Bioinformatics of next generation sequencing approaches: using 454 and Illumina data to look at insect genomes and transcriptomes

Submitted by Ritika Chauhan
To the University of Exeter as a thesis for the degree of
Doctor of Philosophy in Biological Sciences
In April 2013

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: ………………………………………………………….

ABSTRACT

By providing a rapid and cost effective means of generating sequencing resources for almost any organism, 'Next generation sequencing technologies' (NGS) have great potential to help address numerous gene and genome level questions in molecular biology. Progress in NGS is exponentially increasing sequence throughput and large scale studies in the genomics/transcriptomics of non-model organisms are becoming a reality. Therefore the main focus of the work presented in this thesis is on the analysis of the large scale non-model insect datasets generated by NGS technologies and their potential to develop functional genomics tools for these species.

Four different NGS datasets from four very different insects the Greenhouse whitefly (*Trialeurodes vaporariorum*) the Passionvine butterfly (*Heliconius melopmene)*, the blowfly (*Lucilia sericata*) and the Green Dock beetle (*Gastrophysa viridula*) were analysed and annotated. Molecular research in these insects has been hindered in the past due to limited nucleotide sequence information. Transcriptome data generated by 454 pyrosequencing was used as a starting point to study the genomics of these ecologically and economically important non-model insect species. The resulting transcriptomes were annotated for gene families involved in xenobiotic metabolism, namely the glutathione-S-transferases (GSTs), cytochrome P450s (P450s) and the carboxylesterases (CCEs). In each case the number and diversity of gene family members is discussed with those documented in other insects. In the case of *H. melpomone*, the transcriptome data was also used to complement the genomic research by identifying and validating cytochrome P450 gene models in the recently sequenced genome. Furthermore, Illumina generated RNA-seq data was used for SNP characterisation in *L. sericata*.

Transcriptome sequencing is shown to be a useful and cost effective technique to enhance the resources available for non-model organisms as well as for gene discovery in the absence of the reference genomic

resources. By focusing on genes involved in xenobiotic metabolism this thesis has isolated numerous candidate genes potentially involved in important processes such as insecticide resistance (*Lucilia* and *Trialeurodes*) and host plant exploitation (*Gastrophysa* and *Heliconius*). NGS technologies and bioinformatics can thus open up avenues to develop functional genomics resources for diverse species of interest to ecologists and evolutionary biologists.

ACKNOWLEDGEMENTS

I owe sincere and earnest gratitude to all the people who have helped me in various ways during my PhD.

I wish to thank, first and foremost, my supervisor Professor Richard ffrench-Constant for giving me the opportunity to work on this project and for the help and guidance especially during the thesis writing. I would like to sincerely thank my co-supervisor Dr Jamie Stevens for all the support and encouragement not only during my PhD but also during my Master's degree. I gratefully acknowledge the research funding provided by the University of Exeter.

I am truly indebted and thankful to my colleagues for the practical advice and timely help. In particular, Yannick Pauchet for getting me started with the annotations; Nikos Karatolos for helping me with my first project; Dr David Nelson for naming and validating all the cytochrome P450 sequences; Robert Jones for the collaborative work we undertook - even though we did not get to meet we still managed to work as a team. My deepest gratitude to Paul Wilkinson for helping me with the bioinformatics aspect of my work; you have been a massive support throughout my PhD. I am grateful to my friend and colleague MD for helping me with all the server related issues. I would like to share the credit of my work with all of you.

My appreciation also goes out to all my office mates and friends at Tremough, especially Devi, Andrea, Iva, Aruna, and Richa. Thank you Shailja, Khadija, DD, Sabita and all my friends for being a great support system and to help me stay focused in my work. Heartfelt thanks to Amit and Sanchita for being so caring and supportive. A special thank you to Amrita and Antonella for being more than just friends, you have been like my family away from home. To my Dutch friends Hanna and Jelmer, thank you for persuading me to finish my thesis writing.

It would have been impossible to accomplish my PhD without the support of my family. I dedicate this thesis to you all. Heartfelt gratitude to my (late) grandfather, Daadu I know you would have been the happiest person today. Thank you, Mummy and Papa for your endless love and selfless sacrifices. Rohit you are the best brother one can have. Thanks to Ma, Baba, Kunal and Varsha for their overwhelming love and care. It almost seems that I was destined for Cornwall as I retraced the path of my grandpa-in-law (Dilip K Bose) who studied at Camborne School of Mines more than fifty years ago, thank you Nanaji for the constant encouragement and blessings.

Finally, I would like to acknowledge my husband and best friend, Vikram, without whose guidance, encouragement and editing assistance, I would not have finished this thesis. Words are not enough to appreciate your unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS/GLOSSARY

BLAST	Basic Local Alignment Search Tool
bp	base pair
CCEs	carboxyl/cholinesterases
cDNA	complementary DNA
EST	expressed sequence tags
GO	Gene Ontology
GSTs	glutathione-S-transferases
MSA	Multiple Sequence Alignment
NGS	Next Generation Sequencing
nt	nucleotide
OP	organophosphates
P450s	cytochrome P450
SNP	Single Nucleotide Polymorphism

Assembly putting sequenced fragments of DNA into their correct positions

Contig a continuous sequence of DNA that has been assembled from overlapping cloned DNA fragments making up a longer stretch of sequence

Coverage the average number of reads representing a given nucleotide in the assembled sequence

De novo assembly aligning and merging reads into contigs without using previous knowledge of the sequence

Normalisation a procedure to equalise the relative abundance of different cDNA transcripts, thus increasing the overall diversity of transcripts

Preprocessing to perform preliminary processing on the raw data

Reads sequenced cloned DNA fragment

Transcriptome the entire mRNA content of a cell

Xenobiotics compounds foreign to an organism (eg insecticides, pesticides, plant secondary metabolites)

CHAPTER 1

INTRODUCTION

Next generation sequencing and xenobiotic metabolism

1.1 Next generation sequencing (NGS) technologies

Advances in next generation sequencing (NGS) technologies have revolutionised molecular biology by opening fascinating opportunities in genome and genetic research (Ansorge, 2009). This technological boost has significantly reduced the sequencing cost per nucleotide and increased throughput many fold (Paszkiewicz and Studholme, 2010). Prior to the use of NGS, Sanger sequencing methods had been widely used in automated sequencing. Robust characteristics, long reads and relatively low error rates made it the method of choice before the advent of NGS (Metzker, 2010). However one major limitation of the Sanger method was the requirement of *in vivo* amplification of the DNA fragment to be sequenced. This was achieved by cloning it into a bacterial host which was prone to host specific bias, was labour intensive, time consuming and expensive (Morozova and Marra, 2008). These issues have been successfully addressed to varying degrees by all the NGS technologies currently available.

