v1.21) (Lowe and Eddy, 1997) and RNAmmer (v1.2) (Lagesen et al., 2007) respectively. Subsystems were annotated for each strain using RAST and compared to identify strain-specific and shared subsystems across 21 strains. The Cluster of Orthologous Groups (COGs) in all *Wolbachia* strains were identified using Proteinortho (Lechner et al., 2011) with a cut-off of 50% for %identity and %coverage, and an E-value 1E^{-05}. Phylogenetic analysis was performed by MUSCLE (Edgar, 2004) and RAxML (Stamatakis, 2006).

6.2.4 *Arsenophonus* genome from Asia I species

Genomes of two *Arsenophonus* strains: *Arsenophonus nasoniae* (Wilkes et al., 2010) (GenBank accession: NZ_AUCC01000000) and *Arsenophonus* endosymbiont of *N. lugens* (Xue et al., 2014) (GenBank accession: NZ_JRLH01000000) were used to extract *Arsenophonus*-specific reads from the total genomic reads of Asia I species using BWA (v0.7.12) (Li and Durbin, 2009) and the similar approach as described in Siozios et al. (2013). The assembly was performed using DISCOVAR *de novo* (v52488) (Weisenfeld et al., 2014) and the annotation was carried out using RAST (v2.0) (Overbeek et al., 2014). The tRNAs and rRNAs were predicted using tRNAScan-SE (v1.21) (Lowe and Eddy, 1997) and RNAmmer (v1.2) (Lagesen et al., 2007) respectively.

6.3 Results and discussion

6.3.1 Mitogenome of Asia I species

6.3.1.1 Structure and organization of Asia I species mitogenome

The mitogenome of mtAsia I was identified with the 99.0% sequence identity to the published Asia I species mitogenome using BLAST and obtained as a single scaffold from the draft genome assembly. The completeness of mtAsia I mitogenome was assessed by read mapping against the scaffold and the start and end positions were identified from mapped reads. The total length of the mtAsia I mitogenome was 15,453 bp, which is longer than the reported one for Asia I (15,210 bp) (Tay et al., 2016), *B. tabaci* New World I (15,322 bp) (Thao et al., 2004) and *B. afer* (14,968 bp) (Wang et al., 2016b) but shorter than *B. tabaci* MED species (15,632 bp) (Wang et al., 2013). The annotation of the mtAsia I mitogenome by MITOS identified 37 genes in total including 13 protein coding genes (PCGs), two ribosomal RNAs (rRNAs) and 22 transfer RNAs (tRNAs) (Table 6.1), identified in most metazoan mitogenomes (Wolstenholme, 1992). As with other insects, mitogenomes of all three *B. tabaci* species were found to contain a very higher A+T content. The overall A+T content was 75.31% for mtAsia I, 75.67% for New World I, 75.17% for Asia I and 75.21% for MED species. This A+T content was highly similar among mitogenomes of *B. tabaci* species as well as across several dipteran mitogenomes (76-79%) (Oliveira et al., 2008). However, A+T content was found low in *B. afer* (65.66%) in comparison with other the *B. tabaci* species. In insects, the percentage of AT-skew (A-T/A+T) is associated with the codon positions, gene replication and direction, while the GC-skew (G-C/G+C) is affected by reversals in replication orientation (Wei et al., 2010).

The standard metazoan mitogenome contains five membrane-associated protein complexes associated with oxidative phosphorylation, of which four complexes comprise 13 PCGs. NADH ubiquinol oxidoreductase (NADH) complex (I) contain NADH dehydrogenase subunit I (*ND1*), *ND2*, *ND3*, *ND4*, *ND4l*, *ND5* and *ND6*, ubiquinone cytochrome c oxidoreductase complex (III) contain cytochrome b (*CYTB*), cytotochrome c oxidase complex (IV) contain cytochrome c oxidase subunit 1 (*COI*), *COII* and *COIII*, and ATP synthase FO subunit 6 (*ATP6*) and *ATP8* are from ATP synthase complex (V). The annotations of the complete set of 13 PCGs found in mtAsia I mitogenome are listed in Table 6.1. The mtAsia I mitogenome also contains eight overlapping regions across PCGs with the minimum length of 2 bp to maximum length of 23 bp and 22 non-coding regions between PCGs (Table 6.1). The overlapping pairs of PCGs were *ATP6-ATP8* and *ND4-ND4l* by 11 bp and 4 bp respectively on the same strand. This feature has been found commonly in insect mitochondrial genomes (Beckenbach and Stewart, 2009; Negrisolo et al., 2011). These findings were significantly similar to Tay et al. (2016)'s report of the Asia I species mitogenome. The A+T content among 13 PCGs also reflect the A+T bias in all four mitogenomes. The overall A+T content for mtAsia

Gene	Strand	Start-End	Length (bp)	Start	Stop	Intergenic	G + C%	A + T%
COI	Forward	1-1542	1542	ATG	TAA	-3	31.78	68.22
tRNA-Leu2	Forward	1538-1602	65			2	29.23	70.77
COII	Forward	1603-2286	684	ATA	TAA	-18	30.12	69.88
tRNA-Lys	Forward	2267-2334	68			20	25.00	75.00
ATP8	Forward	2353-2589	237	ATA	TAG	-9	27.85	72.15
ATP6	Forward	2579-3229	651	ATG	TAG	35	25.19	74.81
tRNA-Ser1	Reverse	3263-3322	60			11	28.33	71.67
tRNA-Glu	Reverse	3332-3394	63			25	15.87	84.13
tRNA-Phe	Reverse	3418-3486	69			-21	23.19	76.81
ND5	Reverse	3464-5134	1671	ATA	TAA	2	24.45	75.55
tRNA-His	Reverse	5135-5202	68			2	11.76	88.24
ND4	Reverse	5203-6495	1293	ATA	TAA	-2	24.98	75.02
ND4l	Reverse	6492-6776	285	ATG	TAA	3	23.16	76.84
tRNA-Thr	Forward	6778-6841	64			2	14.06	85.94
tRNA-Pro	Reverse	6842-6903	62			36	14.52	85.48
ND6	Forward	6938-7384	447	ATA	TAA	2	18.57	81.43
CYTB	Forward	7385-8518	1134	ATG	TAA	0	29.63	70.37
tRNA-Ser2	Forward	8517-8573	57			19	19.30	80.70
ND1	Reverse	8591-9493	903	ATA	TAG	17	25.03	74.97
tRNA-Leu1	Reverse	9509-9579	71			29	21.13	78.87
rRNA-L	Reverse	9607-10788	1182			1	27.84	72.16
tRNA-Val	Reverse	10788-10854	67			10	20.80	79.10
tRNA-Asp	Reverse	10863-10938	76			7	21.05	78.95
tRNA-Gln	Reverse	10944-11007	64			7	17.19	82.81
rRNA-S	Reverse	11013-11763	751			162	21.57	78.43
tRNA-Asn	Reverse	11924-11987	64			3	15.63	84.38
tRNA-Arg	Reverse	11989-12057	69			5	26.09	73.91
tRNA-Ala	Reverse	12061-12125	65			15	20.00	80.00
ND3	Reverse	12139-12492	354	ATG	TAA	2	23.45	76.55
tRNA-Gly	Reverse	12493-12555	63			46	14.29	85.71
COIII	Reverse	12600-13442	843	ATA	TAA	710	24.32	75.68
Control region		13443-14152	710					
tRNA-Ile	Forward	14153-14218	66			2	27.27	72.73
tRNA-Met	Forward	14219-14288	70			11	25.71	74.29
ND2	Forward	14298-15258	960	ATA	TAA	-1	23.65	76.35
tRNA-Trp	Forward	15256-15324	69			0	11.59	88.41
tRNA-Tyr	Reverse	15323-15385	63			3	20.63	79.37
tRNA-Cys	Reverse	15387-15448	62			5	19.35	80.65

TABLE 6.1: Complete annotation of mtAsia I mitogenome.

I PCGs was 73.80%, highly similar to that of *B. tabaci* species including Asia I (73.76%), MED (73.81%) and New World I (74.26%). This A+T content was also found similar to that of dipteran insect *Trichophthalma punctata* (72.0%) (Wang et al., 2016).

6.3.1.2 Transfer and ribosomal RNA genes

The typical complements of 22 tRNAs were identified in the mtAsia I mitogenome as found in other arthropod insect mitogenomes, with the size ranging from 57 bp (*tRNA-Ser2*) to 76 bp (*tRNA-Asp*) and 1,445 bp in total. Of these, 15 tRNA genes were encoded on the negative strand and the remaining seven tRNAs were encoded on the positive strand. In mtAsia I, most tRNAs overlapped with the adjacent PCGs (Table 6.1). The higher composition of A+T content was also found consistent in mtAsia I tRNAs (79.86%) and also found the same in Asia I (79.86%) and New World I species (79.86%), and highly similar to that of MED species (79.42%). All of the mtAsia I tRNAs folded into the typical "clover-leaf" shaped secondary structure although two tRNAs, *tRNA-Ser1* (60 bp) and *tRNA-Ser2* (57 bp), were found lacking the dihydorouridine (DHU) domain (Appendix B, Figure B6.1) as has been predominantly observed in other metazoan mitogenomes for this particular gene (Wolstenholme, 1992).

The rRNAs of mtAsia I were found on the negative strand with 1,182 bp for *rRNA-L* and 751 bp for *rRNA-S* in size. The *rRNA-L* lies in-between two tRNAs, *tRNA-Leu1* and *tRNA-Val*, whereas the *rRNA-S* lies between *tRNA-Gln* and *tRNA-Asn*. The overall A+T content for rRNAs in mtAsia I was 75.63% and was found highly similar to that in the New World I (75.63%) and MED species (75.57%). This result again confirms the higher composition of A+T content in mtAsia I. Surprisingly, A+T content for rRNAs of Asia I species (78.12%) was higher than the other mitogenomes. The previous studies suggest that the rRNA genes lack functional annotation features, analogous to the start and stop codons of PCGs and therefore it is not possible to determine the boundaries from the sequence alone (Boore, 2001, 2006).

6.3.1.3 The control region

The mitogenome of mtAsia I also contains a non-coding region of 710 bp that is also known as a putative control region as found in mitogenomes of other *B. tabaci* species (Thao et al., 2004; Wang et al., 2013; Tay et al., 2016). According to previous studies on mitochondria, this control region has been described as a putative origin of DNA replication where the initiation of transcription takes place (Boore, 1999; Wolstenholme, 1992). This control region was found to have a higher A+T content (76.24%) than present in the rest of the mitogenome (71.80%). This control region was found longer than that in Asia I (467 bp) and New World I (664 bp) species (Thao et al., 2004) mitogenomes (Tay et al., 2016) but smaller that that in MED (974 bp) species mitogenome (Wang et al., 2013) along with the conserved gene arrangements. The control region lies between *COIII* and *tRNA-Ile* genes in mtAsia I which was found similar to that in Asia I, MED and New World I species (Wang et al., 2013). Additionally, the mtAsia I mitogenome harbours intergenic spacers along with this control region. In the mtAsia I mitogenome, a total of 31 intergenic spacers with size ranging from 1-162 bp and total size of 1,194 bp were found. In general, variation in the

size of mitogenomes is a consequence of variation in the size of repeats within noncoding regions and not in the number of PCGs. Early studies on mitochondrial genomes suggested that a wide range of different organisms from the class Insecta share a highly conserved gene synteny which indicated an ancestral gene synteny for this group (Wolstenholme, 1992; Shao and Barker, 2003). Moreover, conserved gene synteny was observed in the hemipteran insects (including the suborder Sternorrhyncha) but not in other orders including Thysanoptera, Phthiraptera and Psocoptera (Shao et al., 2001a,b; Shao and Barker, 2003; Cranston and Gullan, 2003).

6.3.1.4 Codon usage

All 13 PCGs of mtAsia I mitogenome have identical codons to the recent Asia I mitogenome study published (Tay et al., 2016), with ATG methionine or ATA isoleucine as putative in-frame start codons and TAG or TAA as termination codons. There were five PCGs including *COI*, *ATP6*, *ND4l*, *CYTB* and *ND3* found to have ATG as their start codon and the remaining eight PCGs including *COII*, *ATP8*, *ND5*, *ND4*, *ND6*, *ND1*, *COIII* and *ND2* have ATA (Table 6.2). Among all 13 PCGs, nucleotide composition at the first and second codon positions showed higher usage of A/T than G/C and the synonymous codons ending with A/T were more common in all four species (Table 6.2). The start codon of the *COI* gene was postulated as tetranucleotide (ATAA, TTAA and ATTA) or hexanucleotide (ATTTAA) and has been extensively discussed in various arthropod insects (Caterino and Sperling, 1999; Wilson et al., 2000). However, hemipteran insects including *B. tabaci* (Thao et al., 2004; Wang et al., 2013, 2016b; Tay et al., 2016), *Geisha distinctissima* (Song and Liang, 2009), *Philaenus spumarius* (Stewart and Beckenbach, 2005) and *Triatoma dimidiate* (Dotson and Beard, 2001) typically use the ATG as a start codon, which suggests that the codon bias is taxon specific (Cha et al., 2007).

The canonical stop codons TAG and TAA were found most exclusively used in *ATP8*, *ATP6* and *ND1*, and *COI*, *COII*, *ND5*, *ND4*, *ND4l*, *ND6*, *CYTB*, *ND3*, *COIII* and *ND2* respectively (Table 6.2). The codon usage for 13 PCGs in mitogenomes of mtAsia I, Asia I, New World I and MED species was compared and is listed in Table 6.2. The results indicate that the encoded PCGs in New World I and MED species use different codons compared to mtAsia I and Asia I species PCGs. Two new start codons ATT and ATC were found in MED species for *ND5*, *ND1* and *COIII*, and *ND2* respectively (Wang et al., 2013), which were not found in mtAsia I (Table 6.2). Incomplete stop codons (T) are normally found for *COI* and *COII* genes in metazoan species including hemipteran insects (Crozier and Crozier, 1993; Junqueira et al., 2004; Beckenbach and Stewart, 2009). This phenomenon has been interpreted as the TAA stop codon generated from post-transcriptional polyadenylation (Masta and Boore, 2004). All 13 PCGs of mtAsia I mitogenome have complete translation codons.

	Start codon				Stop codon			
Gene	**mtAsia I**	**Asia I**	**New World I**	**MED**	**mtAsia I**	**Asia I**	**New World I**	**MED**
COI	ATG	ATG	ATG	ATG	TAA	TAA	T	T
COII	ATA	ATA	ATA	ATA	TAA	TAA	T	T
ATP8	ATA	ATA	ATG	ATG	TAG	TAG	TAG	TAG
ATP6	ATG	ATG	ATG	ATG	TAG	TAG	TAA	TAA
ND5	ATA	ATA	ATT	ATT	TAA	TAA	T	T
ND4	ATA	ATA	ATA	ATA	TAA	TAA	TAA	TAA
ND4l	ATG	ATG	ATG	ATG	TAA	TAA	TAA	TAA
ND6	ATA	ATA	ATG	ATG	TAA	TAA	TAA	TAA
CYTB	ATG	ATG	ATG	ATG	TAA	TAA	TAA	TAA
ND1	ATA	ATA	ATG	ATT	TAG	TAG	TAG	TAA
ND3	ATG	ATG	ATG	ATG	TAA	TAA	TAA	TAA
COIII	ATA	ATA	ATG	ATT	TAA	TAA	TAA	TAG
ND2	ATA	ATA	ATA	ATC	TAA	TAA	TAA	TAA

TABLE 6.2: Codon usage comparison across 13 PCGs of mtAsia I, Asia I, New World I, and MED species.

6.3.1.5 Comparison of mtAsia I with other whitefly species

The recent study of the Asia I species mitogenome focused on the comparison of gene features and its arrangement across other species (Tay et al., 2016). For a better understanding of divergence and evolutionary genetic relationships across the whitefly species, the complete mitogenomes from four whitefly species were compared to the mtAsia I mitogenome. For the whole mitogenome comparison, the sequence homology was, as was expected, extremely high between mtAsia I and Asia I species (99%) and lower to New World I (81%), MED species (81%) and *B. afer* (73%) (Figure 6.1). The mitogenomes of mtAsia I and Asia I species shared 99% sequence similarity with only 10 mismatches within coding genes and two deletion mutations within non-coding region in mtAsia I. The only difference in the length of mtAsia I and Asia I species mitogenomes was the length of the putative control region.

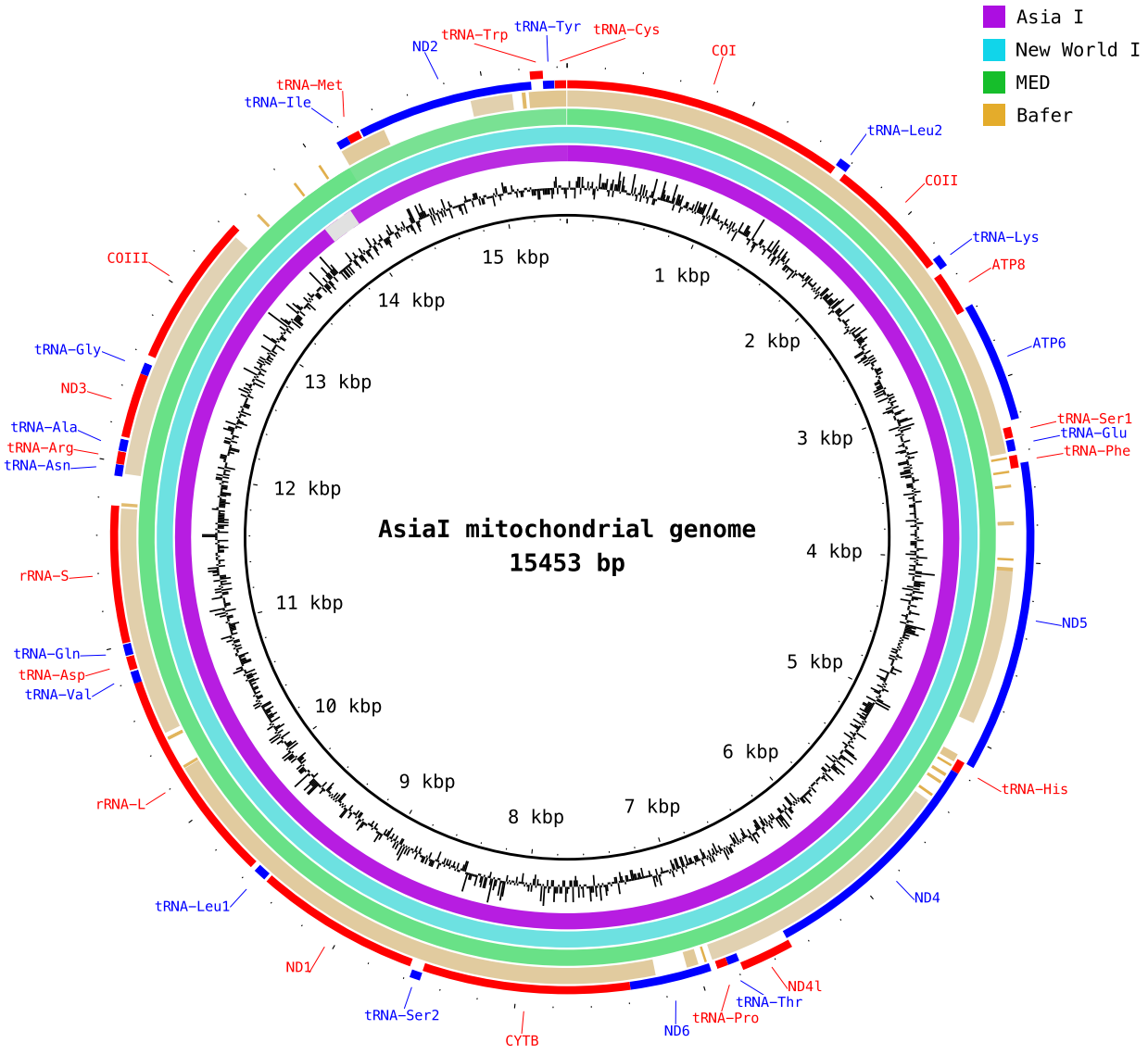


FIGURE 6.1: Sequence comparison of mtAsia I mitogenome with Asia I, New World I, MED species and *B. afer*. From inner circle to outer, the tracks depicts: mtAsia I mitogenome and it's GC content, Asia I, New World I, MED species and *B. afer*. Annotated genes of mtAsia I mitogenome are represented as blocks coloured alternatively as red and blue along with the overlapping genes on outer circle. The respective rings were colour filled according to the aligned region of the mitogenomes with BLAST %identity above 70%. The non-coding and unaligned regions were represented as white gaps.

6.3.1.6 Phylogeny across whitefly species

The molecular marker *mtCOI* has been used to distinguish whitefly species and shared 100% similarity with Asia I species followed by New World I (86%) and MED species (86%). The *B. afer* mitogenome had the lowest %identity to mtAsia I which was expected as the *B. afer* is the most

distantly related species (Chu et al., 2010; Mugerwa et al., 2012; Lee et al., 2013). Mitochondrial PCGs have been used as a reliable molecular marker and become an informative approach to infer phylogenetic relationships across species (Boore et al., 2005). Two different approaches were used to perform phylogenetic analysis and compare the topology across three species of *B. tabaci* such as mtAsia I, New World I and MED, and *B. afer*. As shown in Figure 6.2, partial *mtCOI* and 13 PCGs were used to construct a Bayesian tree for these *B. tabaci* species. The *mtCOI* with 657 bp sequence had 657 aligned sites where as the 13 PCGs dataset (10,961 bp from mtAsiaI) had 9,849 aligned sites across all four species. Bayesian trees from both datasets showed highly similar topologies across four species (Figure 6.2). The monophyly of *B. tabaci* complex was consistently supported for 13 PCGs and partial *mtCOI* datasets with the posterior probabilities 1.00, 0.95 and 0.78 (Figure 6.2). This result support Chowda-Reddy et al. (2012) and Boykin et al. (2013) findings regarding basal branching events within *B. tabaci* complex. This finding suggests that the partial *mtCOI* with 657 bp alone has sufficient significant divergence across these species to classify them in comparison with 13 PCGs which produce the identical classification.

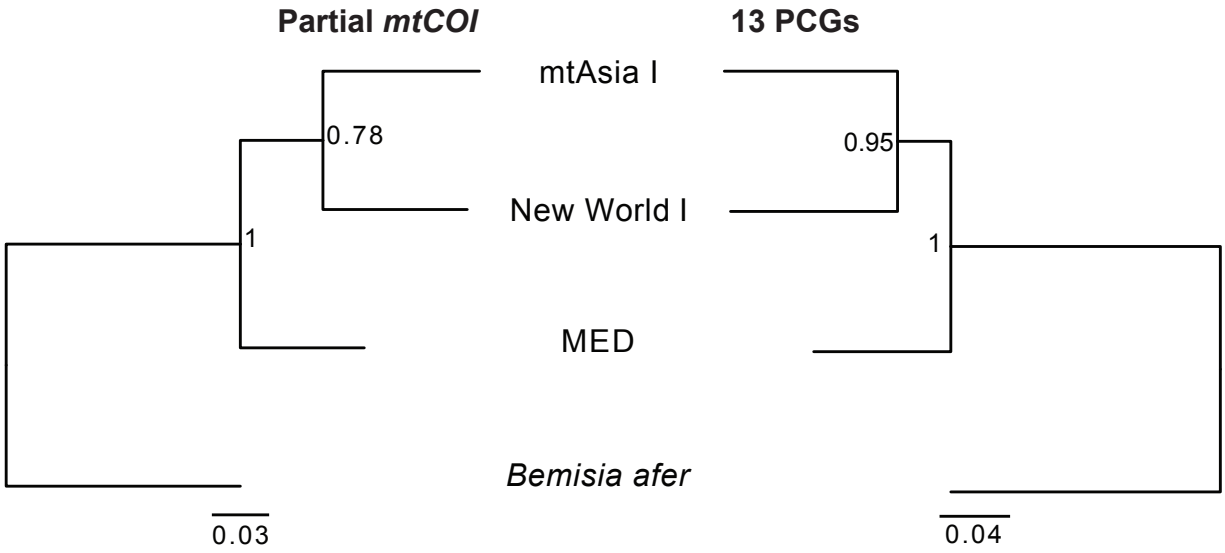


FIGURE 6.2: Phylogenetic tree of partial *mtCOI* (657 bp) and 13 PCGs from three *B. tabaci* species including mtAsia I, New World I and MED, and *B. afer*. The values on the branches represent the posterior probabilities obtained with the Bayesian analysis.

6.3.2 *Portiera* genome from Asia I species

6.3.2.1 *Portiera* genome: assembly and annotation

Using complete genome sequences of *Portiera* from MEAM1 (B, BT-B) and MED species (Q, BT-Q), of the total 886 M Asia I Illumina reads, 24,363,700 (2.75%) reads were retrieved from

two paired end libraries. Of the 2.75% mapped reads, 23,474,234 (96.35%) reads were paired end with average insert size of 200 bp and 450 bp and remaining 889,466 (3.65%) reads were single reads. Preliminary assembly of these reads produced 12 contigs with longest contig size of 91,501 bp. Following gap closure, a complete genome of *Portiera* was retrieved from Asia I yielding a total size of 357,529 bp, 26.1% GC content and a read depth coverage of 50x. Similar to primary endosymbiont of other hemipteran insect, *Portiera* also shares common features including reduced genome size and low GC content (Santos-Garcia et al., 2014b). The *Portiera* genome from Asia I species (357,529 bp) was highly AT-biased (73.88%) and of a similar size to the *Portiera* genome from BT-Q (357,472 bp) and BT-B (358,242 bp) (Table 6.3). Previously reported *Portiera* genomes from MEAM1 (B) and MED species (Q) were smaller in size due to missing a 6.1 kbp region (Jiang et al., 2013).

Feature	**AsiaI**	**B**	**BT-B**	**Q**	**BT-Q**
Accession		CP003868	CP003708	CP003867	CP003835
Size (bp)	357,529	351,658	358,242	350,928	357,472
GC (%)	26.1	26.2	26.2	26.1	26.1
AT (%)	73.88	73.79	73.83	73.84	73.88
PCGs	289	277	256	281	246
tRNAs	33	33	33	33	33
rRNAs	3	3	3	3	3
Coding density (%)	68.68	69.0	67.4	70.3	68.0

TABLE 6.3: General features of *Portiera* genomes from different species of *B. tabaci*.

A total of 289 PCGs were detected in the *Portiera* genome from Asia I, which were higher in numbers than previously reported PCGs in BT-B (256 PCGs) and BT-Q (246 PCGs) (Table 6.3). The genome contains a complete set of genes including 289 PCGs, 33 tRNAs and three copies of each rRNAs (5S, 16S and 23S) (Figure 6.3). Unlike other endosymbiont genomes which contain a higher composition of AT and protein coding densities such *Candidatus* Carsonella ruddii (97.3%) (Tamames et al., 2007), *Candidatus* Evansia muelleri (93.7%) (Santos-Garcia et al., 2014b) and *Buchnera aphidicola* (87.7%) (van Ham et al., 2003), only 68.68% of the *Portiera* genome in Asia I species encodes 289 PCGs (Table 6.3).

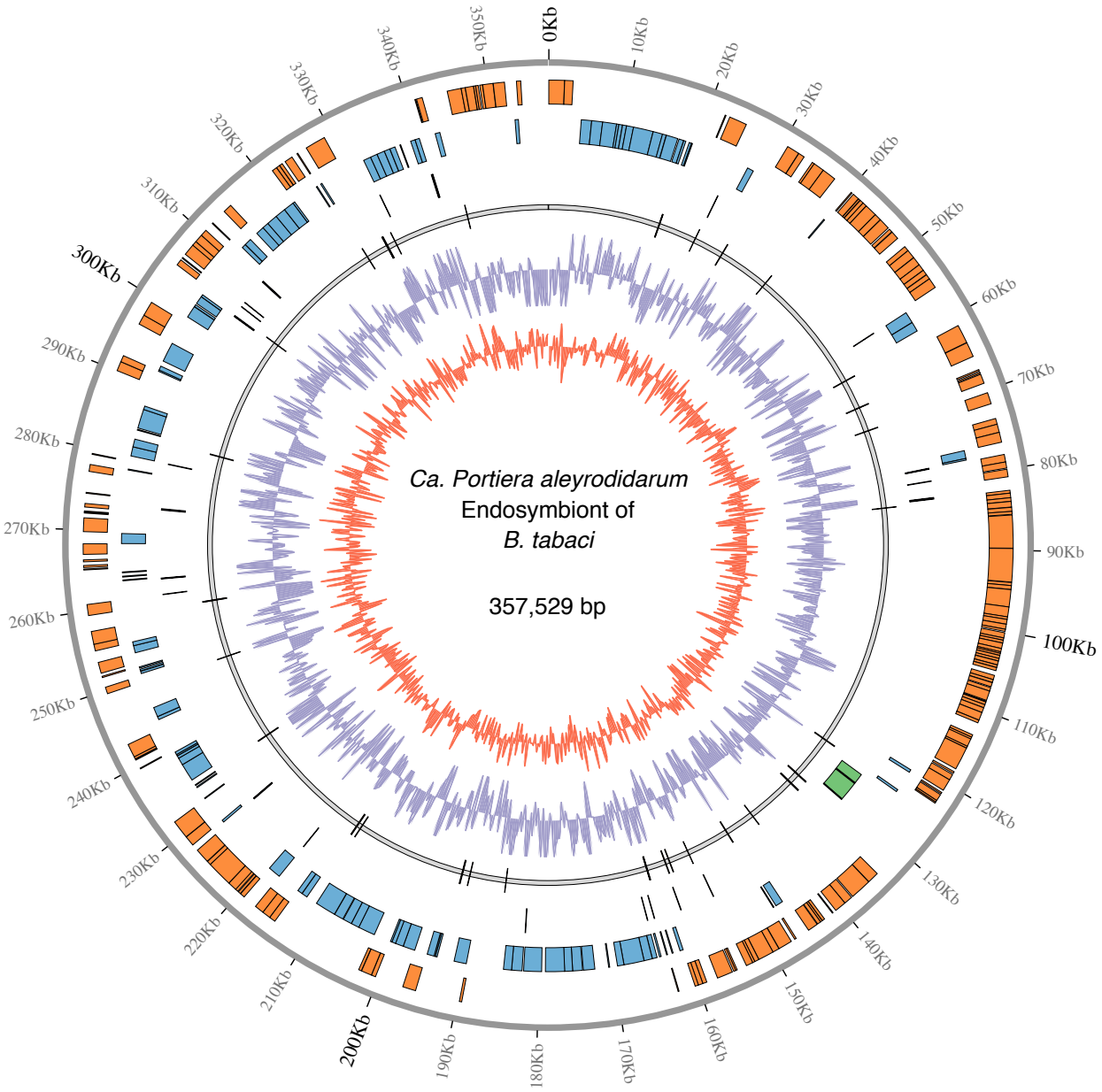


FIGURE 6.3: Complete genome of *Portiera* from inner to outer, the tracks depicts: AT skew, GC skew, tandem repeats, RNAs (rRNA - green and tRNA - purple), coding genes (positive strand - orange, negative strand - blue) and nucleotide positions in kbp.

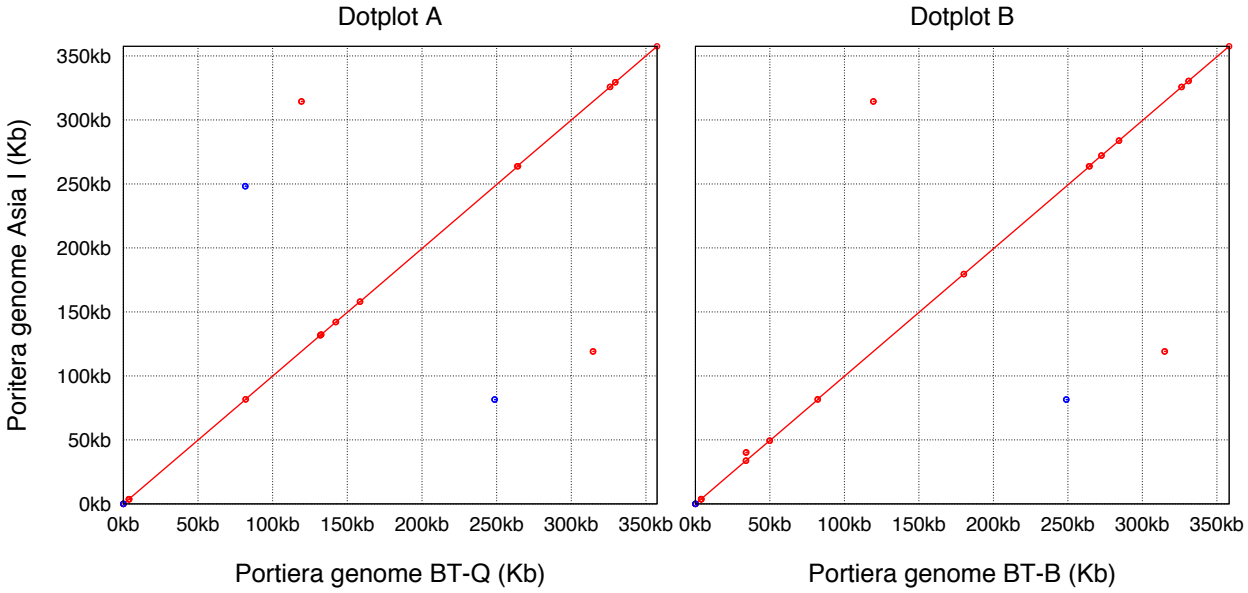
Similar to the primary endosymbiont genomes from MEAM1 and MED species, and other phloem-feeding insects, genes involved in biosynthesis of essential amino acids were identified, while certain genes associated with cellular processes including cell wall and capsule, motility, phages, plasmids, sulfur/iron/potassium/nitrogen/phosphorus metabolism, regulation and cell signaling, and photosynthesis were not detected. These findings support the hypothesis that *Portiera* provides with essential nutrients to their insect hosts. Interestingly, the gene involved in membrane transport was detected in Asia I species, which were not identified in previously reported *Portiera* genomes from MEAM1 and MED species (Jiang et al., 2012, 2013). In addition, B vitamin synthesis pathways

including biotin, riboflavin, folate, pyridoxine and thiamine were found absent in the *Portiera* genome from Asia I species. These findings were consistent with the *Portiera* from MED species where it lacks all the genes involved in the synthesis of cofactors and vitamins (Rao et al., 2015).

6.3.2.2 Comparison of *Portiera* genomes in Asia I, MEAM1 and MED

Comparison of *Portiera* genome sequences from three species of the *B. tabaci* complex, Asia I, MEAM1 and MED, revealed extensive conservation and highly similar gene complements across all three hosts (Figure 6.4A, Figure 6.4B). The *Portiera* genome from Asia I have the highest coding density compared to BT-B and BT-Q, except previously reported *Portiera* genomes B and Q from the same species (Table 6.3). The difference in their coding density reflects the absence of 6.1 kbp region (position 34,181 to 40,300) (Jiang et al., 2013), which may have resulted due to different chromosomal forms within insect hosts (Sloan and Moran, 2013). The 6.2 kbp region (positions 33,766 to 39,997) from *Portiera* genome of Asia I species encodes three genes including a membrane protein insertase (yidC), a GTP-binding protein (trmE) and the tRNA uridine 5-carboxymethylaminomethyl modification enzyme (gidA). GidA and trmE are conserved proteins in eubacteria and are essential for the modification of the wobble uridine of 5-carboxymethylaminomethyluridine tRNA (Osawa et al., 2009; Böhme et al., 2010). These proteins were found highly conserved across *Portiera* strains in Asia I species, BT-B and BT-Q.

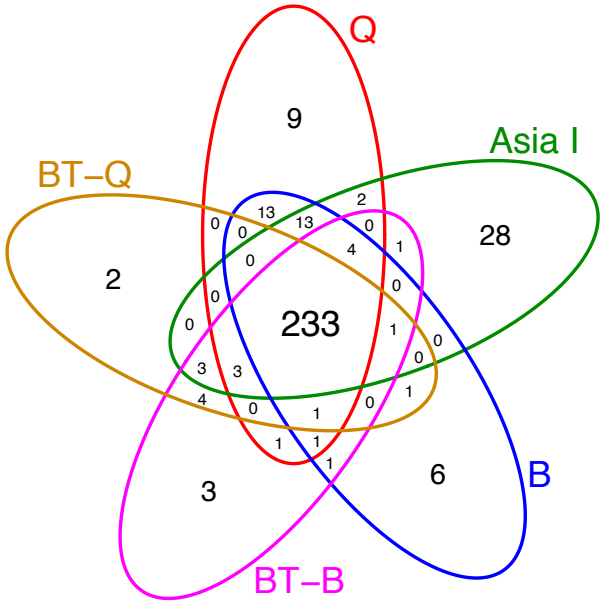
The *Portiera* genome from Asia I species share 98.32% and 98.28% sequence similarity with BT-B and BT-Q respectively (Figure 6.4B). There were 3,459 Single Nucleotide Polymorphisms (SNPs) found between Asia I species and BT-Q, which were lower in number in comparison to Asia I species and BT-B (3,574 SNPs). As the genome is highly AT biased, a higher number of SNPs were seen at duplets formed with A and / or T (Figure 6.4B). Additionally, the *Portiera* genome from Asia I species shared 2,464 indels with BT-Q and 2,537 with BT-B. Similar to SNPs, A and T were highly represented as insertion and deletion mutations across Asia I species, BT-Q and BT-B.



(A)

Feature	Dotplot A		Dotplot A	
	Asia I	BT-Q	Asia I	BT-B
Size (bp)	357,529	357,472	357,529	358,242
Aligned bases (%)	99.86	99.81	99.80	99.87
Aligned blocks	10	10	11	11
Avg. identity (%)	98.32	98.32	98.28	98.28
SNPs	3,459	3,459	3,574	3,574
AT	758	754	762	771
AC	318	313	327	320
AG	137	149	156	142
TA	754	758	771	762
TC	103	121	121	112
TG	347	370	398	389
CA	313	318	320	327
CT	121	103	112	121
CG	34	55	32	44
GA	149	137	142	156
GT	370	347	389	398
GC	55	34	44	32
Indels	2,464	2,464	2,537	2,537

(B)



(C)

FIGURE 6.4: Sequence comparison between *Portiera* genomes: Asia I, MEAM1 and MED species. (A) Dotplot shows pairwise comparison between Asia I species and BT-B (Dotplot A), and Asia I species and BT-Q (Dotplot B). Unaligned regions were highlighted as red circles on the diagonal red line. Repeat regions were also highlighted as red circles (positive strand) and blue circles (negative strand) off the diagonal line. (B) Detailed statistics of comparisons were shown in table. (C) Venn diagram depicts protein orthologs of *Portiera* across different species of *B. tabaci*.

6.3.2.3 Protein orthologs

Whole proteomes of *Portiera* from Asia I species, BT-B, BT-Q, B and Q were clustered to identify shared and unique proteins across three species of *B. tabaci*. A total of 233 proteins were identified as core proteins shared across all proteomes compared (Figure 6.4C). Interestingly, there were 28 proteins from Asia I species that did not share sequence homology with other species. To specifically identify the putative function of these 28 proteins from Asia I species, they were searched against all *Portiera* sequences at NCBI. Of the 28 proteins, only 11 returned with a hit, which includes one copy of dihydrodipicolinate reductase and MFS transporter (sugar porter (SP) family), and nine copies of hypothetical proteins. Additionally, there were 13 proteins shared between Asia I species, B and Q, which were not seen in BT-B and BT-Q, which includes nine copies of hypothetical proteins, two ATP-dependent proteases, and one copy of dihydrodipicolinate reductase and galactose-proton symport of transport system. Despite their larger size, BT-B and BT-Q had less unique proteins in comparison with B and Q.

6.3.2.4 Vertical transmission across *Bemisia* species

Phylogenetic analysis of molecular marker gene *mtCOI* in *B. tabaci* species and 16S rRNA from their *Portiera* endosymbiont revealed an extremely high degree of co-cladogenesis between *B. tabaci* species and its primary endosymbiont *Portiera* (Figure 6.5). The *Portiera* infection in *B. tabaci* species complex is vertically transmitted from ancestor to progeny. This finding was as expected and has been reported in previous studies (Baumann et al., 2004).

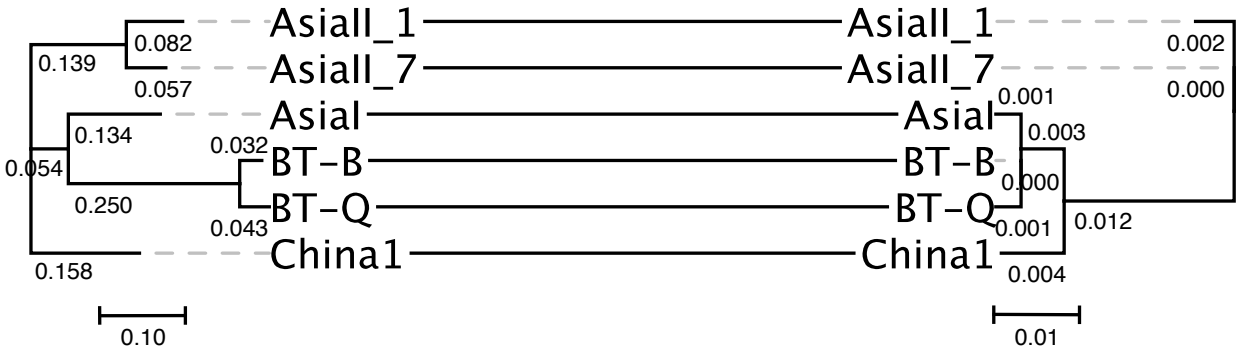


FIGURE 6.5: Phylogenetic tree showing vertical transmission of *Portiera* across *B. tabaci* species was constructed using (A) partial *mtCOI* genes from *B. tabaci* species and (B) partial 16S RNA from their *Portiera* endosymbiont.

6.3.3 *Wolbachia* genome from Asia I species

6.3.3.1 Draft genome: assembly and annotation

Using a read mapping approach, a total of 658,017 genomic reads were retrieved from one paired end Illumina DISCOVAR library with read length of 250 bp. The draft genome assembly of *Wolbachia* from *B. tabaci* Asia I species (wBtab-AsiaI) is 1,181,706 bp in 212 consensus scaffolds with a maximum length of 33,115 bp and an average GC content of 34.05%. The wBtab genome contains 1,030 putative coding genes, 34 tRNA genes and three rRNA genes (5S, 16S and 23S) (Table 6.4). The size of the draft genome sequence of wBtab-AsiaI is smaller than all 10 completed genome sequences of *Wolbachia* strains (Table 6.4).

Ankyrin (ANK) repeat domains have been predominantly identified in *Wolbachia* genomes sequenced to date (Iturbe-Ormaetxe et al., 2005; Walker et al., 2007) and their roles in host interaction and reproduction of phenotypes have been reported (Pan et al., 2008; Papafotiou et al., 2011). A total of 29 ANK coding genes were identified in wBtab-AsiaI, which were higher/similar in numbers than in all *Wolbachia* strains from worm (wOo: 2 ANK, wOv: 0 ANK and wBm: 5 ANK), tsetse fly (wGm: 10 ANK), psyllid (wDi: 29 ANK) and four fruit flies (wMel: 23 ANK, wMelPop: 27 ANK, wRec: 22 ANK and wDs-VAL: 16 ANK) but lower than in all strains from mosquito (wAlbB: 33 ANK, wCp Mol: 32 ANK, wPip Pel: 60 ANK and wPip JHB: 64 ANK), five fruit flies (wDa: 37 ANK, wHa: 30 ANK, wAu: 50 ANK, wRi: 35 and wNo: 54 ANK), butterfly (wBol1-b: 61 ANK) and bed bug (wCle: 49 ANK) (Table 6.4). ANK is one of the most common protein-protein interaction motifs and therefore may play a significant role in interaction with host via many physiological processes including apoptosis, cell signaling and cell cycle control (Al-Khodor et al., 2010). All of the 29 ANK coding genes of wBtab-AsiaI had found their corresponding orthologs in all 20 *Wolbachia* strains with most orthologs in wPip JHB (18 ANK), wDi (16 ANK), wCp-Mol (16 ANK), wBol1-b (16 ANK) and wPip-Pel (16 ANK).

6.3.3.2 COG analysis among *Wolbachia* subgroups

To identify 'Cluster of Orthologous Group (COG)' in all *Wolbachia* strains, whole proteomes from the 21 strains available were compared using Proteinortho (Lechner et al., 2011) with a cut-off of 50% for %identity and %coverage, and an E-value cut-off of $1E^{-05}$. The highest number of COGs for wBtab-AsiaI proteins were found in wBol1-b (856 orthologs), wPip-JHB (836 orthologs), wNo (830 orthologs), wPip-Pel (828 orthologs) and wDi (828 orthologs). A total of 338 COGs were found as core proteins clustered between all 21 strains, and 57 proteins from wBtab-AsiaI were not clustered. The number (338) of clustered core proteins is less than the 621 core proteins of *Wolbachia* predicted previously from microarray-based comparative genome hybridization of

TABLE 6.4: General characteristics of 21 *Wolbachia* genomes.

Group[a]	Host type	Host	Strain	Size (bp)	GC%	PCG[b]	tRNA	rRNA	Ank[c]	Status[d]	GenBank accession
A	Fruit fly	*Drosophila ananassae*	wDa	1,440,750	35.71	1,802	35	3	37	D	AAGB01000001-AAGB01000464
A	Fruit fly	*Drosophila melanogaster*	wMel	1,267,782	35.23	1,195	34	3	23	C	AE017196.1
A	Fruit fly	*Drosophila melanogaster*	wMelPop	1,201,350	35.20	1,111	34	3	27	D	AQQE01000001-AQQE01000080
A	Fruit fly	*Drosophila recens*	wRec	1,126,656	35.17	1,227	34	3	22	D	JQAM01000001-JQAM01000043
A	Fruit fly	*Drosophila simulans*	wHa	1,295,804	35.10	1,009	34	3	30	C	CP003884.1
A	Fruit fly	*Drosophila simulans*	wAu	1,268,461	35.22	1,266	34	3	50	C	LK055284.1
A	Fruit fly	*Drosophila simulans*	wRi	1,445,873	35.16	1,150	34	3	35	C	CP001391.1
A	Fruit fly	*Drosophila suzukii*	wDs-VAL	1,415,342	35.21	1,439	34	3	16	D	CAOU02000001-CAOU02000110
A	Tsetse fly	*Glossina morsitans*	wGm	1,015,675	35.24	800	34	3	10	D	AWUH01000001-AWUH01000201
B	Mosquito	*Aedes albopictus*	wAlbB	1,162,431	33.85	1,058	34	3	33	D	CAGB01000001-CAGB01000165
B	Whitefly	*Bemisia tabaci*	wBtab-AsiaI	1,181,706	34.05	1,030	34	3	29	D	LAEY01000000
B	Mosquito	*Culex pipiens molestus*	wCp Mol	1,479,517	34.32	1,292	34	3	32	D	CACK01000001-CACK01000888
B	Mosquito	*Culex quinquefasciatus*	wPip Pel	1,482,355	34.19	1,275	34	3	60	C	AM999887.1
B	Mosquito	*Culex quinquefasciatus*	wPip JHB	1,542,137	34.19	1,378	34	3	64	D	ABZA01000001-ABZA01000021
B	Psyllid	*Diaphornia citri*	wDi	1,215,679	33.99	1,197	34	3	29	D	AMZJ01000001-AMZJ01000124
B	Fruit fly	*Drosophila simulans*	wNo	1,301,823	34.00	1,040	34	3	54	C	CP003883.1
B	Butterfly	*Hypolimnas bolina*	wBol1-b	1,377,933	33.94	1,319	34	3	61	D	CAOH01000001-CAOH01000144
C	Worm	*Onchocerca ochengi*	wOo	957,990	32.07	647	32	3	2	C	HE660029.1
C	Worm	*Onchocerca volvulus*	wOv	960,618	32.07	646	34	3	0	C	HG810405.1
D	Worm	*Brugua malayi*	wBm	1,080,084	34.18	805	34	3	5	C	AE017321.1
F	Bed bug	*Cimex lectularius*	wCle	1,250,060	36.25	1,216	34	3	49	C	AP013028.1

[a]Group: Based on phylogenetic analysis of 16s rRNA, [b]PCG: Protein coding gene, [c]ANK: Ankyrin repeat domain coding gene, [d]Status: Genome sequencing status as of July 2015, C: complete, D: draft but annotated.

group-A *Wolbachia* strains (Ishmael et al., 2009) and 654 core genes identified by Duplouy et al. (2013) from five *Wolbachia* strains (wBol1-b, wPip, wMel, wRi and wBm).

Similarly to identify whether or not the 57 proteins that did not cluster represented wBtab-AsiaI specific proteins, these proteins were searched against all relevant *Wolbachia* proteins at NCBI using an E-value cut-off of $1E^{-05}$. All of the 57 proteins had significant hits against *Wolbachia* proteins and hence do not represent wBtab-AsiaI specific proteins. Further research is required to determine whether any of the wBtab-AsiaI unclustered proteins confer biological advantages to their whitefly host.

In assessing the completeness of the assembly, it was noted that only three genes present in all seven *Wolbachia* strains previously published, were absent in our wBtab-AsiaI assembly using the alignment cut-off of greater than 50% and an E-value of $1E^{-05}$. However, these three genes were identified in our wBtab-AsiaI assembly when a cut-off of less than 50% alignment coverage was used and it was determined that they had appeared absent, because they had been truncated during the scaffolding process. These findings suggest that the majority of complete protein coding genes (i.e. 973/1030, 973 genes were identified as COG) have been assembled into scaffolds in our assembly. Based on these core ortholog genes, it is considered that the scaffolds for wBtab-AsiaI represent a high quality draft genome. Further sequencing and analysis of *Wolbachia* genome will be required to achieve the complete genome sequence of wBtab-AsiaI. This draft genome of wBtab-AsiaI will enable research on the endosymbiotic relationship between *B. tabaci* and *Wolbachia*.

6.3.3.3 Phylogenetic placement of wBtab-AsiaI

Wolbachia strains were selected where their genome assembly contained a complete 16s rRNA gene for phylogenetic analysis. There were two approaches used in this study to confirm the placement of the wBtab-AsiaI strain across existing subgroups. First, multiple sequence alignment was performed using MUSCLE (Edgar, 2004) on nucleotide sequences of 16s rRNA from 21 *Wolbachia* strains and ML inference methods with 1,000 bootstrap replicates were used to infer phylogenetic relationships (Figure 6.6A). Second, nucleotide sequences of 338 core genes were from the same 21 *Wolbachia* strains used by MUSCLE to produce multiple alignments and the phylogeny was inferred on concatenated alignment in RAxML (Stamatakis, 2006) using the GTRCAT model using 1,000 bootstrap replicates (Figure 6.6B).

ML analysis of 16S rRNA genes indicated that the wBtab-AsiaI strain was placed in subgroup B along with the other existing strains (Figure 6.6A). This result support the previous findings of diversity of *Wolbachia* strain from *B. tabaci* using partial 16S rRNA gene (592 bp) (Bing et al., 2013a). The subgroup B (host: mosquito, whitefly, fruit fly, psyllid and butterfly) is closer to subgroup A (host: fruit fly, tsetse fly) than to subgroup C and D (host: worm), and F (host: bed bug).

To our knowledge, this is the first phylogenetic placement of a *Wolbachia* strain from *B. tabaci* using the complete 16S rRNA gene.

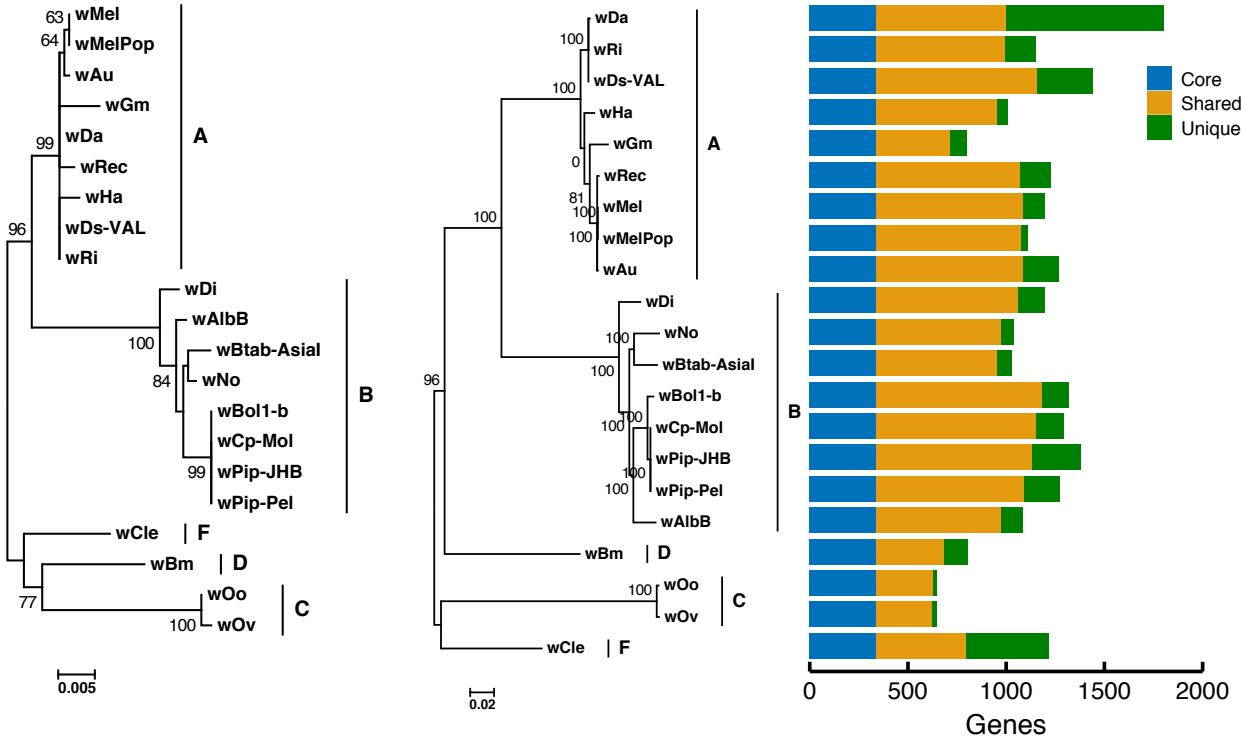


FIGURE 6.6: (A) Phylogenetic analysis based on complete 16s rRNA genes from 21 strains of *Wolbachia* whose genome sequences are available. The subgroups (A, B, C, D and F) were shown on rooted tree and ML bootstrap values were also shown adjacent to each node. (B) Phylogenetic analysis based on concatenated gene sequences of 338 COGs from 21 strains of *Wolbachia*. The subgroups (A, B, C, D and F) were shown on rooted tree along with the ML bootstrap values adjacent to each node. The distribution of ortholog genes across different strains is shown on the right hand side of the tree. Core genes are encoded by all *Wolbachia* genomes, shared genes are found in two or more *Wolbachia* strains but not in all strains and unique genes are found in only one strain.

A second phylogenetic approach using core genes from complete/draft (annotated) genomes of *Wolbachia* strains was applied to test the robustness of the subgroups classification. A total of 338 clustered core proteins (single copy) present in 21 *Wolbachia* strains were selected for phylogenetic analysis. Incongruence across single gene phylogenetic analysis is most often overcome by phylogeny based on concatenated genes (Rokas et al., 2003). MUSCLE was used to produce single-gene alignments of 338 genes (nucleotide sequences) and then concatenated to form a 327,992 bp long alignment. The phylogeny was inferred on this concatenated alignment using RAxML, GTRCAT substitution model and 1,000 bootstrap replicates (Figure 6.6B). Phylogenetic classification of *Wolbachia* subgroups was strongly supported by both, single-gene (16S rRNA) and concatenated (338 single copy genes) phylogenies. These findings confirm the placements of

Wolbachia strains into subgroups and also indicate the robustness of single-gene and concatenated gene analysis approaches. However, placements of *Wolbachia* strains were different within subgroups (subgroup A, B) and also the distance across subgroups (subgroup B, C, D, F) on the rooted trees when using both approaches (Figure 6.6A-B).

6.3.4 *Arsenophonus* genome from Asia I species

The draft genome assembly of *Arsenophonus* endosymbiont of Asia I species (in this chapter named as *Arsenophonus bemisiae*) is 1,860,528 bp (G+C, 37.30%) in 358 consensus scaffolds with an average coverage depth of 90x. The *Arsenophonus bemisiae* genome consist of 1,846 PCGs, 42 tRNAs and 6 copies (5S, 16S and 23S) of rRNA genes. These predicted PCG results were smaller in number than those found in the other two *Arsenophonus* sequences published. For example, 3,588 PCGs, 58 tRNAs and 6 rRNAs were found in *Arsenophonus nasoniae*, where-as 2,646 PCGs, 51 tRNAs and 8 rRNAs were identified in the *Arsenophonus* endosymbiont of *N. lugens*. The draft genome size of the *Arsenophonus bemisiae* was 1.8 Mbp which is also much smaller than the genomes of *Arsenophonus nasoniae* (3.67 Mbp) (Wilkes et al., 2010) and *Arsenophonus* endosymbiont of *N. lugens* (2.95 Mbp) (Xue et al., 2014). These findings indicate that the draft genome assembly of *Arsenophonus bemisiae* is incomplete and missing a significant proportion of the genome (> 1 Mbp). Further sequencing of this endosymbiont is required in order to obtain a complete genome sequence of *Arsenophonus bemisiae* from the Asia I species. Comparative ortholog analysis of proteins from *Arsenophonus bemisiae*, *Arsenophonus nasoniae* and *Arsenophonus* endosymbiont of *N. lugens* revealed 1,245 core proteins. Similar to these two *Arsenophonus* strains, *Arsenophonus bemisiae* had abundant predicted proteins associated with carbohydrate metabolism (n=134), metabolism of cofactors and vitamins (n=96), nucleotide metabolism (n=70), amino acid metabolism (n=53), and energy metabolism (n=39) (Wattam et al., 2014). However, secondary metabolism and, iron acquisition and metabolism associated proteins were found in other two *Arsenophonus* strains but absent in *Arsenophonus bemisiae*. A detailed comparative analysis of *Arsenophonus* genomes will help to better understand the endosymbiotic relationship with their host and their diverse evolution across various hosts.

CHAPTER 7

Using *Drosophila* essential genes to establish a genomic framework for studying pest biology and insecticide discovery

7.1 Introduction

Insect vectors are major threats to food security and human health. Agricultural insect pests such as the aphid, *A. pisum*, and the whitefly, *B. tabaci*, cause devastating crop damage worldwide (Morales, 2006; Dedryver et al., 2010), while parasite vectors such as the *Anopheles* mosquitoes mediate the loss of hundreds of thousands of lives every year (Fidock, 2010).

Despite this importance, the rate of new insecticide discovery, and the arsenal of available insecticides remains small (Kelly-Hope et al., 2008). The few novel insecticides that have been introduced target molecular mechanisms, such as ion channel blockade, that are also present in beneficial pollinators, leading to poor biological selectivity. In addition, the exact mode of action for many insecticides remains ill-defined, making potential human toxicities difficult to predict. Integrated pest management strategies for the chemical control of these insect vectors are being pursued (Bruce, 2010), but, given the intrinsic lack of selectivity amongst available chemical agents, these have little flexibility and currently do not represent sustainable pest control solutions (Chandler et al., 2011).

To add to this gloomy picture, chemical control with the current generation of commercial insecticides is hampered by the rapid development of resistance within treated insect populations. Importantly, resistance development can occur in parallel in both agricultural pests and in insect vectors of human disease such as *A. gambiae* when a single insecticide is used for both purposes. Insecticide resistance is having a serious impact on healthcare strategies where insect vector control is an important treatment arm and insecticide resistance is associated with treatment failure (Ranson et al., 2009; malERA Consultative Group on Vector Control, 2011).

Taken together, these observations emphasize the need for new approaches to insecticide discovery. There is an urgent need to develop new insecticides with lower propensity for resistance development, and greater selectivity for pests over pollinators.

Genomics provides important resources, both to support new target discovery for future insecticides, and to pinpoint molecular mechanisms associated with insecticide resistance development. Complete genomes are available for a variety of hemipteran and coleopteran pests, for example *A. pisum* (IAGC, 2010) and *T. castaneum* (TGSC, 2008), as well as beneficial pollinators such as *A. mellifera* (HGSC, 2006) and insect parasite vectors such as *A. gambiae* (Holt et al., 2002; Mongin et al., 2004). Additionally, the well-characterized *D. melanogaster* genome (Adams et al., 2000) provides a useful genetic tool for new target validation, of central importance in discovery biology programs (Perry et al., 2011). Perhaps surprisingly, genome sequencing studies indicate that insects build their extraordinary phenotypic diversity using a limited set of genes, ranging from some 11,000 genes in the honey bee (HGSC, 2006) to some 15,000 genes in the mosquito and whitefly (Holt et al., 2002; Seal et al., 2012).

In this study, we use insect comparative genomics to identify orthologs of essential genes previously defined in *D. melanogaster*, referred to here as *Drosophila* Essential Genes, or "DEG" genes, within further insect pests and pollinators, focusing on *A. pisum*, *A. gambiae* and *A. mellifera*. Using gene family and chemical biology database approaches (Gaulton et al., 2011; Punta et al., 2012), we triage the orthologs of these essential genes to produce a focussed framework within which to study pest biology and explore novel approaches to insect pest control.

7.2 Methods

7.2.1 Experimental Design

This study centers on the well-characterized model organism the fruit fly, *D. melanogaster*, a species for which experimentally validated, genome-wide information on gene essentiality is available. Fruit fly essential genes were compiled from FlyBase (Miklos and Rubin, 1996; Tweedie et al., 2009) (Appendix B, Figure B7.1). Their counterparts in the insect pest *A. pisum*, the beneficial pollinator *A. mellifera* and the human disease vector *A. gambiae* were determined by BLAST (v+2.2.25) analysis (Altschul et al., 1990, 1997; Camacho et al., 2009). Further comparisons of putative targets within other pests were performed with information taken from the literature. The overall purpose of the study is summarized in Figure 7.1.

7.2.2 Comparative Genomics

FASTA formatted protein sequences for the continuing genome projects for *A. pisum*, *A. mellifera*, *A. gambiae* and *D. melanogaster* were downloaded from AphidBase (36,275 sequences, vAcyr_2.0) (http://www.AphidBase.com) (IAGC, 2010), BeeBase (11,062 sequences, vAmel_2.0) (http://hymen\opteragenome.org/BeeBase/) (Munoz-Torres et al., 2011), VectorBase (14,324 sequences, vAgamP3) (http://www.VectorBase.org/) (Lawson et al., 2009) and FlyBase (5148 sequences, vFB2011_08 Dmel Release 5.40) (http://FlyBase.org/) (Tweedie et al., 2009), respectively, with FlyBase also providing the fully annotated set of fruit fly essential genes used in the study. FASTA formatted protein sequences were further formatted into BLAST searchable databases using the fastacmd command within the BLAST package. To identify orthologs for the DEG genes within *A. pisum*, *A. gambiae* and *A. mellifera*, the set of DEG genes from *D. melanogaster* was used to search Individual Insect protein databases using the BLASTP program from the BLAST package with an E-value cut-off of $1E^{-03}$ to retrieve the top hits.

Protein sequence databases, primarily the NCBI's nr and RefSeq databases (Pruitt et al., 2009) were searched for DEG genes remotely on the NCBI server, using restricted insect search parameters, using standalone BLAST (Altschul et al., 1990, 1997; Camacho et al., 2009), with the BLASTP program operating with E-value cut-off of $1E^{-03}$, retrieving top hits only. A Linux script was written to compare BLAST results across these species to find the most significant orthologous hits for each protein in *D. melanogaster*. A standard set of orthologs was generated for *D. melanogaster* essential proteins, and used with the corresponding orthologs in *A. pisum* (2,546 sequences) throughout this study for structural, chemical and protein family analysis.

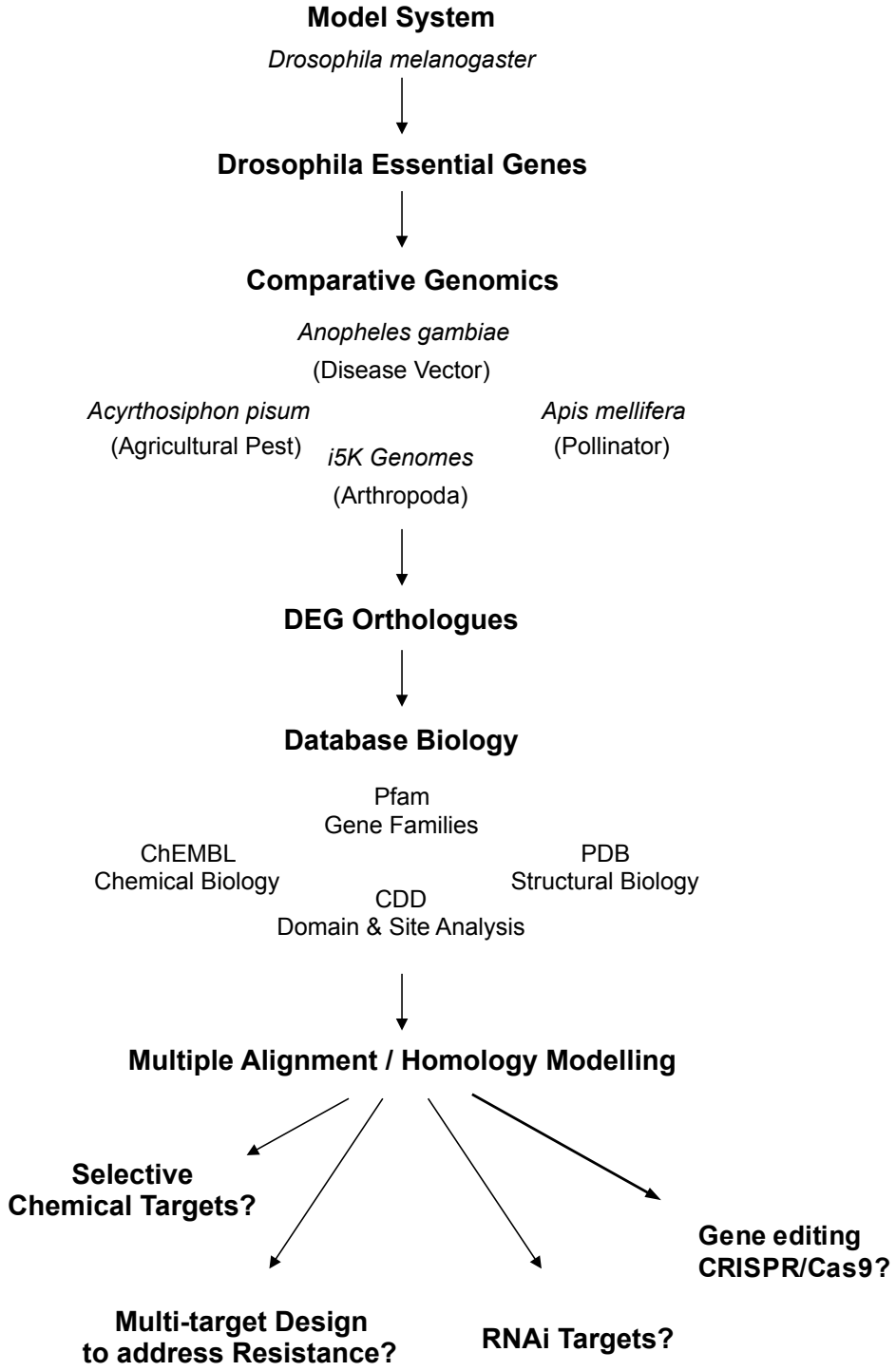


FIGURE 7.1: The diagram outlines a framework for insecticide target discovery, integrating the techniques described in this study within a matrix of genome and structural information becoming available from projects such as the i5K initiative. The three genomes used in this study (*A. pisum*, *A. gambiae*, *A. mellifera*) are indicated, together with future i5K genomes. The databases used in this study, Pfam, ChEMBL, PDB and CDD, are also shown. Resulting molecular targets can either be used on their own, or as part of a family-based approach to multi-target design. The mRNAs encoding the DEG set could also be used as genetic targets in their own right, using techniques such as RNA interference (RNAi) or CRISPR/Cas9.

7.2.3 Structural and chemical informatics

To review structural information for individual proteins, a pre-formatted Protein Data Bank (PDB) database (Wang et al., 2002) was downloaded from the NCBI server and interrogated using standalone BLAST, again using the BLASTP program with an E-value cut-off of $1E^{-03}$, retrieving top hits only. A FASTA formatted version of the ChEMBL database (v15) was downloaded from the European Bioinformatics Institute (EBI) (https://www.ebi.ac.uk/chembldb) (Gaulton et al., 2011) and searched using standalone BLAST, again using the BLASTP program with an E-value cut-off of $1E^{-03}$, retrieving top hits only.

In parallel, the Conserved Domain Database (CDD) was remotely searched at NCBI for those *A. pisum* proteins sharing ChEMBL homology with an E-value cut-off of $1E^{-03}$.

7.2.4 Protein family analysis

The protein family database Pfam (v26, 13,672 families) was downloaded from the Sanger Institute (http://pfam.sanger.ac.uk) (Punta et al., 2012). HEMMER3 software to search the Pfam database locally was downloaded from Janelia Farm (http://hmmer.janelia.org) (Finn et al., 2011). The Pfam databases, Pfam-A and Pfam-B were searched locally for DEG orthologs from the standard dataset for all four insects (*D. melanogaster*, 5,103 sequences; *A. pisum*, 2,546 sequences; *A. gambiae*, 2,670 sequences; and *A. mellifera*, 2,475 sequences) using the pfam_scan.pl script with a cut-off E-value gathering (GA) threshold for Pfam-A and cut-off E-value threshold of $1E^{-03}$ for Pfam-B. Active site prediction was enabled, as was clan overlap to analyze the most significant hits. The Pfam database at the Sanger Institute was searched for the key word "Heat Shock" to fetch heat shock associated protein families. Individual protein families were then populated with insect orthologs, and graphed using R programming (http://www.R-project.org/.).

7.2.5 Multiple alignments and phylogeny

Proteins of interest were selected from the insect orthologs and multiple alignments constructed to study conservation and variation across the insect species and *Homo sapiens*. A JAVA alignment editor (Jalview v2.7) was used to analyze the multiple alignments, which were created using the MAFFT program (v7.182) within Jalview (Waterhouse et al., 2009). The PHYLIP (v3.6) package was used to construct phylogenetic trees using the neighbor-joining method of Felsenstein (2005), edited using FigTree (v1.3.1) (http://tree.bio.ed.ac.uk/software/figtree).

7.3 Results

7.3.1 *Drosophila* essential gene orthologs in pest, pollinator and vector

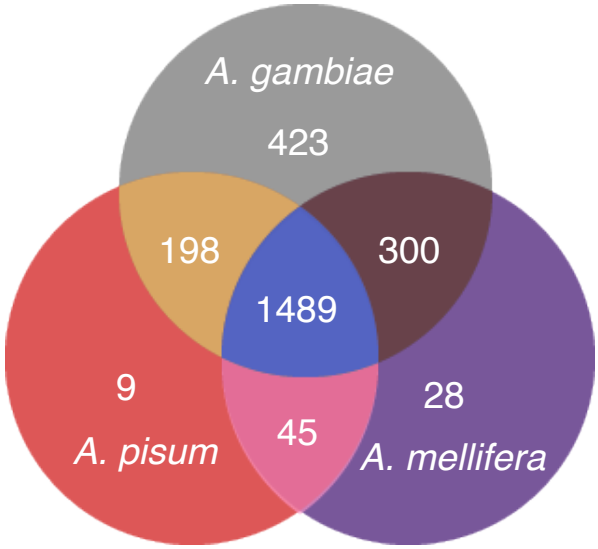
In this study *A. pisum* and *A. gambiae* were used as examples of important insect pests, the former a vector of numerous plant viruses (Hogenhout et al., 2008), the latter a vector of human parasites (Lawson et al., 2009), and *A. mellifera* as an example of a beneficial pollinator. Insecticidal mechanisms of most relevance to environmentally sustainable pest management strategies would target the pests while leaving beneficial pollinators unscathed.

Genetic studies indicate that there are approximately 3,000 essential genes in *D. melanogaster* (Oh et al., 2003). A total of 2,694 DEGs were identified within FlyBase (Miklos and Rubin, 1996; Tweedie et al., 2009), and are appended as an annotated list in Appendix A, Table A7.1.

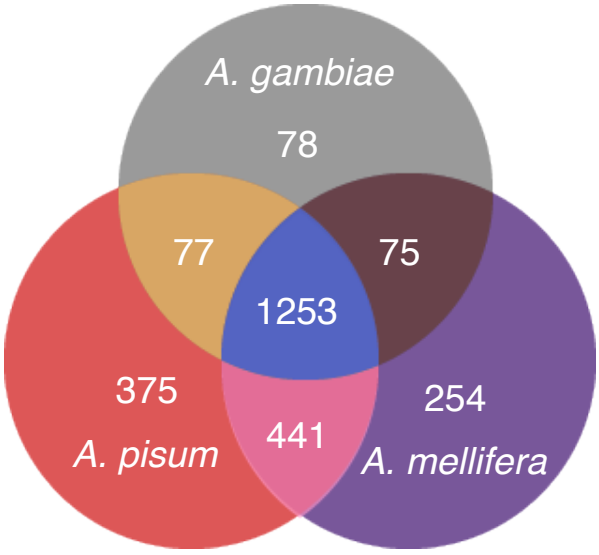
To convert our set of DEG genes into a protein framework from which to explore a broad spectrum of potentially insecticidal mechanisms, the full dataset of 2,694 genes was first used to search for the corresponding *Drosophila* essential proteins. From the original 2,694 DEG genes, a set of 5,148 corresponding non-redundant *Drosophila* proteins were identified, again detailed in Appendix A, Table A7.2.

Orthologs of these proteins were then identified in the genomes of *A. pisum*, *A. gambiae* and *A. mellifera* using both species-specific genome databases (FlyBase, VectorBase, AphidBase and BeeBase), or the more comprehensive NCBI nr database. Systematic analysis across the entire set of DEG gene orthologs within the current insect genome databases is complicated by varying degrees of assembly and annotation within the individual genome projects. In our hands, comparison between the DEG orthologs using the nr database gave a more complete picture of the well-annotated fraction of the genomes than did the genome databases themselves. Nevertheless, in all datasets, between 60 and 70 percent of the DEG orthologs could be unequivocally related to well-annotated *D. melanogaster* or *A. gambiae* counterparts.

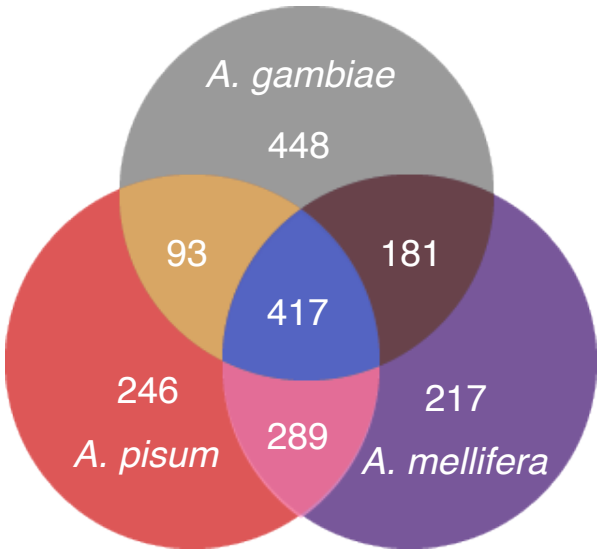
For analysis purposes, DEG proteins were divided into three arbitrary categories: 1) exact matches (E-value 0), whose biochemical and physiological functions are very likely to be conserved between *D. melanogaster* and other insect species; 2) highly similar matches ($1E^{<100}$, hereafter E-value <100) whose biochemical and physiological functions are likely to be conserved; and 3) less similar matches ($1E^{>100}$, hereafter E-value >100), whose annotations are more tentative. The number of DEG orthologs in the pest and pollinator genomes in each of these three categories using data from the nr database is summarized diagrammatically in Figure 7.2. A detailed comparison of the same data with that from individual genome databases, together with a complete inventory of hits, is presented in Appendix B, Figure B7.2.



(A) Panel A (E-value 0)



(B) Panel B (E-value <100)



(C) Panel C (E-value >100)

FIGURE 7.2: The total protein complements of *A. gambiae*, *A. pisum* and *A. mellifera* were compared by BLAST analysis to the complement of *D. melanogaster* essential proteins extracted from FlyBase, using information from the NCBI nr database. Results were classified on the basis of exact match (E-value 0, Panel A), high similarity (E-value <100, Panel B), or lower similarity (E-value >100, Panel C).

7.3.2 DEG orthologs compared

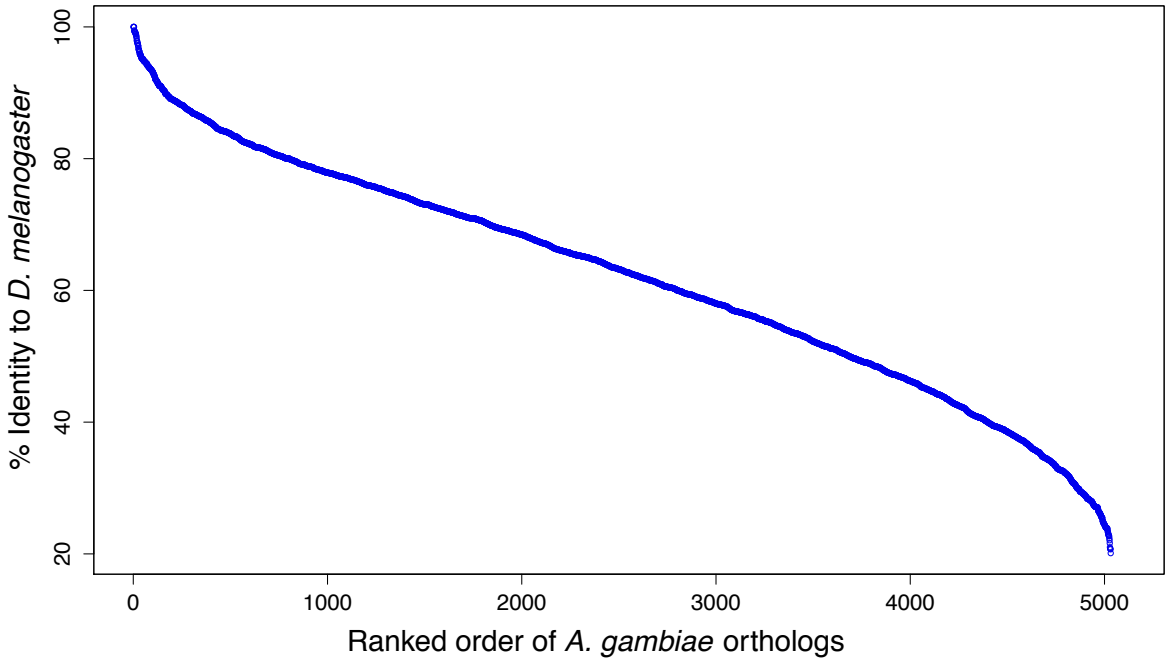
Essential gene orthologs that differ significantly in structure between specific insect pests could provide narrow spectrum, species-specific insecticide targets, while those that are shared between pests but not conserved in pollinators might provide genuinely broad-spectrum targets.

To compare the DEG orthologs in the three insect species, proteins in *A. pisum* and *A. mellifera* were compared to those in *A. gambiae* and *D. melanogaster* by systematically ranking them according to

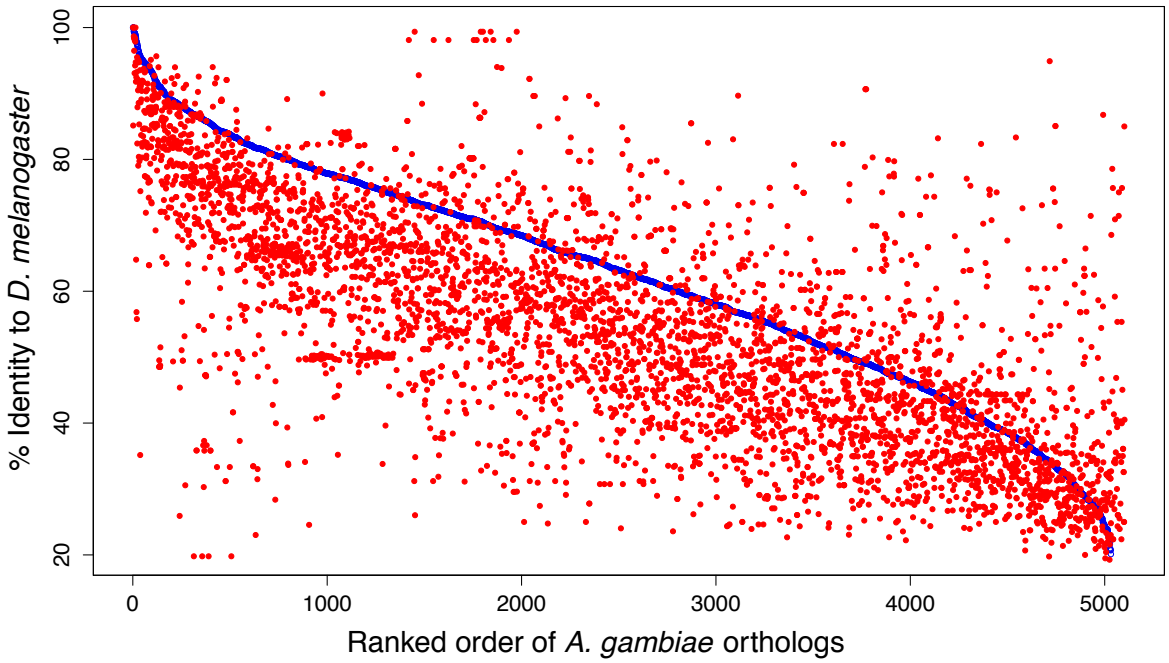
their percentage identity to individual proteins within an ordered list of *A. gambiae* DEG orthologs (Figure 7.3A). Considerable sequence variation between individual DEG orthologs is evident in the four species when they are compared in this way, exemplified by the dataset from *A. pisum* (Figure 7.3B).

To define the target space available for insecticide discovery within the orthologous DEG protein sets, an analysis of the DEG orthologs present in the four insects was carried out using the ChEMBL and PDB databases. When ChEMBL was used, under very conservative thresholds (50% identity score), to analyze the orthologous dataset described for *A. pisum*, 403 chemically tractable orthologs were seen, corresponding to an original set of 851 non-redundant *Drosophila* essential proteins (Figure 7.3C). An annotated list of these DEG proteins, with their ChEMBL identifiers, is provided in Appendix A, Table A7.3.

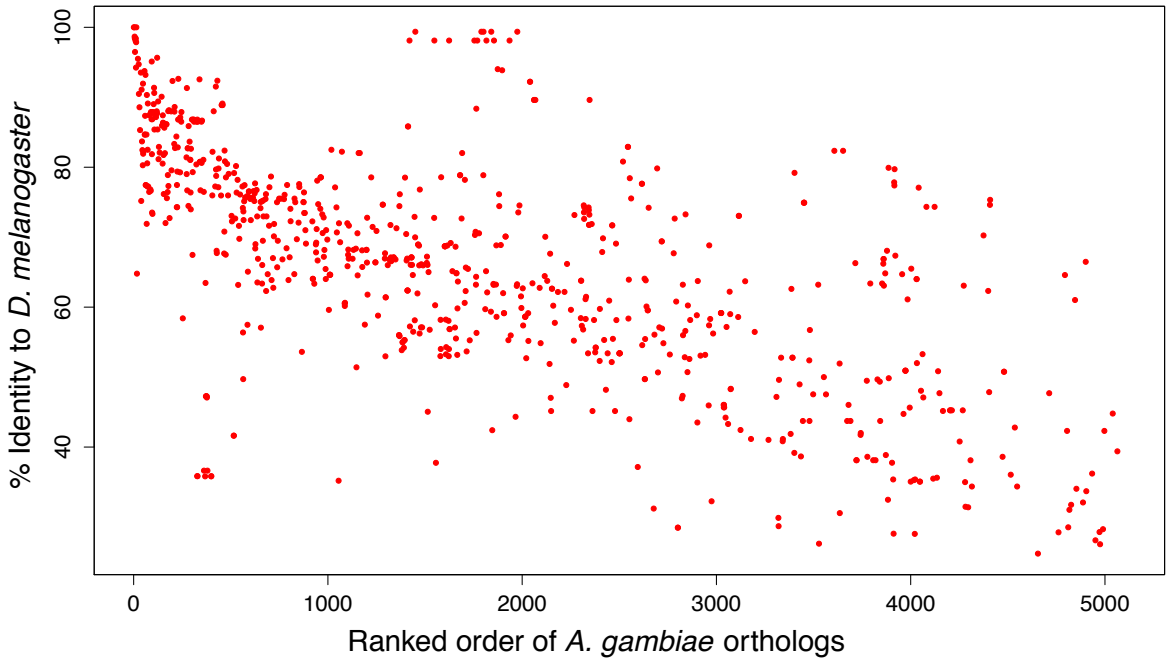
Prioritization of the targets identified using ChEMBL for rational insecticide design requires detailed mapping of available ligand space onto orthologous ligand-binding sites. For this to be achievable, protein structure information suitable for homology modeling is required. Nearly 1,000 (959, 19%) of the *A. pisum* DEG orthologs have closely related structures in the PDB database, indicated in Figure 7.3D (annotated in detail in Appendix A, Table A7.3). While much of this structural information is comparative, and derived from distant relatives such as those in *H. sapiens*, it does provide a measure of the feasibility of mounting structure-based design campaigns against those proteins in which functional sequence variation can be identified. More than three quarters (78%) of the ChEMBL hits have associated PDB structures, while only approximately one third (31%) of the *A. pisum* DEG orthologs with corresponding PDB structures have related ChEMBL hits. Using these conservative thresholds (50% identity score), 314 of the DEG protein orthologs have corresponding entries in both ChEMBL and PDB, and therefore may represent suitable starting points for structure-based insecticide design.



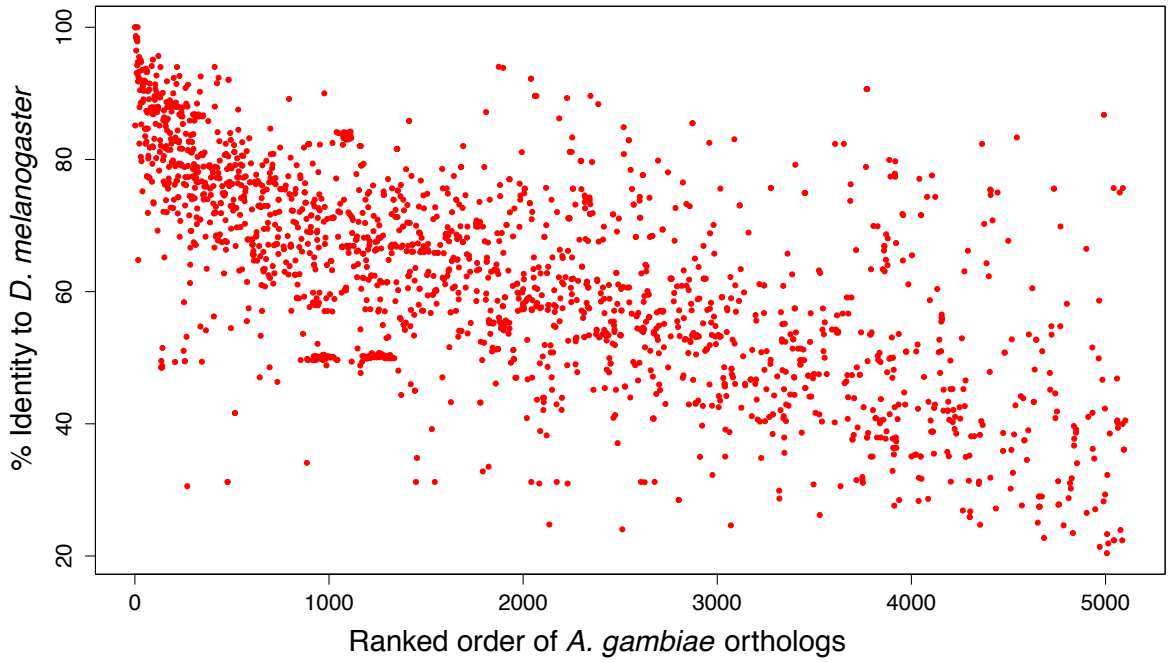
(A) *A. gambiae* DEG protein orthologs ranked against their *D. melanogaster* orthologs



(B) *A. pisum* DEG protein orthologs ranked against those in *A. gambiae*



(C) *A. pisum* DEG protein orthologs with a ChEMBL score above 50% identity



(D) *A. pisum* DEG orthologs with a PDB score above 50% identity

FIGURE 7.3: *A. gambiae* orthologs of the *D. melanogaster* proteins corresponding to the complement of DEG genes were identified and placed in ranked order based on overall protein %identity to their original *D. melanogaster* counterparts (Panel A). The protein %identity of the set of *A. pisum* orthologs was then related to this ranked set of *A. gambiae* proteins (Panel B). Using the conservative 50% identity score in ChEMBL as a threshold, at least 403 proteins amongst the *A. pisum* DEG orthologs appear chemically tractable as discovery targets (Panel C). Using a similar 50% identity score as a threshold for matches in the PDB, structural information was identified for 959 *A. pisum* DEG orthologs (Panel D).

7.3.3 Pfam analysis of the DEG orthologs

The biochemical and pharmacological functions of the DEG orthologs were then explored, to identify attractive mechanistic targets for chemical discovery. To achieve this, DEG orthologs were parsed with the databases Pfam (Finn et al., 2010) and GO (Ashburner et al., 2000) to provide a genome-wide survey of the biochemical pathways in which the orthologs appear.

Appendix B, Figure B7.3 shows a Pfam analysis of the proteins present within one of these families, the extended heat-shock associated family, for all four insects. Heat shock associated proteins represent 273 of the 13,672 protein families in the version (v26) of Pfam used for these studies, and 112 of the 2,902 Pfam protein families we observed amongst the *A. pisum* DEG orthologs. We have performed similar analyses for other Pfam families. The results reveal a number of interesting features.

Related orthologs with similar domain features share similar Pfam scores. Many of the 5,148 DEG proteins fall into this category, reflecting highly conserved domain architectures across the four insect genomes. Conversely, structural variation amongst orthologs within a Pfam family can be indicated by variation in their Pfam scores. Two examples of highly conserved orthologs from the heat shock family, extracted from the complete set shown in the Appendix B, Figure B7.3, the heat shock proteins (Hsp) Hsp70 and Hsp90, are indicated in Figure 7.4.

Certain Pfam families are highly represented amongst the DEG genes. One example is the protein kinase gene family (PF00069, Key 8 in Appendix B, Figure B7.3, also shown in Figure 7.4) in which 471 DEG orthologs are distributed between the four insects. The Pfam scoring system can also identify proteins that occur differentially between the four insects (see for example Keys 54, 63, 81, 83 and 107 in Appendix B, Figure B7.3).

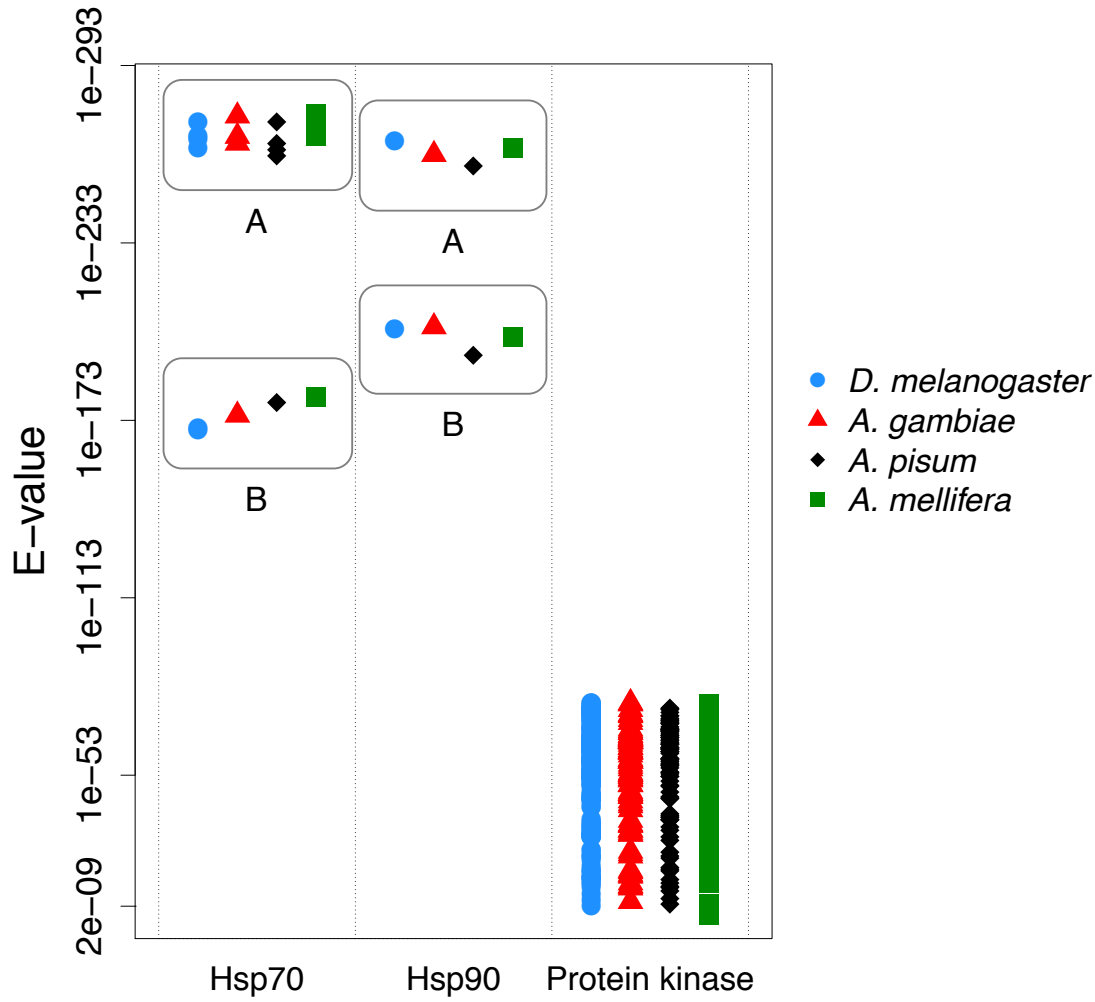


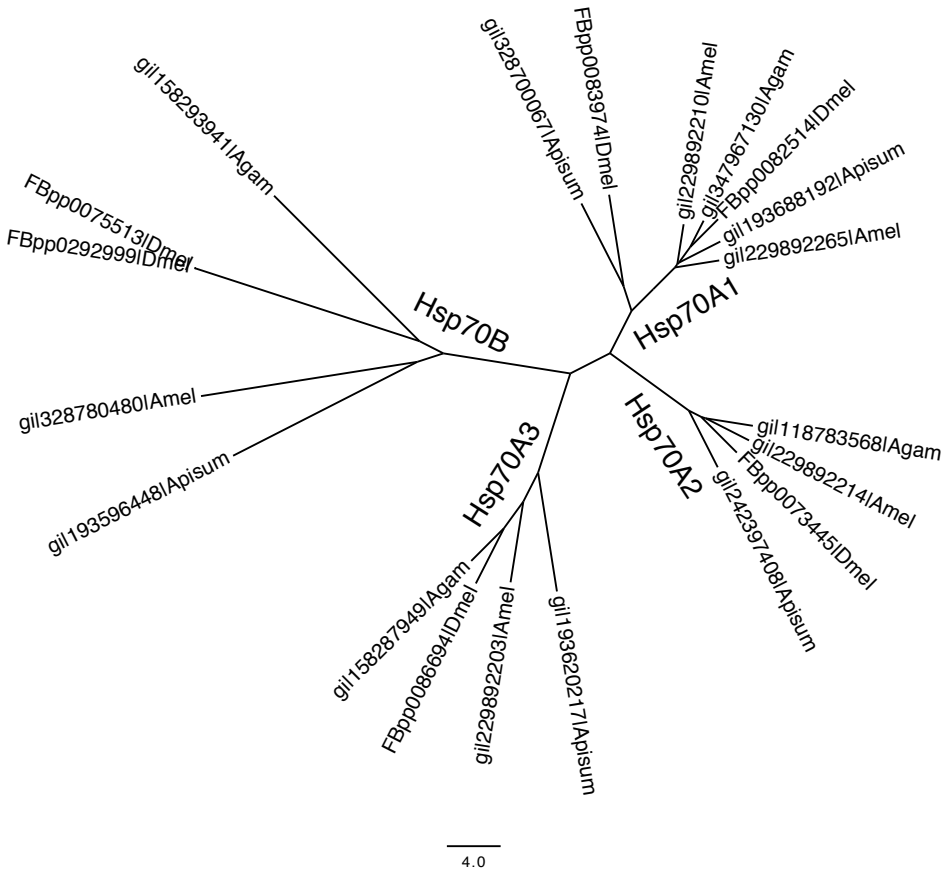
FIGURE 7.4: *D. melanogaster* DEG and their orthologs were subjected to Pfam analysis. Results are shown for two individual heat shock associated protein families, Hsp70 (Pfam family PF00012.15) and Hsp90 (Pfam accession family PF00183.13), from *D. melanogaster*, *A. gambiae*, *A. pisum* and *A. mellifera*, indicated as Hsp70A, Hsp70B, Hsp90A and Hsp90B, respectively. Also shown for comparison are members of the protein kinase family (Pfam family PF00069.20). A full analysis of all the heat shock-associated protein families across the four genomes is provided in Appendix B, Figure B7.3.

7.3.4 Analysis of related Pfam protein clusters

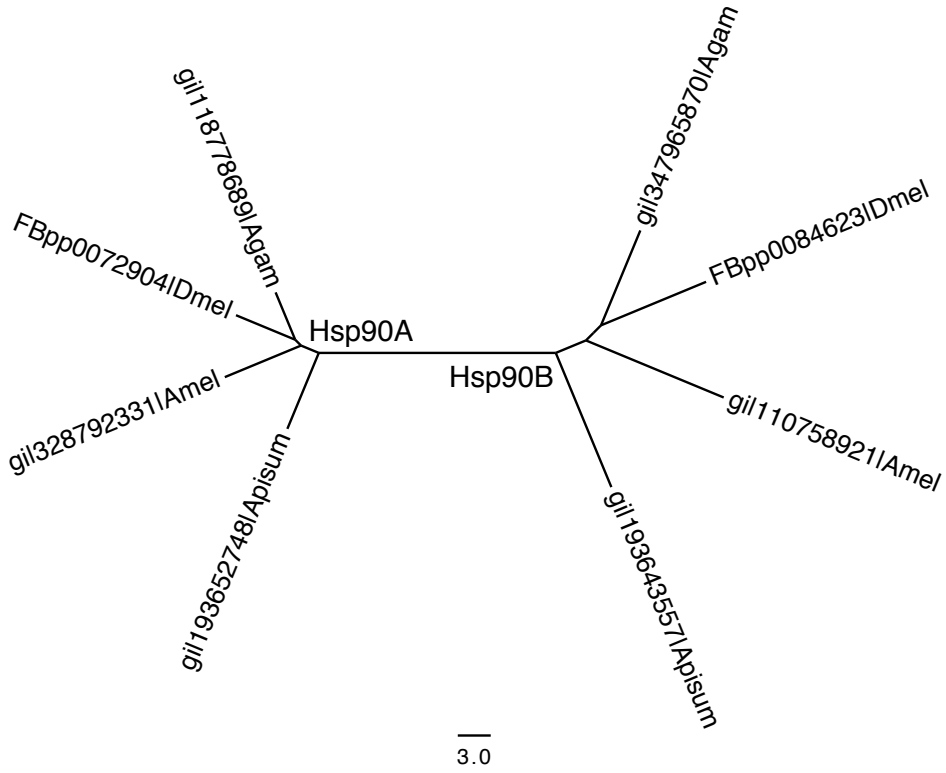
To examine in more detail the clustering within the ortholog subsets observed using Pfam, we performed multiple alignment and phylogenetic analysis on each Pfam cluster. Phylogenetic relationships for the four Hsp70 and Hsp90 clusters (previously addressed in Figure 7.4) are illustrated in Figure 7.5. The corresponding multiple alignments are shown in Appendix B, Figure B7.4.

Importantly for some of the agrochemical design strategies discussed below, clusters sharing Pfam signatures often fall into separate biochemical pathways. This is well exemplified by the molecular clusters within the Hsp70 and Hsp90 families, where family members share a structurally conserved ATPase active site, but are clearly differentiated from each other on the basis of both multiple alignment and phylogenetic analysis, and are known to participate in separate biochemical and developmental networks.

The conservation of ligand-binding sites within suites of functionally differentiated proteins has important discovery implications that can result in unexpected and sometimes valuable off-target effects. Two of the clusters shown here, Hsp90A and Hsp90B, have direct counterparts in the human, Hsp90A and Hsp90B1 (also known as endoplasmin). The N-terminal ATPase site in human Hsp90A is substantially shared by endoplasmin, giving rise to an experimentally observed cross-reactivity in pharmaceutical drugs designed to target Hsp90 (Usmani et al., 2010). Designing compounds to target more than one member of a gene family at a time (multi-target design) can be advantageous in many contexts (for discussion, see Koutsoukas et al., 2011). In insecticide discovery multi-target design may help to address resistance development. Pfam mapping is one way of quickly identifying these relationships.



(A) Hsp70 Pfam cluster



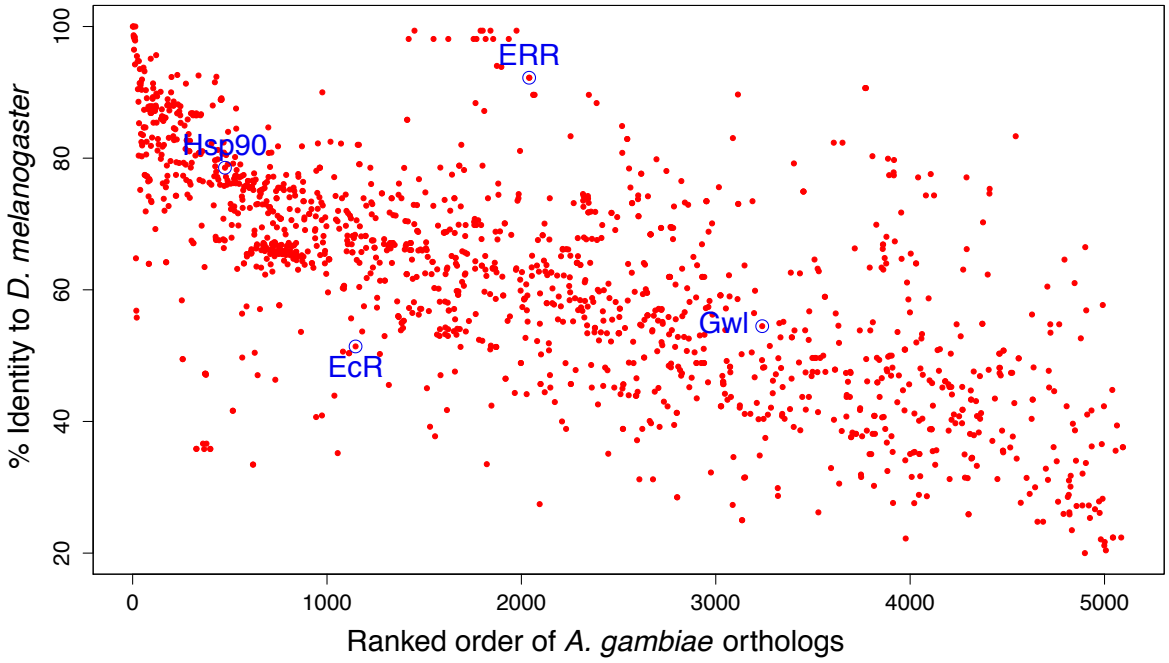
(B) Hsp90 Pfam cluster

FIGURE 7.5: Phylogenetic analyses of Hsp70 and Hsp90 Pfam clusters using the neighbor-joining method. Panel A, Hsp70 Pfam cluster showing three separate phylogenetic groups, labeled A1-A3, derived from Hsp70A, and one phylogenetic group derived from Hsp70B. Panel B, Hsp90 Pfam cluster. Panel C, multiple alignment of Hsp90 proteins from *A. pisum* (Hsp90A and Hsp90B) and *H. sapiens* (Hsp90A and Hsp90B1). The CDD features shown above the alignments are derived from the *H. sapiens* Hsp90A sequence.