Here, the automated Sanger method is referred to as the ‘first generation’ technology and the newer methods are known as the ‘next generation’ technologies or simply NGS (Metzker, 2010). NGS consists of several different technologies each having a distinct set of characteristics which will be briefly discussed in this chapter. The main commercially available NGS platforms are: Roche 454 (GS FLX Titanium/GS Junior), Illumina (Genome Analyzer/HiSeq 2000/MiSeq) and Life Technologies (SOLiD/Ion Torrent PGM) (Liu *et al.*, 2012).

1.1.1 Pyrosequencing on the Roche 454

The Roche 454 was the first NGS system to be commercially introduced in 2005. It uses an alternative sequencing technology known as pyrosequencing which relies on the detection of pyrophosphate released during nucleotide incorporation (Liu *et al.*, 2012). It takes advantage of an *in vitro* DNA method known as 'emulsion' PCR which overcomes the cloning requirements of Sanger sequencing. When first introduced the read length was 100-150 bp, and it produced more than 20,000 individual sequences and generated approximately 20 Mb of output per run (Mardis, 2008). With advancements in sequencing chemistry and the upgrade of the instrument itself, its overall performance has now significantly improved. In the latest GS FLX Titanium system the read length can reach 1000 bp with consensus accuracy reaching 99.99% (Roche).

Comparatively longer reads, high accuracy, short run time and high throughput are the major advantages of the Roche 454 sequencing platform (Table 1.1). This made it a method of choice for addressing a broad range of questions in diverse areas of research, clearly demonstrated by the number of publications (over 2500) in the peer-reviewed journals (Roche). However a higher cost of operation and relatively low read accuracy in homopolymer stretches of more than 6 bp are currently the main limitations of this technology (Rothberg and Leamon, 2008, Ekblom and Galindo, 2011).

1.1.2 Illumina (Solexa)

The Genome Analyser (GA) was the first short read sequencing platform commercialised in 2006 by Solexa (acquired by Illumina in 2007). Sequencing is based on the 'sequencing by synthesis' (SBS), a method that employs reversible terminators for the four bases, each labelled with different fluorescent dyes. The amplification of templates from single molecules is conducted *in situ* via bridge amplification (Zhang and Jeltsch, 2010). Although it was more effective in sequencing homopolymer stretches than 454 pyrosequencing, it generated shorter sequence reads

(Morozova and Marra, 2008). When introduced to the market it generated around 1Gb data/run in 2-3 days having a read length of around 35 bp (single end reads). However, in the latest Illumina sequencers (GA/HiSeq/MiSeq) the read length ranges between 100-250 bp for paired end reads, generating a maximum output between 7.5Gb-600Gb per run, depending on the type of sequencer used (Illumina).

Due to its massively parallel sequencing approach and comparatively lower sequencing cost, it is currently the most widely used platform with over 5000 cited publications in diverse areas of research (Illumina). The shorter sequence length is one of the major drawbacks of this technology; however with the improvement in technology the sequence length has significantly improved (Table 1.1).

1.1.3 SOLiD (Sequence by Oligo Ligation Detection)

The SOLiD platform was developed in 2007 by Applied Biosystems (later merged with Invitrogen to form Life Technologies). It is a short read sequencing platform using chemistry based on ligation. DNA amplification is based on 'emulsion' PCR, similar to 454 sequencing. However DNA ligase, rather than polymerase, is used for sequencing the amplified fragments from single molecules (Zhang and Jeltsch, 2010). The initial read length generated by SOLiD was around 35 bp with an output of 3Gb per run. The recent upgrade to SOLiD 5500xl has improved its read length to 85 bp with accuracy of 99.99% and an output of 30Gb per run (complete run takes around 7 days) (Liu *et al.*, 2012). Lower error rate due to duplicate sequencing of each base pair and generation of large number of reads are the main advantages of this platform (Table 1.1). However due to the shorter read length its preferred current application is still re-sequencing (Ekblom and Galindo, 2011).

Table 1.1 Comparison of the three leading NGS sequencers with Sanger sequencer

Features	**454 GS FLX**	**HiSeq 2000**	**SOLiDv4**	**Sanger 3730xl**
Sequencing mechanism	pyrosequencing	sequencing by synthesis	ligation and two-base coding	dideoxy chain termination
Automated library prep	Yes	Yes	Yes	No
Read length	700 bp	50-150 bp	35-85 bp	400 -900 bp
Accuracy	99.9% (consensus accuracy at 15x coverage)	98% (100% paired end)	99.94% (raw data)	100.00%
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9-84 Kb
Time/run	24 hours	3-10 days	7 -14days	20 mins-3 hours
Instrument price	$500,000	$690,000	$495,000	$95,000
Cost	approx $7000 per run	$6000/(30x) human genome	$15,000/100 Gb	approx $4 per 800 bp reaction
Cost/million bases	$10	$0.07	$0.13	$2400
Advantage	read length, fast	high throughput	accuracy	high quality, long read length
Disadvantage	error rate with polybase more than 6, high cost, low throughput	short read assembly	short read assembly	high cost low throughput

Source (Roche, Illumina, Liu *et al.*, 2012)

1.1.4 Third generation platforms

A massively parallel approach and modest sequencing cost are the main features of the NGS platforms currently being used. However due to the nature of methods used, they are still prone to error and bias (such as bias introduced by PCR amplification). Although ‘second generation’ NGS technologies available today have revolutionised DNA analysis, sequencing technology is evolving rapidly and several new platforms are being developed. A new generation of single molecule sequencing (SMS) technologies is emerging (Schadt *et al.*, 2010). This new approach of sequencing single DNA molecules is called ‘third generation’ sequencing technology. The new generation of sequencing technologies will have the potential to produce longer read length, shorter generation time and lower overall cost. These include platforms like Single-Molecule Real-Time (SMRT) Sequencer, HeliScope Single Molecule Sequencer, and the Ion Personal Genome Machine (Ion Torrent PGM) and Nanopore DNA sequencer (Pareek *et al.*, 2011).

1.2 Applications of NGS technologies

The production of a large number of reads at relatively low cost makes the NGS platforms (described above) useful for many applications (Metzker, 2010). To date, these technologies have been applied in various aspects of biological research including *de novo* sequencing, targeted resequencing, transcriptome sequencing, gene expression profiling, candidate gene finding, variant discovery, metagenomics and epigenetics (Morozova and Marra, 2008). The choice of sequencing platform depends on the type of questions and the genomic resources available for the study species or its closely related species (Ekblom and Galindo, 2011) (Table 1.2).