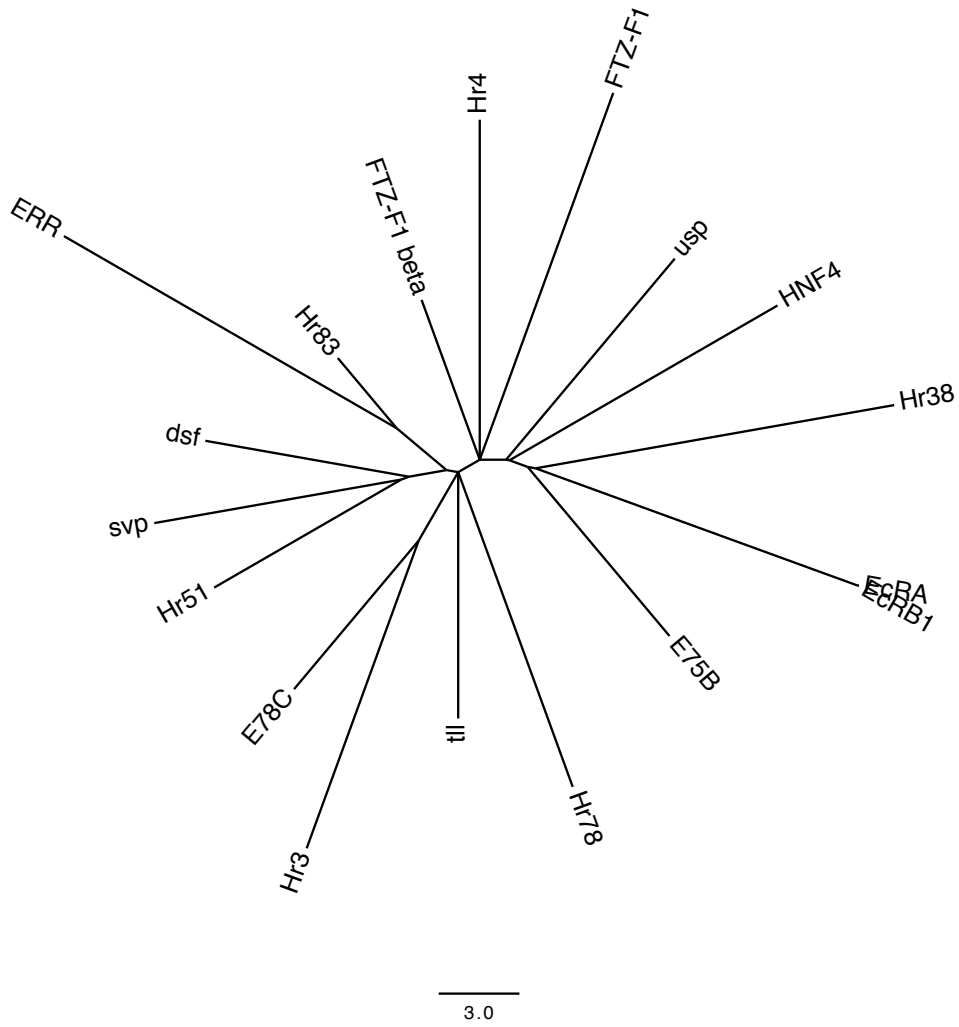
7.3.5 Initial analysis of chemically-tractable DEG orthologs

The majority of the chemically-tractable insect proteins we observe using the ChEMBL database have counterparts within the 40,000 well-characterized protein domains in the NCBI's CDD (Marchler-Bauer et al., 2011). The CDD also contains ligand interaction data for these domains.

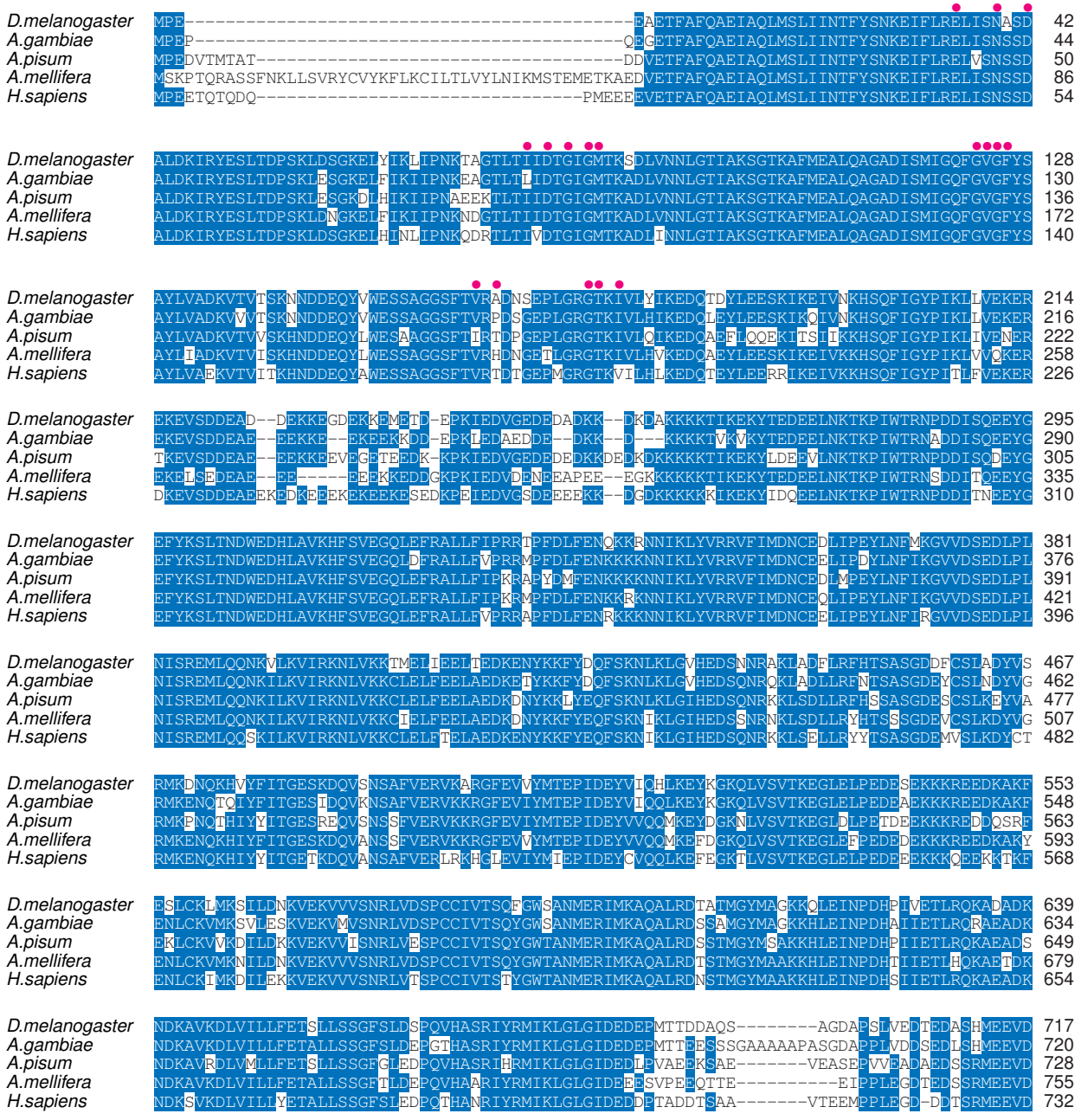
Chemically tractable DEG proteins, identified using ChEMBL and mapped by reference to either the ranked *A. gambiae* or *A. mellifera* protein sets, were examined in detail, using information on domain architecture contained within the CDD. Using a conservative 50% identity parameter within ChEMBL as a threshold for protein similarity, 403 chemically tractable *A. pisum* DEG orthologs were identified (Figure 7.6A). A detailed analysis of the complete set of chemically tractable DEG orthologs in the four insects is beyond the scope of this initial study, but results for four orthologs, Hsp90A, Greatwall (Gwl), Estrogen-related receptor (ERR) and Ecdysone receptor (EcR), across the four insects are shown in Figure 7.6C-F as representative examples.



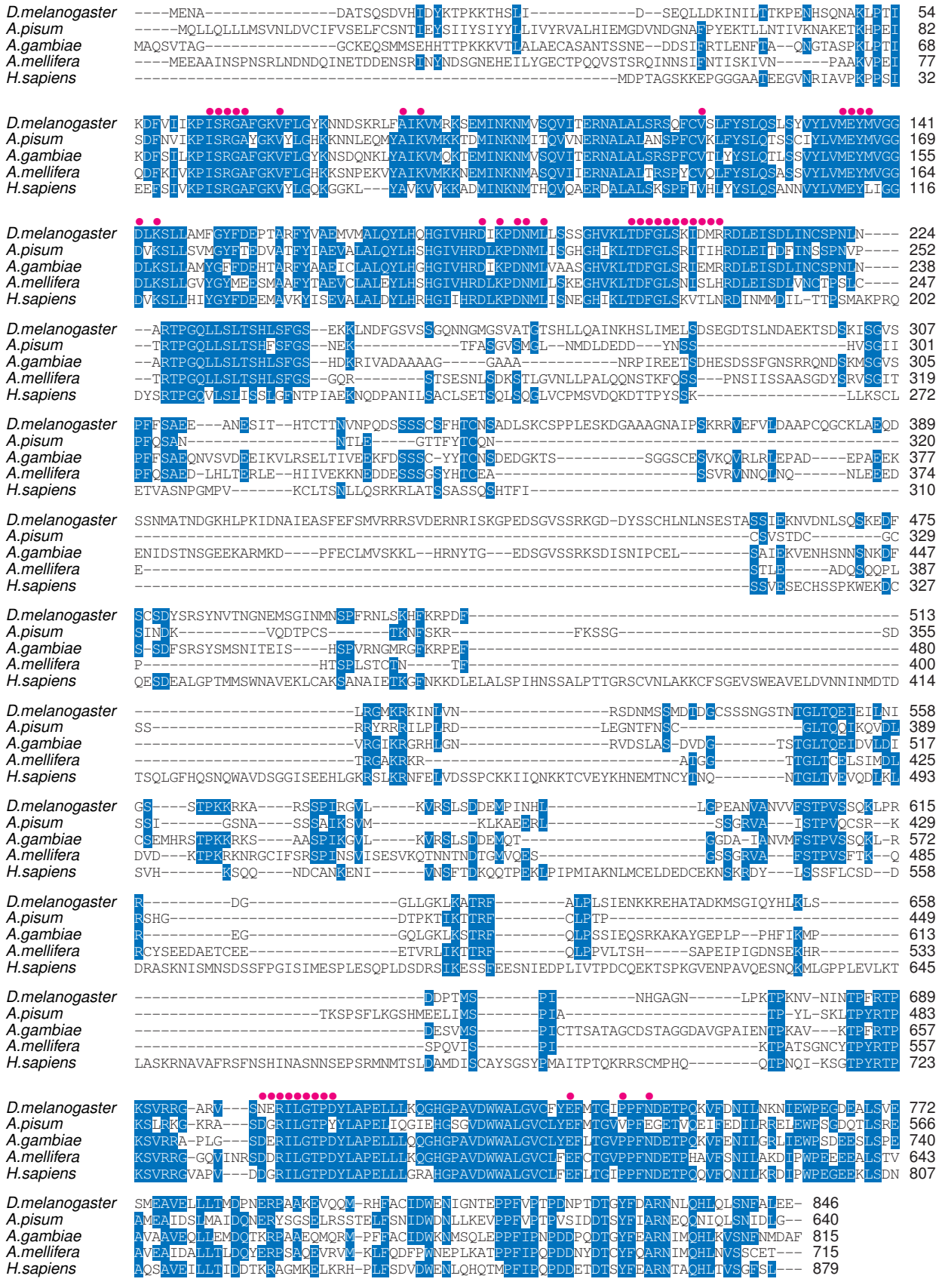
(A) *A. pisum* ChEMBL hits



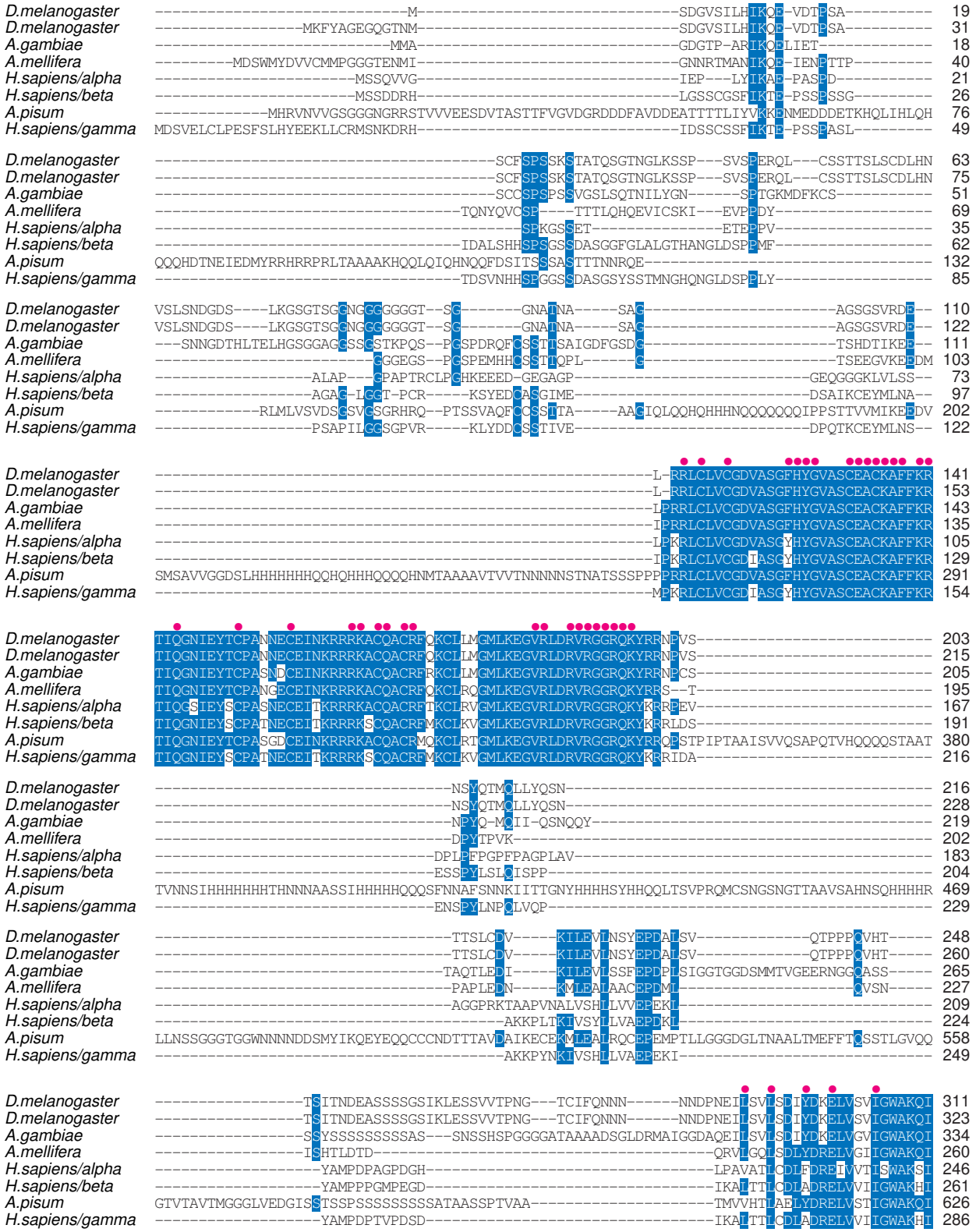
(B) Phylogenetic relationships within the ligand binding domains from the 18 DEG ortholog nuclear receptors of *A. pisum*

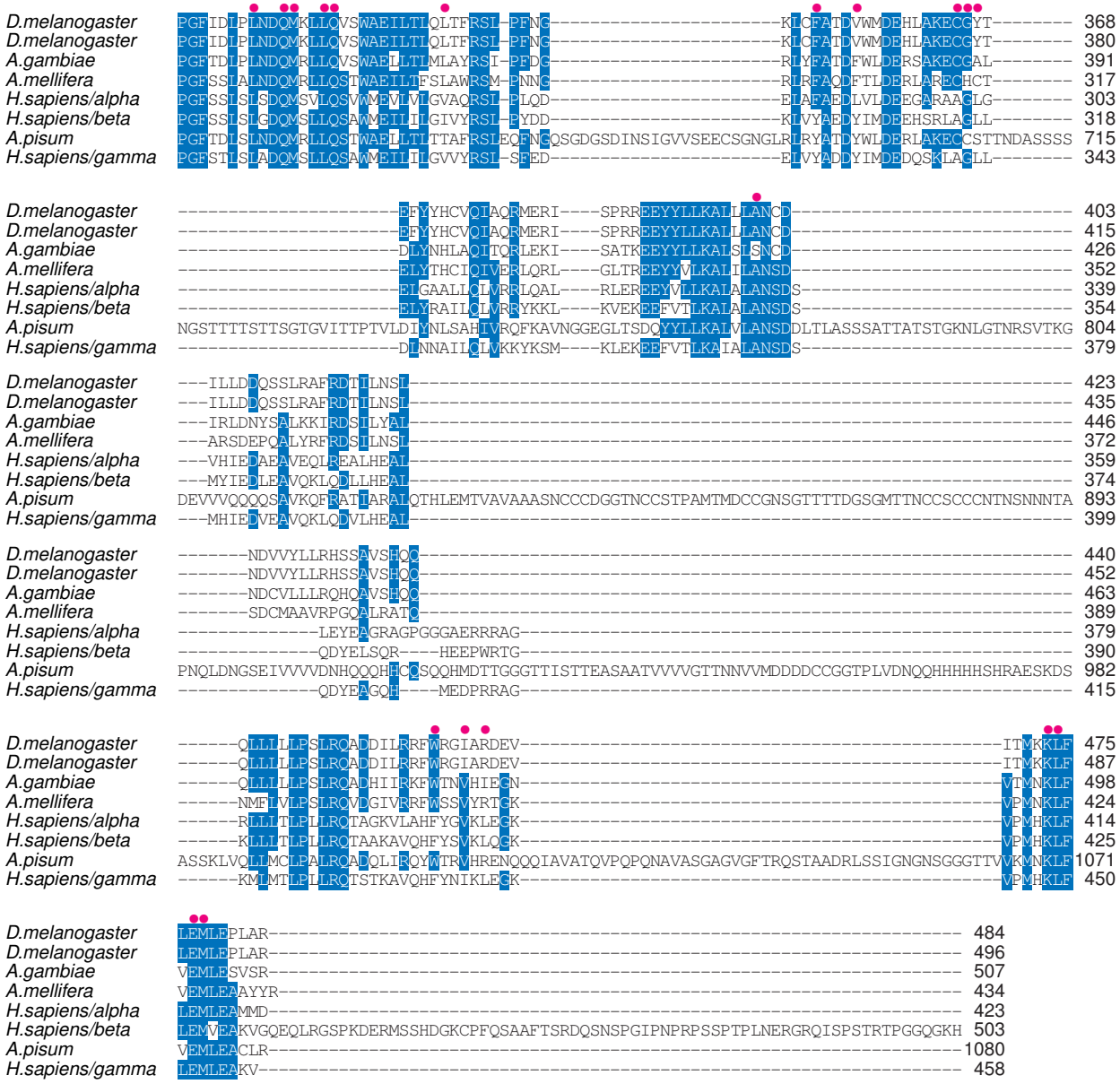


(c) Multiple Alignments of Hsp90A DEG orthologs

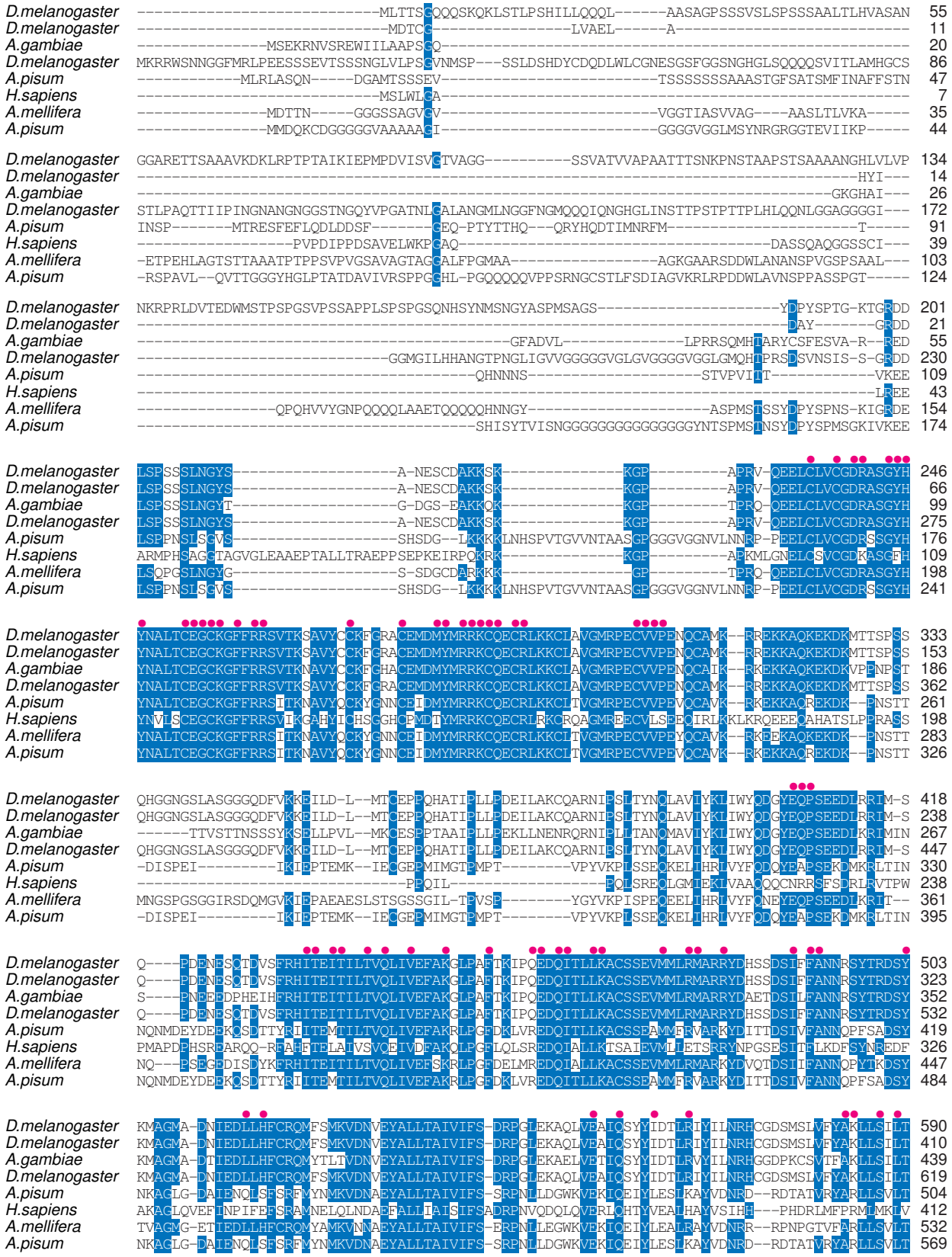


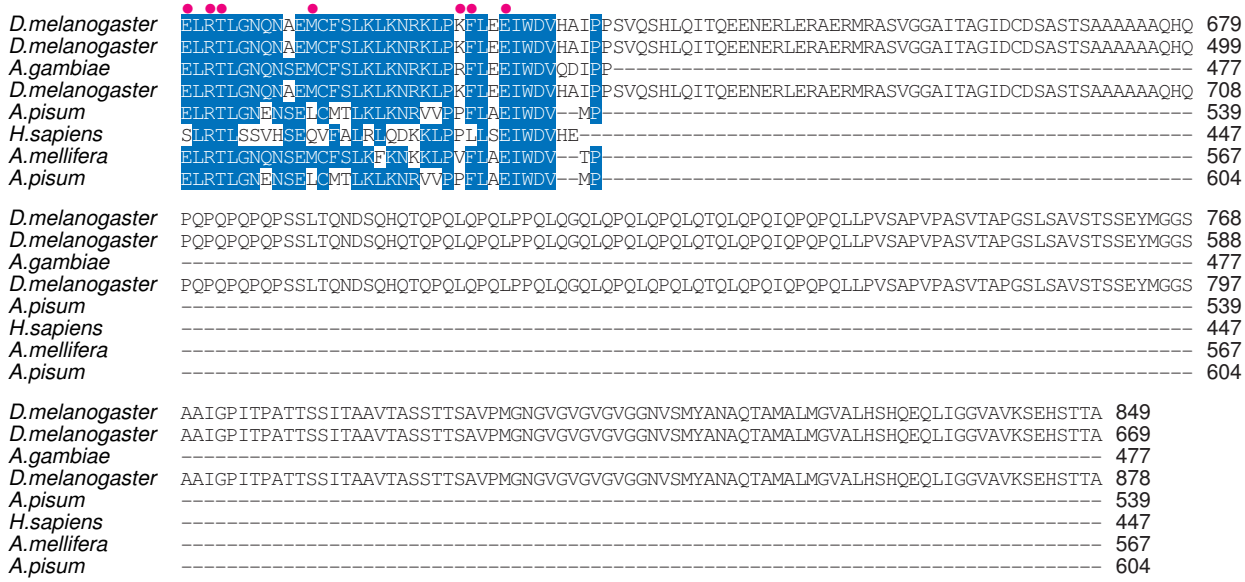
(D) Multiple Alignments of Gwl orthologs


D.melanogaster
D.melanogaster
A.gambiae
A.mellifera
H.sapiens/alpha
H.sapiens/beta
A.pisum
H.sapiens/gamma



(E) Multiple Alignments of ERR orthologs


D.melanogaster
D.melanogaster
A.gambiae
D.melanogaster
A.pisum
H.sapiens
A.mellifera
A.pisum



(F) Multiple Alignments of EcR orthologs

Flybase ID	Protein	Type	Title	Coordinates	Complete size[1]	Mapped size[2]	Source domain[3]
FBpp0292999	HSP70	generic[4]	active site	G339, G340, S341, I344, V369	23	5	198375
FBpp0292999	HSP70	generic[4]	MgATP binding site	G339, G340, S341, I344	21	4	198375
FBpp0292999	HSP70	generic[4]	N- and C-terminal domain interface	V369, S370, R371, A373, A374, Q376, C377	70	7	198375
FBpp0083089	Gwl	specific[5]	active site	I63, S64, R65, G66, A67, V71, A85, K87, V119, M135, E136, M138, D142, K144, D181, K183, D185, N186, L188, T198, D199, L202, I703, L704, G705, T706, P707, D708, E735, P741, N744	31	31	173701
FBpp0083089	Gwl	specific[5]	ATP binding site	I63, S64, R65, G66, A67, V71, A85, K87, V119, E136, Y137, M138, D142, D181, K183, D185, N186, L188, T198, D199	20	20	173701
FBpp0083089	Gwl	specific[5]	substrate binding site	A67, D142, K144, D181, K183, D185, L202, I703, L704, G705, T706, P707, D708, E735, P741, N744	16	16	173701
FBpp0083089	Gwl	specific[5]	activation loop (A-loop)	T198, D199, F200, G201, L202, S203, K204, I205, D206, M207, R208, N700, E701, R702, I703, L704, G705, T706, P707, D708	20	20	173701
FBpp0076447	ERR	specific[5]	ligand binding site	L290, L293, Y297, E300, L338, F351, V355, C365, G366, Y367, W460, I463, R465	13	13	132744
FBpp0076447	ERR	specific[5]	coactivator recognition site	I305, L319, Q322, M323, L326, Q327, K473, L474, E477, M478	10	10	132744
FBpp0076447	ERR	specific[5]	dimer interface	A400	1	1	132744
FBpp0076447	ERR	specific[5]	zinc binding site	C115, C118, C132, C135, C151, C157, C167, C170	8	8	143544
FBpp0076447	ERR	specific[5]	DNA binding site	R113, F125, H126, Y127, G128, E133, A134, K136, A137, F138, K140, R141, Q144, R164, K165, Q168, R171, V185, R186, R189, V190, R191, G192, G193, R194, Q195, K196	27	27	143544
FBpp0085349	EcR	specific[5]	ligand binding site	E406, Q407, P408, I434, T435, I437, T438, T441, M475, R478, M479, R482, I490, F492, A493, Y503, L515, M602	18	18	132736
FBpp0085349	EcR	specific[5]	putative coactivator recognition site	Q443, V446, K450, F455, Q460, E461, Q463, I464, L467, K468, K615, F616, E619	13	13	132736
FBpp0085349	EcR	specific[5]	heterodimer interface	H517, E554, Q557, I561, R565, A583, K584, S587, L589, E591, R593, T594	12	12	132736
FBpp0085349	EcR	specific[5]	zinc binding site	C235, C238, C252, C255, C271, C277, C287, C290	8	8	143535
FBpp0085349	EcR	specific[5]	DNA binding site	G244, Y245, H246, Y247, E253, G254, K256, F258, R260, R261, R284, R285, K286, Q288, R291, C304, V305, V306, P307	19	19	143535
FBpp0085349	EcR	specific[5]	heterodimer interface	D240, R241, M281, Y282, R285	5	5	143535
gi|154146191|	HSP90	specific[5]	ATP binding site	E47, N51, D54, I91, D93, G95, G97, M98, G135, V136, G137, F138, V172, T174, G183, T184, V186	17	17	28956
gi|154146191|	HSP90	specific[5]	Mg2+ binding site	N51	1	1	28956
gi|154146191|	HSP90	specific[5]	G-X-G motif	G95, G97, G135, G137	4	4	28956

[1]Complete size : The total number of residues in the conserved feature/site that has been annotated on the domain model.
[2]Mapped size : The number of residues in the query protein sequence that match residues in the conserved feature/site that was annotated on the domain model.
[3]Source domain : The PSSM ID of the domain model on which the conserved feature/site was annotated.
[4]Specific : conserved features/sites that were mapped onto the set of query sequences from specific hits.
[5]Generic : conserved features/sites that were mapped onto the set of query sequences from non-specific hits, because those non-specific hits belong to a superfamily whose representative is an NCBI curated domain that has such annotations.

(G) Key to the structural features indicated in Panels C-F

FIGURE 7.6: Multiple alignments were performed for four representative chemically tractable DEG orthologs: Hsp90, Gwl, ERR and EcR, together with their corresponding human homologs. Panel A, total dataset of ChEMBL hits, with Hsp90, Gwl, ERR and EcR highlighted; panel B, phylogenetic relationships within the LBDs of the 18 DEG nuclear receptor orthologs in *A. pisum*, a consensus phylogenetic tree was constructed using the neighbor-joining method; panel C, multiple alignment of Hsp90A DEG orthologs; panel D, multiple alignment of Gwl serine/threonine kinase orthologs; panel E, multiple alignment of the steroid receptor ERR orthologs, together with its human homologs ESRR-alpha, ESRR-beta and ESRR-gamma; panel F, multiple alignment of EcR orthologs. Indicated above the alignments are the positions of the conserved amino acids (coloured in blue) that characterize structural features (represented as "•") within the respective *D. melanogaster* proteins, further compiled in the associated Key (panel G).

The first of these examples chosen, Hsp90A, shows little variation in the sequence of its open reading frame between the four insects, and is also highly conserved in the human (Figure 7.6C). The active site amino acids that characterize its ATPase domain, extracted from the CDD, as well as the immediate context of the active site, are highly conserved within the set of four insect proteins and are shared by the corresponding *H. sapiens* sequence (Figure 7.6C). The active site of Hsp90A thus appears a poor target for selective insecticide discovery.

A second representative protein, the *Drosophila* Gwl / MASTL serine-threonine protein kinase, associated with cellular entry into mitosis (Vigneron et al., 2011), shows different properties. Here, although the amino acids of the N-terminal ATP-binding kinase active site are quite highly conserved within the four insects, those within the activation loops and substrate-binding sites, split between the N- and C-terminal domains, show lower conservation (Figure 7.6D). Variations at the latter sites have enabled selective drug design in human serine-threonine kinases (for example the serine/threonine protein kinase CK2, Niefind and Issinger, 2010).

A different pattern of variation is seen in a third essential protein, ERR Figure 7.6E), thought to play an essential role in metabolism during development (Tennessen et al., 2011). This has a very highly conserved N-terminal DNA-binding domain, linked to a more variable C-terminal ligand-binding domain. Variations in the C-terminal steroid-binding domain of ERR occur in all four of the insect species, reflecting variations within the amino acids that mediate the putative interactions with its steroid ligand. ERR is an orphan receptor belonging to the steroid hormone (NR3) subfamily of nuclear receptors (Fahrbach et al., 2012), and has previously been suggested, with certain reservations, as a potential target for insecticide discovery (Ostberg et al., 2003). Designated PF00104.25 within the Pfam database, the family of nuclear receptors to which ERR belongs also contains the ecdysteroid hormone receptor, one of the few already well-characterized insecticide targets (Carmichael et al., 2005; Holmwood and Schindler, 2009). We identified another 36 structurally related members of the nuclear hormone receptor family amongst the 5,148 DEG proteins in *D. melanogaster*, highlighting the considerable functional diversity amongst this family of insect hormone receptors, all of which are essential genes in *Drosophila*.

In the light of the findings with ERR, similar comparative analyses amongst the DEG orthologs for the EcR (Figure 7.6F), another member of the DEG repertoire, were performed. In *D. melanogaster*, the steroid hormone ecdysone triggers larval-to-adult metamorphosis and tissue remodeling (Fahrbach et al., 2012). EcR is a member of the NR1 nuclear receptor family, and binds DNA with high specificity at ecdysone response elements. As with ERR, multiple alignment studies in the four insects shows their EcRs to have a highly conserved N-terminal DNA-binding domain, accompanied by a more variable C-terminal ligand- binding domain. However, unlike the orphan ERR receptor, EcR has been the subject of much molecular biology, structural analysis and chemical design (Kothapalli et al., 1995; Billas and Moras, 2005; Carmichael et al., 2005; Dhadialla et al., 2007; Fujita and Nakagawa, 2007; Beatty et al., 2009; Soin et al., 2010; Harada

et al., 2011; Zotti et al., 2012), and several EcR ligands are currently being commercialized as environmentally-friendly insecticides (Hill et al., 2013).

With selective insecticide design in mind, we further investigated the overlap in structural features displayed by LBD domains within the entire nuclear hormone receptor family (Pfam family PF00104.25). To date, pharmacological studies of ecdysteroid agonists have focused only on single receptor pharmacology (see, for example, Soin et al., 2010; Tohidi-Esfahani et al., 2011). However, phylogenetic analysis (Figure 7.6B) indicate that several features in the EcR ligand binding domain are shared with comparable domains in phylogenetically related nuclear receptor orthologs. This may be an interesting observation to exploit for multi-target polypharmacology.

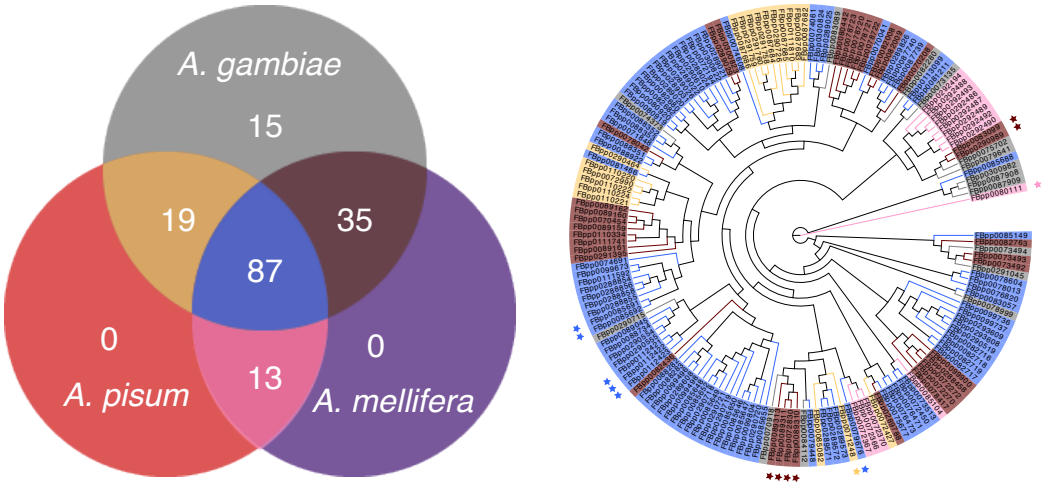
7.3.6 New insecticide targets

The combination of Pfam, ChEMBL and CDD provides a rich resource for target discovery, linking structural and chemical biology to specific gene families. There are many gene families that appear suitable for chemical design emanating from the differentially-expressed essential gene dataset.

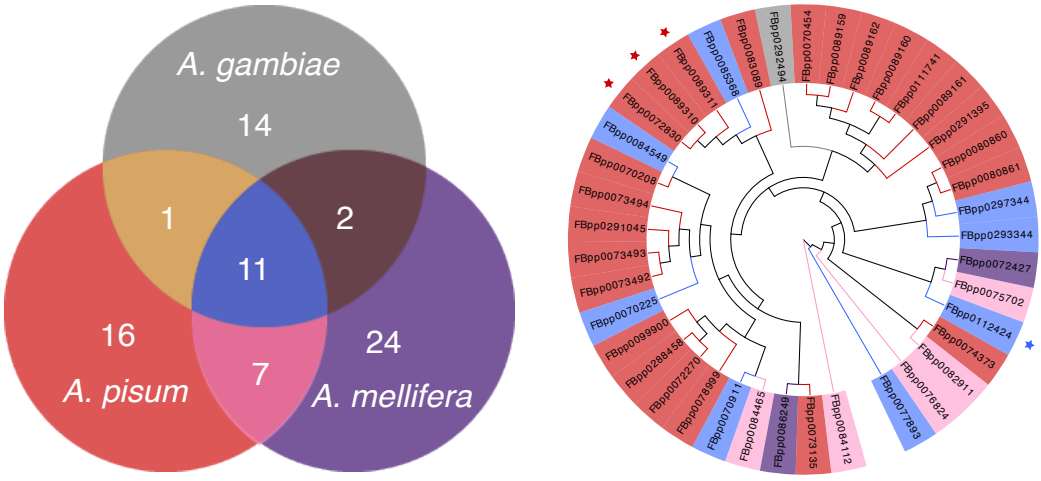
In *H. sapiens*, the protein kinases (Pfam family PF0069) have proven an especially fruitful therapeutic target class for chemical discovery (Bamborough, 2012) because of their extensive involvement in biological development and differentiation (Manning et al., 2002a,b; Manning, 2009). *Drosophila* mutagenesis has made a major contribution to delineating the functional components of these signaling networks, and continues to do so through kinome-wide gene silencing approaches (Bettencourt-Dias et al., 2004; Nybakken et al., 2005). Despite the evident druggability of the protein kinase gene family, to date there have been few systematic studies of this family in insect pests other than *Drosophila*.

To address the "essential" kinome, we extracted the orthologs of *Drosophila* essential kinases in *A. gambiae*, *A. pisum* and *A. mellifera* from the essential gene dataset (Figure 7.7), dividing them into the same three similarity categories as used earlier (as in Figure 7.2). The results confirm the expansion of specific kinases important in the regulation of mitosis in the pea aphid (IAGC, 2010), but also show the presence of a large number of new, differentially expressed essential kinases.

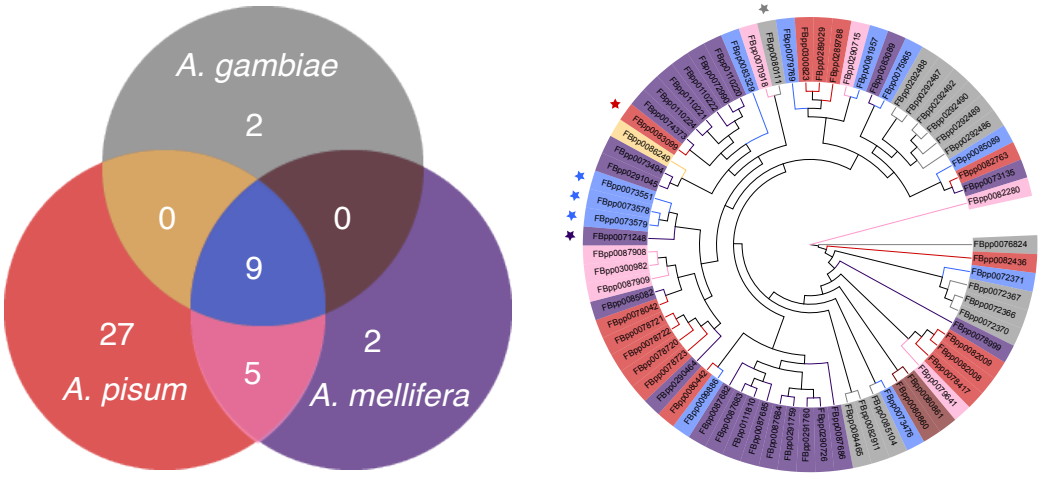
Although well-defined at a structural bioinformatics level, the precise biological functions of the majority of eukaryotic kinases remains to be determined, even for *H. sapiens* (Manning, 2009). While studies in *Drosophila* promise to throw light on their human counterparts, they are of considerable importance and direct relevance to insect biology in their own right. In Figure 7.7, we highlight the essential kinases belonging to the MAP kinase family, implicated in innate immunity (Horton et al., 2011). The differential expression of new members of functionally important families such as the MAP kinases may provide new entry points to pathway-based insecticide development.



(A) Panel A



(B) Panel B



(C) Panel C

FIGURE 7.7: The protein kinase complements of *A. gambiae*, *A. pisum* and *A. mellifera* were extracted from the total set of DEG genes and compared. As in Figure 7.2, hits were classified on the basis of exact match (E-value 0, Panel A, 169 kinases), high similarity (E-value <100, Panel B, 45 kinases), or lower similarity (E-value >100, Panel C, 75 kinases). Phylogenetic analysis was then performed for each class. Kinases unique to specific insects are shown, color-coded, alongside those shared between the 3 insects. The same color-coding is used in the phylogenetic analyses.

7.4 Discussion

This study has examined several methods for analyzing and comparing the protein complements of insect genomes. It has used essential gene sets identified in *Drosophila* to define orthologs in both insect pests and pollinators, triaging these with gene family (Pfam), chemical biology (ChEMBL) and structural biology (PDB, CDD) databases to identify subsets of targets which appear suitable for selective chemical design.

7.4.1 Target triage for agrochemical discovery

The focus on *Drosophila* essential genes as targets for insecticide discovery was driven both theoretically and operationally. As potentially cidal targets, insecticides that target them offer the best possibility of extinguishing pest populations before they can acquire resistance. Operationally, they constitute a clear target set on which to focus available genomic resources. The set of 2,694 essential genes represents 19.2% of the total complement of 14,029 *D. melanogaster* protein-coding genes (Tweedie et al., 2009).

For some of the most interesting future applications of our work (e.g. multi-target chemical design in insects and other organisms), it will be necessary to extend this analysis to the entire *Drosophila* genome and to DEG orthologs in other, non-arthropod genomes, especially the human. We have included some human homologs here for comparison (for example, those shown for ERR in Figure 7.6E). Human safety is of critical importance in insecticide target selection and further detailed comparisons between prospective insecticide targets and their human counterparts will be required before discovery targets can be nominated.

7.4.2 Gene family targets

The power of a gene family focused approach using Pfam for target triage is clear from the data shown in Figure 7.3 and Figure 7.4. The presence of distinct, targetable structural features amongst the four proteins we exemplify (Hsp90, Gwl, ERR and EcR), allows an immediate appreciation of the likely success of structure-based design. These observations are reinforced by the fact that one of the proteins, EcR, has already been the subject of successful insecticide discovery programs (Fahrbach et al., 2012). In addition, structural variation between individual insect orthologs, reflected in their feature matrices, gives an indication of the feasibility of achieving insecticide selectivity by design. A high degree of structural conservation between pest and pollinator targets may also provide an alert for potential ecological threats.

Among the databases deployed in our study, the Pfam database proved especially useful, enabling rapid interrogation and comparative scoring of individual protein domains. Since most proteins contain several Pfam domains, the database will interrogate individual proteins from several perspectives, as seen in the case of the nuclear receptors. Taken together, the 13,672 defined Pfam signatures provide a formidable way of categorizing and comparing target protein sequences.

The availability of a large number of new insect genomes through the i5K initiative (Robinson et al., 2011) promises to dramatically extend our knowledge of insect molecular and chemical biology, providing further pest and pollinator genomes for detailed comparative studies. The studies described here can readily be extended to encompass additional genomes.

A word of caution is, however, merited on the subject of genome-wide target triage. Two of the datasets used in this study, those for *D. melanogaster* (within FlyBase) and *A. gambiae* (within VectorBase), represent highly curated and well annotated resources. The genome databases for *A. pisum* and *A. mellifera*, however, are less completely annotated, with considerable reliance on automated annotation. In using the nr database, another fully curated database, to capture comparative data, coverage of available ortholog space has been maximised, but there remains a pressing need to assemble, curate and carefully annotate newly emerging insect genomes, especially when these are being produced at the speeds projected by the i5k initiative (Robinson et al., 2011).

7.4.3 Target validation

While in silico approaches are valuable in the early stages of target assessment, they need to be followed up with direct target validation using chemical or genetic approaches. Previously published data from functional genomic studies in *Drosophila* were used to support the target discovery studies reported here. *Drosophila* is one of the most highly characterized model systems for eukaryotic genetics (Ashburner et al., 2005; Tweedie et al., 2009). While *D. melanogaster* has many limitations for pest genomics, recapitulating only partially the full gamut of molecular functionality present in highly specialized insect disease vectors, it is nonetheless a valuable tool for hypothesis testing, sharing many features of its biology with other insect pests. Although often under-appreciated, the drosophilids are also a widespread, commercially relevant insect pest population in their own right (Burrack et al., 2013), meriting deeper study as pests in their own right.

To maximize the impact of this approach to target discovery in new pests, further systems to explore the functional details of the physiological pathways available for intervention are urgently required (Mohr et al., 2010). In this regard, the development of large-scale gene knockout platforms for specific target insects such as *B. tabaci* (Ghanim et al., 2007) and T. castaneum (Tomoyasu et al., 2008; Brown et al., 2009), offers further pest-focussed opportunities for target exploration and validation.

7.4.4 Defining an operational framework for insecticide discovery

The approach used for the systematic identification and validation of potential insecticidal targets is depicted schematically in the flow chart in Figure 7.1. Judicious choices between potentially tractable individual genes and gene family targets, made on the basis of increasingly integrated post-genomic approaches, provide a practical framework for the exploration of a new generation of molecular targets for insecticide discovery. The dataset of approximately 20,000 DEG orthologs described provides a rich resource for target triage, also identifying essential targets which might also be appropriate for RNAi-based approaches (Price and Gatehouse, 2008) and gene editing using CRISPR/Cas9 approach (Doudna and Charpentier, 2014; Rathe et al., 2014; Sander and Joung, 2014).

Note that this study only analyzes four target proteins in detail, comprising 1% of the initial set of 403 chemically tractable proteins identified. To extend these studies, large-scale automated homology modeling and site scoring of ligand binding sites will be required to assess the feasibility of selective ligand design. Practical follow-up of promising targets can then be initiated using fragment-based chemical discovery approaches, both to generate probes for chemical biology and to initiate pesticide design studies (see, for example, de Kloe et al., 2009). These chemical probes can then be taken into controlled laboratory trials to assess their impact as new chemical entities on a wider selection of pests, pollinators and host plants.

7.4.5 The question of agrochemical resistance

Biological systems are highly adaptable, and it is not surprising to find that insects possess well-developed mechanisms for counteracting the effects of xenobiotics, including agrochemicals. This study polarizes this question by focusing on those gene targets that are deemed essential for insect growth and development, and whose ablation will present major issues for targeted insects.

Biological resistance to chemicals is governed by several physiological processes, including 1) the induction of metabolic enzymes such as the cytochromes and esterases to detoxify xenobiotic chemicals, and 2) the mutation of specific molecular targets to evade pharmacological toxicities (reviewed for insecticides in Perry et al., 2011). The advent of multi-target design (MTD) approaches using single compounds to interfere at multiple points in signaling cascades, holds considerable promise for chemical intervention in multi-step processes such as growth and development, as exemplified in the human disease area by cancer and infection (Koutsoukas et al., 2011).

Many of the chemically tractable *Drosophila* essential genes identified in our study fall into structurally related protein families which appear amenable to MTD approaches. Focusing such techniques on key members of appropriate gene families, such as the specific protein kinase, ATPase

and nuclear hormone receptor families exemplified in this study, may open the way to developing exquisitely selective chemical entities. Simultaneously attacking multiple points within essential signaling pathways using either a single, carefully optimized compound or a coordinated set of individual agents, may also heighten the barrier for resistance development through site-specific mutation, a resistance mechanism frequently observed in insect populations (Perry et al., 2011).

7.4.6 Discriminating between agriculturally- and therapeutically-relevant pests

The results presented here have focussed on *A. pisum* as a commercially relevant insect pest. However, similar studies have been performed using ranked order analysis with *A. gambiae* and *A. mellifera* that provide parallel datasets to address mosquito-specific targets and honeybee-sparing targets. When commercial insecticides are used as agrochemicals, resistance amongst non-target insects can develop as a bystander effect. Inevitably, when the same insecticides are used in a therapeutic setting, they will prove very much less effective (see Ranson et al., 2009, for further discussion). Target selection to avoid parallel resistance development is therefore of critical importance.

Taken together, the approaches presented here take steps towards addressing some of the challenges of pest management, and provide a route to developing a range of therapeutically relevant and environmentally sustainable insecticides.

CHAPTER 8

Discussion

The genome sequencing project of any species would not be possible without modern sequencing technologies and the methods described in this thesis, transcriptome and genome resources for whitefly can be rapidly developed which enable exciting research. Methods and results described in this thesis are just examples for any genome project, and would suit best to any arthropod particularly those that contain bacterial endosymbiont(s). This concluding chapter summarises the previous chapters and also outlines the possible future directions for improvements and describes the exciting research studies possible.

8.1 Transcriptome provides a useful resource

Chapter 1 began by addressing the research problem and describing whiteflies that have been studied so far and why the genome of this species is needed. As described in Chapter 2, a literature review was conducted to study and understand research carried out on whitefly and its implications so far, and to evaluate the research problem. Although, four transcriptomes of different whitefly species have already been published and the data are publically available. These four transcriptomes were compared in Chapter 3 to identify sequence and functional differences across them. The main aim of Chapter 3 was to obtain a comprehensive transcriptome data set for the Asia I population which could effectively be used for further analysis in this thesis.

In order to capture the complete transcriptome of the Asia I population, two different libraries (Normalized and Unnormalized) were generated and sequenced. These combined data sets from two libraries not only led to an excellent adult transcriptome resource but also to identifying the most abundant and least redundant genes for this population. The transcriptome of the Asia I population was found better in terms of functional annotations and proportion of complete transcripts than that which has been found in other species of the *B. tabaci* complex including MEAM1, MED and Asia

II 3. The strategy and methods described in Chapter 3 also proved that a better transcriptome can be generated with low sequencing coverage, as a lower amount of sequencing data was generated for Asia I species (290 Mbp) than MEAM1 (1.27 GB) (Wang et al., 2011), MED (3.27 GB) (Wang et al., 2010a) and Asia II 3 species (1.24 GB) (Wang et al., 2012). One of the interesting results reported in Chapter 3 was that the termite species, *Z. nevadensis* unexpectedly scored the highest number of top-hits with the Asia I data. This is considered to be due to the higher number of short aligned matches found between Asia I species and *Z. nevadensis*. Further investigations are needed into this, and why these distantly related species share so many short aligned sequences.

Good transcriptomes from different life stages of the Asia I population should be obtained to enable rapid identification of stage-specific genes and their expression levels. It would also enable researchers interested in studying cross-species comparisons to identify species-specific genes. For example, during a variety of hosts shifts between MEAM1 and MED species, significant changes in the activity of cytochrome P450 and COEs was reported (Xu et al., 2014). As the main objective of Chapter 3 was to obtain a transcriptome for the Asia I population, a high coverage Illumina HiSeq 2000 RNA-seq library should in the future be prepared from various developmental stages to generate a comprehensive transcriptome of this Asia I population. Transcriptomes do not only act as a valuable resource for genome annotation (as described in Chapter 4 and Chapter 5) but can also be used to construct scaffolds from genomic contigs using ERANGE (Mortazavi et al., 2010) and SCUBAR (Elsworth, 2016), for which there was not the time or data to do in this study.

Chapter 4 described a new approach to explore gene structures in the Asia I population prior to annotation, and where there was no reference genome available. The methods used in this chapter provide an example applicable to other researchers to wishing to gain insights into the gene structure of their species of interest for which genome information is lacking or poorly annotated. The approach used existing software to align transcriptome (obtained in Chapter 3) to genome assembly for Asia I species. At least 90% of the total span of the transcripts should aligned to a genome assembly when used against the same species. It is unlikely to get 100% of the transcripts aligned to a genome even where a complete genome assembly exists because many transcripts may have been obtained as a single-coverage transcript with sequencing errors or as chimeric transcripts (Parkinson and Blaxter, 2004). In the methods section, specific details were described of how these alignments were filtered and selected for further analysis in Chapter 4. The results of Chapter 4 outlines distinctive gene structures of Asia I in comparison with the other insects of the same Order or even with the distantly related species. The variation in their gene structures was only due to one feature, namely intron size. The introns of Asia I species tend to be much longer than their corresponding orthologs from other insects regardless of their similar CDS size. The accuracy of assembled genomic scaffolds of Asia I were confirmed by the PE reads not only support the coding regions (exons) of the gene but also support the longer non-coding regions (introns) with a minimum of 50x coverage. Further, the presence of the conserved donor splice site (at 5') motif 'GT' and acceptor splice site (at 3') motif 'AG' within the introns added more confidence to the accuracy of

exon-intron-exon structure predictions for the Asia I genome data. These findings helped in the understanding of the complexity of the *B. tabaci* genome and answered the question of why it is a relatively large (640-690 Mbp) genome (Chen et al., 2015; Guo et al., 2015). The findings of Chapter 4 also contributed to evaluation and annotation of genome assembly in Chapter 5, 13% more genes were found complete when used the '-vrt' option was used that allowed for much longer introns (up to 20,000 bp). This result helped to optimise parameters during genome annotation phase in order to obtain more accurate gene predictions for the Asia I population.

Transcript alignment can also be used to assess the contiguity of genome assembly and help to choose a relatively better assembly when comparing more than one assembler, using the approach outlined in this chapter and in Kumar and Blaxter (2010). The quality assessment of genome assembly has already been addressed via comparison studies on different assemblies (IHGSC, 2001; Earl et al., 2011) without using a reference genome, and this could help in further improvements.

8.2 Draft genome of Asia I species at low sequence coverage

Chapter 5 described the methods for assembling the whitely genome using low cost short read sequencing at 70x coverage via partitioning reads to obtain endosymbiont-free assembly for the Asia I population. Only one PE library was generated for Illumina sequencing using the DISCOVAR protocol. Despite using such limited data, a draft genome assembly was essentially obtained with multi-gene sized scaffolds for Asia I population. Prior to the assembly filtering the genomic reads (3.44 million) correspond to endosymbionts helped to achieve Asia I genome-specific reads (Appendix B, Figure B5.1).

The *de novo* assembly strategy described in Chapter 5 addresses a few issues like read quality, read partition for endosymbionts and assembly parameters. Assembling a genome is not a one-button solution, where raw reads can be converted directly into a final genome. From a raw set of reads, many assemblies can be produced using different assemblers and with different parameters for each of them. Most of the genome sequencing projects do not report all assemblies that they have generated using different assemblers and parameters. Reporting of alternative assemblers and their parameters could benefit future genome projects on similar species or similar input data, perhaps like the Assemblathon study (Earl et al., 2011). There were several metrics defined in the Assemblathon study for assembly evaluation and scaffold N50 (the scaffold size N at which 50% of the genome assembly is in scaffolds longer than N) was highly recommended as a measure of contiguity.

To achieve a better quality genome assembly for the Asia I species, several assemblies were produced using different assemblers and the assembly optimality criteria was defined which led to choice of the best assembly for further analysis. Although the scaffold N50 and scaffold numbers are the most widely used assembly metrics for assembly evaluation, they may not always correlate well

with the actual assembly contiguity (Nagarajan and Pop, 2013). Typically, it is assumed that the larger the N50 value the better the assembly, but the higher N50 can be achieved by discarding short sequences from the assembly or by miss-joining scaffolds to make them longer. Another approach that has been used to assess the assembly quality is the cumulative length plot in which the distribution of scaffolds length is sorted from higher to lower to measure the assembly contiguity (Gurevich et al., 2013). The cumulative length plot was used to select the best assembly as it not only showed which assembly had the longest scaffolds but also showed which had short abundance scaffolds along with the greatest span (total length of assembly) (Appendix A, Figure B5.2). The N50 value of the selected assembly (Platanus assembly) was still much lower than the previously sequenced hemipteran species because of the MP and fosmid libraries (IAGC, 2010; Xue et al., 2014). Although the genome assembly of *B. tabaci* in this thesis can become more contiguous when additional high-quality MP libraries of insert sizes 3-20 kbp used. Chapter 5 described that the genome assembly can be obtained from low cost Illumina PE sequencing when the project focuses on capturing multi-gene sized scaffolds.

In addition to the cumulative length plot, other approaches such as CEGMA (Parra et al., 2007, 2009) and BUSCO (Simao et al., 2015) were applied to assess the assembly completeness. It is important to ensure biological accuracy when comparing different assemblies to determine which assembly captures the most biological known sequences. CEGMA and BUSCO pipelines are widely used for this purpose. Assembly with a CEGMA score above 90% would be assumed as a good assembly and higher values (closer to 100%) indicates a more accurate and complete assembly. CEGMA completeness was, however, low in Asia I (81.45%) compared to *A. pisum* (100%) (IAGC, 2010) and *N. lugens* (96.8%) (Xue et al., 2014). These three species belong to the same Order - Hemiptera and the low CEGMA values may be the result of bias in methods or a biological reason which would be interesting to explore further, or more likely due to a poorer assembly. However, CEGMA is no longer being supported for such purposes and the recommendation more recently is to use BUSCO. The BUSCO pipeline starts with a homology search in the given assembly for a BUSCO consensus gene set and then gene structure prediction using Augustus. In the final step, assessment of predicted genes using HMMER and lineage-specific profiles to classify them into four categories: complete (C), duplicated (D), fragmented (F) and missing (M). Both approaches, i.e. CEGMA (81.45%) and BUSCO (C: 35%, D: 1%, F: 32% and M: 32%), scored a low percentage of complete genes in the selected Platanus assembly which indicated that the assembly was incomplete. However, one purpose of using CEGMA and BUSCO was to identify gene structures of highly conserved full-length genes in a given genome assembly. These gene structures can then be used for genome annotation to train the gene predictor programs like Augustus and SNAP as described in this chapter.

After selecting the best genome assembly for Asia I, the sequences were annotated to generate a useful genomic resource for the whitefly community as described in Chapter 5. Any genome sequencing project involves an annotation phase which uses the gold standard annotation of a few

hundred genes to train the gene predictors for the whole genome. Here in this chapter, the putative functions were assigned to protein-coding genes using the two-pass MAKER2 (Holt and Yandell, 2011) strategy, although these gene predictions may be the incomplete set and the missing genes could be the result of poor annotation or poorly assembled scaffolds. The MAKER2 pipeline involves *ab initio* gene predictions using Augustus and SNAP, and evidence based predictions which use ESTs and proteins from the same species or closely related species. These MAKER2 predicted gene models were assigned functions using Blast2GO (Götz et al., 2008) and InterProScan (Jones et al., 2014), while RNA annotation was performed using tRNAscan-SE (Lowe and Eddy, 1997), Rfamscan (Griffiths-Jones et al., 2005) and INFERNAL (Nawrocki et al., 2009).

Similar to the genome assembly programs, genome annotation pipelines need to be evaluated to achieve best set of gene predictions. In future, I would like to develop such a platform which allows systematically evaluation of annotation pipelines. There are two metrics that can be used for producing competing annotations such as homology to known genes or AED score (Eilbeck et al., 2009). Development of such a platform will allow users and developers to submit their annotations in a competition format similar to the Assemblathon (Earl et al., 2011). This competition based platform ("Annotathon") will be very useful to everybody who is working on genome annotation of eukaryotes once the metrics have been established. The submission of annotation results from different annotation pipelines with different parameters and input sources on a range of species will help to establish the best annotation strategy for any eukaryotic genome project.

To further improve the draft genome annotations for Asia I species, additional sequencing data will be required like RNA-seq. The major problem with any automated genome annotation pipeline is to predict alternatively spliced transcripts. However, RNA-seq reads represent the transcriptome and it should be fairly easy to elucidate alternate transcripts when using RNA-seq data by splice alignment tools such as TopHat (Roberts et al., 2011; Trapnell et al., 2012).

8.3 Additional genomes inside Asia I species

In Chapter 6, four additional genomes from the Asia I genome data were reported including a mitogenome and three endosymbiont genomes of *Portiera*, *Wolbachia* and *Arsenophonus*. The mitogenome of Asia I species was obtained as a single scaffold from the genome assembly of Asia I species without any additional assembly. This is because the mitogenome of Asia I species is only 15,453 bp which was expected to be assembled as a single scaffold with significant read coverage. Despite having the same number of genes, the mitogenome of Asia I (15,453 bp) was found slightly longer than that of two published mitogenomes of *B. tabaci* species including Asia I (15,210 bp) (Tay et al., 2016) and New World I (15,322 bp) (Thao et al., 2004) but shorter than that for the MEAM1 species (15,632 bp) (Wang et al., 2013). The variation in the sizes of these

mitogenomes are due to the presence of control regions and non-coding intergenic regions within their mitogenomes. The control region of Asia I species (710 bp) was longer than that in published Asia I (467 bp) (Tay et al., 2016) and New World I (664 bp) species (Thao et al., 2004) but shorter than that in the MED (974 bp) species mitogenome (Wang et al., 2013). The mitogenome of Asia I will be very useful to everyone interested in population studies of *B. tabaci* complex and should help improve understanding of the evolution of this complex.

Chapter 6 also described the strategy used for obtaining draft genomes of endosymbionts from the host genome sequencing reads as previously described in Siozios et al. (2013). The strategy is likely to remain the same for any bacterial species but the programs, quality checks, illustrations and comparative analysis presented here are the best approaches for obtaining low-cost draft genomes. A high quality draft genome of *Portiera*, the primary endosymbiont of Asia I species was assembled and annotated in this chapter. The comparative genome analysis of *Portiera* across three different species including Asia I, MEAM1 and MED of *B. tabaci* complex revealed 98% sequence identity and conserved gene synteny. The phylogenetic analysis of molecular marker genes (16S rRNA of *Portiera* and *mtCOI* gene of Asia I species) from both host and endosymbiont species showed vertical transmission of this endosymbiont. A high quality draft genome of *Wolbachia*, the secondary endosymbiont of *B. tabaci* is also assembled and annotated in this chapter. The genome was found similar in size to the other *Wolbachia* genomes from different hosts. The orthology analysis between the complete proteomes of 21 strains of *Wolbachia* from different hosts revealed 621 core proteins present in all. The most useful innovation in this chapter is using these core proteins to perform phylogenetic analysis across 21 strains and classify them into subgroups (A, B, C, D and F). The draft genome sequence of *Arsenophonus*, another secondary endosymbiont of the Asia I population studied, was also assembled and annotated in this chapter and named *Arsenophonus bemisiae*. The genome assembly of *Arsenophonus bemisiae* (1.8 Mbp) was found much smaller in comparison with the other two strains of *Arsenophonus* including *Arsenophonus nasoniae* (3.67 Mbp) (Wilkes et al., 2010) and *Arsenophonus* endosymbiont of *N. lugens* (2.95 Mbp) (Xue et al., 2014). However, this draft genome assembly of *Arsenophonus bemisiae* contains 1,846 PCGs of which 1,245 PCGs were found as core proteins when compared with the other two *Arsenophonus* strains. These core proteins are conserved proteins and could represent a set of essential proteins required by this bacteria to be able to maintain metabolic homeostasis, evolve and reproduce. This chapter can be used as a guide to assemble endosymbiont genomes fairly straightforwardly from the host without requiring additional laboratory procedures and sequencing. The draft assembly of an endosymbiont genome can naturally be improved with MP sequencing as described in Chapter 5.

8.4 Genomic framework for insecticide discovery

Chapter 7 describes the workflow that was proposed at the beginning of this research study to establish a framework for target selection and insecticide discovery for pests and vectors, and to ascertain how difficult developing selective insecticides might be. The whitefly, main subject of this thesis, lacked genomic information and even transcriptomes were not complete or adequately characterised at the start of this study, and therefore were not included into the framework. The workflow outlined in this chapter uses experimentally-verified 2,694 essential genes of *D. melanogaster* as a reference cidal target dataset to search their corresponding orthologs in the genomes of two important insect disease vectors, the aphid *A. pisum* and the malarial vector *A. gambiae*, for new insecticide target genes, comparing these to the equivalent sequences of a beneficial pollinator, the honeybee, *A. mellifera*. There was a high degree of conservation observed between orthologs in these four insects including the targets of many traditional insecticides. In contrast, through systematic comparative analysis of the entire essential gene dataset by protein family (using the Pfam database) and chemical tractability (using the ChEMBL database), several cohorts of chemically tractable essential protein orthologs were identified in both *A. pisum* and *A. gambiae* that were absent in *A. mellifera*. This essential gene dataset also contains a number of target families, many of which already have putative lead chemicals associated with them, which may be suitable for multi-target discovery approaches. In addition, it was also noted that virtually all members of the essential gene dataset show distinct variability at the RNA sequence level between insects, which suggests that targeting these and other insect essential genes through genetic approaches like RNAi and CRISPR/Cas9 would represent a feasible alternative to chemical insecticide development. Together, these observations support a detailed exploration of essential gene orthologs as new targets for insecticide development.

8.5 Future recommendations

In this thesis, second-generation sequencing technologies like pyrosequencing by Roche 454 and sequencing-by-synthesis by Illumina are described for transcriptome and genome sequencing respectively. The main goal of this study was to obtain and characterise the transcriptome and the genome of Asia I species along with the additional genomes within Asia I species such as endosymbionts or the mitogenome.

Once this primary goal of characterizing transcriptome and the genome of the whitefly, *B. tabaci* Asia I species has been realised, this study could be extended further to carry out three studies:

- ☞ Transcriptome sequencing at different life stages to study their gene expressions as done in previous studies (Wang et al., 2010a, 2011, 2012; Xie et al., 2012). It would be interesting to

identify genes that are expressed or suppressed at different stages and that could provide deeper insight into the host plant resistance mechanism and insecticide resistance mechanism. These stage specific gene expression profile of Asia I species transcriptome can be compared with the other species within the *B. tabaci* complex to get wider overview of whitefly developmental biology. The expression profile of Asia I species genes could lead to development of control strategies for this pest using RNAi (Hannon, 2002; Geley and Müller, 2004) or CRISPR-CAS9 (Doudna and Charpentier, 2014; Rathe et al., 2014; Sander and Joung, 2014).

☞ Achieve "3Cs of Asia I species genome sequencing project": The contiguity, completeness and correctness. There are two ways to achieve this. First, genome sequencing using specific isolines (backcrossed for 7-8 generations) with high coverage (100x) and multiple PE and MP libraries. And second, improve current draft genome assembly via additional genome sequencing and redo annotations with more evidence from transcriptomes. The first option requires a lot of work in the insectary and is time-consuming to generate the isolines. The sequencing and assembly will also be time-consuming as they have to go through all quality checks and evaluation stages. While the second option requires less work and time frame due to the recent advances in sequencing and post-assembly improvements. Now adays the third-generation sequencing (Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) sequencing technology) has been most widely established which can generate much longer reads (1-40 kbp) from the single DNA fragment. These longer reads could overcome previous issues associated with short-read assembly including fragmented assembly, determination of complex genomic regions, extended repetitive regions and gene isoforms. PacBio sequencing along with the short-read sequencing can produce highly-contiguous *de novo* assemblies with closed gaps as already found in previous studies (Conte and Kocher, 2015; Chakraborty et al., 2015; Rhoads and Au, 2015).

After building a highly-contiguous assembly of the Asia I species genome, an interesting study would be to determine the structure of it's DNA molecule. In such a rapidly changing field, several third-generation mapping techniques have been developed to determine the structure of large DNA sequences without reading every single base. One of the most successful mapping technique is the Irys system from BioNano Genomics, launched in 2010. It is an optical mapping system which can produce larger optical maps of a chromosome spanning several megabases in size. This system has limitations such as incomplete nicking of the DNA fragment which cause a proportion of the restriction sites to remain unlabeled, and multiple nick sites in close proximity result into multiple shearing which ultimately limit the overall size of optical map. However, a hybrid approach of using optical mapping with the third-generation sequencing techniques could show improvement in scaffolding and structural resolution (Dong et al., 2013; Pendleton et al., 2015). For example, PacBio reads with BioNano mapping has produced one

of the most contiguous *de novo* assemblies of a human genome with contig N50 of 1.4 Mbp, scaffold N50 of 31.1 Mbp and many hundreds of novel variants (Pendleton et al., 2015).

Assembly ranking as of January 12, 2016

(1). PacBio + BioNano

(2). Illumina + Dovetail

(3). PacBio + Dovetail

(4). Illumina + BioNano

(5). PacBio only

(6). Illumina only

Erich Jarvis (Duke University Medical Center,
personal communication at PAG, 2016)

"Hi-C", another third-generation mapping technique have been developed to construct long-range reads spanned hundreds of kilobases or more and similar to MP from chromatin interactions (Burton et al., 2013). The frequencies of "Hi-C" mappings can be used to infer the relative order and orientation of assembled contigs as the chromatin interactions are highly localized (Kaplan and Dekker, 2013). "Hi-C" technique has not been widely used as it requires a difficult protocol (de Wit and de Laat, 2012) although it is currently under commercial development. Moreover, the "Hi-C" protocol has been optimised more recently by Dovetail Genomics and named "cHiCago" library which is similar to MP and uses second-generation sequencing reads to map long spans (Putnam et al., 2016). This protocol is proprietary to Dovetail, and requires samples to be shipped to their site for processing which could limit its establishment.

☞ Integration of the Asia I species genome into the workflow described in the Chapter 7 to identify Asia I species specific or multi-target genes that could be used for novel insecticide discovery and non-chemical control strategies like RNAi or CRISPR/Cas9. Once the curated set of genes from Asia I species has been obtained, the genomic framework should be able to identify potential target genes by comparative genomic across various insects including model insect, pests and vectors, which then could subject to structural confirmation for insecticide discovery. The target genes could also be used for non-chemical control strategies such as RNAi and or CRISPR/Cas9 which have been most popular these days.

More research work is required in the field of genome annotation evaluation where I would like to contribute by developing automated pipelines for quantifying accuracy of genome annotation. This assessment is necessary as the subsequent analyses rely on the genome annotation. The sequencing and mapping technologies discussed above are I consider the most promising approaches for the future of genomics and combined with the annotation workflow described in this thesis will make it possible for even a graduate student to generate a highly-contiguous genome with accurate annotations for any species.

APPENDIX A

Supplementary tables

Chapter 3 : Transcriptome sequencing of Asia I species

Table A3.1: BLASTX homology against nr database for Unnormalized library of *B. tabaci*.

Table A3.2: BLASTX homology against nr database for Normalized library of *B. tabaci*.

Table A3.3: GO assignments for Unnormalized library of *B. tabaci*.

Table A3.4: GO assignments for Normalized library of *B. tabaci*.

Table A3.5: KEGG pathway annotations for Unnormalized library of *B. tabaci*.

Table A3.6: KEGG pathway annotations for Normalized library of *B. tabaci*.

Table A3.7: MicroSatellites found in *B. tabaci* transcriptome.

Chapter 4: Distinctive gene structure of Asia I species

Table A4.1: Full-length transcripts with annotations from nr and Pfam databases.

Table A4.2: Genome assembly statistics for Asia I (CLC, SSPACE).

Table A4.3: Gene models were produced for a subset of 119/567 genes Asia I gff3.

Table A4.4: Gene orthologs for 119 in 10 insects.

Chapter 5: Genome sequencing of Asia I species

Table A5.1: DISCOVAR *de novo* assembly statistics for Asia I.

Table A5.2: Functional annotation of MAKER2 gene models (OGS v1.0) using BLAST and GO.

Table A5.3: KEGG pathway annotation of MAKER2 gene models (OGS v1.0).

Table A5.4: MAKER2 gene models (OGS v1.0) (without BLAST hit at nr database) aligned against ESTs.

Chapter 7: Genomic framework for insecticide discovery

Table A7.1: *Drosophila* essential genes retrieved from FlyBase.

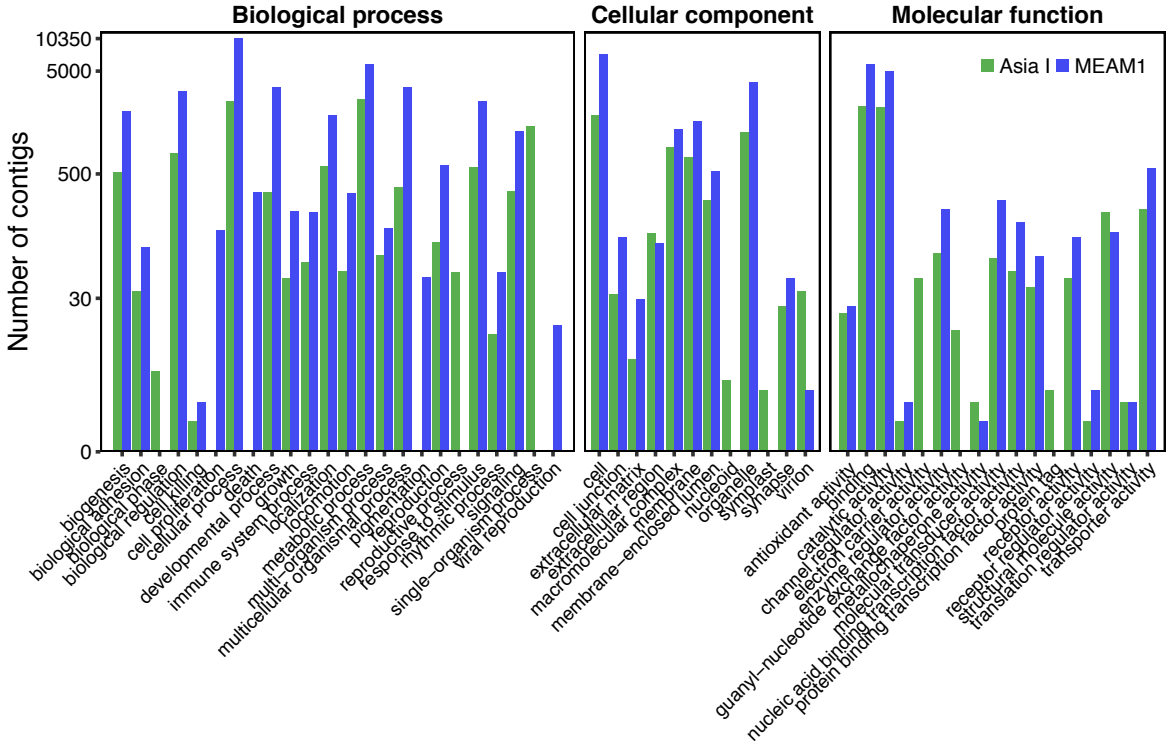
Table A7.2: *Drosophila* essential proteins retrieved from FlyBase.

Table A7.3: DEG orthologs from *A. pisum*, *A. gambiae* and *A. mellifera* with PDB and ChEMBL annotation.

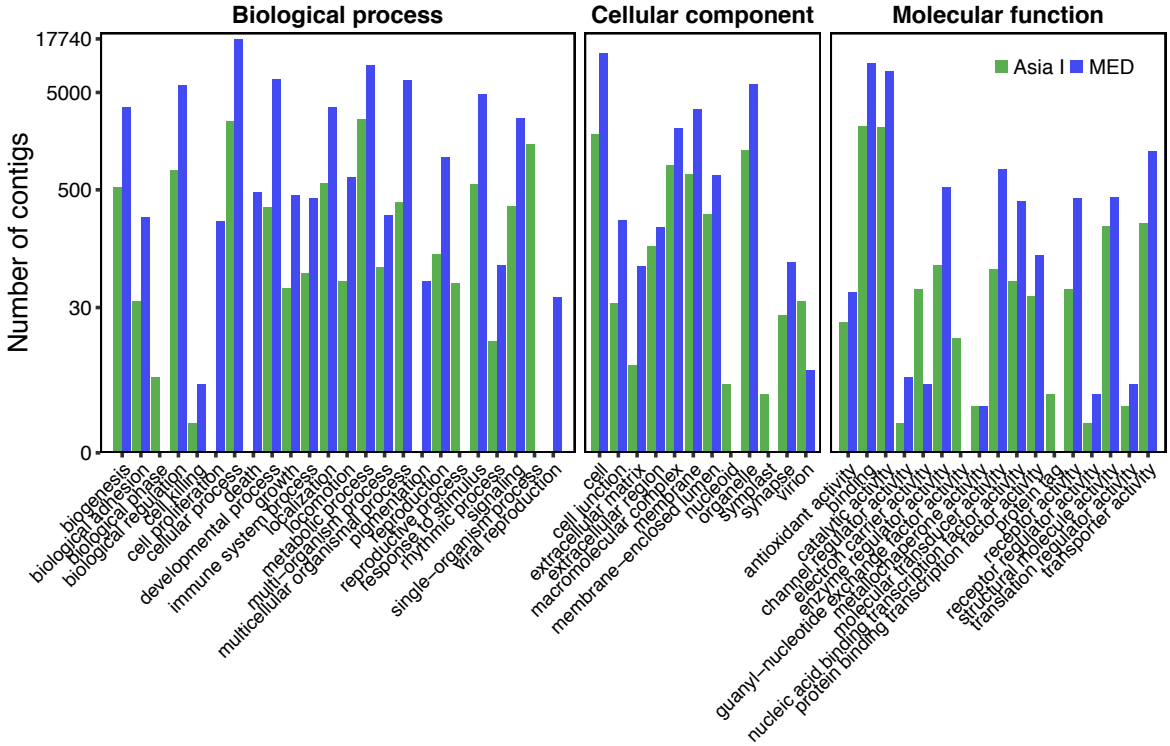
APPENDIX B

Supplementary figures

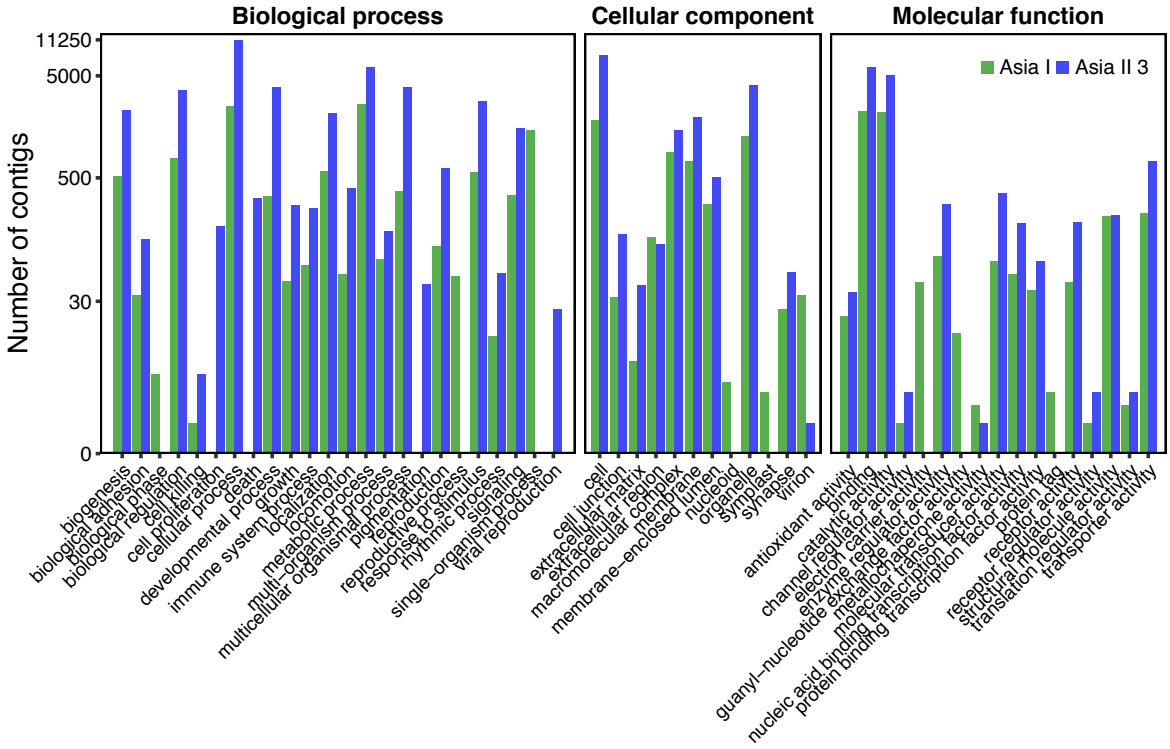
Chapter 3: Transcriptome sequencing of Asia I species



(A)



(B)



(C)

FIGURE B3.1: Gene ontology comparison across whitefly species - Asia I, MEAM1, MED and Asia II 3. (A) Asia I and MEAM1, (B) Asia I and MED, and (C) Asia I and Asia II 3. Each histogram summarized in three main categories: biological process, cellular component and molecular function. The Y-axis on left side indicates the total number of genes belongs to specific subcategory in that main category.

Chapter 5: Genome sequencing of Asia I species

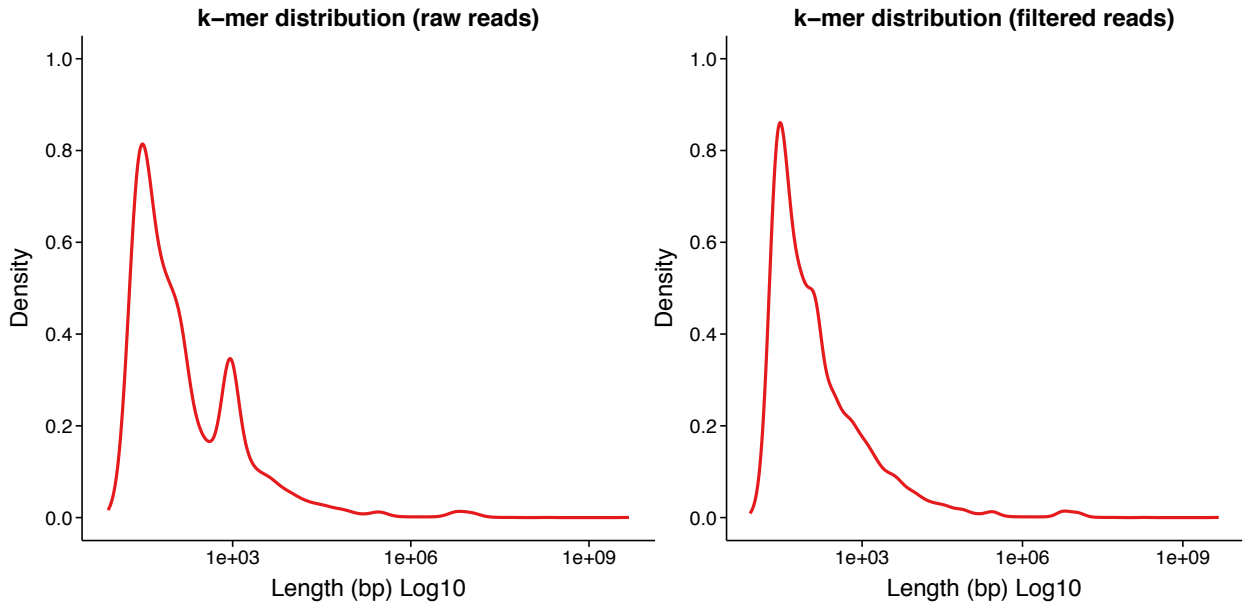


FIGURE B5.1: k-mer distribution for raw and endosymbiont filtered reads from genome sequencing using DISCOVAR protocol. The left plot illustrates k-mer distribution of raw reads where the highest pick denotes to distinct k-mer reads for Asia I genome where the small pick denotes to distinct k-mer reads belongs to endosymbionts. The small pick was disappeared in the k-mer distribution of endosymbiont filtered reads as shown in right plot where only one pick was obtained that belongs to Asia I genome.

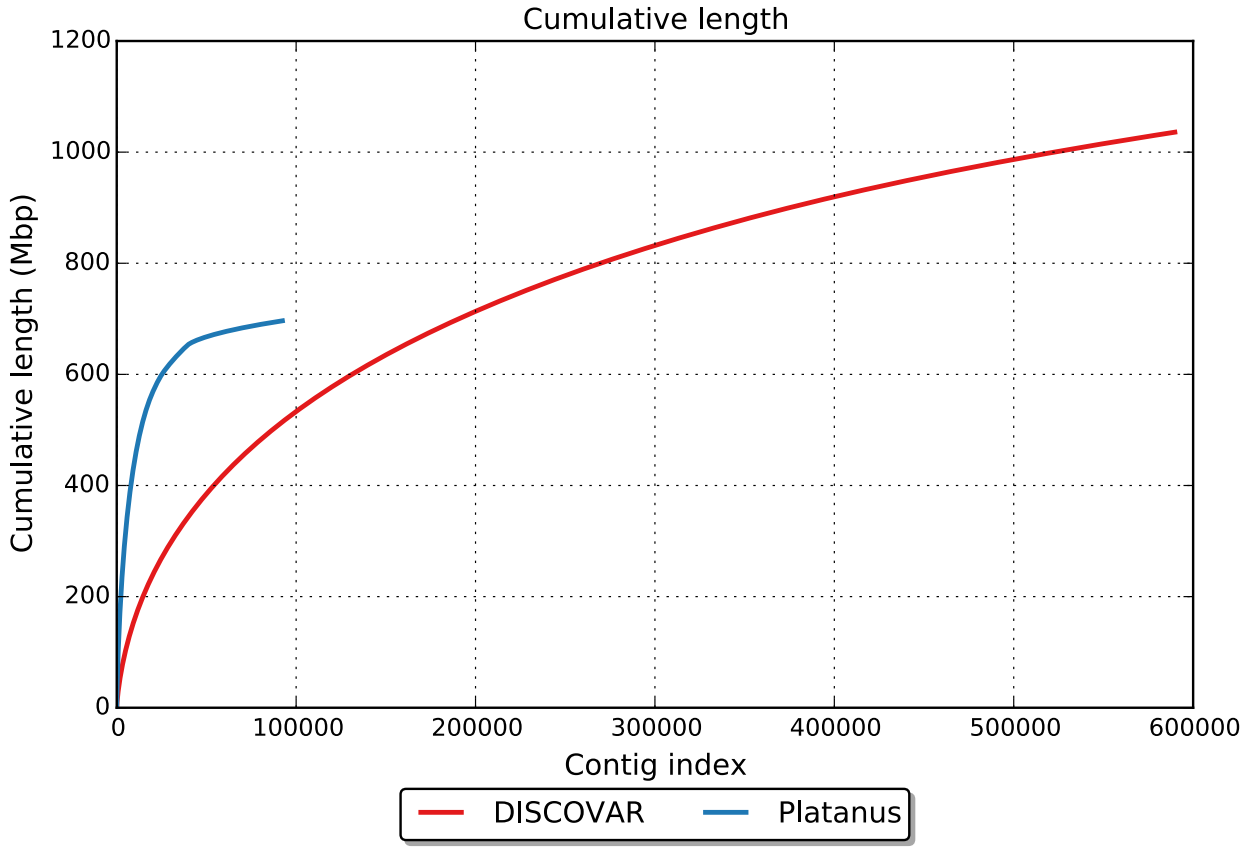


FIGURE B5.2: Cumulative scaffold length distribution for DISCOVAR and Platanus assemblies of Asia I species genome. The curve with the steepest starting curves represents the longest scaffolds. Platanus had the longer scaffolds with larger assembly span in comparison with DISCOVAR.

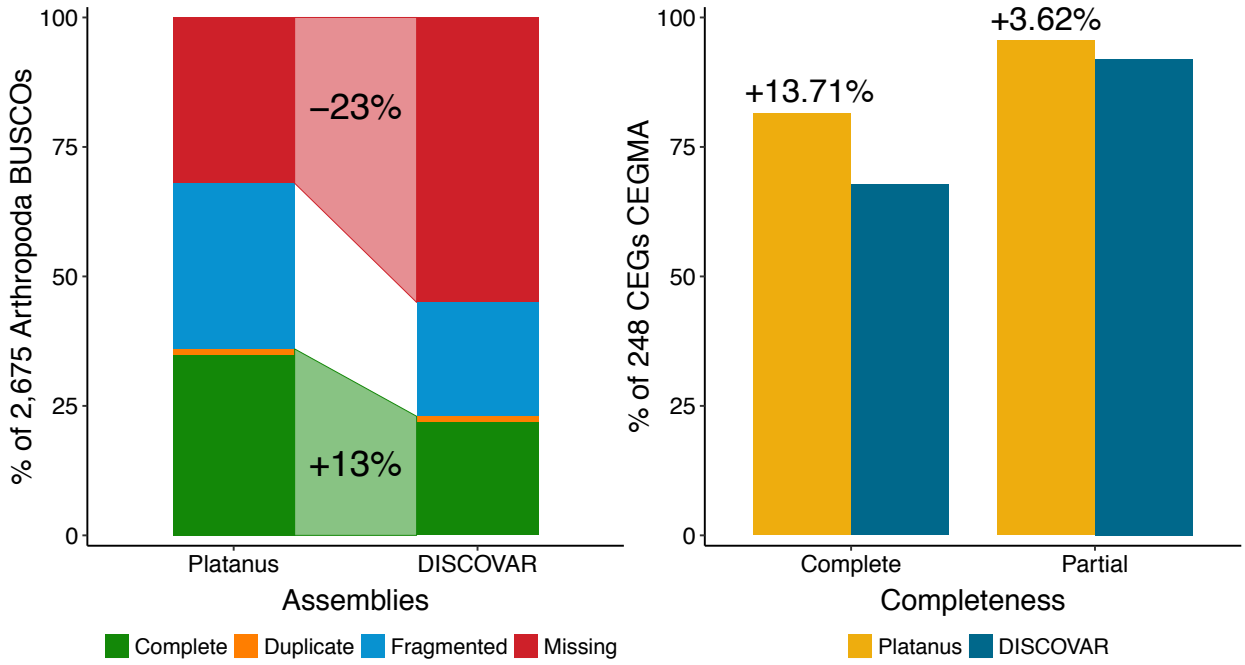


FIGURE B5.3: Assessment of Asia I genome assemblies, Platanus and DISCOVAR *de novo*, using CEGMA and BUSCO. Assembly assessment using BUSCO resulted in four categories like complete, duplicate, fragmented and missing, while CEGMA resulted in only two categories such as complete and partial. BUSCO results on the left side shows Platanus assembly had 13% more complete and 23% less missing genes than that were found in DISCOVAR *de novo* assembly. Similarly, CEGMA results on the right side also shows 13.71% more complete and 3.62% more partial genes in Platanus assembly than DISCOVAR *de novo*.

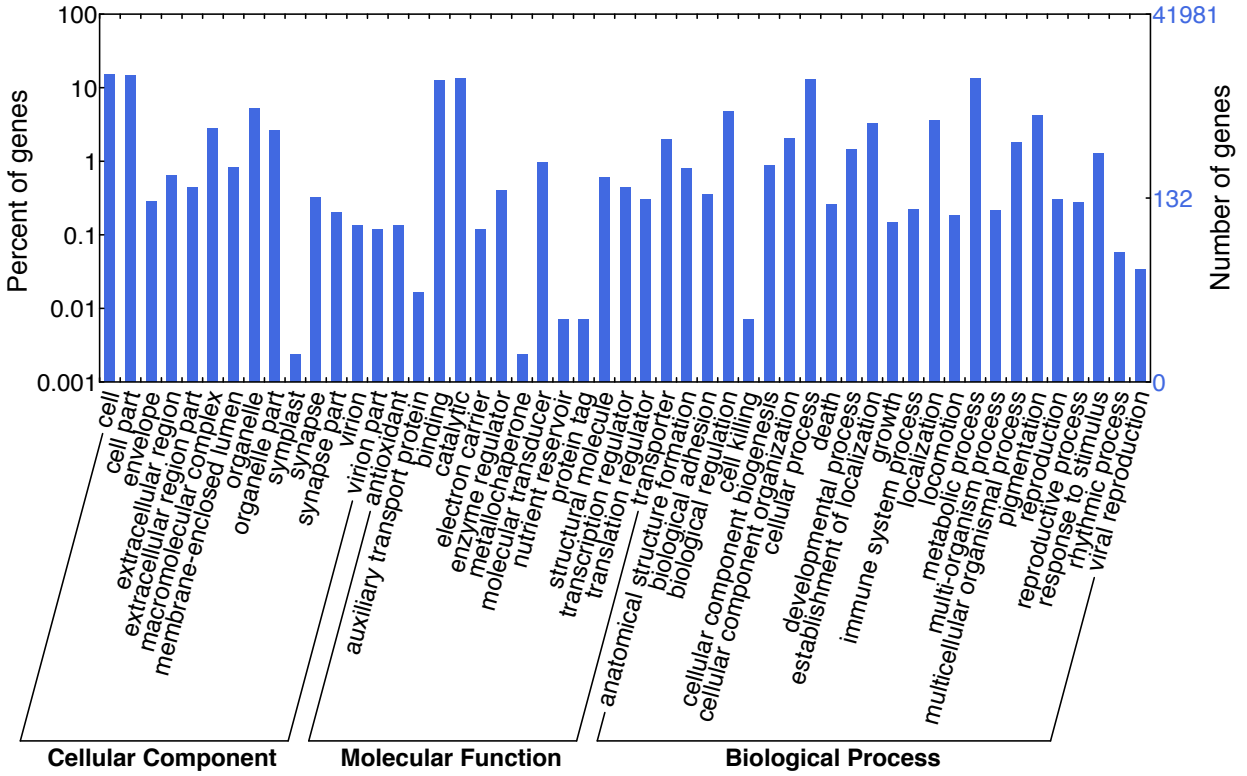


FIGURE B5.4: Gene ontology distribution of MAKER2 predicted genes for Asia I genome. The histogram summarized in three main categories: cellular component, molecular function and biological process. The Y-axis on left side indicates the percentage of genes belongs to specific subcategory in that main category. Whereas the Y-axis on right side indicates the total number of genes belongs to that subcategory.

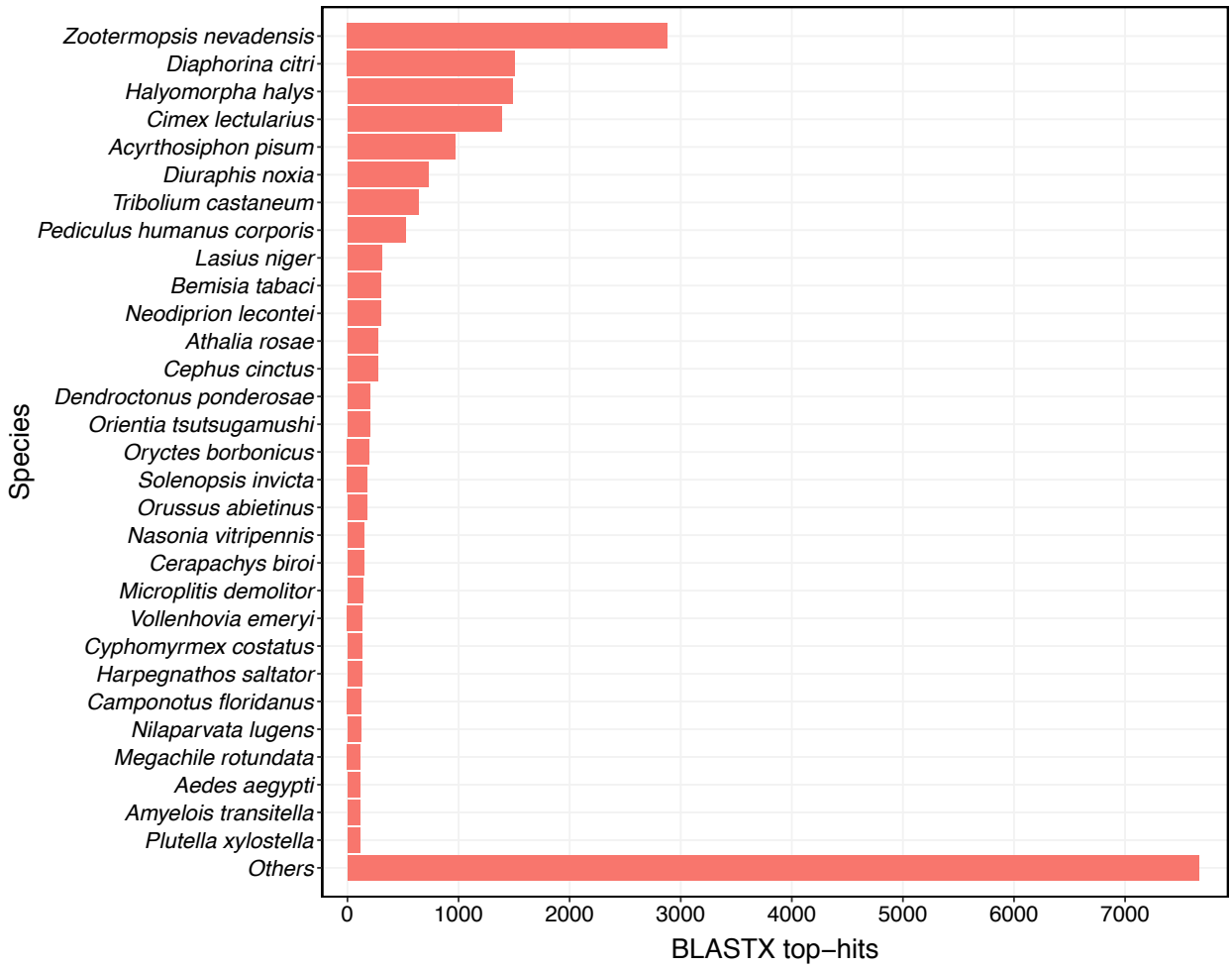


FIGURE B5.5: Taxonomy distribution from BLASTX results of MAKER2 predicted genes for Asia I species. Each bar represents number of BLASTX top-hits from corresponding species listed on X-axis.

Chapter 6: Mitogenome of Asia I species

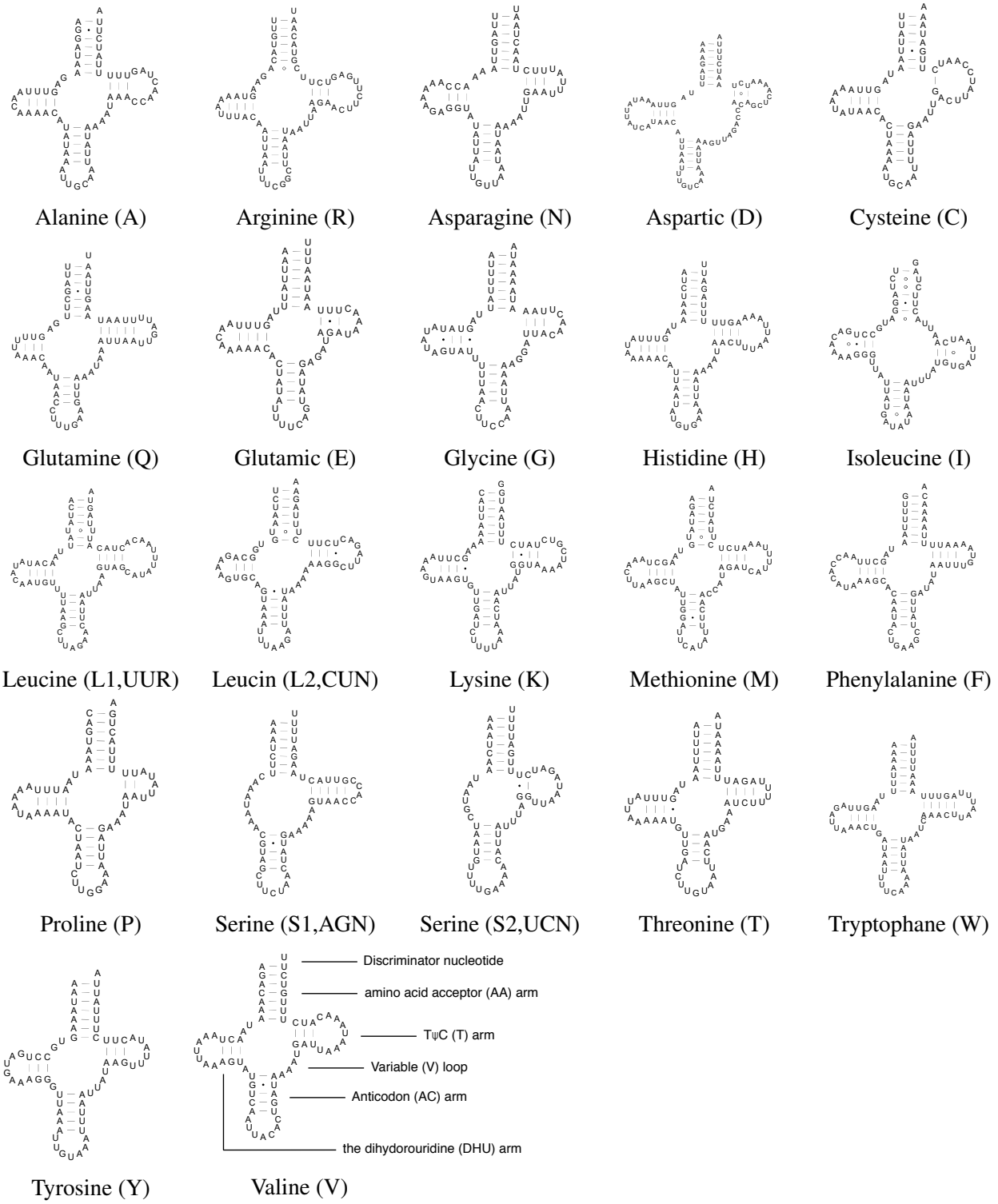


FIGURE B6.1: Putative secondary structures of 22 tRNAs found in the mitogenome of Asia I species. All tRNAs were folded in to the typical clover-leaf secondary structure except two: Serine (S1,AGN) and Serine (S2,UCN) which lacks DHU arm. All tRNAs are labelled with the abbreviations and single letter code of their corresponding amino acids. The Watson-Crick bonds (A-T/U, G-C) are represented by lines, while the G-U bonds are represented by dots.

Chapter 7: Genomic framework for insecticide discovery

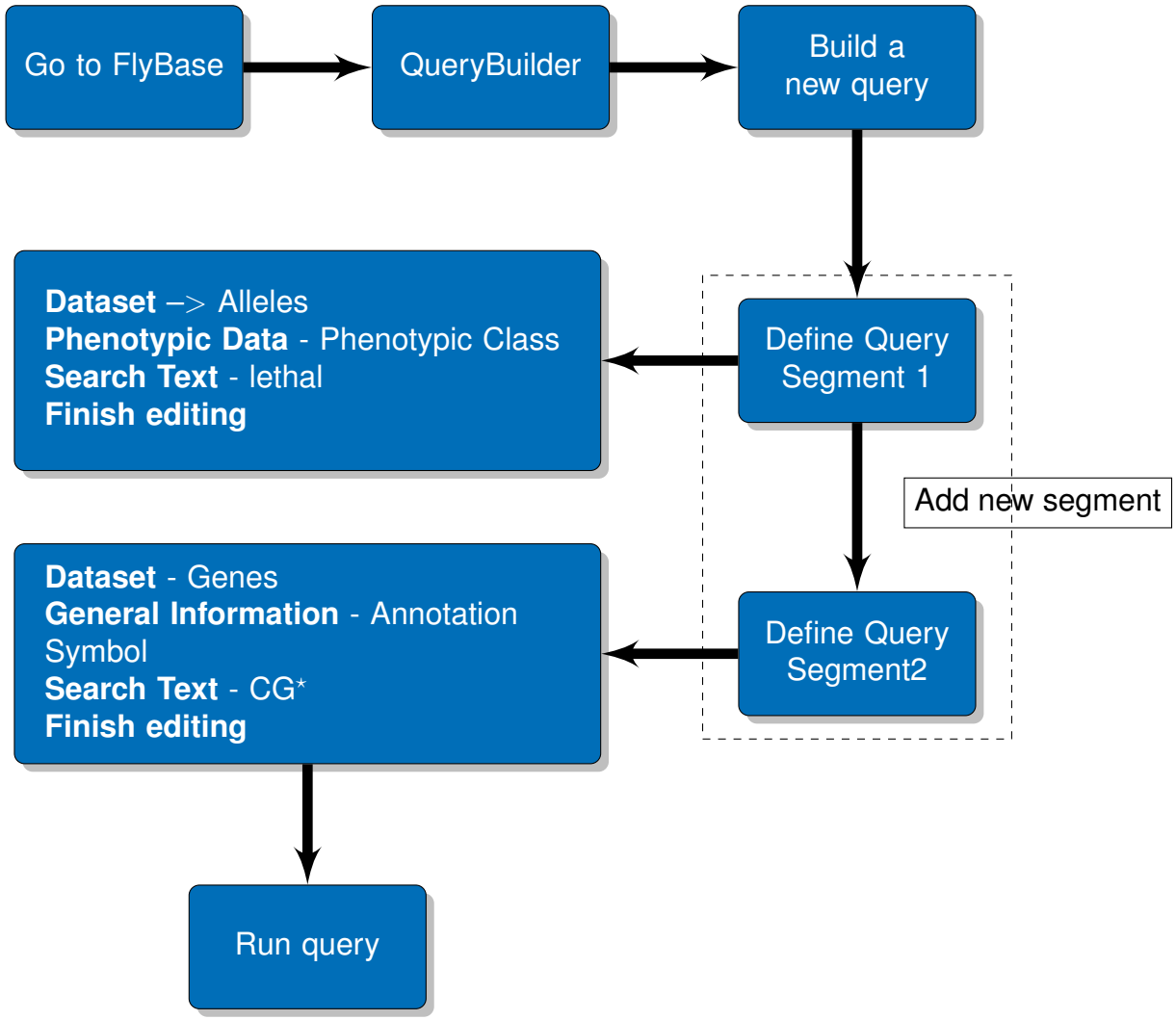


FIGURE B7.1: Flow chart for retrieving *D. melanogaster* essential genes from the FlyBase using Query-Builder tool.