Table 1.2 Applications and characteristics of Roche 454, Illumina and Solid sequencing platforms

Technology	Pros	Cons	Biological applications
Roche 454	longer reads, faster run time	high reagent cost, high error rates in homo-polymer repeats, lower coverage	whole genome *de novo* sequencing, targeted resequencing, metagenomics, transcriptome sequencing, amplicon sequencing
Illumina	massively parallel approach, lower sequencing cost, deep coverage, low error rate	low multiplexing capability of samples, relatively short reads so reference genome desirable	accurate measure of gene exepression level, gene discovery in metagenomics, whole genome sequencing, targeted resequencing, amplicon sequencing, SNP discovery, identification of a copy number variation, chromosomal rearrangement
ABI/SOLiD	deep coverage, lower error rate due to duplicate sequencing of each base pair	long run times, shorter reads so reference genome is desirable	whole genome re-sequencing, targeted genomic re-sequencing, ChiP seq, exome Enrichment

(Illumina, Roche, Ekblom and Galindo, 2011)

Recent advances in NGS technologies have revolutionized the field of genomics, making it possible for even small research groups to generate large amounts of sequence data rapidly and at a substantially lower cost (Rothberg and Leamon, 2008). This has opened up the possibility of generating large scale sequencing data for non-model organisms. Thus shifting the focus from few laboratory-based studies of model organisms to natural populations (Gilad *et al.*, 2009). Non-model species research is undergoing a revolution as experimental protocols for the generation of large -omic datasets are standardized and the costs are decreasing. As a result, data is accumulating at a faster rate than we can process. Therefore the main focus of the work presented in this thesis is on the analysis of the large scale non-model insect datasets generated by next generation sequencing technologies. In order to better understand the annotation of -omics datasets and to mine them for items of interest, NGS data has been used to characterise the genes encoding xenobiotic metabolising enzymes in insects, as an example. Detailed classification of these gene families is presented in the following section.

1.3 Xenobiotic metabolising enzyme superfamilies

Insects encounter numerous potentially toxic compounds (referred to as xenobiotics) in their life. These include allelochemicals, produced naturally by plants and man made insecticides. Exposure to these xenobiotics, pose a constant challenge for organisms to survive. In order to process these environmental chemicals, organisms have evolved various detoxification mechanisms (Misra *et al.*, 2011). Detoxification system consists of three phases that convert xenobiotics into less harmful substances and facilitate their excretion (Xu *et al.*, 2005). In 'phase I' or the 'biotransformation step', various enzymes act to decrease the biological activity of their substrates by introducing reactive and polar groups. Cytochrome P450s are the most abundant class of xenobiotic metabolising enzymes involved in biotransformation (Feyereisen, 1999). Substrates can be readily excreted after phase I or passed on to the second step of detoxification. In 'phase II' or the 'conjugation step', enzymes such as glutathione-S-transferases (GSTs) and carboxyl/cholinesterases (CCEs) act on the toxic by-products of phase I. GSTs add bulky side groups onto these by-products to increase their hydrophilicity, thus facilitating their excretion. CCEs catalyse the hydrolysis of ester-containing xenobiotics, leading to their detoxification. Conjugated forms of xenobiotics, in phase III, can then be sequestered or actively exported out of the cell with the help of transmembrane transporters such as ATP-binding cassette (ABC) transporters (Misra *et al.*, 2011).

Insects, like all eukaryotes, rely primarily on three superfamilies of enzymes for detoxification and metabolism of xenobiotic compounds. These are the glutathione-S-transferases (GSTs), the cytochrome P450s (P450s) and the carboxylesterases (Ranson *et al.*, 2002).

1.3.1 Glutathione-S-transferases (GSTs)

Glutathione-S-transferase superfamily consists of diverse multifunctional enzymes found ubiquitously in most aerobic eukaryotes and prokaryotes (Enayati *et al.*, 2005). They play a key role in the detoxification of the xenobiotic compounds including insecticides and are also involved in biosynthesis of hormones, protection against oxidative stress and intracellular transport (Salinas and Wong, 1999).

1.3.1.1 Classification and nomenclature

There are two distant groups of GSTs in most organisms, microsomal and cytosolic (Che-Mendoza *et al.*, 2009). A third class of GSTs, known as the 'kappa class' are found in mammalian mitochondria and peroxisomes (Lander *et al.*, 2004, Morel *et al.*, 2004). No members of this class have been reported till date in the annotated sequences of insect genomes (Enayati *et al.*, 2005). The cytosolic GSTs are further classified into six sub-classes - delta, epsilon, omega, theta, zeta and sigma (Low *et al.*, 2007). The delta and epsilon classes are specific to insects and are the only GSTs that have been implicated in insecticide resistance (Ranson *et al.*, 2002, Ding *et al.*, 2003). Names are assigned indicating the species they were isolated from followed by the GST class. They are also given a number that may either reflect the genomic organisation or the order of discovery (Enayati *et al.*, 2005). For example, *DmGSTd8* is the eighth member of the *Drosophila melanogaster* belonging to the delta class of GSTs.

1.3.1.2 Role of insect GSTs

GSTs play a key role in the detoxification of both endogenous and xenobiotic compounds either directly or by catalysing the toxic by-products from biotransformation step (phase I). They are also involved in intracellular transport, protection against oxidative stress and various biosynthetic pathways (Wilce and Parker, 1994). The wide range of functions performed by GSTs is supported by the extensive nature of the GST gene superfamily and broad substrate specificity of individual enzymes (Ortelli *et al.*, 2003). Interest in insect GSTs is primarily focused on their role in detoxifying xenobiotics (insecticides and plant allelochemicals) and more recently on their role in mediating oxidative stress response (Enayati *et al.*, 2005).

GSTs can metabolise insecticides by conjugating the electrophilic compounds with the thiol group of reduced glutathione, or by facilitating the dehydrochlorination reaction using the reduced glutathione as a cofactor instead of conjugate (Clark and Shamaan, 1984). Resistance to at least four classes of insecticides is linked to enhanced GST activity either due to gene amplification or more commonly through increase in transcriptional rate (Ranson and Hemingway, 2004). GST based resistance was first identified in organophosphate (OP) resistance. Resistance due to increase in GSTs has now been implicated in OP resistance in many insect species (Hemingway *et al.*, 2004). The dehydrochlorination reaction catalysed by GST is an important mechanism for DDT detoxification. This is thought to be the most common mechanism for DDT resistance in *Aedes aegypti* (Grant *et al.*, 1991) and *Anopheles gambiae* (Prapanthadara *et al.*, 1993). In Insects, GSTs have also been involved in pyrethroid resistance. However they are not implicated in direct pyrethroid metabolism, instead protect against pyrethroid toxicity by sequestering the insecticide (Kostaropoulos *et al.*, 2001).