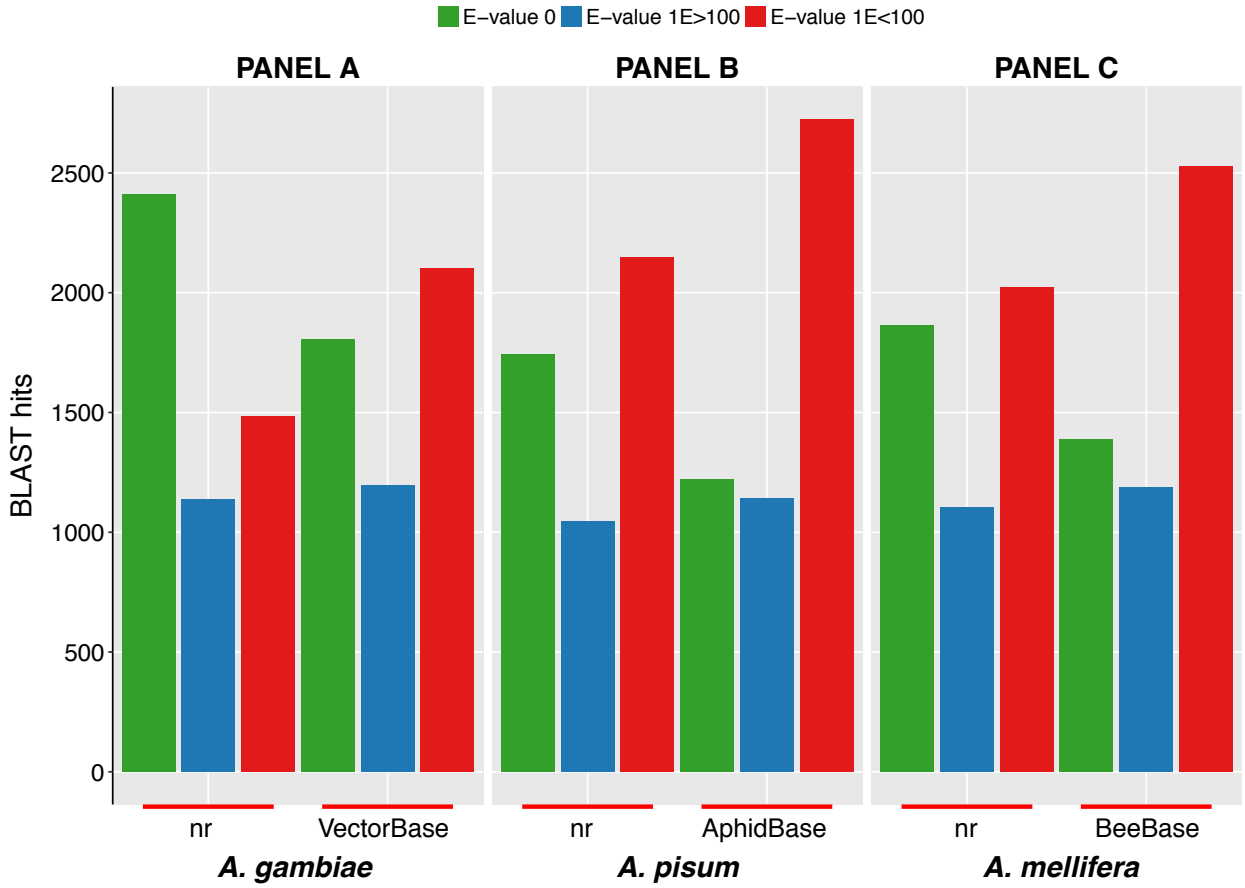
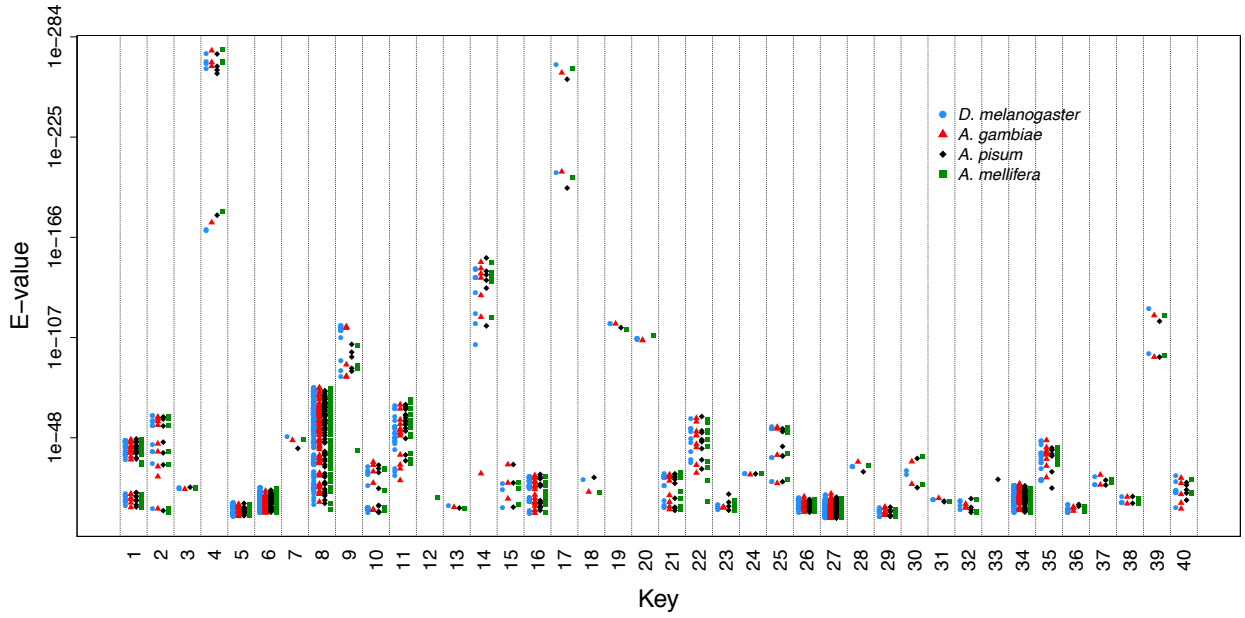
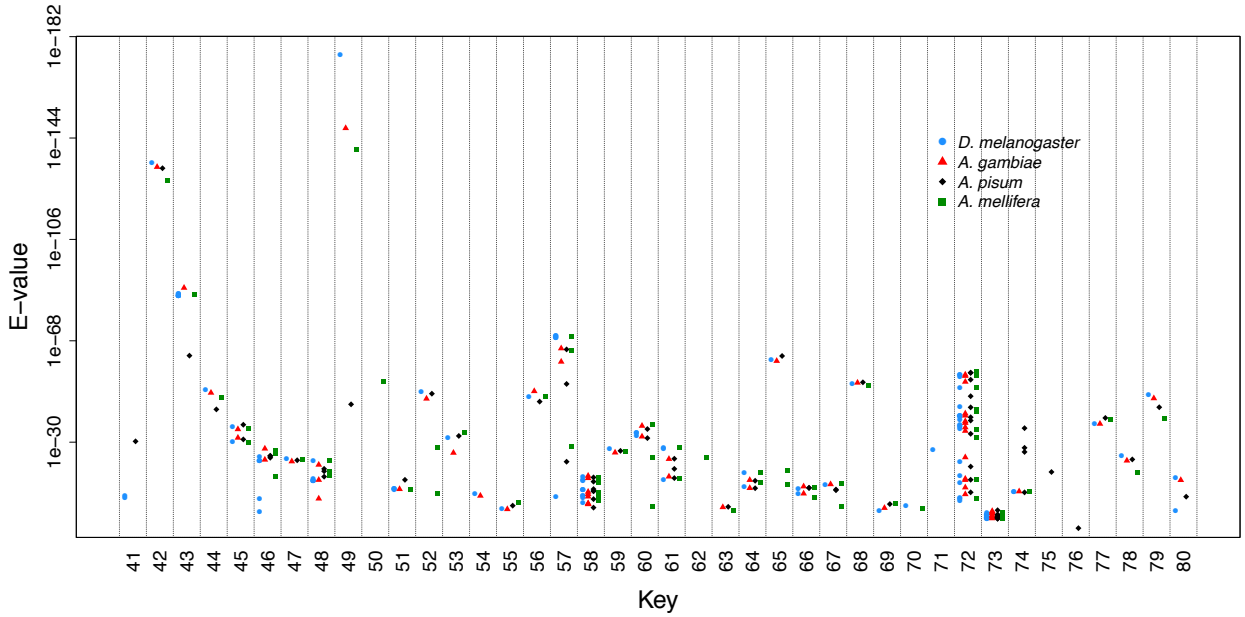


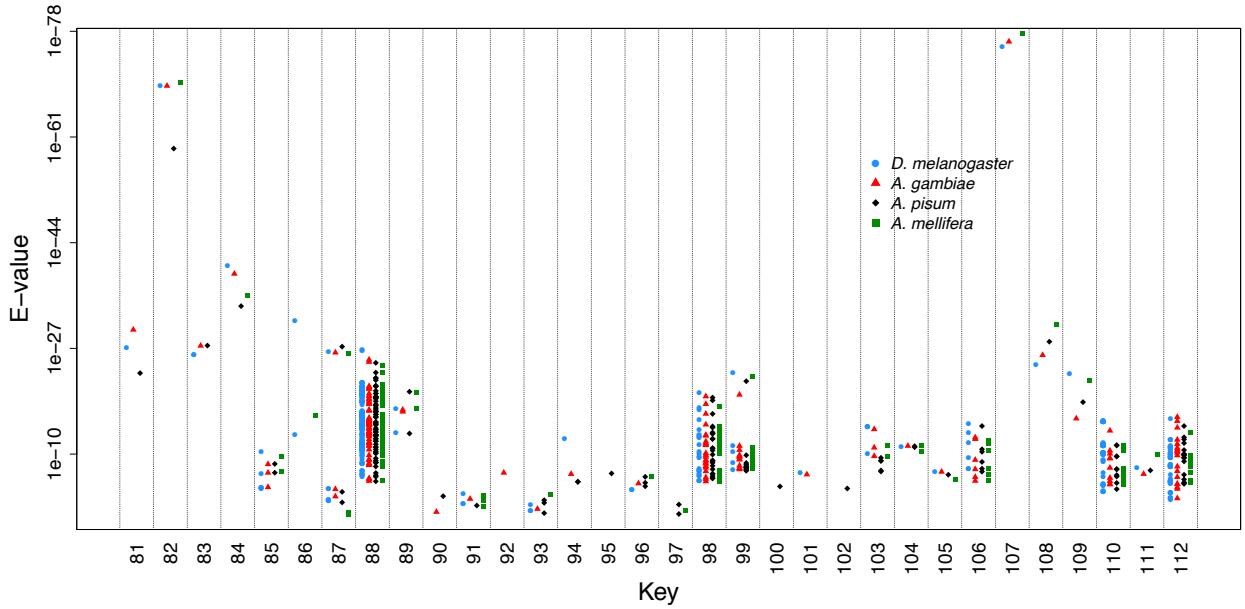
FIGURE B7.2: BLAST hits distribution of *A. pisum*, *A. gambiae* and *A. mellifera* orthologs for *D. melanogaster* essential proteins. The BLAST hits were classified into three arbitrary categories of E-value: 0, $1E^{>100}$ and $1E^{<100}$. BLAST hits were retrieved for each essential protein from nr database and species specific databases including VectorBase (PANEL A: *A. gambiae*), AphidBase (PANEL B: *A. pisum*) and BeeBase (PANEL C: *A. mellifera*).


E−value
1e−284
1e−225
1e−166
1e−107
1e−48
D. melanogaster
A. gambiae
A. pisum
A. mellifera
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
Key

(A)


E−value
1e−182
1e−144
1e−106
1e−68
1e−30
D. melanogaster
A. gambiae
A. pisum
A. mellifera
41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
Key

(B)



(C)

Key	Pfam ID	Name
1	PF00004.24	AAA
2	PF00009.22	GTP_EFTU
3	PF00011.16	HSP20
4	PF00012.15	HSP70
5	PF00023.25	Ank
6	PF00041.16	fn3
7	PF00056.18	Ldh_1_N
8	PF00069.20	Pkinase
9	PF00079.15	Serpin
10	PF00085.15	Thioredoxin
11	PF00089.21	Trypsin
12	PF00092.23	VWA
13	PF00111.22	Fer2
14	PF00118.19	Cpn60_TCP1
15	PF00125.19	Histone
16	PF00168.25	C2
17	PF00183.13	HSP90
18	PF00188.21	CAP
19	PF00206.15	Lyase_1
20	PF00217.14	ATP-gua_Ptrans
21	PF00226.26	DnaJ
22	PF00227.21	Proteasome
23	PF00249.26	Myb_DNA-binding
24	PF00313.17	CSD
25	PF00350.18	Dynamin_N

Key	Pfam ID	Name
26	PF00397.21	WW
27	PF00400.27	WD40
28	PF00447.12	HSF_DNA-bind
29	PF00515.23	TPR_1
30	PF00534.15	Glycos_transf_1
31	PF00564.19	PB1
32	PF00581.15	Rhodanese
33	PF00588.14	SpoU_methylase
34	PF00595.19	PDZ
35	PF00620.22	RhoGAP
36	PF00643.19	zf-B_box
37	PF00679.19	EFG_C
38	PF00684.14	DnaJ_CXXCXGXG
39	PF00735.13	Septin
40	PF00754.20	F5_F8_type_C
41	PF00931.17	NB-ARC
42	PF00982.16	Glyco_transf_20
43	PF00999.16	Na_H_Exchanger
44	PF01025.14	GrpE
45	PF01040.13	UbiA
46	PF01145.20	Band_7
47	PF01521.15	Fe-S_biosyn
48	PF01556.13	DnaJ_C
49	PF01593.19	Amino_oxidase
50	PF01728.14	FtsJ

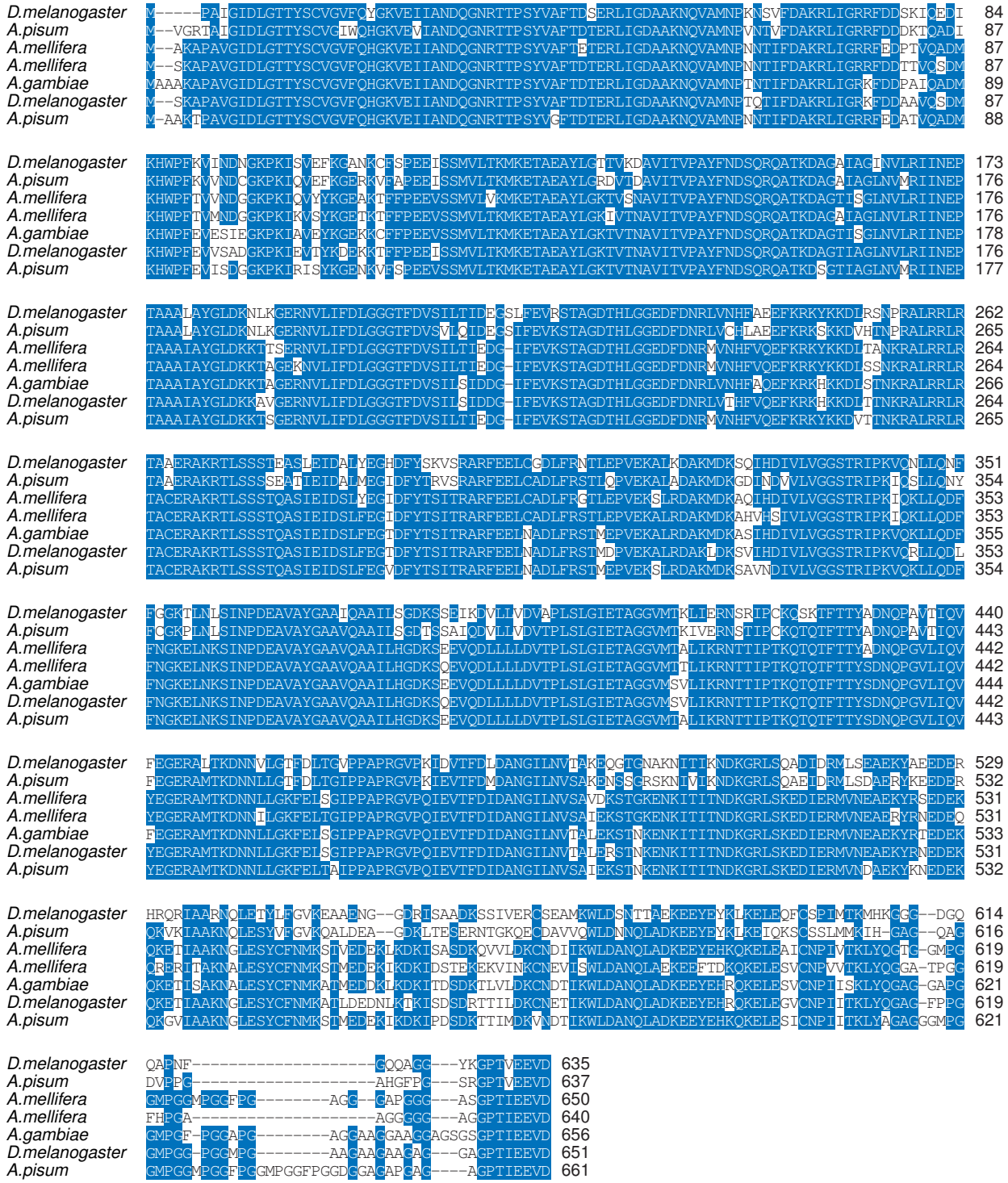
(D)

Key	Pfam ID	Name
51	PF01764.20	Lipase_3
52	PF01951.11	Archease
53	PF01965.19	DJ-1_PfpI
54	PF02518.21	HATPase_c
55	PF02540.12	NAD_synthase
56	PF02866.13	Ldh_1_C
57	PF03114.13	BAR
58	PF03144.20	GTP_EFTU_D2
59	PF03234.9	CDC37_N
60	PF03764.13	EFG_IV
61	PF04055.16	Radical_SAM
62	PF04117.7	Mpv17_PMP22
63	PF04564.10	U-box
64	PF04969.11	CS
65	PF05127.9	Helicase_RecD
66	PF05154.11	TM2
67	PF05485.7	THAP
68	PF06047.6	SynMuv_product
69	PF06325.8	PrmA
70	PF06546.6	Vert_HS_TF
71	PF07240.6	Turandot
72	PF07690.11	MFS_1
73	PF07719.12	TPR_2
74	PF07728.9	AAA_5
75	PF08032.7	SpoU_sub_bind
76	PF08238.7	Sel1
77	PF08351.6	DUF1726
78	PF08564.5	CDC37_C
79	PF08565.6	CDC37_M
80	PF08574.5	DUF1762
81	PF09320.6	DUF1977

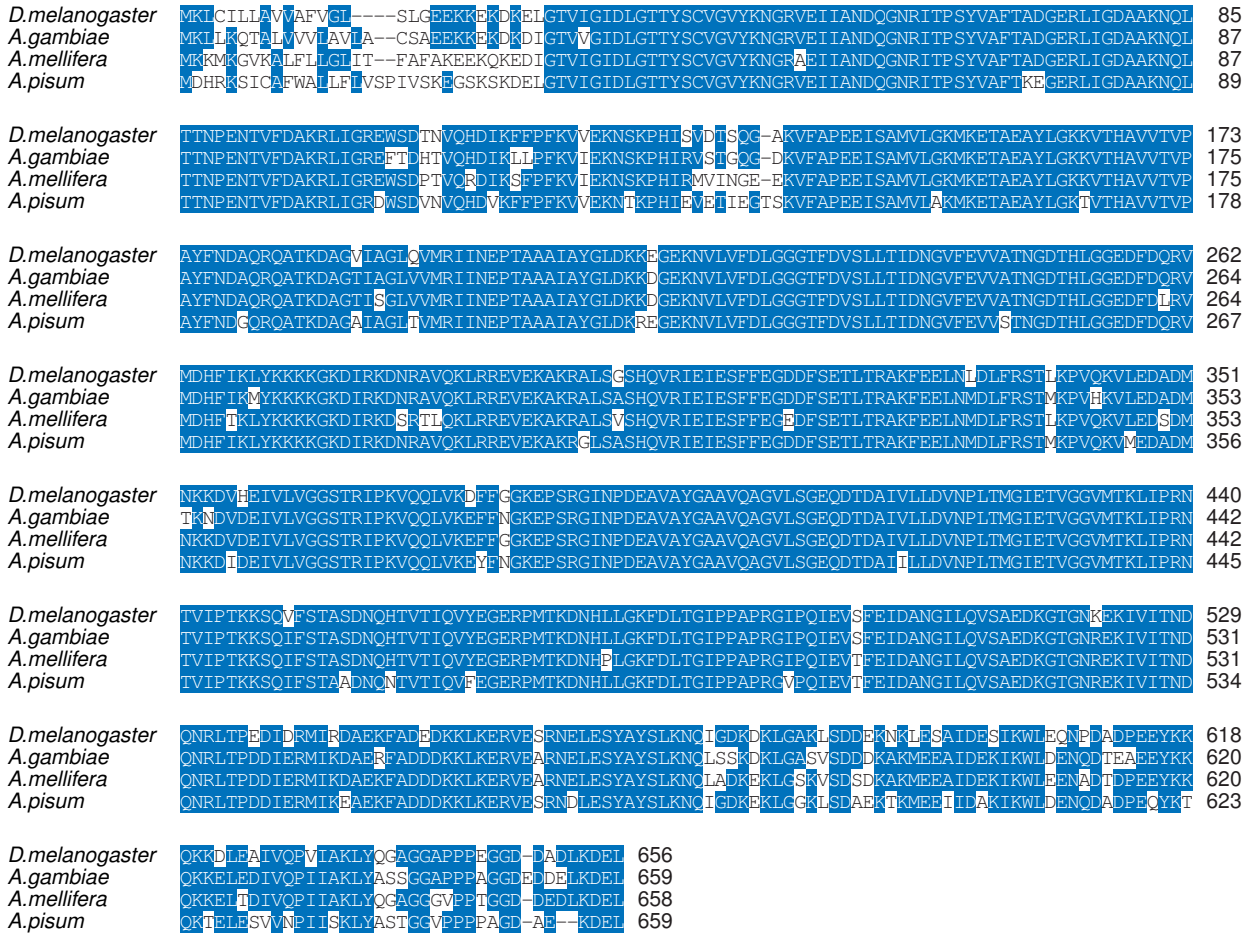
Key	Pfam ID	Name
82	PF10208.4	Armet
83	PF10415.4	FumaraseC_C
84	PF11701.3	UNC45-central
85	PF12171.3	zf-C2H2_jaz
86	PF12718.2	Tropomyosin_1
87	PF12756.2	zf-C2H2_2
88	PF12796.2	Ank_2
89	PF12895.2	Apc3
90	PF13174.1	TPR_6
91	PF13176.1	TPR_7
92	PF13180.1	PDZ_2
93	PF13181.1	TPR_8
94	PF13191.1	AAA_16
95	PF13207.1	AAA_17
96	PF13229.1	Beta_helix
97	PF13371.1	TPR_9
98	PF13414.1	TPR_11
99	PF13424.1	TPR_12
100	PF13428.1	TPR_14
101	PF13432.1	TPR_16
102	PF13450.1	NAD_binding_8
103	PF13519.1	VWA_2
104	PF13589.1	HATPase_c_3
105	PF13621.1	Cupin_8
106	PF13646.1	HEAT_2
107	PF13718.1	GNAT_acetyltr_2
108	PF13725.1	tRNA_bind_2
109	PF13877.1	RPAP3_C
110	PF13893.1	RRM_5
111	PF14237.1	DUF4339
112	PF14259.1	RRM_6

(E)

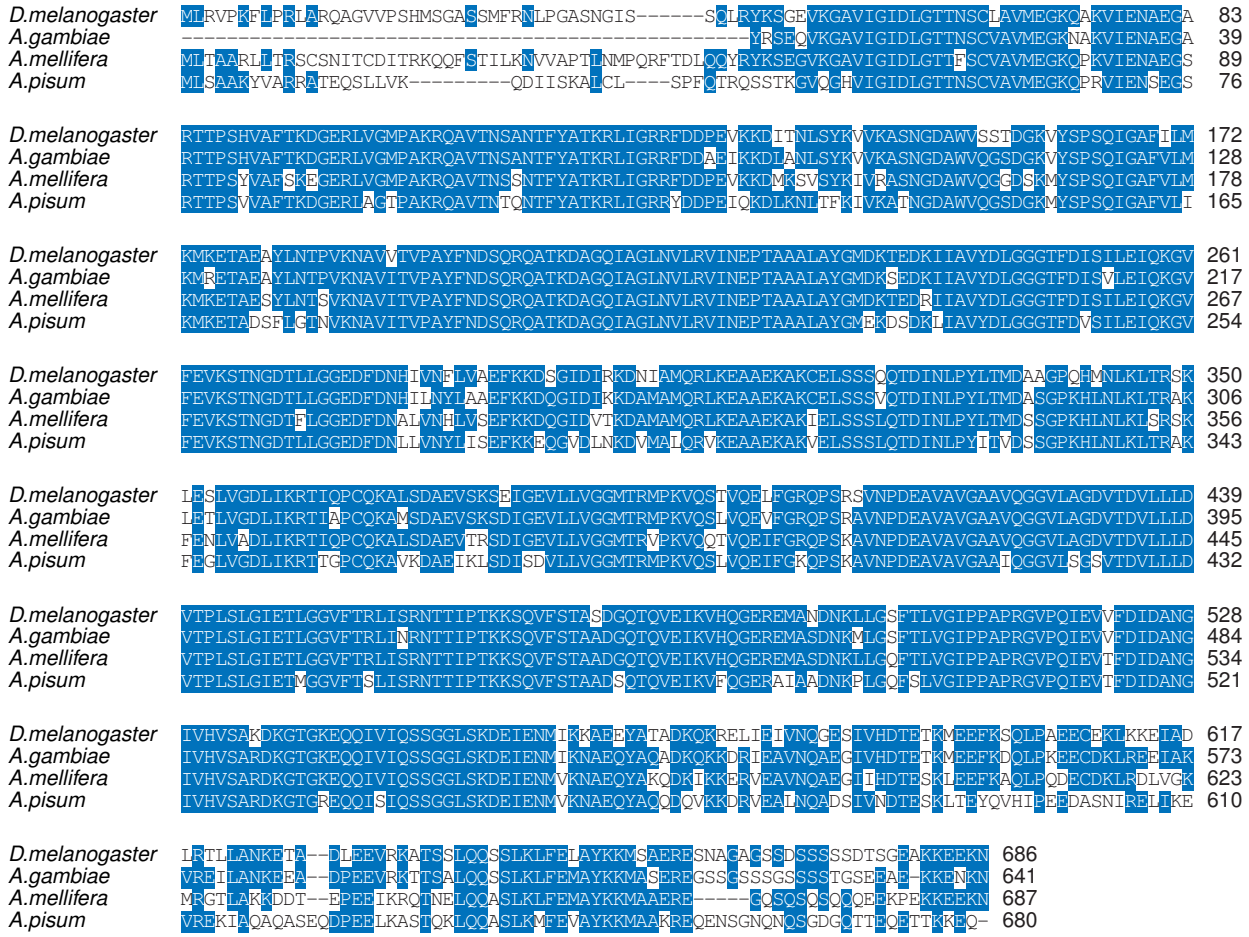
FIGURE B7.3: *D. melanogaster* essential proteins and their orthologs were subjected to Pfam analysis as described in the Chapter 8. Results are shown for an analysis of the heat shock associated proteins from *D. melanogaster*, *A. gambiae*, *A. pisum* and *A. mellifera* (A-C). A total of 112 Pfam families were identified from all four species and listed in the table (E-F) along with their corresponding key.



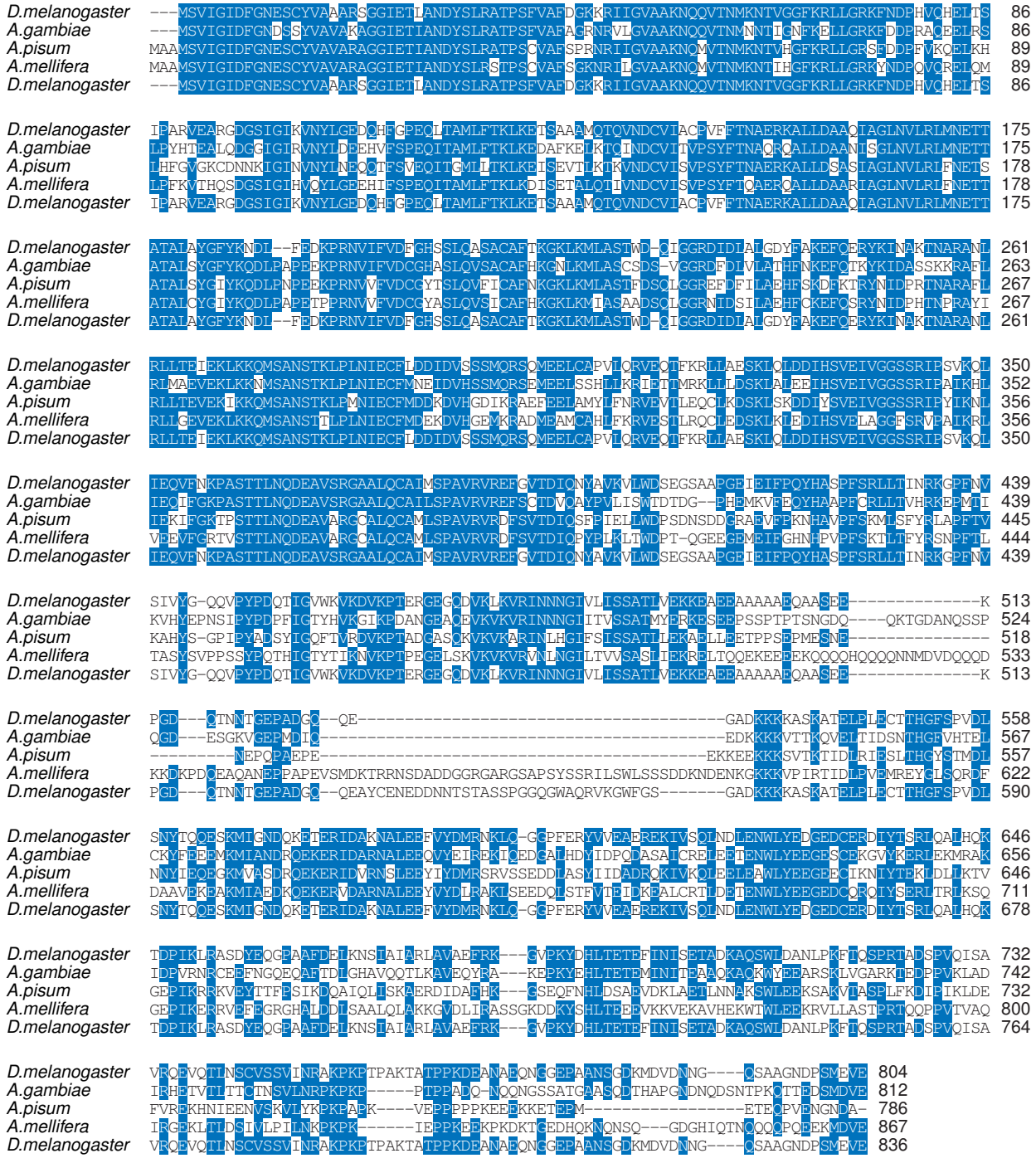
(A) Pfam cluster Hsp70A, sequence group A1



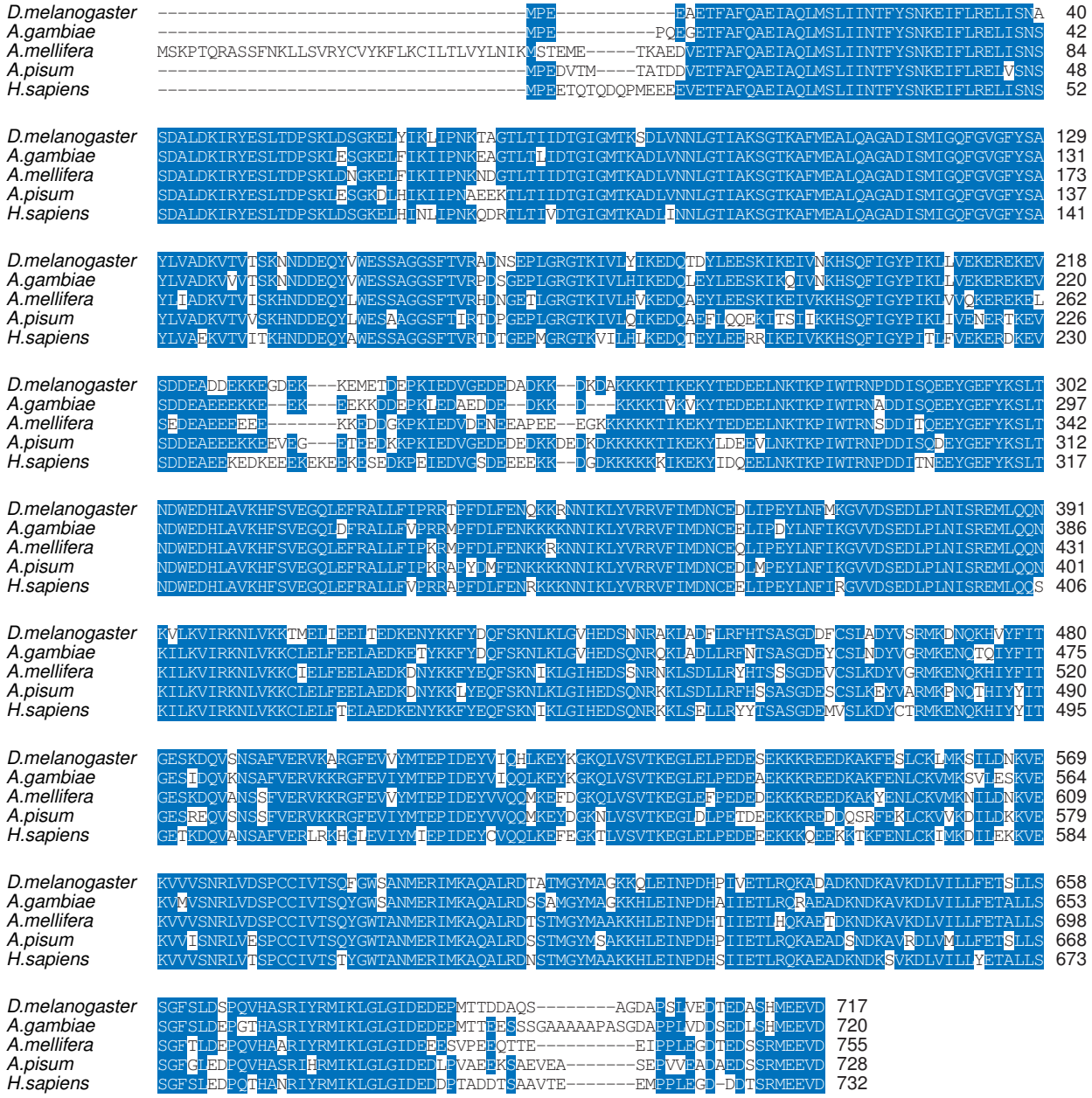
(B) Pfam cluster Hsp70A, sequence group A2



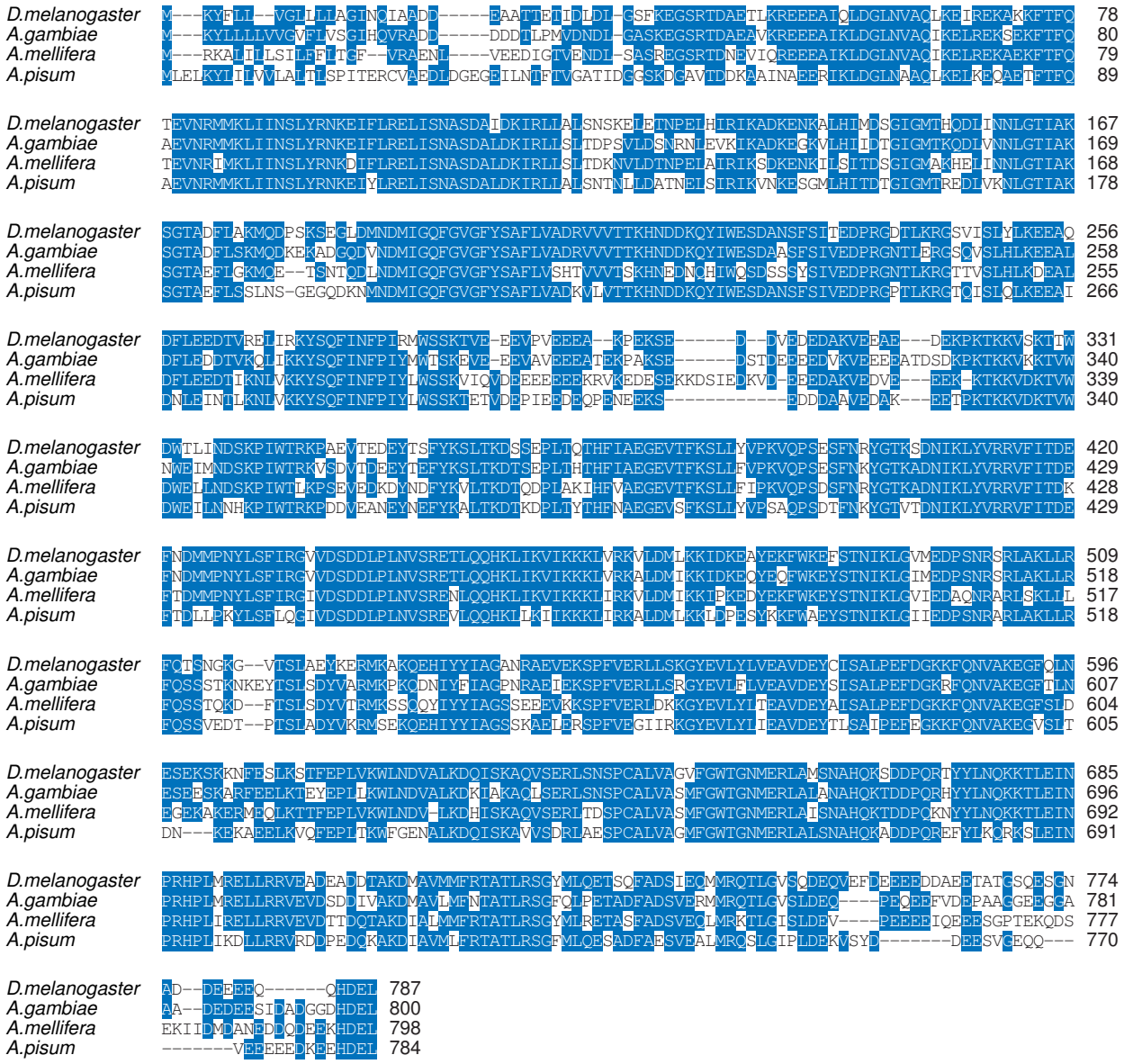
(c) Pfam cluster Hsp70A, sequence group A3



(D) Pfam cluster Hsp70B, sequence group B



(E) Pfam cluster Hsp90A



(F) Pfam cluster Hsp90B

FIGURE B7.4: Multiple alignments for all *D. melanogaster* essential protein orthologs identified within the Hsp70 and Hsp90 Pfam clusters and colored only identical residues (A-F). Note that multiple NCBI nr database entries for these clusters may share identical Pfam scores, appearing as superimposed orthologs, as seen in Figure B7.3. These overlapping relationships are clearly differentiated by multiple alignment.

BIBLIOGRAPHY

Adams M., Celniker S., Holt R., Evans C., Gocayne J., Amanatides P., Scherer S., Li P., Hoskins R., Galle R., George R., and et al (**2000**). "The genome sequence of *Drosophila melanogaster*." *Science*, 287(5461): 2185–2195. URL http://science.sciencemag.org/content/287/5461/2185.

Ahmed M., Ren S., Mandour N., Maruthi M., Naveed M., and Qiu B. (**2010**). "Phylogenetic analysis of *Bemisia tabaci* (Hemiptera: Aleyrodidae) populations from cotton plants in Pakistan, China, and Egypt." *Journal of Pest Science*, 83(2): 135–141. URL http://dx.doi.org/10.1007/s10340-009-0279-4.

Al-Khodor S., Price C., Kalia A., and Kwaik Y. (**2010**). "Ankyrin-repeat containing proteins of microbes: a conserved structure with functional diversity." *Trends in microbiology*, 18(3): 132–139. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2834824/.

Alemandri V., Barro P. D., Bejerman N., Caro E. A., Dumón A., Mattio M., Rodriguez S., and Truol G. (**2012**). "Species within the *Bemisia tabaci* (Hemiptera: Aleyrodidae) Complex in Soybean and Bean Crops in Argentina." *Journal of Economic Entomology*, 105(1): 48–53. URL http://dx.doi.org/10.1603/EC11161.

Alikhan N., Petty N., Zakour N. B., and Beatson S. (**2011**). "BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons." *BMC Genomics*, 12(1): 1–10. URL http://dx.doi.org/10.1186/1471-2164-12-402.

Alioto T. (**2007**). "U12DB: a database of orthologous U12-type spliceosomal introns." *Nucleic Acids Research*, 35(suppl 1): D110–D115. URL http://nar.oxfordjournals.org/content/35/suppl_1/D110.abstract.

Alon M., Alon F., Nauen R., and Morin S. (**2008**). "Organophosphate resistance in the B biotype of *Bemisia tabaci* (Hemiptera: Aleyrodidae) is associated with a point mutation in an ace1-type acetylcholinesterase and overexpression of carboxylesterase." *Insect Biochemistry and Molecular Biology*, 38: 940–949.

Altschul S., Gish W., Miller W., Myers E., and Lipman D. (**1990**). "Basic local alignment search tool." *Journal of Molecular Biology*, 215(3): 403–10.

Altschul S., Madden T., Schaffer A., Zhang J., Zhang Z., Miller W., and Lipman D. (**1997**). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Research*, 25(17): 3389–402.

Amrein H., Hedley M., and Maniatis T. (**1994**). "The role of specific protein-RNA and protein-protein interactions in positive and negative control of pre-mRNA splicing by Transformer 2." *Cell*, 76(4): 735–746. URL http://www.sciencedirect.com/science/article/pii/0092867494905126.

Angiuoli S., Gussman A., Klimke W., Cochrane G., Field D., Garrity G., Kodira C., Kyrpides N., Madupu R., Markowitz V., Tatusova T., Thomson N., and White O. (**2008**). "Toward an Online Repository of Standard Operating Procedures (SOPs) for (Meta)genomic Annotation." *OMICS: A Journal of Integrative Biology*, 12(2): 137–141. URL http://dx.doi.org/10.1089/omi.2008.0017.

Anon (**2001**). *Crop Protection Compendium, Global Module*. CAB International CD-Rom Database, 3rd edition.

Ashburner M., Ball C., Blake J., Botstein D., Butler H., Cherry J., Davis A., Dolinski K., Dwight S., Eppig J., Harris M., Hill D., Issel-Tarver L., Kasarskis A., Lewis S., Matese J., Richardson J., Ringwald M., Rubin G., and Sherlock G. (**2000**). "Gene Ontology: tool for the unification of biology." *Nature Genetics*, 25(1): 25–29. URL http://dx.doi.org/10.1038/75556.

Ashburner M., Golic K., and Hawley R. (**2005**). *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press, New York, 2nd edition.

Avidov Z. and Harpaz I. (**1969**). *Plant Pests of Israel*. Israel Universities Press, Jerusalem.

Azab A., Megahed M., and EI-Mirsawi H. (**1971**). "On the range of host-plants of *Bemisia tabaci* (Genn.)." *Bulletin of Entomolgical Society Egypt*, 54: 319–326.

Bailey L., Searls D., and Overton G. (**1998**). "Analysis of EST-Driven Gene Annotation in Human Genomic?Sequence." *Genome Research*, 8(4): 362–376. URL http://genome.cshlp.org/content/8/4/362.abstract.

Bamborough P. (**2012**). "System-based drug discovery within the human kinome." *Expert Opinion on Drug Discovery*, 7(11): 1053–1070. URL http://dx.doi.org/10.1517/17460441.2012.724056.

Banks G. and Markham P. (**2000**). *Bemisia tabaci - how many biotypes are there? In: Gerling D, Jones W, eds.* 13. Bemisia Newsletter.

Bao Z. and Eddy S. (**2002**). "Automated *De novo* Identification of Repeat Sequence Families in Sequenced Genomes." *Genome Research*, 12(8): 1269–1276. URL http://genome.cshlp.org/content/12/8/1269.abstract.

Baoli Q., Coats S., Shunxiang R., Idris A., Caixia X., and Brown J. (**2007**). "Phylogenetic relationship of native and introduced *Bemisia tabaci* (Homoptera: Aleyrodidae) from China and India based on mtCOI DNA sequencing and host plant comparisons." *Progress in Natural Science*, 17(6): 645–654. URL http://www.tandfonline.com/doi/abs/10.1080/10002007088537453.

Barro P. D., Driver F., Trueman J., and Curran J. (**2000**). "Phylogenetic Relationships of World Populations of *Bemisia tabaci* (Gennadius) Using Ribosomal ITS1." *Molecular Phylogenetics and Evolution*, 16(1): 29–36. URL http://www.sciencedirect.com/science/article/pii/S1055790399907686.

Barro P. D. and Hart P. (**2000**). "Mating interactions between two biotypes of the whitefly, *Bemisia tabaci* (Hemiptera: Aleyrodidae) in Australia." *Bulletin of Entomological Research*, 90: 103–112. URL http://journals.cambridge.org/article_S0007485300000201.

Barro P. D., Liu S., Boykin L., and Dinsdale A. (**2011**). "*Bemisia tabaci*: A Statement of Species Status." *Annual Review of Entomology*, 56(1): 1–19. URL http://www.annualreviews.org/doi/abs/10.1146/annurev-ento-112408-085504. PMID: 20690829.

Basu A. (**1995**). *Bemisia tabaci (Gennadius): Crop Pest and Principal Whitefly Vector of Plant Viruses*. Westview Press, New Delhi.

Baumann L., Thao M., Funk C., Falk B., Ng J., and Baumann P. (**2004**). "Sequence Analysis of DNA Fragments from the Genome of the Primary Endosymbiont of the Whitefly *Bemisia tabaci*." *Current Microbiology*, 48(1): 77–81. URL http://dx.doi.org/10.1007/s00284-003-4132-3.

Baumann P. (**2005**). "Biology of bacteriocyte-associated endosymbionts of plant sap-sucking insects." *Annual Review of Microbiology*, 59: 155–189. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-27144454498&partnerID=40&md5=077f6e9d21b992778afb282b8c7fd3b2.

Beatty J., Smagghe G., Ogura T., Nakagawa Y., Spindler-Barth M., and Henrich V. (**2009**). "Properties of ecdysteroid receptors from diverse insect species in a heterologous cell culture system ? a basis for screening novel insecticidal candidates." *FEBS Journal*, 276(11): 3087–3098. URL http://dx.doi.org/10.1111/j.1742-4658.2009.07026.x.

Beckenbach A. and Stewart J. (**2009**). "Insect mitochondrial genomics 3: the complete mitochondrial genome sequences of representatives from two neuropteroid orders: a dobsonfly (order Megaloptera) and a giant lacewing and an owlfly (order Neuroptera)." *Genome*, 52(1): 31–38. URL http://dx.doi.org/10.1139/G08-098. PMID: 19132069.

Bedford I., Briddon R., Brown J., Rosell R., and Markham P. (**1994**). "Geminivirus transmission and biological characterisation of *Bemisia tabaci* (Gennadius) biotypes from different geographic regions." *Annals of Applied Biology*, 125(2): 311–325. URL http://dx.doi.org/10.1111/j.1744-7348.1994.tb04972.x.

Behere G., Firake D., Tay W., Thakur N. A., and Ngachan S. (**2016**). "Complete mitochondrial genome sequence of a phytophagous ladybird beetle, *Henosepilachna pusillanima* (Mulsant) (Coleoptera: Coccinellidae)." *Mitochondrial DNA*, 27(1): 291–292. URL http://dx.doi.org/10.3109/19401736.2014.892082. PMID: 24617459.

Behura S., Haugen M., Flannery E., Sarro J., Tessier C., Severson D., and Duman-Scheel M. (**2011**). "Comparative Genomic Analysis of *Drosophila melanogaster* and Vector Mosquito Developmental Genes." *PLoS ONE*, 6(7): e21 504. URL http://dx.doi.org/10.1371%2Fjournal.pone.0021504.

Belleghem S. V., Roelofs D., Houdt J. V., and Hendrickx F. (**2012**). "*De novo* Transcriptome Assembly and SNP Discovery in the Wing Polymorphic Salt Marsh Beetle *Pogonus chalceus* (Coleoptera, Carabidae)." *PLoS ONE*, 7(8): e42 605. URL http://dx.doi.org/10.1371%2Fjournal.pone.0042605.

Bennetzen J. (**2005**). "Transposable elements, gene creation and genome rearrangement in flowering plants." *Current Opinion in Genetics and Development*, 15(6): 621–627. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-27444444339&partnerID=40&md5=ce60975f9e4dd137119b521b7e9c9da6.

Benson G. (**1999**). "Tandem repeats finder: a program to analyze DNA sequences." *Nucleic Acids Research*, 27(2): 573–580. URL http://nar.oxfordjournals.org/content/27/2/573.abstract.

Bernt M., Donath A., Jühling F., Externbrink F., Florentz C., Fritzsch G., Pütz J., Middendorf M., and Stadler P. (**2013**). "MITOS: Improved *de novo* metazoan mitochondrial genome annotation." *Molecular Phylogenetics and Evolution*, 69(2): 313 – 319. URL http://www.sciencedirect.com/science/article/pii/S1055790312003326.

Bettencourt-Dias M., Giet R., Sinka R., Mazumdar A., Lock W., Balloux F., Zafiropoulos P., Yamaguchi S., Winter S., Carthew R., Cooper M., Jones D., Frenz L., and Glover D. (**2004**). "Genome-wide survey of protein kinases required for cell cycle progression." *Nature*, 432(7020): 980–987.

Biémont C. and Vieira C. (**2006**). "Genetics: Junk DNA as an evolutionary force." *Nature*, 443(7111): 521–524. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-33749522239&partnerID=40&md5=4a287ac126e56e2f2ee1490b26d10cf1.

Billas I. and Moras D. (**2005**). "Ligand?Binding Pocket of the Ecdysone Receptor." In "Insect Hormones," volume 73 of *Vitamins & Hormones*, pages 101–129. Academic Press. URL http://www.sciencedirect.com/science/article/pii/S0083672905730041.

Bing X., Ruan Y., Rao Q., Wang X., and Liu S. (**2013b**). "Diversity of secondary endosymbionts among different putative species of the whitefly *Bemisia tabaci*." *Insect Science*, 20(2): 194–206. URL http://dx.doi.org/10.1111/j.1744-7917.2012.01522.x.

Bing X., Xia W., Gui J., Yan G., Wang X., and Liu S. (**2014**). "Diversity and evolution of the *Wolbachia* endosymbionts of *Bemisia* (Hemiptera: Aleyrodidae) whiteflies." *Ecology and Evolution*, 4(13): 2714–2737. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4113295/.