1.3.2 Cytochrome P450s

The cytochrome P450s are one of the largest and most ancient category of enzymes comprising a superfamily of heme-thiolate proteins, sometimes referred to as mixed function oxidases or monooxygenases (Alzahrani, 2009). They consist of a diverse class of enzymes, found in virtually all organisms. In insects they are involved in metabolism of xenobiotics, pheromone metabolism, synthesis and degradation of juvenile hormones and ecdysteroids (Feyereisen, 1999). Furthermore, in herbivorous insects they can also detoxify toxic plant allelochemicals (Li *et al.*, 2010).

1.3.2.1 Classification and nomenclature

P450s derive their name from their absorbance peak at 450 nm in the optical spectrum of the carbon monoxide bound reduced form of the enzyme (Feyereisen, 2006). Both mitochondrial and microsomal P450 systems have been described in insects. Mitochondrial P450s are more closely related to the primitive prokaryotic P450s. P450 in the mitochondria use NAD(P)H-ferredoxin reductase and ferredoxin whereas the microsomal P450s interact directly with NADPH-P450 reductase (Alzahrani, 2009). Analysis of the available insect CYP genes indicate that they fall into four major clades – CYP2, CYP3, CYP4 and mitochondrial clades (Feyereisen, 2006).

The nomenclature originally introduced by Nebert (Nebert and Gonzalez, 1987) designates all gene members of the P450 superfamily with a 'CYP' prefix followed by a number for the family, letter for the subfamily and a number for the individual gene (e.g. CYP-6-G-1). All members of the family share >40% identity at the amino acid sequence level and members of subfamily share >55% identity (Feyereisen, 1999). Genes are described in italics (e.g. *CYP6G1*) whereas gene products, mRNA and enzymes are in capitals (e.g. CYP6G1).

1.3.2.2 Role of insect P450s

Each insect genome contains about 100 or so P450 genes, all evolved from a common ancestor and each coding for a different P450 enzyme (Scott, 2008). Insect P450s have wide range of important functional roles which can be divided into two main categories – metabolism of endogenous compounds (including signal molecules involved in growth, development and feeding) and protection against xenobiotics (including resistance to pesticides and tolerance to plant toxins). Regarding the first category, it is well known that insect P450s are involved in biosynthetic pathways of ecdysteroids and juvenile hormones, which are important in insect growth, development, and reproduction (Tijet *et al.*, 2001). For example, the *Drosophila* Halloween genes – *Phantom* (*phm, CYP314A1*), *Disembodied* (*dib, CYP302A1*) and *Shadow* (*sad, CYP315A1*) are involved in the biosynthetic pathway that transforms cholesterol to ecdysone. Another gene called *Shade* (*shd*, *CYP314A1*) catalyses the final hydroxylation step that transforms ecdysone into 20E (20-hydroxyecdysone), the primary insect moulting hormone (Rewitz *et al.*, 2007).

Members of the CYP3 clade of P450s have been implicated in the oxidative detoxification of furanocoumarins, alkaloids, numerous other plant secondary metabolites and synthetic insecticides (Snyder and Glendinning, 1996, Scott, 1999, Mao *et al.*, 2007). Elevated P450 activity is often linked to insecticide resistance and in most cases the CYP gene belongs to the CYP6 family. For example *CYP6D1* is over-produced in pyrethroid resistant *Musca domestica* (Kasai *et al.*, 2000) and *CYP6A1* is associated with organophosphate resistance (Sabourault *et al.*, 2001, Andersen *et al.*, 1994). Up regulation of *CYP6G1* is linked to DDT resistance in *Drosophila melanogaster* (Daborn *et al.*, 2002).

1.3.3 Carboxyl/Cholinesterases (CCEs)

The carboxyl/cholinesterases (CCE) belong to large superfamily of α/β hydrolase proteins. CCE family is comprised of functionally diverse enzymes that hydrolyse carboxylic esters to their component alcohol and acids (Tsubota and Shiotsuki, 2010). Similar to P450s, CCEs can function broadly in xenobiotic detoxification. CCE based resistance has been reported in several veterinary, medical and agriculture insect pests (Hemingway and Karunaratne, 1998).

1.3.3.1 Classification and nomenclature

Due to the overlapping substrate specificity, classification of these enzymes is difficult (Heymann, 1980). Various studies use different classification criteria to classify the same enzymes. According to classification by Aldridge (Aldridge, 1953), esterases inhibited by paraoxon in a progressive and temperature-dependent manner are called B esterases and those not inhibited are A esterases. Another method classifies CCEs into 3 classes and 13 major clades (Ranson *et al.*, 2002, Oakeshott *et al.*, 2005). These classes broadly represent dietary/detoxification, hormone/semiochemical processing and neuro/developmental functions.

There are thousands of ester-containing molecules within insect tissues, making it difficult to identify the native substrate of any particular CCE thus only a few CCEs are named based on their substrate (e.g. acetylcholinesterase and juvenile hormone esterase). Different nomenclatures have been adopted to describe these enzymes in different species or in closely related species (Aldridge, 1953, Ranson *et al.*, 2002, Satoh and Hosokawa, 2006). Despite several efforts to develop an improved system for the nomenclature and classification, a universal standard is yet to be defined (Montella *et al.*, 2012).

1.3.3.2 Role of insect CCEs

CCEs play important role in xenobiotic detoxification and pheromone degradation, neurogenesis and regulating development. Although, the endogenous role of most insect CCEs is unknown, high levels of activity in insect midguts suggest their role in digestion or xenobiotic detoxification (Campbell *et al.*, 2003). Up-regulation of several CCEs in lepidopteran larvae exposed to bacterial infection suggests a role in innate immunity (Shiotsuki and Kato, 1999, Zhu *et al.*, 2003). On the other hand insect CCEs with well-known functions include acetylcholinesterase (AChE) and juvenile hormone esterase (JHE) (Montella *et al.*, 2012).