Bing X., Yang J., Zchori-Fein E., Wang X., and Liu S. (**2013a**). "Characterization of a Newly Discovered Symbiont of the Whitefly *Bemisia tabaci* (Hemiptera: Aleyrodidae)." *Applied and Environmental Microbiology*, 79(2): 569–575. URL http://aem.asm.org/content/79/2/569.abstract.

Birney E. and Durbin R. (**2000**). "Using GeneWise in the Drosophila Annotation Experiment." *Genome Research*, 10(4): 547–548. URL http://genome.cshlp.org/content/10/4/547.abstract.

Board P., Baker R., Chelvanayagam G., and Jermiin L. (**1997**). "Zeta, a novel class of glutathione transferases in a range of species from plants to humans." *Biochemical Journal*, 328: 929–935. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1219006/.

Boetzer M., Henkel C., Jansen H., Butler D., and Pirovano W. (**2011**). "Scaffolding pre-assembled contigs using SSPACE." *Bioinformatics*, 27(4): 578–579. URL http://bioinformatics.oxfordjournals.org/content/27/4/578.abstract.

Böhme S., Meyer S., Krüger A., Steinhoff H., Wittinghofer A., and Klare J. (**2010**). "Stabilization of G Domain Conformations in the tRNA-modifying MnmE-GidA Complex Observed with Double Electron Electron Resonance Spectroscopy." *Journal of Biological Chemistry*, 285(22): 16 991–17 000. URL http://www.jbc.org/content/285/22/16991.abstract.

Bonfield J., Smith K., and Staden R. (**1995**). "A new DNA sequence assembly program." *Nucleic Acids Research*, 23(24): 4992–4999. URL http://nar.oxfordjournals.org/content/23/24/4992.abstract.

Boore J. (**1999**). "Animal mitochondrial genomes." *Nucleic Acids Research*, 27(8): 1767–1780. URL http://nar.oxfordjournals.org/content/27/8/1767.abstract.

Boore J. (**2001**). "Complete Mitochondrial Genome Sequence of the Polychaete Annelid *Platynereis dumerilii*." *Molecular Biology and Evolution*, 18(7): 1413–1416. URL http://mbe.oxfordjournals.org/content/18/7/1413.short.

Boore J. (**2006**). "Requirements and Standards for Organelle Genome Databases." *OMICS: A Journal of Integrative Biology*, 10(2): 119–126. URL http://dx.doi.org/10.1089/omi.2006.10.119.

Boore J., Macey J., and Medina M. (**2005**). "Sequencing and Comparing Whole Mitochondrial Genomes of Animals." In "Molecular Evolution: Producing the Biochemical Data," volume 395 of *Methods in Enzymology*, pages 311 – 348. Academic Press. URL http://www.sciencedirect.com/science/article/pii/S0076687905950192.

Boykin L. (**2014**). "*Bemisia tabaci* nomenclature: lessons learned." *Pest Management Science*, 70(10): 1454–1459. URL http://dx.doi.org/10.1002/ps.3709.

Boykin L., Armstrong K., Kubatko L., and Barro P. D. (**2012**). "Species Delimitation and Global Biosecurity." *Evolutionary Bioinformatics*, 8: 1–37. URL www.la-press.com/species-delimitation-and-global-biosecurity-article-a2954.

Boykin L., Bell C., Evans G., Small I., and Barro P. D. (**2013**). "Is agriculture driving the diversification of the *Bemisia tabaci* species complex (Hemiptera: Sternorrhyncha: Aleyrodidae)?: Dating, diversification and biogeographic evidence revealed." *BMC Evolutionary Biology*, 13(1): 1–10. URL http://dx.doi.org/10.1186/1471-2148-13-228.

Boykin L., Shatters R., Rosell R., McKenzie C., Bagnall R., Barro P. D., and Frohlich D. (**2007**). "Global relationships of *Bemisia tabaci* (Hemiptera: Aleyrodidae) revealed using Bayesian analysis of mitochondrial COI DNA sequences." *Molecular Phylogenetics and Evolution*, 44(3): 1306–1319. URL http://www.sciencedirect.com/science/article/pii/S1055790307001388.

Brar D., Aneja A., Singh J., and Mahal M. (**2005**). "Biology of whitefly, *Bemisia tabaci* (Gennadius) on American cotton, Gossypium hirsutum Linnaeus." *Journal of Insect Science*, 18: 48–59.

Brautigam A., Mullick T., Schliesky S., and Weber A. (**2011**). "Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C3 and C4 species." *Journal of Experimental Botany*, 62(9): 3093–3102. URL http://jxb.oxfordjournals.org/content/62/9/3093.abstract.

Breathnach R. and Chambon P. (**1981**). "Organization and Expression of Eucaryotic Split Genes Coding for Proteins." *Annual Review of Biochemistry*, 50(1): 349–383. URL http://dx.doi.org/10.1146/annurev.bi.50.070181.002025. PMID: 6791577.

Brent M. (**2005**). "Genome annotation past, present, and future: How to define an ORF at each locus." *Genome Research*, 15(12): 1777–1786. URL http://genome.cshlp.org/content/15/12/1777.abstract.

Brock J., Board P., and Oakley A. (**2013**). "Structural Insights into Omega-Class Glutathione Transferases: A Snapshot of Enzyme Reduction and Identification of a Non-Catalytic Ligandin Site." *PLoS ONE*, 8(4): 1–10. URL http://dx.doi.org/10.1371%2Fjournal.pone.0060324.

Broughton S., Harrison J., and Rahman T. (**2013**). "Effect of new and old pesticides on *Orius armatus* (Gross)—an Australian predator of western flower thrips, *Frankliniella occidentalis* (Pergande)." *Pest Management Science*. URL http://dx.doi.org/10.1002/ps.3565.

Brown J. (**2007**). "The *Bemisia tabaci* Complex: Genetic and Phenotypic Variability Drives *Begomovirus* Spread and Virus Diversification." *American Phytopathological Society*. URL http://www.apsnet.org/publications/apsnetfeatures/Pages/BemisiatabaciComplex.aspx. As accessed on 1st September, 2013.

Brown J. (**2010**). "Phylogenetic Biology of the *Bemisia tabaci* Sibling Species Group." In P. A. Stansly and S. E. Naranjo, editors, "Bemisia: Bionomics and Management of a Global Pest," pages 31–67. Springer Netherlands. URL http://dx.doi.org/10.1007/978-90-481-2460-2_2.

Brown J. and Bird J. (**1992**). "Whitefly-transmitted geminiviruses and associated disorders in the Americas and the Caribbean Basin: past and present." *Plant Disease*, 76(3): 220–225.

Brown J., Frohlich D., and Rosell R. (**1995**). "The Sweetpotato or Silverleaf Whiteflies: Biotypes of *Bemisia tabaci* or a Species Complex?" *Annual Review of Entomology*, 40(1): 511–534. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.en.40.010195.002455.

Brown J. and Idris A. (**2005**). "Genetic Differentiation of Whitefly *Bemisia tabaci* Mitochondrial Cytochrome Oxidase I, and Phylogeographic Concordance with the Coat Protein of the Plant Virus Genus *Begomovirus*." *Annals of the Entomological Society of America*, 98(6): 827–837. URL http://aesa.oxfordjournals.org/content/98/6/827.

Brown J., Lambert G., Ghanim M., Czosnek H., and Galbraith D. (**2005**). "Nuclear DNA content of the whitefly *Bemisia tabaci* (Aleyrodidae: Hemiptera) estimated by flow cytometry." *Bulletin of Entomological Research*, 95: 309–312. URL http://journals.cambridge.org/article_S0007485305000295.

Brown S., Shippy T., Miller S., Bolognesi R., Beeman R., Lorenzen M., Bucher G., Wimmer E., and Klingler M. (**2009**). "The Red Flour Beetle, *Tribolium castaneum* (Coleoptera): A Model for Studies of Development and Pest Biology." *Cold Spring Harbor Protocols*, 2009(8): pdb.emo126. URL http://cshprotocols.cshlp.org/content/2009/8/pdb.emo126.abstract.

Bruce T. (**2010**). "Tackling the threat to food security caused by crop pests in the new millennium." *Food Security*, 2(2): 133–141. URL http://dx.doi.org/10.1007/s12571-010-0061-8.

Brumin M., Kontsedalov S., and Ghanim M. (**2011**). "Rickettsia influences thermotolerance in the whitefly *Bemisia tabaci* B biotype." *Insect Science*, 18(1): 57–66. URL http://dx.doi.org/10.1111/j.1744-7917.2010.01396.x.

Buckingham S., Biggin P., Sattelle B., Brown L., and Sattelle D. (**2005**). "Insect GABA Receptors: Splicing, Editing, and Targeting by Antiparasitics and Insecticides." *Molecular Pharmacology*, 68(4): 942–951.

Burge C. and Karlin S. (**1997**). "Prediction of complete gene structures in human genomic DNA1." *Journal of Molecular Biology*, 268(1): 78–94. URL http://www.sciencedirect.com/science/article/pii/S0022283697909517.

Burrack H., Fernandez G., Spivey T., and Kraus D. (**2013**). "Variation in selection and utilization of host crops in the field and laboratory by *Drosophila suzukii* Matsumara (Diptera: Drosophilidae), an invasive frugivore." *Pest Management Science*, 69(10): 1173–1180. URL http://dx.doi.org/10.1002/ps.3489.

Burset M. and Guigo R. (**1996**). "Evaluation of Gene Structure Prediction Programs." *Genomics*, 34(3): 353 – 367. URL http://www.sciencedirect.com/science/article/pii/S0888754396902980.

Burset M., Seledtsov I., and Solovyev V. (**2000**). "Analysis of canonical and non-canonical splice sites in mammalian genomes." *Nucleic Acids Research*, 28(21): 4364–4375. URL http://nar.oxfordjournals.org/content/28/21/4364.abstract.

Burton J., Adey A., Patwardhan R., Qiu R., Kitzman J., and Shendure J. (**2013**). "Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions." *Nature Biotechnology*, 31(12): 1119–1125. URL http://dx.doi.org/10.1038/nbt.2727.

Byrne D. and Bellows T. (**1991**). "Whitefly Biology." *Annual Review of Entomology*, 36(1): 431–457. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.en.36.010191.002243.

Byrne F. and Devonshire A. (**1996**). "Biochemical evidence of haplodiploidy in the whitefly *Bemisia tabaci*." *Biochemical Genetics*, 34: 93–107.

Byrne F., Gorman K., Cahill M., Denholm I., and Devonshire A. (**2000**). "The role of B-type esterases in conferring insecticide resistance in the tobacco whitefly, *Bemisia tabaci* (Genn.)." *Pest Management Science*, 56: 867–874.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., and Madden T. (**2009**). "BLAST+: architecture and applications." *BMC Bioinformatics*, 10: 421.

Cameron S., Beckenbach A., Dowton M., and Whiting M. (**2006**). "Evidence from mitochondrial genomics on interordinal relationships in insects." *Arthropod Systematics and Phylogeny*, 64(1): 27–34.

Carmichael J., Lawrence M., Graham L., Pilling P., Epa V., Noyce L., Lovrecz G., Winkler D., Pawlak-Skrzecz A., Eaton R., Hannan G., and Hill R. (**2005**). "The X-ray structure of a hemipteran ecdysone receptor ligand-binding domain: comparison with a lepidopteran ecdysone receptor ligand-binding domain and implications for insecticide design." *Journal of Biological Chemistry*, 280(23): 22 258–22 269. URL http://www.jbc.org/content/280/23/22258.abstract.

Carriere Y. (**2003**). "Haplodiploidy, sex, and the evolution of pesticide resistance." *Journal of Economic Entomology*, 96: 1626–1640.

Caspi-Fluger A., Inbar M., Mozes-Daube N., Katzir N., Portnoy V., Belausov E., Hunter M., and Zchori-Fein E. (**2012**). "Horizontal transmission of the insect symbiont *Rickettsia* is plant-mediated." *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1734): 1791–1796. URL http://rspb.royalsocietypublishing.org/content/279/1734/1791.

Castle S., Henneberry T., Prabhaker N., and Toscano N. (**1996**). "Trends in relative susceptibilities of whiteflies to insecticides through the cotton season in the Imperial Valley, CA." *Proceedings*, 2: 1032–1035. URL http://europepmc.org/abstract/AGR/IND20575080.

Caterino M. and Sperling F. (**1999**). "Papilio Phylogeny Based on Mitochondrial Cytochrome Oxidase I and II Genes." *Molecular Phylogenetics and Evolution*, 11(1): 122–137. URL http://www.sciencedirect.com/science/article/pii/S1055790398905498.

Cha S., Yoon H., Lee E., Yoon M., Hwang J., Jin B., Han Y., and Kim I. (**2007**). "The complete nucleotide sequence and gene organization of the mitochondrial genome of the bumblebee, *Bombus ignitus* (Hymenoptera: Apidae)." *Gene*, 392(1Äì2): 206 – 220. URL http://www.sciencedirect.com/science/article/pii/S0378111907000029.

Chakraborty M., Baldwin-Brown J., Long A., and Emerson J. (**2015**). "A practical guide to *de novo* genome assembly using long reads." *bioRxiv*. URL http://biorxiv.org/content/early/2015/10/16/029306.

Chan E., Rowe H., Hansen B., and Kliebenstein D. (**2010**). "The Complex Genetic Architecture of the Metabolome." *PLoS Genetics*, 6(11): e1001 198. URL http://dx.doi.org/10.1371%2Fjournal.pgen.1001198.

Chandler D., Bailey A., Tatchell G., Davidson G., Greaves J., and Grant W. (**2011**). "The development, regulation and use of biopesticides for integrated pest management." *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1573): 1987–1998. URL http://rstb.royalsocietypublishing.org/content/366/1573/1987.

Chapman A. (**2009**). *Numbers of living species in Australia and the World*. Canberra: Australian Biological Resources Study, 2nd edition. Report for the Australian Biological Resources Study.

Chen N. (**2004**). "Using RepeatMasker to identify repetitive elements in genomic sequences." *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ...* [*et al.*], Chapter 4: Unit 4.10. URL http://europepmc.org/abstract/MED/18428725.

Chen W., Hasegawa D., Arumuganathan K., Simmons A., Wintermantel W., Fei Z., and Ling K. (**2015**). "Estimation of the Whitefly *Bemisia tabaci* Genome Size Based on k-mer and Flow Cytometric Analyses." *Insects*, 6(3): 704–715. URL http://www.mdpi.com/2075-4450/6/3/704.

Chevreux B., Pfisterer T., Drescher B., Driesel A., Müller W., Wetter T., and Suhai S. (**2004**). "Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs." *Genome Research*, 14(6): 1147–1159. URL http://genome.cshlp.org/content/14/6/1147.abstract.

Chiel E., Gottlieb Y., Zchori-Fein E., Mozes-Daube N., Katzir N., Inbar M., and Ghanim M. (**2007**). "Biotype-dependent secondary symbiont communities in sympatric populations of *Bemisia tabaci*." *Bulletin of Entomological Research*, 97(4): 407–413. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-34547094542&partnerID=40&md5=5221e0e26e759cbbffdcffb2790bd1ee. Cited By 108.

Chougule N. and Bonning B. (**2012**). "Toxins for Transgenic Resistance to Hemipteran Pests." *Toxins*, 4(6): 405–429. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3398418/.

Chowda-Reddy R., Kirankumar M., Seal S., Muniyappa V., Valand G., Govindappa M., and Colvin J. (**2012**). "*Bemisia tabaci* Phylogenetic Groups in India and the Relative Transmission Efficacy of *Tomato leaf curl Bangalore virus* by an Indigenous and an Exotic Population." *Journal of Integrative Agriculture*, 11(2): 235–248. URL http://www.sciencedirect.com/science/article/pii/S2095311912600082.

Chu D., Gao C., Barro P. D., Zhang Y., Wan F., and Khan I. (**2011**). "Further insights into the strange role of bacterial endosymbionts in whitefly, *Bemisia tabaci*: Comparison of secondary symbionts from biotypes B and Q in China." *Bulletin of Entomological Research*, 101: 477–486. URL http://journals.cambridge.org/article_S0007485311000083.

Chu D., Liu G., Wan F., Tao Y., and Gill R. (**2010**). "Phylogenetic analysis and rapid identification of the whitefly, *Bemisia afer*, in China." *Journal of Insect Science*, 10(1). URL //jinsectscience.oxfordjournals.org/content/10/1/86.

Chu D., Pan H., Li X., Guo D., Tao Y., Liu B., and Zhang Y. (**2013**). "Spatial Genetic Heterogeneity in Populations of a Newly Invasive Whitefly in China Revealed by a Nation-Wide Field Survey." *PLoS ONE*, 8(11): e79 997. URL http://dx.doi.org/10.1371%2Fjournal.pone.0079997.

Clancy M. and Hannah L. (2002). "Splicing of the Maize Sh1 First Intron Is Essential for Enhancement of Gene Expression, and a T-Rich Motif Increases Expression without Affecting Splicing." *Plant Physiology*, 130(2): 918–929. URL http://www.plantphysiol.org/content/130/2/918.abstract.

Clary D. and Wolstenholme D. (1985). "The mitochondrial DNA molecule of *Drosophila yakuba*: Nucleotide sequence, gene organization, and genetic code." *Journal of Molecular Evolution*, 22(3): 252–271. URL http://dx.doi.org/10.1007/BF02099755.

Claudianos C., Ranson H., Johnson R., Biswas S., Schuler M., Berenbaum M., Feyereisen R., and Oakeshott J. (2006). "A deficit of detoxification enzymes: Pesticide sensitivity and environmental response in the honeybee." *Insect Molecular Biology*, 15(5): 615–636. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-33750432880&partnerID=40&md5=7382965e9a620f9e7d321509ed3f4113. Cited By 228.

Clayton J., Cripps R., Sparrow J., and Bullard B. (1998). "Interaction of troponin-H and glutathione S-transferase-2 in the indirect flight muscles of *Drosophila melanogaster*." *Journal of Muscle Research Cell Motility*, 19(2): 117–127. URL http://dx.doi.org/10.1023/A:1005304527563.

Comeron J. (2001). "What controls the length of noncoding DNA?" *Current Opinion in Genetics Development*, 11(6): 652 – 659. URL http://www.sciencedirect.com/science/article/pii/S0959437X00002495.

Comeron J. (2004). "Selective and Mutational Patterns Associated With Gene Expression in Humans." *Genetics*, 167(3): 1293–1304. URL http://www.genetics.org/content/167/3/1293.

Concha C., Li F., and Scott M. (2010). "Conservation and sex-specific splicing of the *doublesex* gene in the economically important pest species *Lucilia cuprina*." *Journal of Genetics*, 89(3): 279–285. URL http://dx.doi.org/10.1007/s12041-010-0039-5.

Conte M. and Kocher T. (2015). "An improved genome reference for the African cichlid, *Metriaclima zebra*." *BMC Genomics*, 16(1): 1–13. URL http://dx.doi.org/10.1186/s12864-015-1930-5.

Costa H., Brown J., Sivasupramaniam S., and Bird J. (1993). "Regional distribution, insecticide resistance, and reciprocal crosses between the A and B biotypes of *Bemisia tabaci*." *International Journal of Tropical Insect Science*, 14: 255–266. URL http://journals.cambridge.org/article_S1742758400014703.

Cranston P. and Gullan P. (2003). *In Encyclopedia of Insects*, chapter Phylogeny of insects, pages 882–898. New York: Academic Press/Elsevier Science.

Crowder D., Horowitz A., Barro P. D., Liu S., Showalter A., Kontsedalov S., Khasdan V., Shargal A., Liu J., and Carriere Y. (**2010**). "Mating behaviour, life history and adaptation to insecticides determine species exclusion between whiteflies." *Journal of Animal Ecology*, 79(3): 563–570. URL http://dx.doi.org/10.1111/j.1365-2656.2010.01666.x.

Crowder D., Horowitz A., Tabashnik B., Dennehy T., Denholm I., Gorman K., and Carriere Y. (**2009**). "Analyzing haplodiploid inheritance of insecticide resistance in whitefly biotypes." *Bulletin of Entomological Research*, 99: 307–315.

Crozier R. and Crozier Y. (**1993**). "The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization." *Genetics*, 133(1): 97–117. URL http://www.genetics.org/content/133/1/97.

Cuthbertson A., Blackburn L., Eyre D., Cannon R., Miller J., and Northing P. (**2011**). "*Bemisia tabaci*: The current situation in the UK and the prospect of developing strategies for eradication using entomopathogens." *Insect Science*, 18(1): 1–10. URL http://dx.doi.org/10.1111/j.1744-7917.2010.01383.x.

Cuthbertson A., Walters K., Northing P., and Luo W. (**2007**). "Efficacy of the entomopathogenic nematode, *Steinernema feltiae*, against sweetpotato whitefly *Bemisia tabaci* (Homoptera: Aleyrodidae) under laboratory and glasshouse conditions." *Bulletin of Entomological Research*, 97: 9–14. URL http://journals.cambridge.org/article_S0007485307004701.

Czosnek H. and Brown J. (**2010**). *The Whitefly genome - white paper: proposal to sequence multiple genomes of Bemisia tabaci*. Springer, The Netherlands.

Czosnek H., Ghanim M., and Ghanim M. (**2002**). "The circulative pathway of begomoviruses in the whitefly vector *Bemisia tabaci*Äî insights from studies with Tomato yellow leaf curl virus." *Annals of Applied Biology*, 140(3): 215–231. URL http://dx.doi.org/10.1111/j.1744-7348.2002.tb00175.x.

Czosnek H. and Laterrot H. (**1997**). "A worldwide survey of tomato yellow leaf curl viruses." *Archives of Virology*, 142: 1391–1406.

Dalton R. (**2006**). "Whitefly infestations: The Christmas Invasion." *Nature*, 443: 898–900.

Dasgupta I., Malathi V., and Mukherjee S. (**2003**). "Genetic engineering for virus resistance." *Current Science*, 84(3): 340–354.

de Kloe G., Bailey D., Leurs R., and de Esch I. (**2009**). "Transforming fragments into candidates: small becomes big in medicinal chemistry." *Drug Discovery Today*, 14(13-14): 630–646. URL http://www.sciencedirect.com/science/article/pii/S1359644609001111.

de Koning A., Gu W., Castoe T., Batzer M., and Pollock D. (**2011**). "Repetitive Elements May Comprise Over Two-Thirds of the Human Genome." *PLoS Genetics*, 7(12): 1–12. URL http://dx.doi.org/10.1371%2Fjournal.pgen.1002384.

de Wit E. and de Laat W. (**2012**). "A decade of 3C technologies: insights into nuclear organization." *Genes & Development*, 26(1): 11–24. URL http://genesdev.cshlp.org/content/26/1/11.abstract.

Dedryver C., Ralec A. L., and Fabre F. (**2010**). "The conflicting relationships between aphids and men: A review of aphid damage and control strategies." *Comptes Rendus Biologies*, 333(6-7): 539–553. URL http://www.sciencedirect.com/science/article/pii/S1631069110001150. Les pucerons : mod?es biologiques et ravageurs des cultures.

Degnan P., Yu Y., Sisneros N., Wing R., and Moran N. (**2009**). "*Hamiltonella defensa*, genome evolution of protective bacterial endosymbiont from pathogenic ancestors." *Proceedings of the National Academy of Sciences*, 106(22): 9063–9068. URL http://www.pnas.org/content/106/22/9063.abstract.

Denholm I., Cahill M., Dennehy T., and Horowitz A. (**1998**). "Challenges with managing insecticide resistance in agricultural pests, exemplified by the whitefly *Bemisia tabaci*." *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 353: 1757–1767.

Dennehy T., Williams L., Russell J., Li X., and Wigert M. (**1996**). "Monitoring and Management of Whitefly Resistance to Insecticides in Arizona."

Despres L., David J., and Gallet C. (**2007**). "The evolutionary ecology of insect resistance to plant chemicals." *Trends in Ecology and Evolution*, 22(6): 298–307. URL http://www.sciencedirect.com/science/article/pii/S0169534707000651.

Deutsch M. and Long M. (**1999**). "Intron-exon structures of eukaryotic model organisms." *Nucleic Acids Research*, 27(15): 3219–3228. URL http://nar.oxfordjournals.org/content/27/15/3219.abstract.

Dhadialla T., Le D., Palli S., Raikhel A., and Carlson G. (**2007**). "A photoaffinity, non-steroidal, ecdysone agonist, bisacylhydrazine compound, RH-131039: Characterization of binding and functional activity." *Insect Biochemistry and Molecular Biology*, 37(8): 865–875. URL http://www.sciencedirect.com/science/article/pii/S0965174807001142. Special Issue in Honour of Lynn M. Riddiford.

Ding Y., Ortelli F., Rossiter L., Hemingway J., and Ranson H. (**2003**). "The *Anopheles gambiae* glutathione transferase supergene family: annotation, phylogeny and expression profiles." *BMC Genomics*, 4(1): 1–16. URL http://dx.doi.org/10.1186/1471-2164-4-35.

Dinsdale A., Cook L., Riginos C., Buckley Y., and Barro P. D. (**2010**). "Refined Global Analysis of *Bemisia tabaci* (Hemiptera: Sternorrhyncha: Aleyrodoidea: Aleyrodidae) Mitochondrial Cytochrome Oxidase 1 to Identify Species Level Genetic Boundaries." *Annals of the Entomological Society of America*, 103(2): 196–208. URL http://dx.doi.org/10.1603/AN09061.

Djègbè I., Agossa F., Jones C., Poupardin R., Cornelie S., Akogbéto M., Ranson H., and Corbel V. (**2014**). "Molecular characterization of DDT resistance in *Anopheles gambiae* from Benin." *Parasites Vectors*, 7(1): 1–9. URL http://dx.doi.org/10.1186/1756-3305-7-409.

Dobson S., Marsland E., and Rattanadechakul W. (**2002**). "Mutualistic *Wolbachia* Infection in *Aedes albopictus*: Accelerating Cytoplasmic Drive." *Genetics*, 160(3): 1087–1094. URL http://www.genetics.org/content/160/3/1087.

Dolling W. (**1991**). *The Hemiptera*. Oxford University Press, Incorporated.

Dong Y., Xie M., Jiang Y., Xiao N., Du X., Zhang W., Tosser-Klopp G., Wang J., Yang S., Liang J., Chen W., Chen J., Zeng P., Hou Y., Bian C., Pan S., Li Y., Liu X., Wang W., Servin B., Sayre B., Zhu B., Sweeney D., Moore R., Nie W., Shen Y., Zhao R., Zhang G., Li J., Faraut T., Womack J., Zhang Y., Kijas J., Cockett N., Xu X., Zhao S., Wang J., and Wang W. (**2013**). "Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus)." *Nature Biotechnology*, 31(2): 135–141. URL http://dx.doi.org/10.1038/nbt.2478.

Donnelly A. and Blagg B. (**2008**). "Novobiocin and Additional Inhibitors of the Hsp90 C-Terminal Nucleotide-binding Pocket." *Current medicinal chemistry*, 15(26): 2702–2717.

Dotson E. and Beard C. (**2001**). "Sequence and organization of the mitochondrial genome of the Chagas disease vector, *Triatoma dimidiata*." *Insect Molecular Biology*, 10(3): 205–215. URL http://dx.doi.org/10.1046/j.1365-2583.2001.00258.x.

Doudna J. and Charpentier E. (**2014**). "The new frontier of genome engineering with CRISPR-Cas9." *Science*, 346(6213). URL http://science.sciencemag.org/content/346/6213/1258096.

Douglas A. (**1998**). "Nutritional Interactions in Insect-Microbial Symbioses: Aphids and Their Symbiotic Bacteria *Buchnera*." *Annual Review of Entomology*, 43(1): 17–37. URL http://dx.doi.org/10.1146/annurev.ento.43.1.17. PMID: 15012383.

Drummond A. and Rambaut A. (**2007**). "BEAST: Bayesian evolutionary analysis by sampling trees." *BMC Evolutionary Biology*, 7(1): 1–8. URL http://dx.doi.org/10.1186/1471-2148-7-214.

Duffus J. (**1987**). "Whitefly Transmission of Plant Viruses." In K. Harris, editor, "Current Topics in Vector Research," volume 4 of *Current Topics in Vector Research*, pages 73–91. Springer New York. URL http://dx.doi.org/10.1007/978-1-4612-4712-8_3.

Duffus J. (**1996**). *Whitefly borne viruses. In: Gerling, D., Mayer, R.T. (Eds.), Bemisia: 1995 Taxonomy, Biology, Damage, Control and Management*. Intercept, United Kingdom.

Duncker B., Davies P., and Walker V. (**1997**). "Introns boost transgene expression in *Drosophila melanogaster*." *Molecular and general genetics*, 254(3): 291–296.

Duplouy A., Iturbe-Ormaetxe I., Beatson S., Szubert J., Brownlie J., McMeniman C., McGraw E., Hurst G., Charlat S., O'Neill S., and Woolfit M. (**2013**). "Draft genome sequence of the male-killing *Wolbachia* strain wBol1 reveals recent horizontal gene transfers from diverse sources." *BMC Genomics*, 14(1): 1–13. URL http://dx.doi.org/10.1186/1471-2164-14-20.

Duron O., Bouchon D., Boutin S., Bellamy L., Zhou L., Engelstädter J., and Hurst G. (**2008**). "The diversity of reproductive parasites among arthropods: *Wolbachia* do not walk alone." *BMC Biology*, 6(1): 1–12. URL http://dx.doi.org/10.1186/1741-7007-6-27.

Duron O., Wilkes T., and Hurst G. (**2010**). "Interspecific transmission of a male-killing bacterium on an ecological timescale." *Ecology Letters*, 13(9): 1139–1148. URL http://dx.doi.org/10.1111/j.1461-0248.2010.01502.x.

Earl D., Bradnam K., John J. S., Darling A., Lin D., Faas J., Yu H., Vince B., Zerbino D., Diekhans M., Nguyen N., Nuwantha P., and et al A. S. (**2011**). "Assemblathon 1: A competitive assessment of *de novo* short read assembly methods." *Genome Research*, 21: 2224–2241. URL http://genome.cshlp.org/content/early/2011/09/16/gr.126599.111.abstract.

Edgar R. (**2004**). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic Acids Research*, 32(5): 1792–1797. URL http://nar.oxfordjournals.org/content/32/5/1792.abstract.

Edlund A., Ek K., Breitholtz M., and Gorokhova E. (**2012**). "Antibiotic-Induced Change of Bacterial Communities Associated with the Copepod *Nitocra spinipes*." *PLoS ONE*, 7(3): 1–9. URL http://dx.doi.org/10.1371%2Fjournal.pone.0033107.

Eilbeck K., Moore B., Holt C., and Yandell M. (**2009**). "Quantitative measures for the management and comparison of annotated genomes." *BMC Bioinformatics*, 10(1): 1–15. URL http://dx.doi.org/10.1186/1471-2105-10-67.

Elbaz M., Halon E., Malka O., Malitsky S., Blum E., Aharoni A., and Morin S. (**2012**). "Asymmetric adaptation to indolic and aliphatic glucosinolates in the B and Q sibling species of *Bemisia tabaci* (Hemiptera: Aleyrodidae)." *Molecular Ecology*, 21(18): 4533–4546. URL http://dx.doi.org/10.1111/j.1365-294X.2012.05713.x.

Elbert A., Overbeck H., Iwaya H., and Tsuboi S. (**1990**). "Imidacloprid, a novel systemic nitromethylene analogue insecticide for crop protection."

Ellsworth P. and Martinez-Carrillo J. (**2001**). "IPM for *Bemisia tabaci*: a case study from North America." *Crop Protection*, 20(9): 853–869. URL http://www.sciencedirect.com/science/article/pii/S0261219401001168.

Elsworth B. (**2016**). "SCUBAT (Scaffolding Contigs Using BLAT And Transcripts)." URL https://github.com/elswob/SCUBAT. [Online: accessed 09/02/2016].

Enayati A., Ranson H., and Hemingway J. (**2005**). "Insect glutathione transferases and insecticide resistance." *Insect Molecular Biology*, 14(1): 3–8. URL http://dx.doi.org/10.1111/j.1365-2583.2004.00529.x.

Enright A., Dongen S. V., and Ouzounis C. (**2002**). "An efficient algorithm for large-scale detection of protein families." *Nucleic Acids Research*, 30(7): 1575–1584. URL http://nar.oxfordjournals.org/content/30/7/1575.abstract.

Everett K., Thao M., Horn M., Dyszynski G., and Baumann P. (**2005**). "Novel chlamydiae in whiteflies and scale insects: endosymbionts "*Candidatus* Fritschea bemisiae" strain Falk and "*Candidatus* Fritschea eriococci" strain Elm." *International Journal of Systematic and Evolutionary Microbiology*, 55(4): 1581–1587. URL http://ijs.sgmjournals.org/content/55/4/1581.abstract.

Ewen-Campen B., Shaner N., Panfilio K., Suzuki Y., Roth S., and Extavour C. (**2011**). "The maternal and early embryonic transcriptome of the milkweed bug *Oncopeltus fasciatus*." *BMC Genomics*, 12(1): 61. URL http://www.biomedcentral.com/1471-2164/12/61.

Fahrbach S., Smagghe G., and Velarde R. (**2012**). "Insect Nuclear Receptors." *Annual Review of Entomology*, 57(1): 83–106. URL http://www.annualreviews.org/doi/abs/10.1146/annurev-ento-120710-100607. PMID: 22017307.

Fast E., Toomey M., Panaram K., Desjardins D., Kolaczyk E., and Frydman H. (**2011**). "*Wolbachia* Enhance *Drosophila* Stem Cell Proliferation and Target the Germline Stem Cell Niche." *Science*, 334(6058): 990–992. URL http://science.sciencemag.org/content/334/6058/990.

Fauquet C., Briddon R., Brown J., Moriones E., Stanley J., Zerbini M., and Zhou X. (**2008**). "Geminivirus strain demarcation and nomenclature." *Archives of Virology*, 153(4): 783–821.

Fauquet C. and Stanley J. (**2003**). "Geminivirus classification and nomenclature: progress and problems." *Annals of Applied Biology*, 142(2): 165–189. URL http://dx.doi.org/10.1111/j.1744-7348.2003.tb00241.x.

Faustino N. and Cooper T. (**2003**). "Pre-mRNA splicing and human disease." *Genes & Development*, 17(4): 419–437. URL http://genesdev.cshlp.org/content/17/4/419.short.

Felsenstein J. (**2005**). "PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author."

Feng Y., Wu Q., Wang S., Chang X., Xie W., Xu B., and Zhang Y. (**2010**). "Cross-resistance study and biochemical mechanisms of thiamethoxam resistance in B-biotype *Bemisia tabaci* (Hemiptera: Aleyrodidae)." *Pest Management Science*, 66: 313–318.

Fernández-Medina R., Struchiner C., and Ribeiro J. (**2011**). "Novel transposable elements from *Anopheles gambiae*." *BMC Genomics*, 12(1): 1–18. URL http://dx.doi.org/10.1186/1471-2164-12-260.

Feschotte C. and Pritham E. (**2007**). "DNA transposons and the evolution of eukaryotic genomes." *Annual Review of Genetics*, 41: 331–368. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-37549029474&partnerID=40&md5=2dd7e96adfb8e42680a4228d7f1b1c71. Cited By 403.

Fidock D. (**2010**). "Drug discovery: Priming the antimalarial pipeline." *Nature*, 465(7296): 297–298. URL http://dx.doi.org/10.1038/465297a.

Finn R., Clements J., and Eddy S. (**2011**). "HMMER web server: interactive sequence similarity searching." *Nucleic Acids Research*, 39: 29–37.

Finn R., Mistry J., Tate J., Coggill P., Heger A., Pollington J., Gavin O., Gunasekaran P., Ceric G., Forslund K., Holm L., Sonnhammer E., Eddy S., and Bateman A. (**2010**). "The Pfam protein families database." *Nucleic Acids Research*, 38.

Firdaus S., Vosman B., N. Hidayati N., Supena E. J., Visser R., and van Heusden A. (**2013**). "The *Bemisia tabaci* species complex: Additions from different parts of the world." *Insect Science*, 20(6): 723–733. URL http://dx.doi.org/10.1111/1744-7917.12001.

Fisher S., Barry A., Abreu J., Minie B., Nolan J., Delorey T., Young G., Fennell T., Allen A., Ambrogio L., Berlin A., Blumenstiel B., Cibulskis K., Friedrich D., Johnson R., Juhn F., Reilly B., Shammas R., Stalker J., Sykes S., Thompson J., Walsh J., Zimmer A., Zwirko Z., Gabriel S., Nicol R., and Nusbaum C. (**2011**). "A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries." *Genome Biology*, 12(1): R1–R1. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3091298/.

Fiume E., Christou P., Giani S., and Breviario D. (**2004**). "Introns are key regulatory elements of rice tubulin expression." *Planta*, 218(5): 693–703. URL http://www.jstor.org/stable/23388343.

Florea L., Hartzell G., Zhang Z., Rubin G., and Miller W. (**1998**). "A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence." *Genome Research*, 8(9): 967–974. URL http://genome.cshlp.org/content/8/9/967.abstract.

Fraser M. (**2012**). "Insect Transgenesis: Current Applications and Future Prospects." *Annual Review of Entomology*, 57(1): 267–289. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.ento.54.110807.090545. PMID: 22149266.

Friedman R. (**2011**). "Genomic organization of the glutathione S-transferase family in insects." *Molecular Phylogenetics and Evolution*, 61(3): 924 – 932. URL http://www.sciencedirect.com/science/article/pii/S1055790311003861.

Frohlich D., Torres-Jerez I., Bedford I., Markham P., and Brown J. (**1999**). "A phylogeographical analysis of the *Bemisia tabaci* species complex based on mitochondrial DNA markers." *Molecular Ecology*, 8(10): 1683–1691. URL http://dx.doi.org/10.1046/j.1365-294x.1999.00754.x.

Fu L., Niu B., Zhu Z., Wu S., and Li W. (**2012**). "CD-HIT: accelerated for clustering the next-generation sequencing data." *Bioinformatics*, 28(23): 3150–3152. URL http://bioinformatics.oxfordjournals.org/content/28/23/3150.abstract.

Fujita T. and Nakagawa Y. (**2007**). "QSAR and mode of action studies of insecticidal ecdysone agonists." *SAR & QSAR in environmental research*, 18: 77–88.

Gameel O. (**1972**). "A new description, distribution and hosts of cotton whitefly, *Bemisia tabaci* (Gennadius) (Homoptera: Aleyrodidae)." *Review of Zoology and Botany of Africa*, 86: 50–64.

Gaulton A., Bellis L., Bento A., Chambers J., Davies M., Hersey A., Light Y., McGlinchey S., Michalovich D., Al-Lazikani B., and Overington J. (**2011**). "ChEMBL: a large-scale bioactivity database for drug discovery." *Nucleic Acids Research*, 40.

Gawel N. and Bartlett A. (**1993**). "Characterization of differences between whiteflies using RAPD-PCR." *Insect Molecular Biology*, 2(1): 33–38. URL http://dx.doi.org/10.1111/j.1365-2583.1993.tb00123.x.

Geley S. and Müller C. (**2004**). "RNAi: ancient mechanism with a promising future." *Experimental Gerontology*, 39(7): 985–998. URL http://www.sciencedirect.com/science/article/pii/S0531556504001445.

Gelman D., Blackburn M., and Hu J. (**2005**). "Identification of the molting hormone of the sweet potato (*Bemisia tabaci*) and greenhouse (*Trialeurodes vaporariorum*) whitefly." *Journal of Insect Physiology*, 51: 47–53.

Genest O., Hoskins J., Camberg J., Doyle S., and Wickner S. (**2011**). "Heat shock protein 90 from *Escherichia coli* collaborates with the DnaK chaperone system in client protein remodeling." *Proceedings of the National Academy of Sciences*, 108(20): 8206–8211. URL http://www.pnas.org/content/108/20/8206.abstract.

Gennadius P. (**1889**). *Disease of the tobacco plantations in the Trikonia. The aleurodid of tobacco*. 5. Ellenike Georgia.

Gerling D., Alomar O., and Arnò J. (**2001**). "Biological control of *Bemisia tabaci* using predators and parasitoids." *Crop Protection*, 20(9): 779 – 799. URL http://www.sciencedirect.com/science/article/pii/S0261219401001119.

Gerling D. and Mayer R. (**1996**). *Bemisia: 1995, Taxonomy, Biology, Damage, Control and Management*. Intercept. URL http://books.google.co.uk/books?id=EzQgAQAAMAAJ.

Gerstein M., Lu Z., Nostrand E. V., Cheng C., Arshinoff B., Liu T., Yip K., Robilotto R., Rechtsteiner A., Ikegami K., Alves P., Chateigner A., Perry M., Morris M., Auerbach R., Feng X., Leng J., Vielle A., and et al W. N. (**2010**). "Integrative Analysis of the *Caenorhabditis elegans* Genome by the modENCODE Project." *Science*, 330(6012): 1775–1787. URL http://science.sciencemag.org/content/330/6012/1775.

Ghanim M. and Kontsedalov S. (**2007**). "Gene expression in pyriproxyfen- resistant *Bemisia tabaci* Q biotype." *Pest Management Science*, 63(8): 776–783. URL http://dx.doi.org/10.1002/ps.1410.

Ghanim M. and Kontsedalov S. (**2009**). "Susceptibility to insecticides in the Q biotype of *Bemisia tabaci* is correlated with bacterial symbiont densities." *Pest Management Science*, 65(9): 939–942. URL http://dx.doi.org/10.1002/ps.1795.

Ghanim M., Kontsedalov S., and Czosnek H. (**2007**). "Tissue-specific gene silencing by RNA interference in the whitefly *Bemisia tabaci* (Gennadius)." *Insect Biochemistry and Molecular Biology*, 37(7): 732–738. URL http://www.sciencedirect.com/science/article/pii/S0965174807000884.

Gherna R., Werren J., Weisburg W., Cote R., Woese C., Mandelco L., and Brenner D. (**1991**). "NOTES: Arsenophonus nasoniae gen. nov., sp. nov., the Causative Agent of the Son-Killer Trait in the Parasitic Wasp Nasonia vitripennis." *International Journal of Systematic and Evolutionary Microbiology*, 41(4): 563–565. URL http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-41-4-563.

Gibbons J., Janson E., Hittinger C., Johnston M., Abbot P., and Rokas A. (**2009**). "Benchmarking Next-Generation Transcriptome Sequencing for Functional and Evolutionary Genomics." *Molecular Biology and Evolution*, 26(12): 2731–2744. URL http://mbe.oxfordjournals.org/content/26/12/2731.abstract.

Gilbertson R., Rojas M., and Natwick E. (**2011**). *The Whitefly, Bemisia tabaci (Homoptera: Aleyrodidae) Interaction with Geminivirus-Infected Host Plants: Bemisia tabaci, Host Plants and Geminiviruses*, chapter Development of Integrated Pest Management (IPM) Strategies for Whitefly (*Bemisia tabaci*)-Transmissible Geminiviruses, pages 323–356. Springer Netherlands, Dordrecht. URL http://dx.doi.org/10.1007/978-94-007-1524-0_12.

Glaser R. and Meola M. (**2010**). "The Native *Wolbachia* Endosymbionts of *Drosophila melanogaster* and *Culex quinquefasciatus* Increase Host Resistance to West Nile Virus Infection." *PLoS ONE*, 5(8): 1–11. URL http://dx.doi.org/10.1371%2Fjournal.pone.0011977.

Gorman K., Slater R., Blande J., Clarke A., Wren J., McCaffery A., and Denholm I. (**2010**). "Cross-resistance relationships between neonicotinoids and pymetrozine in *Bemisia tabaci* (Hemiptera: Aleyrodidae)." *Pest Management Science*, 66: 1186–1190.

Gotoh O. (**2008**). "Direct mapping and alignment of protein sequences onto genomic sequence." *Bioinformatics*, 24(21): 2438–2444. URL http://bioinformatics.oxfordjournals.org/content/24/21/2438.abstract.

Gottlieb Y., Ghanim M., Chiel E., Gerling D., Portnoy V., Steinberg S., Tzuri G., Horowitz A., Belausov E., Mozes-Daube N., Kontsedalov S., Gershon M., Gal S., Katzir N., and Zchori-Fein E. (**2006**). "Identification and Localization of a *Rickettsia sp.* in *Bemisia tabaci* (Homoptera: Aleyrodidae)." *Applied and Environmental Microbiology*, 72(5): 3646–3652. URL http://aem.asm.org/content/72/5/3646.abstract.

Gottlieb Y., Ghanim M., Gueguen G., Kontsedalov S., Vavre F., Fleury F., and Zchori-Fein E. (**2008**). "Inherited intracellular ecosystem: symbiotic bacteria share bacteriocytes in whiteflies." *The FASEB Journal*, 22(7): 2591–2599. URL http://www.fasebj.org/content/22/7/2591.abstract.

Gottlieb Y., Zchori-Fein E., Mozes-Daube N., Kontsedalov S., Skaljac M., Brumin M., Sobol I., Czosnek H., Vavre F., Fleury F., and Ghanim M. (**2010**). "The Transmission Efficiency of *Tomato Yellow Leaf Curl Virus* by the Whitefly *Bemisia tabaci* Is Correlated with the Presence of a Specific Symbiotic Bacterium Species." *Journal of Virology*, 84(18): 9310–9317. URL http://jvi.asm.org/content/84/18/9310.abstract.

Götz S., García-Gómez J., Terol J., Williams T., Nagaraj S., Nueda M., Robles M., Talón M., Dopazo J., and Conesa A. (**2008**). "High-throughput functional annotation and data mining with the Blast2GO suite." *Nucleic Acids Research*, 36(10): 3420–3435. URL http://nar.oxfordjournals.org/content/36/10/3420.abstract.

Greathead A. (**1986**). *Host Plants. Chapter 3, pp. 17-25. In: Bemisia tabaci - a literature survey on the cotton whitefly with an annotated bibliography (Ed. M.J.W. Cock).* CAB International Institute of Biological Control, Ascot, United Kingdom.

Griffiths-Jones S., Moxon S., Marshall M., Khanna A., Eddy S., and Bateman A. (**2005**). "Rfam: annotating non-coding RNAs in complete genomes." *Nucleic Acids Research*, 33(suppl 1): D121–D124. URL http://nar.oxfordjournals.org/content/33/suppl_1/D121.abstract.

Grimaldi D. and EngeL M. (**2005**). *Evolution of the insects*. Cambridge University Press, New York, USA.

Gruwell M., Wu J., and Normark B. (2009). "Diversity and Phylogeny of *Cardinium* (Bacteroidetes) in Armored Scale Insects (Hemiptera: Diaspididae)." *Annals of the Entomological Society of America*, 102(6): 1050–1061. URL http://dx.doi.org/10.1603/008.102.0613.

Gueguen G., Vavre F., Gnankine O., Peterschmitt M., Charif D., Chiel E., Gottlieb Y., Zchori-Fein M. G. E., and Fleury F. (2010). "Endosymbiont metacommunities, mtDNA diversity and the evolution of the *Bemisia tabaci* (Hemiptera: Aleyrodidae) species complex." *Molecular Ecology*, 19(19): 4365–4376. URL http://dx.doi.org/10.1111/j.1365-294X.2010.04775.x.

Guengerich F. (2008). "Cytochrome P450 and Chemical Toxicology." *Chemical Research in Toxicology*, 21(1): 70–83. URL http://pubs.acs.org/doi/abs/10.1021/tx700079z.

Guigo R., Agarwal P., Abril J., Burset M., and Fickett J. (2000). "An Assessment of Gene Prediction Accuracy in Large DNA Sequences." *Genome Research*, 10(10): 1631–1642. URL http://genome.cshlp.org/content/10/10/1631.abstract.

Gullan P. and Cranston P. (2010). *The Insects: An Outline of Entomology*. Wiley. URL http://books.google.co.uk/books?id=S7yGZasJ7nEC.

Guo L., Wang S., Wu Q., Zhou X., Xie W., and Zhang Y. (2015). "Flow cytometry and K-mer analysis estimates of the genome sizes of *Bemisia tabaci* B and Q (Hemiptera: Aleyrodidae)." *Frontiers in Physiology*, 6(144). URL http://www.frontiersin.org/integrative_physiology/10.3389/fphys.2015.00144/abstract.

Guo X., Rao Q., Zhang F., Luo C., Zhang H., and Gao X. (2012). "Diversity and Genetic Differentiation of the Whitefly *Bemisia tabaci* Species Complex in China Based on mtCOI and cDNA-AFLP Analysis." *Journal of Integrative Agriculture*, 11(2): 206–214. URL http://www.sciencedirect.com/science/article/pii/S2095311912600057.

Gurevich A., Saveliev V., Vyahhi N., and Tesler G. (2013). "QUAST: quality assessment tool for genome assemblies." *Bioinformatics*, 29(8): 1072–1075. URL http://bioinformatics.oxfordjournals.org/content/29/8/1072.abstract.

Guruprasad N., Mouton L., and Puttaraju H. (2011). "Effect of *Wolbachia* infection and temperature variations on the fecundity of the uzi fly *Exorista sorbillans* (Diptera: Tachinidae)." *Symbiosis*, 54: 151–158.

Haas B., Volfovsky N., Town C., Troukhan M., Alexandrov N., Feldmann K., Flavell R., White O., and Salzberg S. (2002). "Full-length messenger RNA sequences greatly improve genome annotation." *Genome Biology*, 3(6): research0029.1–research0029.12. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC116726/.

Hahn M., Han M., and Han S.-G. (**2007**). "Gene Family Evolution across 12 *Drosophila* Genomes." *PLoS Genetics*, 3(11).

Hale M., McCormick C., Jackson J., and DeWoody J. (**2009**). "Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery." *BMC Genomics*, 10(1): 203. URL http://www.biomedcentral.com/1471-2164/10/203.

Hannon G. (**2002**). "RNA interference." *Nature*, 418(6894): 244–251. URL http://dx.doi.org/10.1038/418244a.

Hansen A., Jeong G., Paine T., and Stouthamer R. (**2007**). "Frequency of Secondary Symbiont Infection in an Invasive Psyllid Relates to Parasitism Pressure on a Geographic Scale in California." *Applied and Environmental Microbiology*, 73(23): 7531–7535. URL http://aem.asm.org/content/73/23/7531.abstract.

Harada T., Nakagawa Y., Ogura T., Yamada Y., Ohe T., and Miyagawa H. (**2011**). "Virtual Screening for Ligands of the Insect Molting Hormone Receptor." *Journal of Chemical Information and Modeling*, 51(2): 296–305. URL http://dx.doi.org/10.1021/ci100400k.

Hayes J., Flanagan J., and Jowsey I. (**2005**). "Glutathione transferases." *Annual Review of Pharmacology and Toxicology*, 45(1): 51–88. URL http://dx.doi.org/10.1146/annurev.pharmtox.45.120403.095857. PMID: 15822171.

He H., Liu Z., Dong B., Zhang J., Shu X., Zhou J., and Ji Y. (**2011**). "Localization of Receptor Site on Insect Sodium Channel for Depressant B-toxin BmK IT2." *PLoS ONE*, 6(1).

He Y., Huang J., Yang X., and Weng Q. (**2007**). "Pyrethroid resistance mechanisms in *Bemisia tabaci* (Gennadius)." *Acta Entomologica Sinica*, 50: 241–247.

Hedges L., Brownlie J., O'Neill S., and Johnson K. (**2008**). "*Wolbachia* and Virus Protection in Insects." *Science*, 322(5902): 702–702. URL http://science.sciencemag.org/content/322/5902/702.

Hemingway J., Field L., and Vontas J. (**2002**). "An Overview of Insecticide Resistance." *Science*, 298(5591): 96–97.

HGSC H. G. S. C. (**2006**). "Insights into social insects from the genome of the honeybee *Apis mellifera*." *Nature*, 443(7114): 931–949.

Hilgenboecker K., Hammerstein P., Schlattmann P., Telschow A., and Werren J. (**2008**). "How many species are infected with *Wolbachia*? –a statistical analysis of current data." *Fems Microbiology Letters*, 281(2): 215–220. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2327208/.

Hilje L., Costa H., and Stansly P. (**2001**). "Cultural practices for managing *Bemisia tabaci* and associated viral diseases." *Crop Protection*, 20(9): 801 – 812. URL http://www.sciencedirect.com/science/article/pii/S0261219401001120.

Hill R., Billas I., Bonneton F., Graham L., and Lawrence M. (**2013**). "Ecdysone Receptors: From the Ashburner Model to Structural Biology." *Annual Review of Entomology*, 58(1): 251–271. URL http://dx.doi.org/10.1146/annurev-ento-120811-153610.

Himler A., Adachi-Hagimori T., Bergen J., Kozuch A., Kelly S., Tabashnik B., Chiel E., Duckworth V., Dennehy T., Zchori-Fein E., and Hunter M. (**2011**). "Rapid Spread of a Bacterial Symbiont in an Invasive Whitefly Is Driven by Fitness Benefits and Female Bias." *Science*, 332(6026): 254–256. URL http://www.sciencemag.org/content/332/6026/254.abstract.

Ho S., So G., and Chow K. (**2001**). "Postembryonic expression of *Caenorhabditis elegans* mab-21 and its requirement in sensory ray differentiation." *Developmental Dynamics*, 221(4): 422–430. URL http://dx.doi.org/10.1002/dvdy.1161.

Hogenhout S., Ammar E.-D., Whitfield A., and Redinbaugh M. (**2008**). "Insect Vector Interactions with Persistently Transmitted Viruses." *Annual Review of Phytopathology*, 46(1): 327–359. URL http://dx.doi.org/10.1146/annurev.phyto.022508.092135.

Holmwood G. and Schindler M. (**2009**). "Protein structure based rational design of ecdysone agonists." *Bioorganic and Medicinal Chemistry*, 17(12): 4064–4070.

Holt C. (**2015**). "[maker-devel] training SNAP with ests and cegma proteins - Google Groups." URL https://groups.google.com/d/msg/maker-devel/wbILWRVQ7r0/K9jngyzMLuEJ. [Online: accessed 04/11/2015].

Holt C. and Yandell M. (**2011**). "MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects." *BMC Bioinformatics*, 12(1): 1–14. URL http://dx.doi.org/10.1186/1471-2105-12-491.

Holt R., Subramanian G., Halpern A., Sutton G., Charlab R., Nusskern D., Wincker P., Clark A., Ribeiro J., Wides R., Salzberg S., Loftus B., Yandell M., Majoros W., Rusch D., Lai Z., Kraft C., Abril J., Anthouard V., Arensburger P., and et al (**2002**). "The genome sequence of the malaria mosquito *Anopheles gambiae*." *Science*, 298(5591): 129–149.

Horowitz A., Gorman K., Ross G., and Denholm I. (**2003**). "Inheritance of pyriproxyfen resistance in the whitefly, *Bemisia tabaci* (Q biotype)." *Archives of Insect Biochemical Physiology*, 54: 177–186.

Horowitz A. and Ishaaya I. (**1996**). *Chemical control of Bemisia management and application*. Intercept Ltd., Andover, Hants, UK.

Horowitz A., Kontsedalov S., Khasdan V., and Ishaaya I. (**2005**). "Biotypes B and Q of *Bemisia tabaci* and their relevance to neonicotinoid and pyriproxyfen resistance." *Archives of Insect Biochemistry and Physiology*, 58: 216–225.

Horowitz A., Toscano N., Youngman R., and Georghiou G. (**1988**). "Synergism of insecticides with DEF in sweetpotato whitefly (Homoptera: Aleyrodidae)." *Journal of Economic Entomology*, 81: 110–114.

Horton A., Wang B., Camp L., Price M., Arshi A., Nagy M., Nadler S., Faeder J., and Luckhart S. (**2011**). "The mitogen-activated protein kinome from *Anopheles gambiae*: identification, phylogeny and functional characterization of the ERK, JNK and p38 MAP kinases." *BMC Genomics*, 12(1): 1–13. URL http://dx.doi.org/10.1186/1471-2164-12-574.

Hosokawa T., Koga R., Kikuchi Y., Meng X., and Fukatsu T. (**2010**). "*Wolbachia* as a bacteriocyte-associated nutritional mutualist." *Proceedings of the National Academy of Sciences*, 107(2): 769–774. URL http://www.pnas.org/content/107/2/769.abstract.

Hsieh C., Ko C., Chung C., and Wang H. (**2014**). "Multilocus approach to clarify species status and the divergence history of the *Bemisia tabaci* (Hemiptera: Aleyrodidae) species complex." *Molecular Phylogenetics and Evolution*, 76: 172–180. URL http://www.sciencedirect.com/science/article/pii/S1055790314001213.

Hu J., Barro P. D., Zhao H., Wang J., Nardi F., and Liu S. (**2011a**). "An Extensive Field Survey Combined with a Phylogenetic Analysis Reveals Rapid and Widespread Invasion of Two Alien Whiteflies in China." *PLoS ONE*, 6(1): e16 061. URL http://dx.doi.org/10.1371%2Fjournal.pone.0016061.

Hu J., Jiang Z., Nardi F., Liu Y., Luo X., Li H., and Zhang Z. (**2014**). "Members of *Bemisia tabaci* (Hemiptera: Aleyrodidae) Cryptic Species and the Status of Two Invasive Alien Species in the Yunnan Province (China)." *Journal of Insect Science*, 14(1). URL //jinsectscience.oxfordjournals.org/content/14/1/281.

Huang H., Yao H., Liu J., Samra A., Kamita S., Cornel A., and Hammock B. (**2012**). "Development of pyrethroid-like fluorescent substrates for glutathione S-transferase." *Analytical Biochemistry*, 431(2): 77 – 83. URL http://www.sciencedirect.com/science/article/pii/S0003269712004587.

Huang X., Adams M., Zhou H., and Kerlavage A. (**1997**). "A Tool for Analyzing and Annotating Genomic Sequences." *Genomics*, 46(1): 37 – 45. URL http://www.sciencedirect.com/science/article/pii/S0888754397949843.

Huang X. and Madan A. (**1999**). "CAP3: A DNA Sequence Assembly Program." *Genome Research*, 9(9): 868–877. URL http://genome.cshlp.org/content/9/9/868.abstract.

Huelsenbeck J., Larget B., and Alfaro M. (2004). "Bayesian Phylogenetic Model Selection Using Reversible Jump Markov Chain Monte Carlo." *Molecular Biology and Evolution*, 21(6): 1123–1133. URL http://mbe.oxfordjournals.org/content/21/6/1123.abstract.

Huger A., Skinner S., and Werren J. (1985). "Bacterial infections associated with the son-killer trait in the parasitoid wasp Nasonia (= Mormoniella) vitripennis (Hymenoptera: Pteromalidae)." *Journal of Invertebrate Pathology*, 46(3): 272 – 280. URL http://www.sciencedirect.com/science/article/pii/0022201185900692.

Hunter W., Reese J., and Consortium I. P. G. (2014). "The Asian Citrus Psyllid Genome (*Diaphorina citri*, Hemiptera)." *Journal of Citrus Pathology*, 1: 1.

Husnik F., Nikoh N., Koga R., Ross L., Duncan R., Fujie M., Tanaka M., Satoh N., Bachtrog D., Wilson A., Dohlen C., Fukatsu T., and McCutcheon J. (2013). "Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis." *Cell*, 153(7): 1567–1578. URL http://www.sciencedirect.com/science/article/pii/S0092867413006466.

Hussain M. and Trehan K. (1933). "Observations on the life-history, bionomics and control of the whitefly of cotton (*Bemisia gossypiperda* M. L.)." *Indian Journal of Agricultural Science*, 3: 701–753.

IAGC I. A. G. C. (2010). "Genome Sequence of the Pea Aphid *Acyrthosiphon pisum*." *PLoS Biology*, 8(2): e1000 313. URL http://dx.doi.org/10.1371%2Fjournal.pbio.1000313.

IHGSC I. H. G. S. C. (2001). "Initial sequencing and analysis of the human genome." *Nature*, 409(6822): 860–921. URL http://dx.doi.org/10.1038/35057062.

Ishmael N., Hotopp J., Ioannidis P., Biber S., Sakamoto J., Siozios S., Nene V., Werren J., Bourtzis K., Bordenstein S., and Tettelin H. (2009). "Extensive genomic diversity of closely related *Wolbachia* strains." *Microbiology*, 155(7): 2211–2222. URL http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.027581-0.

Iturbe-Ormaetxe I., Burke G., Riegler M., and O'Neill S. (2005). "Distribution, Expression, and Motif Variability of Ankyrin Domain Genes in *Wolbachia pipientis*." *Journal of Bacteriology*, 187(15): 5136–5145. URL http://jb.asm.org/content/187/15/5136.abstract.

Janice J., Pande A., Weiner J., Lin C., and Makałowski W. (2012). "U12-type Spliceosomal Introns of Insecta." *International Journal of Biological Sciences*, 8(3): 344–352. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3291851/.

Jiang Z., Xia F., Johnson K., Bartom E., Tuteja J., Stevens R., Grossman R., Brumin M., White K., and Ghanim M. (2012). "Genome Sequences of the Primary Endosymbiont

"*Candidatus* Portiera aleyrodidarum" in the Whitefly *Bemisia tabaci* B and Q Biotypes." *Journal of Bacteriology*, 194(23): 6678–6679. URL http://jb.asm.org/content/194/23/6678.abstract.

Jiang Z., Xia F., Johnson K., Brown C., Bartom E., Tuteja J., Stevens R., Grossman R., Brumin M., White K., and Ghanim M. (**2013**). "Comparison of the Genome Sequences of "*Candidatus* Portiera aleyrodidarum"Primary Endosymbionts of the Whitefly *Bemisia tabaci* B and Q Biotypes." *Applied and Environmental Microbiology*, 79(5): 1757–1759. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3591977/.

Jiu M., Zhou X., and Liu S. (**2006**). "Acquisition and Transmission of two Begomoviruses by the B and a non-B Biotype of *Bemisia tabaci* from Zhejiang, China." *Journal of Phytopathology*, 154(10): 587–591. URL http://dx.doi.org/10.1111/j.1439-0434.2006.01151.x.

Johnson F., Short D., and Castner J. (**2005**). "Sweetpotato/Silverleaf Whitefly Life Stages and Damage. Entomology and Nematology Department document SP90." URL http://edis.ifas.ufl.edu/pdffiles/IN/IN00400.pdf.

Johnson K. (**1999**). "Comparative Detoxification of Plant (Magnolia virginiana) Allelochemicals by Generalist and Specialist Saturniid Silkmoths." *Journal of Chemical Ecology*, 25(2): 253–269. URL http://dx.doi.org/10.1023/A%3A1020890628279.

Jones P., Binns D., Chang H., Fraser M., Li W., McAnulla C., McWilliam H., Maslen J., Mitchell A., Nuka G., Pesseat S., Quinn A., Sangrador-Vegas A., Scheremetjew M., Yong S., Lopez R., and Hunter S. (**2014**). "InterProScan 5: genome-scale protein function classification." *Bioinformatics*, 30: 1236–1240. URL http://bioinformatics.oxfordjournals.org/content/early/2014/01/29/bioinformatics.btu031.abstract.

Juneau K., Miranda M., Hillenmeyer M., Nislow C., and Davis R. (**2006**). "Introns Regulate RNA and Protein Abundance in Yeast." *Genetics*, 174(1): 511–518. URL http://www.genetics.org/content/174/1/511.

Junqueira A., Lessinger A., Torres T., da Silva F., Vettore A., Arruda P., and Espin A. A. (**2004**). "The mitochondrial genome of the blowfly *Chrysomya chloropyga* (Diptera: Calliphoridae)." *Gene*, 339: 7–15. URL http://www.sciencedirect.com/science/article/pii/S0378111904003622.

Jurka J., Kapitonov V., Pavlicek A., Klonowski P., Kohany O., and Walichiewicz J. (**2005**). "Repbase Update, a database of eukaryotic repetitive elements." *Cytogenetic and Genome Research*, 110(1-4): 462–467. URL http://www.karger.com/DOI/10.1159/000084979.

Jurka J. and Pethiyagoda C. (**1995**). "Simple repetitive DNA sequences from primates: Compilation and analysis." *Journal of Molecular Evolution*, 40: 120–126. URL http://dx.doi.org/10.1007/BF00167107.

Kaiser W., Huguet E., Casas J., Commin C., and Giron D. (**2010**). "Plant green-island phenotype induced by leaf-miners is mediated by bacterial symbionts." *Proceedings of the Royal Society of London B: Biological Sciences*, 277(1692): 2311–2319. URL http://rspb.royalsocietypublishing.org/content/early/2010/03/24/rspb.2010.0214.

Kajitani R., Toshimoto K., Noguchi H., Toyoda A., Ogura Y., Okuno M., Yabana M., Harada M., Nagayasu E., Maruyama H., Kohara Y., Fujiyama A., Hayashi T., and Itoh T. (**2014**). "Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads." *Genome Research*, 24(8): 1384–1395. URL http://genome.cshlp.org/content/24/8/1384.abstract.

Kaplan N. and Dekker J. (**2013**). "High-throughput genome scaffolding from in vivo DNA interaction frequency." *Nat Biotech*, 31(12): 1143–1147. URL http://dx.doi.org/10.1038/nbt.2768.

Karatolos N., Pauchet Y., Wilkinson P., Chauhan R., Denholm I., Gorman K., Nelson D., Bass C., Ffrench-Constant R., and Williamson M. (**2011**). "Pyrosequencing the transcriptome of the greenhouse whitefly, *Trialeurodes vaporariorum* reveals multiple transcripts encoding insecticide targets and detoxifying enzymes." *BMC Genomics*, 12(1): 56. URL http://www.biomedcentral.com/1471-2164/12/56.

Karunker I., Benting J., Lueke B., Ponge T., Nauen R., Roditakis E., Vontas J., Gorman K., Denholm I., and Morin S. (**2008**). "Over-expression of cytochrome P450 CYP6CM1 is associated with high resistance to imidacloprid in the B and Q biotypes of *Bemisia tabaci* (Hemiptera: Aleyrodidae)." *Insect Biochemistry and Molecular Biology*, 38(6): 634–644. URL http://www.sciencedirect.com/science/article/pii/S096517480800060X.

Karunker I., Morou E., Nikou D., Nauen R., Sertchook R., Stevenson B., Paine M., Morin S., and Vontas J. (**2009**). "Structural model and functional characterization of the *Bemisia tabaci* CYP6CM1vQ, a cytochrome P450 associated with high levels of imidacloprid resistance." *Insect Biochemistry and Molecular Biology*, 39: 697–706.

Katoh K. and Standley D. (**2013**). "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution*, 30(4): 772–780. URL http://mbe.oxfordjournals.org/content/30/4/772.abstract.

Kaur S., Cogan N., Pembleton L., Shinozuka M., Savin K., Materne M., and Forster J. (**2011**). "Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery." *BMC Genomics*, 12(1): 265. URL http://www.biomedcentral.com/1471-2164/12/265.

Kean J., Rainey S., McFarlane M., Donald C., Schnettler E., Kohl A., and Pondeville E. (**2015**). "Fighting Arbovirus Transmission: Natural and Engineered Control of Vector

Competence in *Aedes Mosquitoes*." *Insects*, 6(1): 236. URL http://www.mdpi.com/2075-4450/6/1/236.

Kelly-Hope L., Ranson H., and Hemingway J. (**2008**). "Lessons from the past: managing insecticide resistance in malaria control and eradication programmes." *The Lancet Infectious Diseases*, 8(6): 387–389. URL http://dx.doi.org/10.1016/S1473-3099(08)70045-8.

Kent W. (**2002**). "BLAT - The BLAST-Like Alignment Tool." *Genome Research*, 12(4): 656–664. URL http://genome.cshlp.org/content/12/4/656.abstract.

Kikuchi Y. (**2009**). "Endosymbiotic bacteria in insects: Their diversity and culturability." *Microbes and Environments*, 24: 195–204.

Kirkness E., Haas B., Sun W., Braig H., Perotti M., Clark J., Lee S., Robertson H., Kennedy R., Elhaik E., Gerlach D., Kriventseva E., Elsik C., Graur D., Hill C., and et al (**2010**). "Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle." *Proceedings of the National Academy of Sciences of the United States of America*, 107(27): 12 168–12 173. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-77955459130&partnerID=40&md5=922c62fdd1d72798b2d1c84f54ec69ca.

Kobayashi K., Ehrlich S., Albertini A., Amati G., Andersen K., Arnaud M., Asai K., Ashikaga S., Aymerich S., and et al P. B. (**2003**). "Essential Bacillus subtilis genes." *Proceedings of the National Academy of Sciences of the United States of America*, 100: 4678–4683.

Kondo N., Shimada M., and Fukatsu T. (**2005**). "Infection density of Wolbachia endosymbiont affected by co-infection and host genotype." *Biology Letters*, 1(4): 488–491. URL http://rsbl.royalsocietypublishing.org/content/1/4/488.abstract.

Kontsedalov S., Zchori-Fein E., Chiel E., Gottlieb Y., Inbar M., and Ghanim M. (**2008**). "The presence of *Rickettsia* is associated with increased susceptibility of *Bemisia tabaci* (Homoptera: Aleyrodidae) to insecticides." *Pest Management Science*, 64(8): 789–792. URL http://dx.doi.org/10.1002/ps.1595.

Koonin E. (**2000**). "How many genes can make a cell: the minimalgene-set concept." *Annual Review of Genomics and Human Genetics*, 1: 99–116.

Koonin E. (**2003**). "Comparative genomics, minimal gene-sets and the last universal common ancestor." *Nature reviews. Microbiology*, 1: 127–136.

Korf I. (**2004**). "Gene finding in novel genomes." *BMC Bioinformatics*, 5(1): 59. URL http://www.biomedcentral.com/1471-2105/5/59.

Kothapalli R., Palli S., Ladd T., Sohi S., Cress D., Dhadialla T., Tzertzinis G., and Retnakaran A. (**1995**). "Cloning and developmental expression of the ecdysone receptor gene from the spruce budworm, *choristoneura fumiferana*." *Developmental Genetics*, 17(4): 319–330. URL http://dx.doi.org/10.1002/dvg.1020170405.

Koutsoukas A., Simms B., Kirchmair J., Bond P., Whitmore A., Zimmer S., Young M., Jenkins J., Glick M., Glen R., and Bender A. (**2011**). "From in silico target prediction to multi-target drug design: Current databases, methods and applications." *Journal of Proteomics*, 74(12): 2554–2574. URL http://www.sciencedirect.com/science/article/pii/S1874391911002028. Pharmacoproteomics and Toxicoproteomics.

Kristensen N. (**1991**). *Chapter 5. Phylogeny of extant Hexapods. In: CSIRO, Division of Entomology. Insects of Australia*. Cornell University Press, Ithaca, USA, 2nd edition.

Kriventseva E., Tegenfeldt F., Petty T., Waterhouse R., Simão F., Pozdnyakov I., Ioannidis P., and Zdobnov E. (**2015**). "OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software." *Nucleic Acids Research*, 43(D1): D250–D256. URL http://nar.oxfordjournals.org/content/43/D1/D250.abstract.

Kuechler S., Dettner K., and Kehl S. (**2011**). "Characterization of an Obligate Intracellular Bacterium in the Midgut Epithelium of the Bulrush Bug Chilacis typhae (Heteroptera, Lygaeidae, Artheneinae)." *Applied and Environmental Microbiology*, 77(9): 2869–2876. URL http://aem.asm.org/content/77/9/2869.abstract.

Kumar S. and Blaxter M. (**2010**). "Comparing *de novo* assemblers for 454 transcriptome data." *BMC Genomics*, 11(1): 571. URL http://www.biomedcentral.com/1471-2164/11/571.

Kurtz S., Phillippy A., Delcher A., Smoot M., Shumway M., Antonescu C., and Salzberg S. (**2004**). "Versatile and open software for comparing large genomes." *Genome Biology*, 5(2): 1–9. URL http://dx.doi.org/10.1186/gb-2004-5-2-r12.

Lagesen K., Hallin P., Rodland E., Staerfeldt H., Rognes T., and Ussery D. (**2007**). "RNAmmer: consistent and rapid annotation of ribosomal RNA genes." *Nucleic Acids Research*, 35(9): 3100–3108. URL http://nar.oxfordjournals.org/content/35/9/3100.abstract.

Lamond A. (**1993**). "The Spliceosome." *BioEssays*, 15(9): 595–603. URL http://dx.doi.org/10.1002/bies.950150905.

Lang B., Gray M., and Burger G. (**1999**). "Mitochondrial Genome Evolution and the Origin of Eukaryotes." *Annual Review of Genetics*, 33(1): 351–397. URL http://dx.doi.org/10.1146/annurev.genet.33.1.351. PMID: 10690412.

Larsen F., Gundersen G., Lopez R., and Prydz H. (**1992**). "CpG islands as gene markers in the human genome." *Genomics*, 13(4): 1095–1107. URL http://www.sciencedirect.com/science/article/pii/088875439290024M.

Lawson D., Arensburger P., Atkinson P., Besansky N., Bruggner R., Butler R., Campbell K., Christophides G., Christley S., Dialynas E., Hammond M., Hill C., Konopinski N., Lobo N., MacCallum R., Madey G., Megy K., Meyer J., Redmond S., Severson D., Stinson E., Topalis P., Birney E., Gelbart W., Kafatos F., Louis C., and Collins F. (**2009**). "VectorBase: a data resource for invertebrate vector genomics." *Nucleic Acids Research*, 37.

Lechner M., Findeiß S., Steiner L., Marz M., Stadler P., and Prohaska S. (**2011**). "Proteinortho: Detection of (Co-)orthologs in large-scale analysis." *BMC Bioinformatics*, 12(1): 1–9. URL http://dx.doi.org/10.1186/1471-2105-12-124.

Lee W., Park J., Lee G., Lee S., and Akimoto S. (**2013**). "Taxonomic Status of the *Bemisia tabaci* Complex (Hemiptera: Aleyrodidae) and Reassessment of the Number of Its Constituent Species." *PLoS ONE*, 8(5): 1–10. URL http://dx.doi.org/10.1371%2Fjournal.pone.0063817.

Leshkowitz D., Gazit S., Reuveni E., Ghanim M., Czosnek H., McKenzie C., Shatters R., and Brown J. (**2006**). "Whitefly (*Bemisia tabaci*) genome project: analysis of sequenced clones from egg, instar, and adult (viruliferous and non-viruliferous) cDNA libraries." *BMC Genomics*, 7: 79.

Letunic I. and Bork P. (**2011**). "Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy." *Nucleic Acids Research*, 39(suppl 2): W475–W478. URL http://nar.oxfordjournals.org/content/39/suppl_2/W475.abstract.

Levine A. and Durbin R. (**2001**). "A computational scan for U12-dependent introns in the human genome sequence." *Nucleic Acids Research*, 29(19): 4006–4013. URL http://nar.oxfordjournals.org/content/29/19/4006.abstract.

Li H. and Durbin R. (**2009**). "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics*, 25(14): 1754–1760. URL http://bioinformatics.oxfordjournals.org/content/25/14/1754.abstract.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., and Subgroup . G. P. D. P. (**2009**). "The Sequence Alignment/Map format and SAMtools." *Bioinformatics*, 25(16): 2078–2079. URL http://bioinformatics.oxfordjournals.org/content/25/16/2078.abstract.

Li L., Jr C. S., and Roos D. (**2003**). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." *Genome Research*, 13(9): 2178–2189.

Li R., Zhu H., Ruan J., Qian W., Fang X., Shi Z., Li Y., Li S., Shan G., Kristiansen K., Li S., Yang H., Wang J., and Wang J. (**2010**). "*De novo* assembly of human genomes with massively parallel short read sequencing." *Genome Research*, 20(2): 265–272. URL http://genome.cshlp.org/content/20/2/265.abstract.

Li S., Xue X., Ahmed M., Ren S., Du Y., Wu J., Cuthbertson A., and Qiu B. (**2011**). "Host plants and natural enemies of *Bemisia tabaci* (Hemiptera: Aleyrodidae) in China." *Insect Science*, 18: 101–120.

Lisch D. and Bennetzen J. (**2011**). "Transposable element origins of epigenetic gene regulation." *Current Opinion in Plant Biology*, 14(2): 156–161. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-79955082486&partnerID=40&md5=8e82a0a24d4de7fb3751486ac867c097.

Lisha V., Antony B., Palaniswami M., and Henneberry T. (**2003**). "*Bemisia tabaci* (Homoptera: Aleyrodidae) Biotypes in India." *Journal of Economic Entomology*, 96(2): 322–327. URL http://jee.oxfordjournals.org/content/96/2/322.

Liu S., Barro P. D., Xu J., Luan J., Zang L., Ruan Y., and Wan F. (**2007**). "Asymmetric Mating Interactions Drive Widespread Invasion and Displacement in a Whitefly." *Science*, 318(5857): 1769–1772. URL http://www.sciencemag.org/content/318/5857/1769.abstract.

Liu S., Colvin J., and Barro P. D. (**2012**). "Species Concepts as Applied to the Whitefly *Bemisia tabaci* Systematics: How Many Species Are There?" *Journal of Integrative Agriculture*, 11(2): 176–186. URL http://www.sciencedirect.com/science/article/pii/S2095311912600021.

Lo N., Casiraghi M., Salati E., Bazzocchi C., and Bandi C. (**2002**). "How Many *Wolbachia* Supergroups Exist?" *Molecular Biology and Evolution*, 19(3): 341–346. URL http://mbe.oxfordjournals.org/content/19/3/341.short.

Loman N., Misra R., Dallman T., Constantinidou C., Gharbia S., Wain J., and Pallen M. (**2012**). "Performance comparison of benchtop high-throughput sequencing platforms." *Nature Biotechnology*, 30: 434–439.

López-Madrigal S., Latorre A., Porcar M., Moya A., and Gil R. (**2013**). "Mealybugs nested endosymbiosis: going into the 'matryoshka'system in *Planococcus citri* in depth." *BMC Microbiology*, 13(1): 1–12. URL http://dx.doi.org/10.1186/1471-2180-13-74.

Lorenzi H., Thiagarajan M., Haas B., Wortman J., Hall N., and Caler E. (**2008**). "Genome wide survey, discovery and evolution of repetitive elements in three Entamoeba species." *BMC Genomics*, 9(1): 1–15. URL http://dx.doi.org/10.1186/1471-2164-9-595.

Lowe T. and Eddy S. (**1997**). "tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence." *Nucleic Acids Research*, 25(5): 955–964. URL http://nar.oxfordjournals.org/content/25/5/955.abstract.

Lu Z. and Wan F. (**2008**). "Differential gene expression in whitefly (*Bemisia tabaci*) B-biotype females and males under heat-shock condition." *Comparative biochemistry and physiology. Part D, Genomics and proteomics*, 3(4): 257–262. URL http://dx.doi.org/10.1016/j.cbd.2008.06.003.

Lukashin A. and Borodovsky M. (**1998**). "GeneMark.hmm: New solutions for gene finding." *Nucleic Acids Research*, 26(4): 1107–1115. URL http://nar.oxfordjournals.org/content/26/4/1107.abstract.

Lumjuan N., McCarroll L., Prapanthadara L., Hemingway J., and Ranson H. (**2005**). "Elevated activity of an Epsilon class glutathione transferase confers DDT resistance in the dengue vector, *Aedes aegypti* White star." *Insect Biochemistry and Molecular Biology*, 35(8): 861–871. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-23844499810&partnerID=40&md5=1e5009dcda0434676f45725be71e7ad7. Cited By 100.

Luo C., Jones C., Devine G., Zhang F., Denholm I., and Gorman K. (**2010**). "Insecticide resistance in *Bemisia tabaci* biotype Q (Hemiptera: Aleyrodidae) from China." *Crop Protection*, 29: 429–434. URL http://www.sciencedirect.com/science/article/pii/S026121940900249X.

Lynch M. and Conery J. (**2003**). "The Origins of Genome Complexity." *Science*, 302(5649): 1401–1404. URL http://science.sciencemag.org/content/302/5649/1401.

Ma D., Gorman K., Devine G., Luo W., and Denholm I. (**2007**). "The biotype and insecticide-resistance status of whiteflies, *Bemisia tabaci* (Hemiptera: Aleyrodidae), invading cropping systems in Xinjiang Uygur Autonomous Region, northwestern China." *Crop Protection*, 26(4): 612–617.

Mahadav A., Gerling D., Gottlieb Y., Czosnek H., and Ghanim M. (**2008**). "Parasitization by the wasp *Eretmocerus mundus* induces transcription of genes related to immune response and symbiotic bacteria proliferation in the whitefly *Bemisia tabaci*." *BMC Genomics*, 9(1): 1–11. URL http://dx.doi.org/10.1186/1471-2164-9-342.

Majoros W., Pertea M., and Salzberg S. (**2004**). "TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders." *Bioinformatics*, 20(16): 2878–2879. URL http://bioinformatics.oxfordjournals.org/content/20/16/2878.abstract.

malERA Consultative Group on Vector Control T. (**2011**). "A Research Agenda for Malaria Eradication: Vector Control." *PLoS Med*, 8(1): 1–8. URL http://dx.doi.org/10.1371%2Fjournal.pmed.1000401.

Malka O., Shekhov A., Reichelt M., Gershenzon J., Vassão D., and Morin S. (**2016**). "Glucosinolate Desulfation by the Phloem-Feeding Insect *Bemisia tabaci*." *Journal of Chemical Ecology*, 42(3): 230–235. URL http://dx.doi.org/10.1007/s10886-016-0675-1.

Mann R., Sidhu J., and Butter N. (**2009**). "Settling preference of the whitefly *Bemisia tabaci* (Hemiptera: Aleyrodidae) on healthy versus cotton leaf curl virus-infected cotton plants." *International Journal of Tropical Insect Science*, 29: 57–61. URL http://journals.cambridge.org/article_S1742758409990142.

Manning B. (**2009**). "Challenges and Opportunities in Defining the Essential Cancer Kinome." *Science Signaling*, 2(63): pe15. URL http://stke.sciencemag.org/cgi/content/abstract/sigtrans;2/63/pe15.

Manning G., Plowman G., Hunter T., and Sudarsanam S. (**2002b**). "Evolution of protein kinase signaling from yeast to man." *Trends in Biochemical Sciences*, 27(10): 514–520.

Manning G., Whyte D., Martinez R., Hunter T., and Sudarsanam S. (**2002a**). "The protein kinase complement of the human genome." *Science*, 298(5600): 1912–1934.

Marçais G. and Kingsford C. (**2011**). "A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers." *Bioinformatics*, 27(6): 764–770. URL http://bioinformatics.oxfordjournals.org/content/27/6/764.abstract.

Marchler-Bauer A., Lu S., Anderson J., Chitsaz F., Derbyshire M., DeWeese-Scott C., Fong J., Geer L., Geer R., Gonzales N., Gwadz M., Hurwitz D., Jackson J., Ke Z., Lanczycki C., Lu F., Marchler G., Mullokandov M., Omelchenko M., Robertson C., Song J., Thanki N., Yamashita R., Zhang D., Zhang N., Zheng C., and Bryant S. (**2011**). "CDD: a Conserved Domain Database for the functional annotation of proteins." *Nucleic Acids Research*, 39(suppl 1): D225–D229. URL http://nar.oxfordjournals.org/content/39/suppl_1/D225.abstract.

Martin J., Mifsud D., and Rapisarda C. (**2000**). "The whiteflies (Hemiptera: Aleyrodidae) of Europe and the Mediterranean basin." *Bulletin of Entomological Research*, 90: 407–448.

Marubayashi J., Kliot A., Yuki V., Rezende J., Krause-Sakate R., Pavan M., and Ghanim M. (**2014**). "Diversity and Localization of Bacterial Endosymbionts from Whitefly Species Collected in Brazil." *PLoS ONE*, 9(9): 1–10. URL http://dx.doi.org/10.1371%2Fjournal.pone.0108363.

Maruthi M., Colvin J., and Seal S. (**2001**). "Mating compatibility, life-history traits, and RAPD-PCR variation in *Bemisia tabaci* associated with the cassava mosaic disease pandemic in East Africa." *Entomologia Experimentalis et Applicata*, 99: 13–23.

Maruthi M., Colvin J., Thwaites R., Banks G., Gibson G., and Seal S. (**2004**). "Reproductive incompatibility and cytochrome oxidase I gene sequence variability amongst host-adapted and geographically separate *Bemisia tabaci* populations (Hemiptera: Aleyrodidae)." *Systematic Entomology*, 29(4): 560–568. URL http://dx.doi.org/10.1111/j.0307-6970.2004.00272.x.

Mascarenhas D., Mettler I., Pierce D., and Lowe H. (**1990**). "Intron-mediated enhancement of heterologous gene expression in maize." *Plant Molecular Biology*, 15(6): 913–920. URL http://dx.doi.org/10.1007/BF00039430.

Masta S. and Boore J. (**2004**). "The Complete Mitochondrial Genome Sequence of the Spider *Habronattus oregonensis* Reveals Rearranged and Extremely Truncated tRNAs." *Molecular Biology and Evolution*, 21(5): 893–902. URL http://mbe.oxfordjournals.org/content/21/5/893.abstract.

Matsuda K., Buckingham S., Kleier D., Rauh J., Grauso M., and Sattelle D. (**2001**). "Neonicotinoids: insecticides acting on insect nicotinic acetylcholine receptors." *Trends in Pharmacological Sciences*, 22: 573–580.

McCutcheon J. and vonDohlen C. (**2011**). "An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs." *Current Biology*, 21(16): 1366–1372. URL http://www.sciencedirect.com/science/article/pii/S096098221100724X.

McLean K., Hans M., and Munro A. (**2012**). "Cholesterol, an essential molecule: diverse roles involving cytochrome P450 enzymes." *Biochemical Society Transactions*, 40(3): 587–593. URL http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=85229128&site=ehost-live.

Miklos G. and Rubin G. (**1996**). "The Role of the Genome Project in Determining Gene Function: Insights from Model Organisms." *Cell*, 86(4): 521–529. URL http://dx.doi.org/10.1016/S0092-8674(00)80126-9.

Min X., Butler G., Storms R., and Tsang A. (**2005**). "OrfPredictor: predicting protein-coding regions in EST-derived sequences." *Nucleic Acids Research*, 33(suppl 2): W677–W680. URL http://nar.oxfordjournals.org/content/33/suppl_2/W677.abstract.

Mohanty A. and Basu A. (**1987**). "Biology of Whitefly Vector, *Bemisia tabaci* (Genn.) on Four Host Plants throughout the Year." *Journal of Entomological Research*, 11: 15–18.

Mohr S., Bakal C., and Perrimon N. (**2010**). "Genomic Screening with RNAi: Results and Challenges." *Annual Review of Biochemistry*, 79(1): 37–64. URL http://dx.doi.org/10.1146/annurev-biochem-060408-092949.

Mongin E., Louis C., Holt R., Birney E., and Collins F. (**2004**). "The *Anopheles gambiae* genome: an update." *Trends in Parasitology*, 20(2): 49–52.

Morales F. (**2006**). "Tropical Whitefly IPM Project." volume 69 of *Advances in Virus Research*, pages 249–311. Academic Press. URL http://www.sciencedirect.com/science/article/pii/S0065352706690064.

Moran N. (**2001**). "The Coevolution of Bacterial Endosymbionts and Phloem-Feeding Insects." *Annals of the Missouri Botanical Garden*, 88: 35–44. URL http://biostor.org/reference/13130.

Moreira L., Iturbe-Ormaetxe I., Jeffery J., Lu G., Pyke A., Hedges L., Rocha B., Hall-Mendelin S., Day A., Riegler M., Hugo L., Johnson K., Kay B., McGraw E., van den Hurk A., Ryan P., and O'Neill S. (**2009**). "A *Wolbachia* Symbiont in *Aedes aegypti* Limits Infection with Dengue, Chikungunya, and Plasmodium." *Cell*, 139(7): 1268–1278. URL http://www.sciencedirect.com/science/article/pii/S0092867409015001.

Moreno-Hagelsieb G. and Latimer K. (**2008**). "Choosing BLAST options for better detection of orthologs as reciprocal best hits." *Bioinformatics*, 24(3): 319–324. URL http://bioinformatics.oxfordjournals.org/content/24/3/319.abstract.

Morin S., Ghanim M., Zeidan M., Czosnek H., Verbeek M., and van den Heuvel J. (**1999**). "A GroEL Homologue from Endosymbiotic Bacteria of the Whitefly *Bemisia tabaci* Is Implicated in the Circulative Transmission of Tomato Yellow Leaf Curl Virus." *Virology*, 256(1): 75 – 84. URL http://www.sciencedirect.com/science/article/pii/S0042682299996319.

Morin S., Williamson M., Goodsons J., Tabashnik B., and Dennehy T. (**2002**). "Mutations in the *Bemisia tabaci* para sodium channel gene associated with resistance to a pyrethroid plus organophosphate mixture." *Insect Biochemistry and Molecular Biology*, 32: 1781–1791.

Moriya Y., Itoh M., Okuda S., Yoshizawa A., and Kanehisa M. (**2007**). "KAAS: an automatic genome annotation and pathway reconstruction server." *Nucleic Acids Research*, 35(suppl 2): W182–W185. URL http://nar.oxfordjournals.org/content/35/suppl_2/W182.abstract.

Mortazavi A., Schwarz E., Williams B., Schaeffer L., Antoshechkin I., Wold B., and Sternberg P. (**2010**). "Scaffolding a Caenorhabditis nematode genome with RNA-seq." *Genome Research*, 20(12): 1740–1747. URL http://genome.cshlp.org/content/20/12/1740.abstract.

Mortazavi A., Williams B., McCue K., Schaeffer L., and Wold B. (**2008**). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nature Methods*, 5(7): 621–628. URL http://dx.doi.org/10.1038/nmeth.1226.

Mound L. and Halsey S. (**1978**). *Whitefly of the World. A Systematic Catalogue of the Aleyrodidae (Homoptera) with Host Plant and Natural Enemy Data*. British Museum of Natural History and John Wiley and Sons, New York.

Mugerwa H., Rey M., Alicai T., Ateka E., Atuncha H., Ndunguru J., and Sseruwagi P. (**2012**). "Genetic diversity and geographic distribution of *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidae) genotypes associated with cassava in East Africa." *Ecology and Evolution*, 2(11): 2749–2762. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3501627/.

Mullins J. (**1993**). "midacloprid: a new nitroguanidine insecticide."

Munhenga G., Masendu H., Brooke B., Hunt R., and Koekemoer L. (**2008**). "Pyrethroid resistance in the major malaria vector *Anopheles arabiensis* from Gwave, a malaria-endemic area in Zimbabwe." *Malaria Journal*, 7: 247.

Munoz-Torres M., Reese J., Childers C., Bennett A., Sundaram J., Childs K., Anzola J., Milshina N., and Elsik C. (**2011**). "Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera." *Nucleic Acids Research*, 39.

Nagai K., Oubridge C., Jessen T., Li J., and Evans P. (**1990**). "Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A." *Nature*, 348(6301): 515–520. URL http://dx.doi.org/10.1038/348515a0.

Nagarajan N. and Pop M. (**2013**). "Sequence assembly demystified." *Nature Reviews Genetics*, 14(3): 157–167. URL http://dx.doi.org/10.1038/nrg3367.

Nakamura Y., Kawai S., Yukuhiro F., Ito S., Gotoh T., Kisimoto R., Yanase T., Matsumoto Y., Kageyama D., and Noda H. (**2009**). "Prevalence of *Cardinium* Bacteria in Planthoppers and Spider Mites and Taxonomic Revision of "*Candidatus* Cardinium hertigii" Based on Detection of a New *Cardinium* Group from Biting Midges." *Applied and Environmental Microbiology*, 75(21): 6757–6763. URL http://aem.asm.org/content/75/21/6757.abstract.

Naranjo S. (**2001**). "Conservation and evaluation of natural enemies in IPM systems for *Bemisia tabaci*." *Crop Protection*, 20(9): 835 – 852. URL http://www.sciencedirect.com/science/article/pii/S0261219401001156.

Naranjo S., Castle S., Barro P., and Liu S. (**2010**). "Population Dynamics, Demography, Dispersal and Spread of *Bemisia tabaci*." In P. A. Stansly and S. E. Naranjo, editors, "*Bemisia*: Bionomics and Management of a Global Pest," pages 185–226. Springer Netherlands. URL http://dx.doi.org/10.1007/978-90-481-2460-2_6.

Naranjo S. and Ellsworth P. (**2009**). "The contribution of conservation biological control to integrated control of *Bemisia tabaci* in cotton." *Biological Control*, 51(3): 458 – 470. URL http://www.sciencedirect.com/science/article/pii/S104996440900214X.

Nardini L., Christian R., Coetzer N., Ranson H., Coetzee M., and Koekemoer L. (**2012**). "Detoxification enzymes associated with insecticide resistance in laboratory strains of *Anopheles arabiensis* of different geographic origin." *Parasites and Vectors*, 5: 113.

Nauen R., Stumpf N., and Elbert A. (**2002**). "Toxicological and mechanistic studies on neonicotinoid cross resistance in Q-type *Bemisia tabaci* (Hemiptera: Aleyrodidae)." *Pest Management Science*, 58(9): 868–875. URL http://dx.doi.org/10.1002/ps.557.

Nawrocki E., Kolbe D., and Eddy S. (**2009**). "Infernal 1.0: inference of RNA alignments." *Bioinformatics*, 25(10): 1335–1337. URL http://bioinformatics.oxfordjournals.org/content/25/10/1335.abstract.

Negrisolo E., Babbucci M., and Patarnello T. (**2011**). "The mitochondrial genome of the ascalaphid owlfly *Libelloides macaronius* and comparative evolutionary mitochondriomics of neuropterid insects." *BMC Genomics*, 12(1): 1–26. URL http://dx.doi.org/10.1186/1471-2164-12-221.

Nene V., Wortman J., Lawson D., Haas B., Kodira C., Tu Z., Loftus B., Xi Z., Megy K., Grabherr M., Ren Q., Zdobnov E., Lobo N., Campbell K., and et al S. B. (**2007**). "Genome Sequence of *Aedes aegypti*, a Major Arbovirus Vector." *Science*, 316(5832): 1718–1723. URL http://science.sciencemag.org/content/316/5832/1718.

New T. (**2011**). *Origins, Distributions and Diversity. In Considerable Variety: Introducing the Diversity of Australias Insects*. Springer, Netherlands.

Ng P. and Kirkness E. (**2010**). "Whole Genome Sequencing." In M. Barnes and G. Breen, editors, "Genetic Variation," volume 628 of *Methods in Molecular Biology*, pages 215–226. Humana Press. URL http://dx.doi.org/10.1007/978-1-60327-367-1_12.

Niefind K. and Issinger O. (**2010**). "Conformational plasticity of the catalytic subunit of protein kinase CK2 and its consequences for regulation and drug design." *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1804(3): 484–492. URL http://www.sciencedirect.com/science/article/pii/S1570963909002702.

Nirgianaki A., Banks G., Frohlich D., Veneti Z., Braig H., Miller T., Bedford I., Markham P., Savakis C., and Bourtzis K. (**2003**). "Wolbachia Infections of the Whitefly *Bemisia tabaci*." *Current Microbiology*, 47(2): 0093–0101. URL http://dx.doi.org/10.1007/s00284-002-3969-1.

Nissen I., Müller M., and Beye M. (**2012**). "The Am-tra2 Gene Is an Essential Regulator of Female Splice Regulation at Two Levels of the Sex Determination Hierarchy of the Honeybee." *Genetics*, 192(3): 1015–1026. URL http://www.genetics.org/content/192/3/1015.

Nováková E., Hypša V., and Moran N. (**2009**). "*Arsenophonus*, an emerging clade of intracellular symbionts with a broad host distribution." *BMC Microbiology*, 9(1): 1–14. URL http://dx.doi.org/10.1186/1471-2180-9-143.

Nybakken K., Vokes S., Lin T., McMahon A., and Perrimon N. (**2005**). "A genome-wide RNA interference screen in *Drosophila melanogaster* cells for new components of the Hh signaling pathway." *Nature Genetics*, 37(12): 1323–1332.

Oakeshott J., Horne I., Sutherland T., and Russell R. (**2003**). "The genomics of insecticide resistance." *Genome Biology*, 4: 202.

Oakeshott J., Johnson R., Berenbaum M., Ranson H., Cristino A., and Claudianos C. (**2010**). "Metabolic enzymes associated with xenobiotic and chemosensory responses in *Nasonia vitripennis*." *Insect Molecular Biology*, 19: 147–163. URL http://dx.doi.org/10.1111/j.1365-2583.2009.00961.x.

Oh S., Kingsley T., Shin H., Zheng Z., Chen H., Chen X., Wang H., Ruan P., Moody M., and Hou S. (**2003**). "A P-Element Insertion Screen Identified Mutations in 455 Novel Essential Genes in *Drosophila*." *Genetics*, 163(1): 195–201. URL http://www.genetics.org/content/163/1/195.

Oliveira M., Barau J., Junqueira A., Feijao P., Rosa A., Abreu C., Azeredo-Espin A., and Lessinger A. (**2008**). "Structure and evolution of the mitochondrial genomes of *Haematobia irritans* and *Stomoxys calcitrans*: The Muscidae (Diptera: Calyptratae) perspective." *Molecular Phylogenetics and Evolution*, 48(3): 850–857. URL http://www.sciencedirect.com/science/article/pii/S1055790308002674.

Oliveira M., Henneberry T., and Anderson P. (**2001**). "History, current status, and collaborative research projects for *Bemisia tabaci*." *Crop Protection*, 20: 709–723.

Oliver K., Russell J., Moran N., and Hunter M. (**2003**). "Facultative bacterial symbionts in aphids confer resistance to parasitic wasps." *Proceedings of the National Academy of Sciences of the United States of America*, 100: 1803–1807.

O'Neil S., Dzurisin J., Carmichael R., Lobo N., Emrich S., and Hellmann J. (**2010**). "Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*." *BMC Genomics*, 11(1): 1–15. URL http://dx.doi.org/10.1186/1471-2164-11-310.

Osawa T., Ito K., Inanaga H., Nureki O., Tomita K., and Numata T. (**2009**). "Conserved Cysteine Residues of GidA Are Essential for Biogenesis of 5-Carboxymethylaminomethyluridine at tRNA Anticodon." *Structure*, 17(5): 713 – 724. URL http://www.sciencedirect.com/science/article/pii/S0969212609001609.

Ostberg T., Jacobsson M., Attersand A., de Urquiza A. M., and Jendeberg L. (**2003**). "A Triple Mutant of the Drosophila ERR Confers Ligand-Induced Suppression of Activity." *Biochemistry*, 42(21): 6427–6435. URL http://dx.doi.org/10.1021/bi027279b.

Östlund G., Schmitt T., Forslund K., Köstler T., Messina D., Roopra S., Frings O., and Sonnhammer E. (**2010**). "InParanoid 7: new algorithms and tools for eukaryotic orthology analysis." *Nucleic Acids Research*, 38(Database issue): D196–D203. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2808972/.

Overbeek R., Olson R., Pusch G., Olsen G., Davis J., Disz T., Edwards R., Gerdes S., Parrello B., Shukla M., Vonstein V., Wattam A., Xia F., and Stevens R. (**2014**). "The SEED and the

Rapid Annotation of microbial genomes using Subsystems Technology (RAST)." *Nucleic Acids Research*, 42(D1): D206–D214. URL http://nar.oxfordjournals.org/content/42/D1/D206.abstract.

Owor B., Legg J., Okao-Okuja G., Obonyo R., and Ogenga-Latigo M. (**2004**). "The effect of cassava mosaic geminiviruses on symptom severity, growth and root yield of a cassava mosaic virus disease-susceptible cultivar in Uganda." *Annals of Applied Biology*, 145(3): 331–337. URL http://dx.doi.org/10.1111/j.1744-7348.2004.tb00390.x.

Palumbo J., Horowitz A., and Prabhaker N. (**2001**). "Insecticidal control and resistance management for *Bemisia tabaci*." *Crop Protection*, 20(9): 739 – 765. URL http://www.sciencedirect.com/science/article/pii/S026121940100117X.

Pan H., Chu D., Liu B., Xie W., Wang S., Wu Q., Xu B., and Zhang Y. (**2013**). "Relative Amount of Symbionts in Insect Hosts Changes with Host-Plant Adaptation and Insecticide Resistance." *Environmental Entomology*, 42(1): 74–78. URL http://www.bioone.org/doi/abs/10.1603/EN12114.

Pan H., Li X., Ge D., Wang S., Wu Q., Xie W., Jiao X., Chu D., Liu B., Xu B., and Zhang Y. (**2012**). "Factors Affecting Population Dynamics of Maternally Transmitted Endosymbionts in *Bemisia tabaci*." *PLoS ONE*, 7(2): e30 760. URL http://dx.doi.org/10.1371%2Fjournal.pone.0030760.

Pan X., Lührmann A., Satoh A., Laskowski-Arce M., and Roy C. (**2008**). "Ankyrin Repeat Proteins Comprise a Diverse Family of Bacterial Type IV Effectors." *Science*, 320(5883): 1651–1654. URL http://science.sciencemag.org/content/320/5883/1651.

Pang Y. (**2006**). "Novel Acetylcholinesterase Target Site for Malaria Mosquito Control." *PLoS ONE*, 1(1).

Papafotiou G., Oehler S., Savakis C., and Bourtzis K. (**2011**). "Regulation of *Wolbachia* ankyrin domain encoding genes in *Drosophila* gonads." *Research in Microbiology*, 162(8): 764 – 772. URL http://www.sciencedirect.com/science/article/pii/S0923250811001173.

Parchman T., Geist K., Grahnen J., Benkman C., and Buerkle C. (**2010**). "Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery." *BMC Genomics*, 11(1): 180. URL http://www.biomedcentral.com/1471-2164/11/180.

Parkinson J. and Blaxter M. (**2004**). "Expressed sequence tags: analysis and annotation." *Methods in Molecular Biology*, 270: 93–126. URL http://dx.doi.org/10.1385/1-59259-793-9:093.

Parra G., Bradnam K., and Korf I. (**2007**). "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes." *Bioinformatics*, 23(9): 1061–1067. URL http://bioinformatics.oxfordjournals.org/content/23/9/1061.abstract.

Parra G., Bradnam K., Ning Z., Keane T., and Korf I. (**2009**). "Assessing the gene space in draft genomes." *Nucleic Acids Research*, 37(1): 289–297. URL http://nar.oxfordjournals.org/content/37/1/289.abstract.

Parrella G., Scassillo L., and Giorgini M. (**2012**). "Evidence for a new genetic variant in the *Bemisia tabaci* species complex and the prevalence of the biotype Q in southern Italy." *Journal of Pest Science*, 85(2): 227–238. URL http://dx.doi.org/10.1007/s10340-012-0417-2.

Pavy N., Rombauts S., Dehais P., Mathe C., Ramana D., Leroy P., and Rouze P. (**1999**). "Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences." *Bioinformatics*, 15(11): 887–899. URL http://bioinformatics.oxfordjournals.org/content/15/11/887.abstract.

Pendleton M., Sebra R., Pang A., Ummat A., Franzen O., Rausch T., Stutz A., Stedman W., Anantharaman T., Hastie A., Dai H., Fritz M., Cao H., Cohain A., Deikus G., Durrett R., Blanchard S., Altman R., Chin C., Guo Y., Paxinos E., Korbel J., Darnell R., McCombie W., Kwok P., Mason C., Schadt E., and Bashir A. (**2015**). "Assembly and diploid architecture of an individual human genome via single-molecule technologies." *Nature Methods*, 12(8): 780–786. URL http://dx.doi.org/10.1038/nmeth.3454.

Pérez-Brocal V., Gil R., Ramos S., Lamelas A., Postigo M., Michelena J., Silva F., Moya A., and Latorre A. (**2006**). "A Small Microbial Genome: The End of a Long Symbiotic Relationship?" *Science*, 314(5797): 312–313. URL http://science.sciencemag.org/content/314/5797/312.

Permpoon R., Aketarawong N., and Thanaphum S. (**2011**). "Isolation and characterization of *Doublesex* homologues in the Bactrocera species: *B. dorsalis* (Hendel) and *B. correcta* (Bezzi) and their putative promoter regulatory regions." *Genetica*, 139(1): 113–127. URL http://dx.doi.org/10.1007/s10709-010-9508-2.

Perring T. (**2001**). "The *Bemisia tabaci* species complex." *Crop Protection*, 20: 725–737.

Perry T., Batterham P., and Daborn P. (**2011**). "The biology of insecticidal activity and resistance." *Insect Biochemistry and Molecular Biology*, 41(7): 411–422. URL http://www.sciencedirect.com/science/article/pii/S0965174811000622.

Pertea G., Huang X., Liang F., Antonescu V., Sultana R., Karamycheva S., Lee Y., White J., Cheung F., Parvizi B., Tsai J., and Quackenbush J. (**2003**). "TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets." *Bioinformatics*, 19(5): 651–652. URL http://bioinformatics.oxfordjournals.org/content/19/5/651.abstract.

Pimentel D., Zuniga R., and Morrison D. (**2005**). "Update on the environmental and economic costs associated with alien-invasive species in the United States." *Ecological Economics*,

52(3): 273 – 288. URL http://www.sciencedirect.com/science/article/pii/S0921800904003027. Integrating Ecology and Economics in Control BioinvasionsIEECB S.I.