In insect species, CCEs contribute to insecticide degradation and are involved in insecticide resistance. According to Peiris and Hemingway (Peiris and Hemingway, 1993), it is the primary mechanism for organophosphate (OP) resistance and secondary mechanism for carbamate resistance. However, pyrethroid resistance is also conferred in some other insect species (Devonshire and Moores, 1982). CCEs impart resistance through gene amplification, up-regulation and coding sequence mutation. For example, two separate gene mutations were found to be responsible for diazinon and malathion OP resistance in *Musca domestica (*Taskin *et al.*, 2004*)*. Sibling species *Lucilia cuprina* and *Lucilia serciata* have also been found to be polymorphic for these mutations, two single nucleotide changes in *aE7* gene encoding esterase 3 (E3) confer the two forms of OP resistance. Point mutation G137D is associated with diazinon resistance whereas mutation W251L is associated with malathion resistance (Hartley *et al.*, 2006). On the other hand overexpression of CCE, caused by increased copy number within genome, in *Myzus persicae* has been shown to contribute to resistance. Gene amplification in E4 of *M. persicae* imparts resistance by sequestration of the OP substrates (Devonshire, 1998).

Xenobiotic metabolising enzyme superfamilies appear to be rapidly evolving and representatives of these three families are generally quite numerous within each insect species. However surprising diversity can be seen in the representation of these three families in the annotated insect genomes that have been published to date (Adams *et al.*, 2000, Holt *et al.*, 2002, Consortium, 2006, Nene *et al.*, 2007, Oakeshott *et al.*, 2010, Consortium, 2010) (Table 1.3)

Table 1.3 Number of annotated glutathione-S-transferases (GSTs), cytochrome P450s (P450s) and carboxyl/cholinesterases (CCEs) in the insect genomes

	GSTs	**P450s**	**CCEs**
Drosophila melanogsater	37	85	35
Anopheles gambiae	28	106	51
Aedes aegypti	27	164	54
Apis mellifera	8	46	24
Nasonia vitripennis	19	92	41
Acyrthosiphon pisum	20	83	29
Tribolium castaneum	35	131	49
Bombyx Mori	23	86	69

Data taken from (Adams *et al.*, 2000, Holt *et al.*, 2002, Consortium, 2006, Oakeshott *et al.*, 2010, Ramsey *et al.*, 2010, Tsubota and Shiotsuki, 2010)

With an increasing number and diversity of insect genome sequences becoming available, the diversity of these three enzyme families can be better appreciated and their evolution in insects can be further understood. The availability of sequencing information from a wide range of insects may reveal more functions of these large superfamilies, which in turn can allow better classification and nomenclature of CCE genes within the main insect clades. Furthermore, on going sequencing projects can improve our understanding of these enzyme families and their importance in the adaptation of insects to new ecological niches.

However in the case of non-model species it is not feasible to invest heavily in genome sequence resource for every species or natural population (Bouck and Vision, 2007). Instead, transcriptome data can be used as starting point for deducing coding sequences and function of genes that have not been previously isolated or sequenced (Sims *et al.*, 2004). Up until now, 454 technology has dominated next generation application in transcriptomics of non-model organisms (Morozova and Marra, 2008). Due to the longer read length compared to that produced by other NGS techniques, sequence reads generated by 454 can be effectively used for *de novo* analysis, including assembly of transcriptomes (Morozova *et al.*, 2009). Thus, transcriptome data obtained by 454 pyrosequencing has been used for building genomic reference and for manual curation of the above mentioned gene superfamiles in the present study.

1.4 Thesis overview

The primary aim of the work described in this thesis was the analysis of large scale non-model insect datasets generated by next generation sequencing technology in order to better understand how to mine the datasets for specific research queries. The first objective was the discovery and annotation of the three gene superfamilies responsible for the xenobiotic metabolism in insects. The second objective was single nucleotide polymorphism (SNP) discovery for population genetics studies using a *Lucilia sericata* dataset. The third and final objective was to come up with a strategy to separate mixed sequences from insect and microorganisms, which sometimes cause problems, when the insects are obtained from the field.

Specifically, the objective of Chapter 2 was to utilise the *Trialeurodes vaporariorum* transcriptome data, generated by 454 pyrosequencing, to identify and annotate enzymes involved in xenobiotic metabolism. And to further identify trancripts with potential role in conferring insecticide resistance in this species. The cDNA libraries were constructed by Yannick Pauchet. Paul Wilkinson conducted the transcriptome assembly. I manually curated the P450s, GSTs and CCEs with Nikos Karatoloas. David Nelson named the manually curated P450s. The data generated forms part of the paper published in the journal BMC Genomics (Karatolos *et al.*, 2011)

The aim of Chapter 3 was to identify all the CYP sequences found in the genome and transcriptome of *Heliconius melpomene*. The results of the *H. melpomene* genome wide analysis of the cytochrome P450s is reported in this chapter. Paul Wilkinson assembled the 454 sequenced reads. I manually curated the *H. melpomene* transcriptome and validated the P450 gene models. Robert Jones manually curated the *H. melpomene* genome scaffolds and further helped with the CYP4 gene model validation. David Nelson named the manually curated P450s. This analysis was carried out for the *H. melpomene* genome consortium (Consortium, 2012).

The main objective of Chapter 4 was to improve the -omics resource available for *Lucilia sericata*. 454 pyrosequencing technology was used to generate a reliable transcriptome reference for *L. sericata* which was utilised to identify gene families associated with detoxification. The second objective was to carry out SNP analysis using Illumina RNA-seq data generated for the UK population. Yannick Pauchet generated the cDNA libraries and Konrad Paszkiewicz conducted the preprocessing of the Illumina data. I conducted the 454 transcriptome assembly, Blast2GO analysis, manual curation of the enzymes families and SNP discovery. David Nelson named the manually curated P450s.

The objective of Chapter 5 was to survey the 454 transcriptomic data from five beetle species (*Gastrophysa viridula*, *Chyrsomela tremulae*, *Leptinotarsa decemlineata*, *Sitophilus oryzae* and *Callosobrunchus maculatus*) for associated microorganisms. The second objective was to come up with a strategy to separate the sequences of the host and the associated microorganisms. cDNA libraries for the five beetle datasets were generated by Yannick Pauchet. Paul Wilkinson conducted the preliminary assembly. I conducted the survey, separated the mixed sequences, conducted re-assemblies and Blast2GO analysis.

CHAPTER 2

Identifying putative xenobiotic metabolising enzyme superfamilies in the *Trialeurodes vaporariorum* transcriptome

2.1 INTRODUCTION

Whiteflies are small phloem-feeding insects belonging to the family Aleyrodidae of the order Hemiptera. Around 1500 species are found worldwide, including major pests of many agricultural crops (Chougule and Bonning, 2012). Whiteflies feed on the nutrients essential for plant growth and development, thus affecting the physiology, development, biochemistry and anatomy of the infected plant (Henneberry *et al.*, 2000). Some species are known to transmit viruses causing damaging plant diseases (Brown and Czosnek, 2002). Only whiteflies in the *Trialeurodes* and *Bemisia* genera are virus vectors (Wisler and Duffus, 2001). The most damaging species are the Greenhouse whitefly, *Trialeurodes vaporariorum* (Westwood), and the Cotton, Sweet-potato or Tobacco whitefly, *Bemisia tabaci* (Gennadius) (Faria and Wraight, 2001, Oliveira *et al.*, 2001).

Trialeurodes vaporariorum, also known as the 'Glasshouse' or the 'Greenhouse whitefly' is a common sap-feeding pest found in the temperate regions. Although widely distributed in Europe, in UK and northern European countries it is primarily found in crops grown in greenhouses *(Martin et al.*, 2000). It is broadly polyphagous and is a major pest of ornamental and edible crops (Brødsgaard and Albajes, 1999). All life stages, except egg, can cause damage to plants by direct feeding on leaves and inducing physiological disorders (Wardlow *et al.*, 1976). They can also affect the host by excretion of 'honeydew' (excreted sugars) which prevents leaves from functioning properly, and encourages sooty moulds (Gorman *et al.*, 2007). Heavy infestation can lead to leaf necrosis, stunted growth and induce physiological disorders (Byrne *et al.*, 1990). *T. vaporariorum* is known to transmit a handful of plant viruses belonging to the genus Crinivirus. These include disease causing plant viruses such as

Tomato infectious chlorosis virus (TICV), Tomato chlorosis virus (ToCV), Strawberry pallidosis associated virus (SPaV) and Beet pseudo-yellows virus (BPYV) (Wintermantel, 2004, Jones, 2003).

In recent years the Greenhouse whitefly has developed resistance against several classes of insecticides used in whitefly control*, such as* neonicotinoids (Karatolos *et al.*, 2010). Mechanisms of insecticide resistance include mutation within the target site of the insecticide or an alteration in the rate of detoxification of the toxic compounds. The latter involves enhanced metabolism or sequestration of insecticides by three enzymes superfamilies – cytochrome P450 (P450s), glutathione-S-transferases (GSTs) and carboxyl/cholinesterases (CCEs) (Ranson *et al.*, 2002).

Although *T. vaporariorum* is an economically important pest, it had limited genomic resources available. Prior to this study only 43 nucleotide sequences were publically available. On the other hand, much of the genomic work is focused on *B. tabaci* with significant amount of sequencing data and an ongoing genome project (Leshkowitz *et al.*, 2006). Identifying and characterising the diverse genes encoding the detoxification enzymes cannot be achieved by traditional methods and requires significant amount of sequencing data. The most cost effective method to gain access to the sequencing information of an organism is by utilising next generation sequencing technologies, such as 454 pyrosequencing. Transcriptome sequencing is a desirable alternative to engage in the genomics of organisms that lack a fully sequenced genome (Morozova and Marra, 2008).

The current study utilises the *T. vaporariorum* transcriptome data generated by 454 pyrosequencing. The main objective of this chapter was to identify and annotate enzymes involved in xenobiotic metabolism in *T. vaporariorum.* Trancripts encoding the three enzymes superfamilies (P450s, GSTs and CCEs) were identified from the transcriptomic data. The second objective was to identify transcripts that have potential role in conferring insecticide resistance in this species. These candidates would further facilitate the research involving the characterisation of molecular mechanism underlying insecticide resistance in *T. vaporariorum*. The data generated forms part of the paper published in the journal BMC genomics (Karatolos *et al.*, 2011)

2.2 METHODOLOGY

2.2.1 Libraries and sequencing

In order to identify as many genes encoding detoxification as possible and to generate a reliable transcriptome reference, two different strains of *T. vaporariorum* were used to generate cDNA libraries. One was insecticide resistant strain (TV6) obtained from Turkey and the other was insecticide susceptible strain (TV1). Two separate cDNA libraries were prepared from adults of TV1 strain and TV6 strain. Enriched, full length cDNAs were generated from 2μg RNA using SMART PCR cDNA synthesis kit (BD Clonetech). PrimeScript reverse transcription enzyme (Takara) was used to perform reverse transcription. In order to reduce over abundant transcripts, the double stranded cDNA was further normalised using the Trimmer cDNA normalisation kit (Evrogen). The two cDNA libraries were tagged prior to 454 pyrosequencing using molecular barcodes (Multiplex identifiers, Roche). The tagged and normalised cDNA libraries were 454 sequenced at the Advanced Genomic facility at the University of Liverpool (Karatolos *et al.*, 2011).

2.2.2 454 transcriptome assembly

The raw reads obtained by 454 pyrosequencing were preprocessed (removal of Poly-A tails and SMART adaptors) and assembled using our in house pipeline called est2assembly (Papanicolaou *et al.*, 2009). A pool of processed reads from both cDNA libraries (TV1 and TV6) were clustered using the MIRA v2.9.26x3 assembler with the *de novo*, normal, EST, 454 parameters. Specifying a minimum read length of 40 nt, a minimum sequence overlap of 40 nt and a minimum percentage overlap identity of 80% (Karatolos *et al.*, 2011).

2.2.3 Manual curation of sequences encoding xenobiotic metabolising enzyme superfamilies

Sequences encoding cytochrome P450s (P450s), glutathione-S-transferases (GSTs) and carboxy/cholinesterases (CCEs) were identified using previously annotated *Acyrthosiphon pisum* (Pea aphid) P450, GSTs and CCEs protein sequences. The *A. pisum* protein sequences were checked for the presence of the domain for the corresponding superfamily using interproscan. Each full length *A. pisum* reference protein sequence was used as input for BLAST search against the assembled *T. vaporariorum* transcriptome. TBLASTn searches were conducted using standalone BLAST. To ensure exhaustive search, each contig was further BLAST searched to obtain the reads corresponding to the particular contig. All the reads obtained were reassembled using geneious software (Drummond *et al.*, 2009). The consensus sequence was checked for frame shifts using NCBI BLASTx. Each contig was manually corrected for any misassembly, where possible. The final corrected consensus sequence was translated using EXPASY translate tool. Longest open reading frame was selected to carry out BLASTp searches to confirm the presence of the P450s, GSTs and CCEs superfamilies and to check if the sequence has complete or partial protein domain.