Prabhaker N., Toscano N., and Henneberry T. (1998). "Evaluation of insecticide rotations and mixtures as resistance management strategies for *Bemisia argentifolii* (Homoptera: Aleyrodidae)." *Journal of Economic Entomology*, 91.

Price A., Jones N., and Pevzner P. (2005). "*De novo* identification of repeat families in large genomes." *Bioinformatics*, 21(suppl 1): i351–i358. URL http://bioinformatics.oxfordjournals.org/content/21/suppl_1/i351.abstract.

Price D. and Gatehouse J. (2008). "RNAi-mediated crop protection against insects." *Trends in Biotechnology*, 26(7): 393–400. URL http://dx.doi.org/10.1016/j.tibtech.2008.04.004.

Price M., Dehal P., and Arkin A. (2010). "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments." *PLoS ONE*, 5(3): 1–10. URL http://dx.doi.org/10.1371%2Fjournal.pone.0009490.

Pruitt K., Tatusova T., Klimke W., and Maglott D. (2009). "NCBI Reference Sequences: current status, policy and new initiatives." *Nucleic Acids Research*, 37: D32–D36. URL http://nar.oxfordjournals.org/content/37/suppl_1/D32.abstract.

Punta M., Coggill P., Eberhardt R., Mistry J., Tate J., Boursnell C., Pang N., Forslund K., Ceric G., Clements J., Heger A., Holm L., Sonnhammer E., Eddy S., Bateman A., and Finn R. (2012). "The Pfam protein families database." *Nucleic Acids Research*, 40.

Putnam N., OĆonnell B., Stites J., Rice B., Blanchette M., Calef R., Troll C., Fields A., Hartley P., Sugnet C., Haussler D., Rokhsar D., and Green R. (2016). "Chromosome-scale shotgun assembly using an in vitro method for long-range linkage." *Genome Research*. URL http://genome.cshlp.org/content/early/2016/02/08/gr.193474.115.abstract.

Quaintance A. (1900). *Contribution towards a monographof the American Aleurodidae*. 8. Washington: Government Printing Office.

Quaintance A. and Baker A. (1914). "Classification of the aleyrodidae part II." *Technical Series, Bureau of Entomology, United States Department of Agriculture*, 27: 95–109. Cited By (since 1996)1.

Rabello A., Queiroz P., oes K. S., Hiragi C., Lima L., Oliveira M., and Mehta A. (2008). "Diversity analysis of *Bemisia tabaci* biotypes: RAPD, PCR-RFLP and sequencing of the ITS1 rDNA region." *Genetics and Molecular Biology*, 31: 585–590. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-47572008000300029&nrm=iso.

Ramsey J., Rider D., Walsh T., Vos M., Gordon K., Ponnala L., MacMil S., Roe B., and Jander G. (**2010**). "Comparative analysis of detoxification enzymes in *Acyrthosiphon pisum* and *Myzus persicae*." *Insect Molecular Biology*, 19(SUPPL. 2): 155–164. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-77953231229&partnerID=40&md5=9ca0b786d06adb9950a6941d95e4f67e. Cited By 71.

Rana V., Singh S., Priya N., Kumar J., and Rajagopal R. (**2012**). "*Arsenophonus* GroEL Interacts with CLCuV and Is Localized in Midgut and Salivary Gland of Whitefly *B. tabaci*." *PLoS ONE*, 7(8): e42 168. URL http://dx.doi.org/10.1371%2Fjournal.pone.0042168.

Ranson H., Abdallah H., Badolo A., Guelbeogo W., Kerah-Hinzoumbé C., Yangalbé-Kalnoné E., Sagnon N., Simard F., and Coetzee M. (**2009**). "Insecticide resistance in *Anopheles gambiae*: data from the first year of a multi-country study highlight the extent of the problem." *Malaria Journal*, 8(1): 1–12. URL http://dx.doi.org/10.1186/1475-2875-8-299.

Ranson H., Jensen B., Wang X., Prapanthadara L., Hemingway J., and Collins F. (**2000**). "Genetic mapping of two loci affecting DDT resistance in the malaria vector *Anopheles gambiae*." *Insect Molecular Biology*, 9(5): 499–507. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0033760462&partnerID=40&md5=db9f2e705182e59713415f50b9d5f2d7. Cited By 77.

Rao Q., Rollat-Farnier P., Zhu D., Santos-Garcia D., Silva F., Moya A., Latorre A., Klein C., Vavre F., Sagot M., Liu S., Mouton L., and Wang X. (**2015**). "Genome reduction and potential metabolic complementation of the dual endosymbionts in the whitefly *Bemisia tabaci*." *BMC Genomics*, 16(1): 1–13. URL http://dx.doi.org/10.1186/s12864-015-1379-6.

Rao Q., Wang S., Su Y., Bing X., Liu S., and Wang X. (**2012a**). "Draft Genome Sequence of "*Candidatus* Hamiltonella defensa, an Endosymbiont of the Whitefly *Bemisia tabaci*." *Journal of Bacteriology*, 194(13): 3558. URL http://jb.asm.org/content/194/13/3558.abstract.

Rao Q., Wang S., Zhu D., Wang X., and Liu S. (**2012b**). "Draft Genome Sequence of *Rickettsia sp.* Strain MEAM1, Isolated from the Whitefly *Bemisia tabaci*." *Journal of Bacteriology*, 194(17): 4741–4742. URL http://jb.asm.org/content/194/17/4741.abstract.

Rathe S., Moriarity B., Stoltenberg C., Kurata M., Aumann N., Rahrmann E., Bailey N., Melrose E., Beckmann D., Liska C., and Largaespada D. (**2014**). "Using RNA-seq and targeted nucleases to identify mechanisms of drug resistance in acute myeloid leukemia." *Scientific Reports*, 4: 6048. URL http://dx.doi.org/10.1038/srep06048.

Rauch N. and Nauen R. (**2003**). "Identification of biochemical markers linked to neonicotinoid cross resistance in *Bemisia tabaci* (Hemiptera: Aleyrodidae)." *Archives of Insect Biochemistry and Physiology*, 54: 165–176.

Reeve A. and Lightowlers R. (**2012**). *Mitochondrial Dysfunction in Neurodegenerative Disorders*, chapter An Introduction to Mitochondria, pages 3–18. Springer London, London. URL http://dx.doi.org/10.1007/978-0-85729-701-3_1.

Rekha A., Maruthi M., Muniyappa V., and Colvin J. (**2005**). "Occurrence of three genotypic clusters of *Bemisia tabaci* and the rapid spread of the B biotype in south India." *Entomologia Experimentalis et Applicata*, 117(3): 221–233. URL http://dx.doi.org/10.1111/j.1570-7458.2005.00352.x.

Rhoads A. and Au K. (**2015**). "PacBio Sequencing and Its Applications." *Genomics, Proteomics & Bioinformatics*, 13(5): 278 – 289. URL http://www.sciencedirect.com/science/article/pii/S1672022915001345.

Roberts A., Pimentel H., Trapnell C., and Pachter L. (**2011**). "Identification of novel transcripts in annotated genomes using RNA-Seq." *Bioinformatics*, 27(17): 2325–2329. URL http://bioinformatics.oxfordjournals.org/content/27/17/2325.abstract.

Robertson H., Warr C., and Carlson J. (**2003**). "Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*." *Proceedings of the National Academy of Sciences*, 100(suppl 2): 14 537–14 542. URL http://www.pnas.org/content/100/suppl_2/14537.abstract.

Robinson G., Hackett K., Purcell-Miramontes M., Brown S., Evans J., Goldsmith M., Lawson D., Okamuro J., Robertson H., and Schneider D. (**2011**). "Creating a Buzz About Insect Genomes." *Science*, 331(6023): 1386–1386. URL http://science.sciencemag.org/content/331/6023/1386.

Robinson M., McCarthy D., and Smyth G. (**2010**). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, 26(1): 139–140. URL http://bioinformatics.oxfordjournals.org/content/26/1/139.abstract.

Roditakis E., Grispou M., Morou E., Kristoffersen J., Roditakis N., Nauen R., Vontas J., and Tsagkarakou A. (**2009**). "Current status of insecticide resistance in Q biotype *Bemisia tabaci* populations from Crete." *Pest Management Science*, 65: 313–322.

Rogers M., Jani M., and Vogt R. (**1999**). "An olfactory-specific glutathione-S-transferase in the sphinx moth *Manduca sexta*." *Journal of Experimental Biology*, 202(12): 1625–1637. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0033153407&partnerID=40&md5=6db00719eb8d1ed296655f54d9dc4d46. Cited By 74.

Rokas A., Williams B., King N., and Carroll S. (**2003**). "Genome-scale approaches to resolving incongruence in molecular phylogenies." *Nature*, 425(6960): 798–804. URL http://dx.doi.org/10.1038/nature02053.

Ronquist F. and Huelsenbeck J. (**2003**). "MrBayes 3: Bayesian phylogenetic inference under mixed models." *Bioinformatics*, 19(12): 1572–1574. URL http://bioinformatics.oxfordjournals.org/content/19/12/1572.abstract.

Rose A. (**2008**). *Nuclear pre-mRNA Processing in Plants*, chapter Intron-Mediated Regulation of Gene Expression, pages 277–290. Springer Berlin Heidelberg, Berlin, Heidelberg. URL http://dx.doi.org/10.1007/978-3-540-76776-3_15.

Russell J. and Moran N. (**2006**). "Costs and benefits of symbiont infection in aphids: variation among symbionts and across temperatures." *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1586): 603–610. URL http://rspb.royalsocietypublishing.org/content/273/1586/603.

Saccone G., Salvemini M., Pane A., and Polito L. (**2008**). "Masculinization of XX *Drosophila* transgenic flies expressing the Ceratitis capitata DoublesexM isoform." *The International journal of developmental biology*, 52(8): 1051–1057.

Saitou N. and Nei M. (**1987**). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution*, 4(4): 406–425. URL http://mbe.oxfordjournals.org/content/4/4/406.abstract.

Salamov A. and Solovyev V. (**2000**). "Ab initio Gene Finding in Drosophila Genomic DNA." *Genome Research*, 10(4): 516–522. URL http://genome.cshlp.org/content/10/4/516.abstract.

Saldanha R., Mohr G., Belfort M., and Lambowitz A. (**1993**). "Group I and group II introns." *The FASEB Journal*, 7(1): 15–24. URL http://www.fasebj.org/content/7/1/15.abstract.

Sander J. and Joung J. (**2014**). "CRISPR-Cas systems for editing, regulating and targeting genomes." *Nat Biotech*, 32(4): 347–355. URL http://dx.doi.org/10.1038/nbt.2842.

Santos-Garcia D., Farnier P., Beitia F., Zchori-Fein E., Vavre F., Mouton L., Moya A., Latorre A., and Silva F. (**2012**). "Complete Genome Sequence of "*Candidatus* Portiera aleyrodidarum" BT-QVLC, an Obligate Symbiont That Supplies Amino Acids and Carotenoids to *Bemisia tabaci*." *Journal of Bacteriology*, 194(23): 6654–6655. URL http://jb.asm.org/content/194/23/6654.abstract.

Santos-Garcia D., Latorre A., Moya A., Gibbs G., Hartung V., Dettner K., Kuechler S., and Silva F. (**2014b**). "Small but powerful, the primary endosymbiont of moss bugs, *Candidatus* Evansia muelleri, holds a reduced genome with large biosynthetic capabilities." *Genome Biology and Evolution*, 6: 1875–1893. URL http://gbe.oxfordjournals.org/content/early/2014/07/10/gbe.evu149.abstract.

Santos-Garcia D., Rollat-Farnier P., Beitia F., Zchori-Fein E., Vavre F., Mouton L., Moya A., Latorre A., and Silva F. (**2014a**). "The Genome of *Cardinium* cBtQ1 Provides Insights into Genome Reduction, Symbiont Motility, and Its Settlement in *Bemisia tabaci*." *Genome Biology and Evolution*, 6(4): 1013–1030. URL http://gbe.oxfordjournals.org/content/6/4/1013.abstract.

Schnable P., Ware D., Fulton R., Stein J., Wei F., Pasternak S., Liang C., Zhang J., Fulton L., Graves T., Minx P., Reily A., Courtney L., Kruchowski S., Tomlinson C., Strong C., Delehaunty K., Fronick C., Courtney B., Rock S., Belter E., Du F., Kim K., and et al (**2009**). "The B73 maize genome: Complexity, diversity, and dynamics." *Science*, 326(5956): 1112–1115. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-70450202132&partnerID=40&md5=c429257f62279b271a0d03bc9ede1a30. Cited By 1456.

Schuh R. and Slater J. (**1995**). *rue bugs of the world (Hemiptera: Heteroptera). Classification and natural history*. Cornell University Press, Ithaca, USA.

Schuler M. (**2011**). "P450s in plant?insect interactions." *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1814(1): 36–45. URL http://www.sciencedirect.com/science/article/pii/S157096391000261X. Cytochrome P450: Structure, biodiversity and potential for application.

Seal S., Patel M., Collins C., Colvin J., and Bailey D. (**2012**). "Next Generation Transcriptome Sequencing and Quantitative Real-Time PCR Technologies for Characterisation of the *Bemisia tabaci* Asia 1 mtCOI Phylogenetic Clade." *Journal of Integrative Agriculture*, 11(2): 281 – 292. URL http://www.sciencedirect.com/science/article/pii/S2095311912600124.

Seal S., vandenBosch F., and Jeger M. (**2006**). "Factors Influencing Begomovirus Evolution and Their Increasing Global Significance: Implications for Sustainable Control." *Critical Reviews in Plant Sciences*, 25(1): 23–46. URL http://dx.doi.org/10.1080/07352680500365257.

Seki M., Narusaka M., Kamiya A., Ishida J., Satou M., Sakurai T., Nakajima M., Enju A., Akiyama K., Oono Y., Muramatsu M., Hayashizaki Y., Kawai J., Carninci P., Itoh M., Ishii Y., Arakawa T., Shibata K., Shinagawa A., and Shinozaki K. (**2002**). "Functional Annotation of a Full-Length *Arabidopsis* cDNA Collection." *Science*, 296(5565): 141–145. URL http://science.sciencemag.org/content/296/5565/141.

Shankarappa K., Rangaswamy K., Narayana D. A., Rekha A., Raghavendra N., Reddy C. L., Chancellor T., and Maruthi M. (**2007**). "Development of silverleaf assay, protein and nucleic acid-based diagnostic techniques for the quick and reliable detection and monitoring of biotype B of the whitefly, *Bemisia tabaci* (Gennadius)." *Bulletin of Entomological Research*, 97(5): 503–513. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-34948863840&partnerID=40&md5=52217447323eb3d00ba2354bd78459b2.

Shao R. and Barker S. (**2003**). "The Highly Rearranged Mitochondrial Genome of the Plague Thrips, *Thrips imaginis* (Insecta: Thysanoptera): Convergence of Two Novel Gene Boundaries

and an Extraordinary Arrangement of rRNA Genes." *Molecular Biology and Evolution*, 20(3): 362–370. URL http://mbe.oxfordjournals.org/content/20/3/362.abstract.

Shao R., Campbell N., and Barker S. (**2001a**). "Numerous Gene Rearrangements in the Mitochondrial Genome of the Wallaby Louse, *Heterodoxus macropus* (Phthiraptera)." *Molecular Biology and Evolution*, 18(5): 858–865. URL http://mbe.oxfordjournals.org/content/18/5/858.abstract.

Shao R., Campbell N., Schmidt E., and Barker S. (**2001b**). "Increased Rate of Gene Rearrangement in the Mitochondrial Genomes of Three Orders of Hemipteroid Insects." *Molecular Biology and Evolution*, 18(9): 1828–1832. URL http://mbe.oxfordjournals.org/content/18/9/1828.short.

Sharp P. and Burge C. (**1997**). "Classification of Introns: U2-Type or U12-Type." *Cell*, 91(7): 875 – 879. URL http://www.sciencedirect.com/science/article/pii/S0092867400804791.

Shatters R., Powell C., Boykin L., Liansheng H., and McKenzie C. (**2009**). "Improved DNA barcoding method for *Bemisia tabaci* and related Aleyrodidae: development of universal and *Bemisia tabaci* biotype-specific mitochondrial cytochrome c oxidase I polymerase chain reaction primers." *Journal of economic entomology*, 102(2): 750–758. URL http://www.biomedsearch.com/nih/Improved-DNA-barcoding-method-Bemisia/19449657.html.

Shigenobu S., Watanabe H., Hattori M., Sakaki Y., and Ishikawa H. (**2000**). "Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS." *Nature*, 407(6800): 81–86. URL http://dx.doi.org/10.1038/35024074.

Shukla J. and Palli S. (**2012**). "Sex determination in beetles: Production of all male progeny by Parental RNAi knockdown of transformer." *Scientific Reports*, 2: 602. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3426794/.

Simao F., Waterhouse R., Ioannidis P., Kriventseva E., and Zdobnov E. (**2015**). "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs." *Bioinformatics*. URL http://bioinformatics.oxfordjournals.org/content/early/2015/06/08/bioinformatics.btv351.abstract.

Simpson J., Wong K., Jackman S., Schein J., Jones S., and Birol I. (**2009**). "ABySS: A parallel assembler for short read sequence data." *Genome Research*, 19(6): 1117–1123. URL http://genome.cshlp.org/content/19/6/1117.abstract.

Singh S., Coronella J., Benes H., Cochrane B., and Zimniak P. (**2001**). "Catalytic function of *Drosophila melanogaster* glutathione S-transferase DmGSTS1-1 (GST-2) in conjugation of lipid peroxidation end products." *European Journal of Biochemistry*, 268(10): 2912–2923. URL http://dx.doi.org/10.1046/j.1432-1327.2001.02179.x.

Siozios S., Cestaro A., Kaur R., Pertot I., Rota-Stabelli O., and Anfora G. (**2013**). "Draft Genome Sequence of the *Wolbachia* Endosymbiont of *Drosophila suzukii*." *Genome Announcements*, 1(1). URL http://genomea.asm.org/content/1/1/e00032-13.abstract.

Skaljac M., Zanic K., Ban S., Kontsedalov S., and Ghanim M. (**2010**). "Co-infection and localization of secondary symbionts in two whitefly species." *BMC Microbiology*, 10(1): 1–15. URL http://dx.doi.org/10.1186/1471-2180-10-142.

Slater G. and Birney E. (**2005**). "Automated generation of heuristics for biological sequence comparison." *BMC Bioinformatics*, 6(1): 1–11. URL http://dx.doi.org/10.1186/1471-2105-6-31.

Sloan D. and Moran N. (**2012**). "Genome Reduction and Co-evolution between the Primary and Secondary Bacterial Symbionts of Psyllids." *Molecular Biology and Evolution*, 29(12): 3781–3792. URL http://mbe.oxfordjournals.org/content/29/12/3781.abstract.

Sloan D. and Moran N. (**2012a**). "Endosymbiotic bacteria as a source of carotenoids in whiteflies." *Biology Letters*, 8(6): 986–989. URL http://rsbl.royalsocietypublishing.org/content/8/6/986.

Sloan D. and Moran N. (**2013**). "The Evolution of Genomic Instability in the Obligate Endosymbionts of Whiteflies." *Genome Biology and Evolution*, 5(5): 783–793. URL http://gbe.oxfordjournals.org/content/5/5/783.abstract.

Sloan D., Nakabachi A., Richards S., Qu J., Murali S., Gibbs R., and Moran N. (**2014**). "Parallel Histories of Horizontal Gene Transfer Facilitated Extreme Reduction of Endosymbiont Genomes in Sap-Feeding Insects." *Molecular Biology and Evolution*, 31(4): 857–871. URL http://mbe.oxfordjournals.org/content/31/4/857.abstract.

Soin T., Swevers L., Kotzia G., Iatrou K., Janssen C., Rougé P., Harada T., Nakagawa Y., and Smagghe G. (**2010**). "Comparison of the activity of non-steroidal ecdysone agonists between dipteran and lepidopteran insects, using cell-based EcR reporter assays." *Pest Management Science*, 66(11): 1215–1229. URL http://dx.doi.org/10.1002/ps.1998.

Sonah H., Deshmukh R., Sharma A., Singh V., Gupta D., Gacche R., Rana J., Singh N., and Sharma T. (**2011**). "Genome-Wide Distribution and Organization of Microsatellites in Plants: An Insight into Marker Development in *Brachypodium*." *PLoS ONE*, 6(6): e21 298. URL http://dx.doi.org/10.1371%2Fjournal.pone.0021298.

Song N. and Liang A. (**2009**). "The complete mitochondrial genome sequence of *Geisha distinctissima* (Hemiptera: Flatidae) and comparison with other hemipteran insects." *Acta Biochimica et Biophysica Sinica*, 41(3): 206–216. URL http://abbs.oxfordjournals.org/content/41/3/206.abstract.

Spradling A., Stern D., Beaton A., Rhem E., Laverty T., Mozden N., Misra S., and Rubin G. (**1999**). "The Berkeley *Drosophila* Genome Project Gene Disruption Project: Single P-Element Insertions Mutating 25% of Vital *Drosophila* Genes." *Genetics*, 153(1): 135–177.

Stamatakis A. (**2006**). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." *Bioinformatics*, 22(21): 2688–2690. URL http://bioinformatics.oxfordjournals.org/content/22/21/2688.abstract.

Stanke M., Keller O., Gunduz I., Hayes A., Waack S., and Morgenstern B. (**2006**). "AUGUSTUS: *ab initio* prediction of alternative transcripts." *Nucleic Acids Research*, 34: W435–W439. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1538822/.

Stark A., Lin M., Kheradpour P., Pedersen J., Parts L., Carlson J., Crosby M., Rasmussen M., Roy S., Deoras A., Ruby J., Brennecke J., Matthews B., Schroeder A., Gramates L., Pierre S. S., Roark M., Jr K. W., Kulathinal R., Zhang P., Myrick K., Antone J., Gelbart W., Yu C., Park S., Wan K., Celniker S., Hodges E., Hinrichs A., Caspi A., Paten B., Park S., Han M., Maeder M., Polansky B., Robson B., Aerts S., van Helden J., Hassan B., Gilbert D., Eastman D., Rice M., Weir M., Hahn M., Park Y., Dewey C., Pachter L., Kent W., Haussler D., Lai E., Bartel D., Hannon G., Kaufman T., Eisen M., Clark A., Smith D., and Kellis M. (**2007**). "Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures." *Nature*, 450(7167): 219–232. URL http://dx.doi.org/10.1038/nature06340.

Stein L., Bao Z., Blasiar D., Blumenthal T., Brent M., Chen N., Chinwalla A., Clarke L., Clee C., Coghlan A., Coulson A., D'Eustachio P., Fitch D., Fulton L., Fulton R., Griffiths-Jones S., Harris T., Hillier L., Kamath R., Kuwabara P., and et al (**2003**). "The Genome Sequence of *Caenorhabditis briggsae*: A Platform for Comparative Genomics." *PLoS Biology*, 1(2): e45. URL http://dx.doi.org/10.1371%2Fjournal.pbio.0000045.

Stewart J. and Beckenbach A. (**2005**). "Insect mitochondrial genomics: the complete mitochondrial genome sequence of the meadow spittlebug *Philaenus spumarius* (Hemiptera: Auchenorrhyncha: Cercopoidae)." *Genome*, 48(1): 46–54. URL http://dx.doi.org/10.1139/g04-090. PMID: 15729396.

Stouthamer R., Breeuwer J., and Hurst G. (**1999**). "*Wolbachia Pipientis*: Microbial Manipulator of Arthropod Reproduction." *Annual Review of Microbiology*, 53(1): 71–102. URL http://dx.doi.org/10.1146/annurev.micro.53.1.71. PMID: 10547686.

Su Q., Oliver K., Pan H., Jiao X., Liu B., Xie W., Wang S., Wu Q., Xu B., White J., Zhou X., and Zhang Y. (**2013**). "Facultative Symbiont *Hamiltonella* Confers Benefits to *Bemisia tabaci* (Hemiptera: Aleyrodidae), an Invasive Agricultural Pest Worldwide." *Environmental Entomology*, 42(6): 1265–1271. URL http://ee.oxfordjournals.org/content/42/6/1265.

Suzuki Y., Tsunoda T., Sese J., Taira H., Mizushima-Sugano J., Hata H., Ota T., Isogai T., Tanaka T., Nakamura Y., Suyama A., Sakaki Y., Morishita S., Okubo K., and Sugano S. (**2001**). "Identification and Characterization of the Potential Promoter Regions of 1031 Kinds of Human Genes." *Genome Research*, 11(5): 677–684. URL http://genome.cshlp.org/content/11/5/677.abstract.

Tae H., Ryu D., Sureshchandra S., and Choi J. (**2012**). "ESTclean: a cleaning tool for next-gen transcriptome shotgun sequencing." *BMC Bioinformatics*, 13(1): 247. URL http://www.biomedcentral.com/1471-2105/13/247.

Takahashi R. (**1936**). "Some Aleyrodidae, Aphididae, Coccidae (Homoptera) and Thysanoptera from Micronesia." *Tenthredo*, 1: 109–120.

Tamames J., Gil R., Latorre A., Peretó J., Silva F., and Moya A. (**2007**). "The frontier between cell and organelle: genome analysis of *Candidatus* Carsonella ruddii." *BMC Evolutionary Biology*, 7: 181–181. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2175510/.

Tariq M., Kim H., Jejelowo O., and Pourmand N. (**2011**). "Whole-transcriptome RNAseq analysis from minute amount of total RNA." *Nucleic Acids Research*, 39(18): e120. URL http://nar.oxfordjournals.org/content/early/2011/07/06/nar.gkr547.abstract.

Tay W., Elfekih S., Court L., Gordon K., and Barro P. D. (**2016**). "Complete mitochondrial DNA genome of *Bemisia tabaci* cryptic pest species complex Asia I (Hemiptera: Aleyrodidae)." *Mitochondrial DNA*, 27(2): 972–973. URL http://dx.doi.org/10.3109/19401736.2014.926511. PMID: 24960562.

Tazi J., Bakkour N., and Stamm S. (**2009**). "Alternative splicing and disease." *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1792(1): 14 – 26. URL http://www.sciencedirect.com/science/article/pii/S0925443908001932.

Teixeira L., Ferreira A., and Ashburner M. (**2008**). "The Bacterial Symbiont *Wolbachia* Induces Resistance to RNA Viral Infections in *Drosophila melanogaster*." *PLoS Biol*, 6(12): 1–11. URL http://dx.doi.org/10.1371%2Fjournal.pbio.1000002.

Tene B. F., Poupardin R., Costantini C., Awono-Ambene P., Wondji C., Ranson H., and Antonio-Nkondjio C. (**2013**). "Resistance to DDT in an Urban Setting: Common Mechanisms Implicated in Both M and S Forms of *Anopheles gambiae* in the City of Yaounde Cameroon." *PLoS ONE*, 8(4): 1–9. URL http://dx.doi.org/10.1371%2Fjournal.pone.0061408.

Tennessen J., Baker K., Lam G., Evans J., and Thummel C. (**2011**). "The *Drosophila* Estrogen-Related Receptor Directs a Metabolic Switch that Supports Developmental Growth." *Cell Metabolism*, 13(2): 139–148. URL http://dx.doi.org/10.1016/j.cmet.2011.01.005.

Terrapon N., Li C., Robertson H., Ji L., Meng X., Booth W., Chen Z., Childers C., Glastad K., Gokhale K., Gowin J., Gronenberg W., Hermansen R., Hu H., Hunt B., Huylmans A., Khalil S., Mitchell R., Munoz-Torres M., Mustard J., Pan H., Reese J., Scharf M., Sun F., Vogel H., Xiao J., Yang W., Yang Z., Yang Z., Zhou J., Zhu J., Brent C., Elsik C., Goodisman M., Liberles D., Roe R., Vargo E., Vilcinskas A., Wang J., Bornberg-Bauer E., Korb J., Zhang G., and Liebig J. (**2014**). "Molecular traces of alternative social organization in a termite genome." *Nature Communications*, 5: 3636. URL http://dx.doi.org/10.1038/ncomms4636.

TGSC T. G. S. C. (**2008**). "The genome of the model beetle and pest *Tribolium castaneum*." *Nature*, 452(7190): 949–955. URL http://dx.doi.org/10.1038/nature06784.

Thao M., Baumann L., and Baumann P. (**2004**). "Organization of the mitochondrial genomes of whiteflies, aphids, and psyllids (Hemiptera, Sternorrhyncha)." *BMC Evolutionary Biology*, 4(1): 1–13. URL http://dx.doi.org/10.1186/1471-2148-4-25.

Thao M., Baumann L., Hess J., Falk B., Ng J., Gullan P., and Baumann P. (**2003**). "Phylogenetic Evidence for Two New Insect-Associated Chlamydia of the Family *Simkaniaceae*." *Current Microbiology*, 47(1): 46–50. URL http://dx.doi.org/10.1007/s00284-002-3953-9.

Thao M. and Baumann P. (**2004**). "Evidence for Multiple Acquisition of *Arsenophonus* by Whitefly Species (Sternorrhyncha: Aleyrodidae)." *Current Microbiology*, 48(2): 140–144. URL http://dx.doi.org/10.1007/s00284-003-4157-7.

Thiel T., Michalek W., Varshney R., and Graner A. (**2003**). "Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.)." *Theoretical and applied genetics*, 106(3): 411–422.

Thierry M., Becker N., Hajri A., Reynaud B., Lett J., and Delatte H. (**2011**). "Symbiont diversity and non-random hybridization among indigenous (Ms) and invasive (B) biotypes of *Bemisia tabaci*." *Molecular Ecology*, 20(10): 2172–2187. URL http://dx.doi.org/10.1111/j.1365-294X.2011.05087.x.

Thompson W. (**2011**). *The Whitefly, Bemisia tabaci (Homoptera: Aleyrodidae) Interaction with Geminivirus-Infected Host Plants: Bemisia tabaci, Host Plants and Geminiviruses*. SpringerLink : Bücher. Springer Science, Business Media B.V. URL http://books.google.co.uk/books?id=u1XaNf1agnEC.

Thresh J., Ottim-Nape G., Thankappan M., and Muniyappa V. (**1998**). "The mosaic diseases of cassava in Africa and India caused by whitefly-borne geminivirus." *Annual Review of Plant Pathology*, 77: 935–945.

Tohidi-Esfahani D., Lawrence M., Graham L., Hannan G., Simpson A., and Hill R. (**2011**). "Isoforms of the heteropteran *Nezara viridula* ecdysone receptor: protein characterisation, RH5992

insecticide binding and homology modelling." *Pest Management Science*, 67(11): 1457–1467. URL http://dx.doi.org/10.1002/ps.2200.

Tomoyasu Y., Miller S., Tomita S., Schoppmeier M., Grossmann D., and Bucher G. (**2008**). "Exploring systemic RNA interference in insects: a genome-wide survey for RNAi genes in *Tribolium*." *Genome Biology*, 9(1): 1–22. URL http://dx.doi.org/10.1186/gb-2008-9-1-r10.

Torres T., Dolezal M., Schlötterer C., and Ottenwälder B. (**2009**). "Expression profiling of Drosophila mitochondrial genes via deep mRNA sequencing." *Nucleic Acids Research*, 37(22): 7509–7518. URL http://nar.oxfordjournals.org/content/37/22/7509.abstract.

Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D., Pimentel H., Salzberg S., Rinn J., and Pachter L. (**2012**). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." *Nature Protocols*, 7(3): 562–578. URL http://dx.doi.org/10.1038/nprot.2012.016.

Trautwein M., Wiegmann B., Beutel R., Kjer K., and Yeates D. (**2012**). "Advances in Insect Phylogeny at the Dawn of the Postgenomic Era." *Annual Review of Entomology*, 57(1): 449–468. URL http://dx.doi.org/10.1146/annurev-ento-120710-100538.

Trowbridge R. E., Dittmar K., and Whiting M. F. (**2006**). "Identification and phylogenetic analysis of *Arsenophonus*- and *Photorhabdus*-type bacteria from adult Hippoboscidae and Streblidae (Hippoboscoidea)." *Journal of Invertebrate Pathology*, 91(1): 64–68. URL http://www.sciencedirect.com/science/article/pii/S0022201105001710.

Tweedie S., Ashburner M., Falls K., Leyland P., McQuilton P., Marygold S., Millburn G., Osumi-Sutherland D., Schroeder A., Seal R., Zhang H., and Consortium T. F. (**2009**). "FlyBase: enhancing Drosophila Gene Ontology annotations." *Nucleic Acids Research*, 37: D555–D559. URL http://nar.oxfordjournals.org/content/37/suppl_1/D555.abstract.

Upadhyay S., Sharma S., Singh H., Dixit S., Kumar J., Verma P., and Chandrashekar K. (**2015**). "Whitefly Genome Expression Reveals Host-Symbiont Interaction in Amino Acid Biosynthesis." *PLoS ONE*, 10(5): e0126 751. URL http://dx.doi.org/10.1371%2Fjournal.pone.0126751.

Usmani S., Bona R., Chiosis G., and Li Z. (**2010**). "The anti-myeloma activity of a novel purine scaffold HSP90 inhibitor PU-H71 is via inhibition of both HSP90A and HSP90B1." *Journal of Hematology & Oncology*, 3(1): 1–8. URL http://dx.doi.org/10.1186/1756-8722-3-40.

van der Giezen M. (**2011**). "Mitochondria and the Rise of Eukaryotes." *BioScience*, 61(8): 594–601. URL http://bioscience.oxfordjournals.org/content/61/8/594.abstract.

van Ham R., Kamerbeek J., Palacios C., Rausell C., Abascal F., Bastolla U., Fernández J., Jiménez L., Postigo M., Silva F., Tamames J., Viguera E., Latorre A., Valencia A., Morán F., and Moya A. (**2003**). "Reductive genome evolution in *Buchnera aphidicola*." *Proceedings of the National Academy of Sciences*, 100(2): 581–586. URL http://www.pnas.org/content/100/2/581.abstract.

Vea I. and Grimaldi D. (**2016**). "Putting scales into evolutionary time: the divergence of major scale insect lineages (Hemiptera) predates the radiation of modern angiosperm hosts." *Scientific Reports*, 6: 23 487 EP–. URL http://dx.doi.org/10.1038/srep23487.

Venter J., Adams M., Myers E., Li P., Mural R., Sutton G., Smith H., Yandell M., Evans C., Holt R., Gocayne J., Amanatides P., Ballew R., Huson D., Wortman J., Zhang Q., Kodira C., Zheng X., Chen L., Skupski M., Subramanian G., Thomas P., and et al (**2001**). "The Sequence of the Human Genome." *Science*, 291(5507): 1304–1351. URL http://science.sciencemag.org/content/291/5507/1304.

Vigneron S., Gharbi-Ayachi A., Raymond A., Burgess A., Labbé J., Labesse G., Monsarrat B., Lorca T., and Castro A. (**2011**). "Characterization of the Mechanisms Controlling Greatwall Activity." *Molecular and Cellular Biology*, 31(11): 2262–2275. URL http://mcb.asm.org/content/31/11/2262.abstract.

Wahl M., Will C., and Lührmann R. (**2009**). "The Spliceosome: Design Principles of a Dynamic RNP Machine." *Cell*, 136(4): 701–718. URL http://dx.doi.org/10.1016/j.cell.2009.02.009.

Walker T., Johnson P., Moreira L., Iturbe-Ormaetxe I., Frentiu F., McMeniman C., Leong Y., Dong Y., Axford J., Kriesner P., Lloyd A., Ritchie S., O/'Neill S., and Hoffmann A. (**2011**). "The wMel Wolbachia strain blocks dengue and invades caged Aedes aegypti populations." *Nature*, 476(7361): 450–453. URL http://dx.doi.org/10.1038/nature10355.

Walker T., Klasson L., Sebaihia M., Sanders M., Thomson N., Parkhill J., and Sinkins S. (**2007**). "Ankyrin repeat domain-encoding genes in the wPip strain of *Wolbachia* from the *Culex pipiens* group." *BMC Biology*, 5(1): 1–9. URL http://dx.doi.org/10.1186/1741-7007-5-39.

Wan F., Zhang G., Liu S., Luo C., Chu D., Zhang Y., Zang L., Jiu M., Lü Z., Cui X., Zhang L., Zhang F., Zhang Q., Liu W., Liang P., Lei Z., and Zhang Y. (**2009**). "Invasive mechanism and management strategy of *Bemisia tabaci* (Gennadius) biotype B: Progress report of 973 Program on invasive alien species in China." *Science in China Series C: Life Sciences*, 52(1): 88–95. URL http://dx.doi.org/10.1007/s11427-008-0135-4.

Wang G. and Cooper T. (**2007**). "Splicing in disease: disruption of the splicing code and the decoding machinery." *Nature Reviews Genetics*, 8(10): 749–761. URL http://dx.doi.org/10.1038/nrg2164.

Wang H., Xiao N., Yang J., Wang X., Colvin J., and Liu S. (**2016b**). "The complete mitochondrial genome of *Bemisia afer* (Hemiptera: Aleyrodidae)." *Mitochondrial DNA*, 27(1): 98–99. URL http://dx.doi.org/10.3109/19401736.2013.873921. PMID: 24438292.

Wang H., Yang J., Boykin L., Zhao Q., Li Q., Wang X., and Liu S. (**2013**). "The characteristics and expression profiles of the mitochondrial genome for the Mediterranean species of the *Bemisia tabaci* complex." *BMC Genomics*, 14(1): 1–15. URL http://dx.doi.org/10.1186/1471-2164-14-401.

Wang H., Yang J., Boykin L., Zhao Q., Wang Y., Liu S., and Wang X. (**2014a**). "Developing conversed microsatellite markers and their implications in evolutionary analysis of the *Bemisia tabaci* complex." *Scientific Reports*, 4. URL http://dx.doi.org/10.1038/srep06351.

Wang K., Li X., Ding S., Wang N., Mao M., Wang M., and Yang D. (**2016**). "The complete mitochondrial genome of the *Atylotus miser* (Diptera: Tabanomorpha: Tabanidae), with mitochondrial genome phylogeny of lower Brachycera (Orthorrhapha)." *Gene*, 586(1): 184–196. URL http://www.sciencedirect.com/science/article/pii/S037811191630261X.

Wang S., Lorenzen M., Beeman R., and Brown S. (**2008**). "Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome." *Genome Biology*, 9(3): 1–14. URL http://dx.doi.org/10.1186/gb-2008-9-3-r61.

Wang X., Luan J., Li J., Bao Y., Zhang C., and Liu S. (**2010a**). "*De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development." *BMC Genomics*, 11(1): 400. URL http://www.biomedcentral.com/1471-2164/11/400.

Wang X., Luan J., Li J., Su Y., Xia J., and Liu S. (**2011**). "Transcriptome analysis and comparison reveal divergence between two invasive whitefly cryptic species." *BMC Genomics*, 12(1): 458. URL http://www.biomedcentral.com/1471-2164/12/458.

Wang X., Zhao Q., Luan J., Wang Y., Yan G., and Liu S. (**2012**). "Analysis of a native whitefly transcriptome and its sequence divergence with two invasive whitefly species." *BMC Genomics*, 13(1): 1–13. URL http://dx.doi.org/10.1186/1471-2164-13-529.

Wang Y., Anderson J., Chen J., Geer L., He S., Hurwitz D., Liebert C., Madej T., Marchler G., Marchler-Bauer A., Panchenko A., Shoemaker B., Song J., Thiessen P., Yamashita R., and Bryant S. (**2002**). "MMDB: Entrez's 3D-structure database." *Nucleic Acids Research*, 30(1): 249–252. URL http://nar.oxfordjournals.org/content/30/1/249.abstract.

Wang Z. and Burge C. (**2008**). "Splicing regulation: From a parts list of regulatory elements to an integrated splicing code." *RNA*, 14(5): 802–813. URL http://rnajournal.cshlp.org/content/14/5/802.abstract.

Wang Z., Yan H., Yang Y., and Wu Y. (2010b). "Biotype and insecticide resistance status of the whitefly *Bemisia tabaci* from China." *Pest Management Science*, 66: 1360–1366.

Waterhouse A., Procter J., Martin D., Clamp M., and Barton G. (2009). "Jalview Version 2-a multiple sequence alignment editor and analysis workbench." *Bioinformatics*, 25(9): 1189–1191. URL http://bioinformatics.oxfordjournals.org/content/25/9/1189.abstract.

Wattam A., Abraham D., Dalay O., Disz T., Driscoll T., Gabbard J., Gillespie J., Gough R., Hix D., Kenyon R., Machi D., Mao C., Nordberg E., Olson R., Overbeek R., Pusch G., Shukla M., Schulman J., Stevens R., Sullivan D., Vonstein V., Warren A., Will R., Wilson M., Yoo H., Zhang C., Zhang Y., and Sobral B. (2014). "PATRIC, the bacterial bioinformatics database and analysis resource." *Nucleic Acids Research*, 42(D1): D581–D591. URL http://nar.oxfordjournals.org/content/42/D1/D581.abstract.

Weeks A., Velten R., and Stouthamer R. (2003). "Incidence of a new sex–ratio–distorting endosymbiotic bacterium among arthropods." *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1526): 1857–1865. URL http://rspb.royalsocietypublishing.org/content/270/1526/1857.abstract.

Wei S., Shi M., Chen X., Sharkey M., van Achterberg C., Ye G., and He J. (2010). "New Views on Strand Asymmetry in Insect Mitochondrial Genomes." *PLoS ONE*, 5(9): 1–10. URL http://dx.doi.org/10.1371%2Fjournal.pone.0012708.

Weisenfeld N., Yin S., Sharpe T., Lau B., Hegarty R., Holmes L., Sogoloff B., Tabbaa D., Williams L., Russ C., Nusbaum C., Lander E., MacCallum I., and Jaffe D. (2014). "Comprehensive variation discovery in single human genomes." *Nat Genet*, 46(12): 1350–1355. URL http://dx.doi.org/10.1038/ng.3121.

Werren J., Baldo L., and Clark M. (2008). "*Wolbachia*: master manipulators of invertebrate biology." *Nature Reviews Microbiology*, 6(10): 741–751. URL http://dx.doi.org/10.1038/nrmicro1969.

Werren J., Skinner S., and Huger A. (1986). "Male-killing bacteria in a parasitic wasp." *Science*, 231(4741): 990–992. URL http://science.sciencemag.org/content/231/4741/990.

Wheelan S., Church D., and Ostell J. (2001). "Spidey: A Tool for mRNA-to-Genomic Alignments." *Genome Research*, 11(11): 1952–1957. URL http://genome.cshlp.org/content/11/11/1952.abstract.

White J., Kelly S., Cockburn S., Perlman S., and Hunter M. (2011). "Endosymbiont costs and benefits in a parasitoid infected with both *Wolbachia* and *Cardinium*." *Heredity*, 106(4): 585–591. URL http://dx.doi.org/10.1038/hdy.2010.89.

Wilkes T., Darby A., Choi J., Colbourne J., Werren J., and Hurst G. (**2010**). "The draft genome sequence of *Arsenophonus nasoniae*, son-killer bacterium of *Nasonia vitripennis*, reveals genes associated with virulence and symbiosis." *Insect Molecular Biology*, 19: 59–73. URL http://dx.doi.org/10.1111/j.1365-2583.2009.00963.x.

Will C. and Lührmann R. (**2011**). "Spliceosome Structure and Function." *Cold Spring Harbor Perspectives in Biology*, 3(7). URL http://cshperspectives.cshlp.org/content/3/7/a003707.abstract.

Wille B. and Hartman G. (**2009**). "Two Species of Symbiotic Bacteria Present in the Soybean Aphid (Hemiptera: Aphididae)." *Environmental Entomology*, 38(1): 110–115. URL http://ee.oxfordjournals.org/content/38/1/110.

Wilson A., Ashton P., Calevro F., Charles H., Colella S., Febvay G., Jander G., Kushlan P., Macdonald S., Schwartz J., Thomas G., and Douglas A. (**2010**). "Genomic insight into the amino acid relations of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera aphidicola*." *Insect Molecular Biology*, 19: 249–258. URL http://dx.doi.org/10.1111/j.1365-2583.2009.00942.x.

Wilson E. (**2009**). "Threats to Global Diversity."

Wilson K., Cahill V., Ballment E., and Benzie J. (**2000**). "The Complete Sequence of the Mitochondrial Genome of the Crustacean *Penaeus monodon*: Are Malacostracan Crustaceans More Closely Related to Insects than to Branchiopods?" *Molecular Biology and Evolution*, 17(6): 863–874. URL http://mbe.oxfordjournals.org/content/17/6/863.abstract.

Wiśniewski J., Ostasiewicz P., Duś K., Zielińska D., Gnad F., and Mann M. (**2012**). "Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma." *Molecular systems biology*, 8: 611. URL http://europepmc.org/abstract/MED/22968445.

Wolstenholme D. (**1992**). *Animal Mitochondrial DNA: Structure and Evolution*, volume 141 of *International Review of Cytology*. Academic Press. URL http://www.sciencedirect.com/science/article/pii/S0074769608620665.

Wu T. and Watanabe C. (**2005**). "GMAP: a genomic mapping and alignment program for mRNA and EST sequences." *Bioinformatics*, 21(9): 1859–1875. URL http://bioinformatics.oxfordjournals.org/content/21/9/1859.abstract.

Xie W., Guo L., Jiao X., Yang N., Yang X., Wu Q., Wang S., Zhou X., and Zhang Y. (**2014**). "Transcriptomic dissection of sexual differences in *Bemisia tabaci*, an invasive agricultural pest worldwide." *Scientific Reports*, 4: 4088. URL http://dx.doi.org/10.1038/srep04088.

Xie W., Meng Q., Wu Q., Wang S., Yang X., Yang N., Liu R., Jiao X., Pan H., Liu B., Su Q., Xu B., Hu S., Zhou X., and Zhang Y. (**2012**). "Pyrosequencing the *Bemisia tabaci* Transcriptome

Reveals a Highly Diverse Bacterial Community and a Robust System for Insecticide Resistance." *PLoS ONE*, 7(4).

Xu J. and Gong Z. (**2003**). "Intron requirement for AFP gene expression in *Trichoderma viride*." *Microbiology*, 149(11): 3093–3097. URL http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.26514-0.

Xu J., Ji P., Wang B., Zhao L., Wang J., Zhao Z., Zhang Y., Li J., Xu P., and Sun X. (**2013**). "Transcriptome Sequencing and Analysis of Wild Amur Ide (*Leuciscus waleckii*) Inhabiting an Extreme Alkaline-Saline Lake Reveals Insights into Stress Adaptation." *PLoS ONE*, 8(4): e59 703. URL http://dx.doi.org/10.1371%2Fjournal.pone.0059703.

Xu Q., Chai F., An X., and Han S. (**2014**). "Comparison of Detoxification Enzymes of *Bemisia tabaci* (Hemiptera: Aleyrodidae) Biotypes B and Q After Various Host Shifts." *Florida Entomologist*, 97(2): 715–723. URL http://dx.doi.org/10.1653/024.097.0253.

Xue J., Zhou X., Zhang C., Yu L., Fan H., Wang Z., Xu H., Xi Y., Zhu Z., Zhou W., Pan P., Li B., Colbourne J., Noda H., Suetsugu Y., Kobayashi T., Zheng Y., Liu S., Zhang R., Liu Y., Luo Y., Fang D., Chen Y., Zhan D., Lv X., Cai Y., Wang Z., Huang H., Cheng R., Zhang X., Lou Y., Yu B., Zhuo J., Ye Y., Zhang W., Shen Z., Yang H., Wang J., Wang J., Bao Y., and Cheng J. (**2014**). "Genomes of the rice pest brown planthopper and its endosymbionts reveal complex complementary contributions for host adaptation." *Genome Biology*, 15(12): 1–20. URL http://dx.doi.org/10.1186/s13059-014-0521-0.

Xue X., Li S., Ahmed M., Barro P. D., Ren S., and Qiu B. (**2012**). "Inactivation of *Wolbachia* Reveals Its Biological Roles in Whitefly Host." *PLoS ONE*, 7(10): 1–11. URL http://dx.doi.org/10.1371%2Fjournal.pone.0048148.

Yandell M., Mungall C., Smith C., Prochnik S., Kaminker J., Hartzell G., Lewis S., and Rubin G. (**2006**). "Large-Scale Trends in the Evolution of Gene Structures within 11 Animal Genomes." *PLoS Comput Biol*, 2(3): 1–13. URL http://dx.plos.org/10.1371%2Fjournal.pcbi.0020015.

Yang P., Li X., Shipp M., Shockey J., and Cahoon E. (**2010**). "Mining the bitter melon (*momordica charantia* l.) seed transcriptome by 454 analysis of non-normalized and normalized cDNA populations for conjugated fatty acid metabolism-related genes." *BMC Plant Biology*, 10(1): 250. URL http://www.biomedcentral.com/1471-2229/10/250.

Yang X., He C., Xie W., Liu Y., Xia J., Yang Z., Guo L., Wen Y., Wang S., Wu Q., Yang F., Zhou X., and Zhang Y. (**2016**). "Glutathione S-transferases are involved in thiamethoxam resistance in the field whitefly *Bemisia tabaci* Q (Hemiptera: Aleyrodidae)." *Pesticide Biochemistry and Physiology*, pages –. URL http://www.sciencedirect.com/science/article/pii/S0048357516300311.

Ye X., Su Y., Zhao Q., Xia W., Liu S., and Wang X. (**2014**). "Transcriptomic analyses reveal the adaptive features and biological differences of guts from two invasive whitefly species." *BMC Genomics*, 15(1): 1–12. URL http://dx.doi.org/10.1186/1471-2164-15-370.

Yeh R., Lim L., and Burge C. (**2001**). "Computational Inference of Homologous Gene Structures in the Human Genome." *Genome Research*, 11(5): 803–816. URL http://genome.cshlp.org/content/11/5/803.abstract.

Yuan L., Wang S., Zhou J., Du Y., Zhang Y., and Wang J. (**2012**). "Status of insecticide resistance and associated mutations in Q-biotype of whitefly, *Bemisia tabaci*, from eastern China." *Crop Protection*, 31: 67–71.

Zchori-Fein E. and Brown J. (**2002**). "Diversity of Prokaryotes Associated with *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidae)." *Annals of the Entomological Society of America*, 95(6): 711–718. URL http://dx.doi.org/10.1603/0013-8746(2002)095[0711:DOPAWB]2.0.CO;2.

Zchori-Fein E. and Perlman S. (**2004**). "Distribution of the bacterial symbiont *Cardinium* in arthropods." *Molecular Ecology*, 13(7): 2009–2016. URL http://dx.doi.org/10.1111/j.1365-294X.2004.02203.x.

Zchori-Fein E., Perlman S., Kelly S., Katzir N., and Hunter M. (**2004**). "Characterization of a 'Bacteroidetes'Äô symbiont in *Encarsia* wasps (Hymenoptera: Aphelinidae): proposal of 'Äò*Candidatus* Cardinium hertigii'Äô." *International Journal of Systematic and Evolutionary Microbiology*, 54(3): 961–968. URL http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.02957-0.

Zdobnov E. and Bork P. (**2007**). "Quantification of insect genome divergence." *Trends in Genetics*, 23(1): 16–20.

Zhang M. (**2002**). "Computational prediction of eukaryotic protein-coding genes." *Nature Reviews Genetics*, 3: 698–709.

Zhang R. and Lin Y. (**2009**). "DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes." *Nucleic Acids Research*, 37.

Zhu L., Zhang Y., Zhang W., Yang S., Chen J., and Tian D. (**2009**). "Patterns of exon-intron architecture variation of genes in eukaryotic genomes." *BMC Genomics*, 10(1): 1–12. URL http://dx.doi.org/10.1186/1471-2164-10-47.

Zhu Y., Machleder E., Chenchik A., Li R., and Siebert P. (**2001**). "Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction." *Biotechniques*, 30(4): 892–897.

Zotti M., Christiaens O., Rougé P., Grutzmacher A., Zimmer P., and Smagghe G. (**2012**). "Sequencing and structural homology modeling of the ecdysone receptor in two chrysopids used in biological control of pest insects." *Ecotoxicology*, 21(3): 906–918. URL http://dx.doi.org/10.1007/s10646-012-0852-0.

Zreik L., Bove J., and Garnier M. (**1998**). "Phylogenetic characterization of the bacterium-like organism associated with marginal chlorosis of strawberry and proposition of a *Candidatus* taxon for the organism, '*Candidatus* phlomobacter fragariae'." *International journal of systematic bacteriology*, 1: 257–261.

Zuccolo A., Sebastian A., Talag J., Yu Y., Kim H., Collura K., Kudrna D., and Wing R. (**2007**). "Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*." *BMC Evolutionary Biology*, 7(1): 1–15. URL http://dx.doi.org/10.1186/1471-2148-7-152.