The curated cytochrome P450 sequences were sent to Dr David Nelson (cytochrome P450 consortium) for further analysis and to assign them names. In case of GSTs and CCEs classification was done on the basis of phylogenetic analysis result.

2.2.4 Phylogenetic analysis

Consensus phylogenetic trees were constructed for each superfamily. All the putative full length and partial protein sequences were further analysed to classify them into individual gene families/clades. The curated *T. vaporariorum* sequences for each superfamily along with the corresponding reference protein sequences from other insect genome were aligned. MUSCLE software (Edgar, 2004) was used to perform multiple sequence alignment. This alignment was manually refined, where required. For the construction of the phylogenetic tree, the Neighbor-Joining (NJ) method with bootstrap analysis of 1000 replicates was applied to the multiple sequence alignment using MEGA 4.0 (Tamura *et al.*, 2007).

2.3 RESULTS

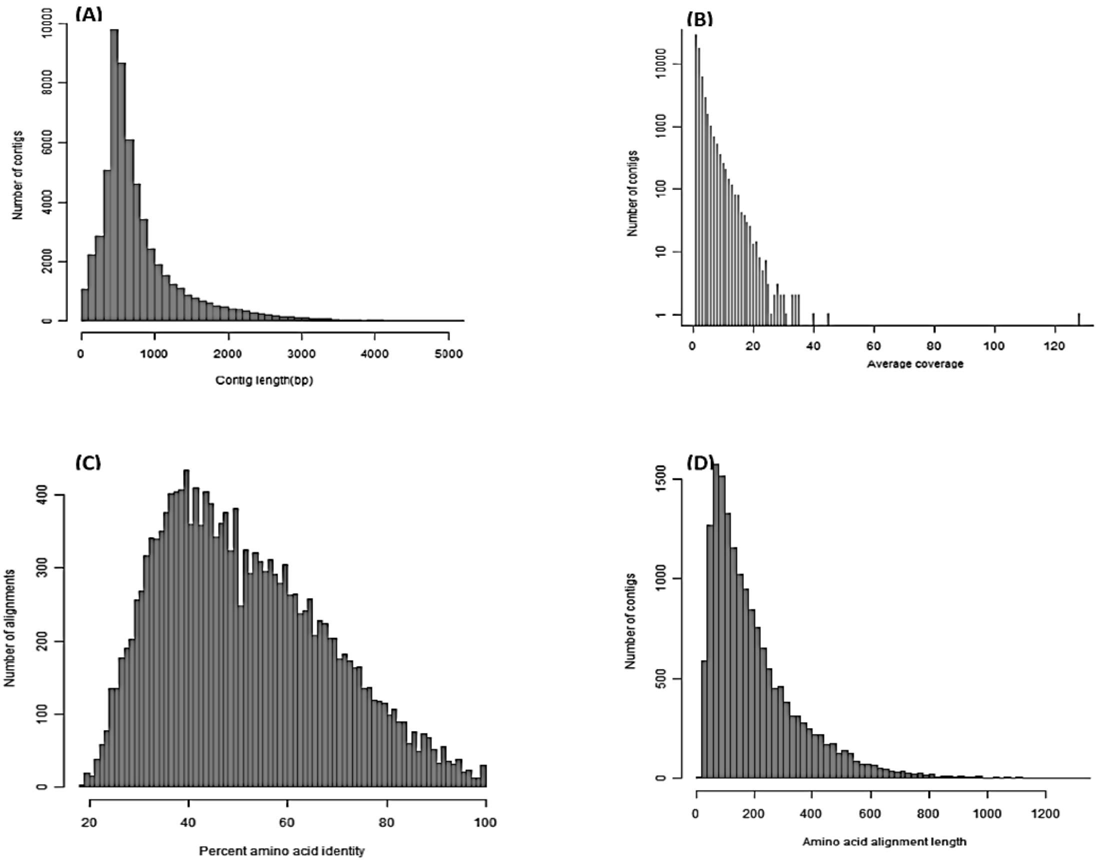
2.3.1 454 transcriptome assembly

In total, 1,104,651 reads were obtained by sequencing the two normalised cDNA libraries. After pre-processing 990,945 high quality reads with an average length of 362 bp were selected for the assembly. Reads from both libraries were pooled to generate a single assembly which resulted in 54,748 contigs with an average length of 965 bp (Table 2.1). Figure 2.1 demonstrates the characteristics of the assembled *T. vaporariorum* 454 contigs and BLASTx alignments against the *D. melanogaster* uniprot database.

Table 2.1 Summary statistics for *Trialeurodes vaporariorum* transcriptome assembly

Total number of reads	1,104,651
Number of reads after preprocessing	990,945
Average read length after preprocessing	362 bp
Total number of contigs	54,748
Average contig length	965 bp
Number of bases	52,832,938 bp
Average read coverage per contig	4.34
Percent GC content of contigs	37.77%

Figure 2.1 Characteristics of assembled *Trialeurodes vaporariorum* 454 contigs and BLASTx alignments against *Drosophila melanogaster*. (A, B) length and coverage of contigs, (C, D) percent identity and deduced amino acid alignment length for all blast hits against *D. melanogaster* predicted proteins.



2.3.2 Manual curation of xenobiotic metabolising enzyme superfamilies

Representatives of all the three enzyme superfamilies (P450s, GSTs and CCEs) were identified in the *T. vaporariorum* transcriptome.

2.3.2.1 Putative cytochrome P450 transcripts

Analysis of *T. vaporariorum* transcriptome identified a total of 123 P450 related contigs. 57 contigs were manually curated and the rest were excluded as they were found to be either allelic variants of the same P450 gene or contained a large number of sequencing errors. All the curated sequences were sent to Dr David Nelson (Cytochrome P450 naming consortium) to assign them correct names. 54 sequences were named in accordance with the P450 nomenclature committee convention and three were assigned partial names (Appendix A Table S2.1). The sequences with partial names were truncated and contained incomplete P450 domains. 37 P450s were found to represent full length ORFs. Additionally, 20 P450-related sequences in the *T. vaporariorum* EST data had incomplete P450 domains. Representatives of all the four major insect clades (clade2, 3, 4 and mitochondrial clade) were found in the *T. vaporariorum* dataset (Figure 2.2). The majority of *T. vaporariorum* P450 sequences belonged to CYP3 clade (34), followed by those related to the CYP4 clade (13) and 3 and 7 members in CYP2 and mitochondrial clades, respectively (Table 2.2).

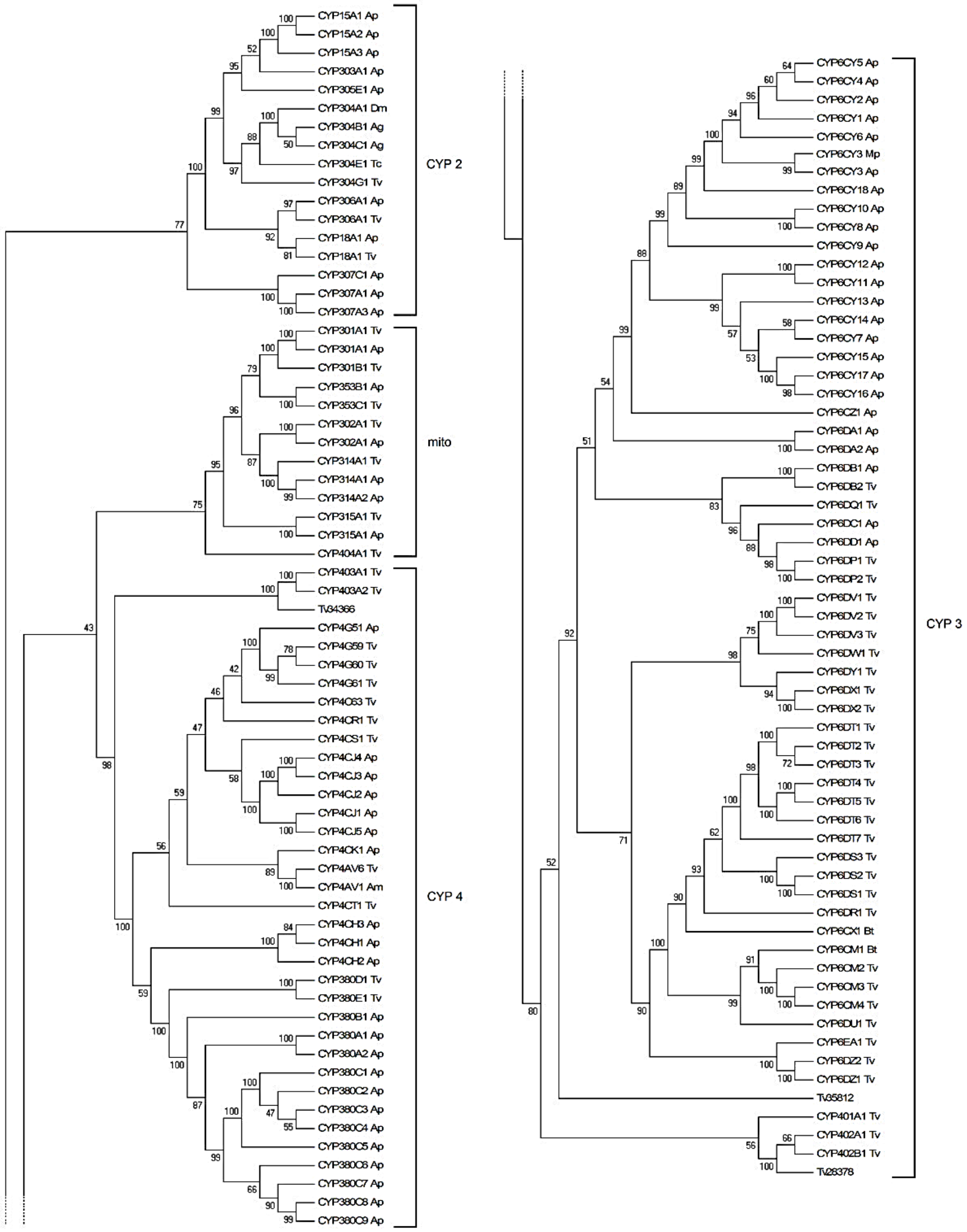
Table 2.2 Number of annotated insect cytochrome P450s in the order Hemiptera and their distribution across different clades in *Trialeurodes vaporariorum*, *Myzus persicae* and *Acyrthosiphon pisum*

	HEMIPTERA		
CLADE	*Acyrthosiphon pisum*	**Myzus persicae*	* ***Trialeurodes vaporariorum***
CYP4	32	48	**13**
CYP3	33	63	**34**
CYP2	10	3	**3**
Mitochondrial CYP	8	1	**7**
P450 Total	83	115	**57**

(* numbers based on transcriptome data)

Data taken from the this study and (Ramsey *et al.*, 2010)

Figure 2.2 Unrooted Neighbour joining tree showing the phylogenetic analysis of cytochrome P450s (P450s) proteins *of Trialeurodes vaporariorum* (Tv) in relation to P450 proteins from *Drosophila melanogaster* (Dm), *Anopheles gambiae* (Ag), *Apis mellifera* (Am), *Acyrthosiphon pisum* (Ap), *Myzus persicae* (Mp), *Bemisia tabaci* (Bt) and *Tribolium castaneum* (Tc). Bootstrap values next to the nodes represent the percentage of 1000 replicate trees that preserved the corresponding clade. Positions containing alignment gaps and missing data were eliminated with pairwise deletion.



2.3.2.2 Putative glutathione-S-transferase transcripts

Analysis of *T. vaporariorum* transcriptome identified a total of 44 GST related contigs. 17 unique contigs were manually curated of which 13 sequences were found to be full length. The number of putative GSTs found in *T. vaporariorum* was comparable to the other sequenced Hemipterans (*A. pisum*-20 and *M. persicae*-20) (Table 2.3). Insects generally have microsomal GSTs and six different classes of cytosolic GSTs (Chelvanayagam *et al.*, 2001). These contigs were assigned to various GST classes by using phylogenetic analysis and closest BLAST hits in the NCBI nr database (Figure 2.3). The *T. vaporariorum* transcriptome dataset provided evidence for one microsomal GST and 16 cytosolic GSTs in the delta (9), epsilon (1), sigma (5) and zeta (1) classes. However, it lacked evidence for the omega and zeta classes (Table 2.3). The majority of the identified GSTs were assigned to the delta class (9). In other insects, members of this class are known to play a significant role in insecticide detoxification (Claudianos *et al.*, 2006). The number of delta class GSTs in *T. vaporariorum* is similar to the number found in *A. pisum.* Furthermore, one member each for epsilon and zeta classes was found, these classes are absent in other sequenced Hemipterans (*A. pisum* and *M. persicae*).

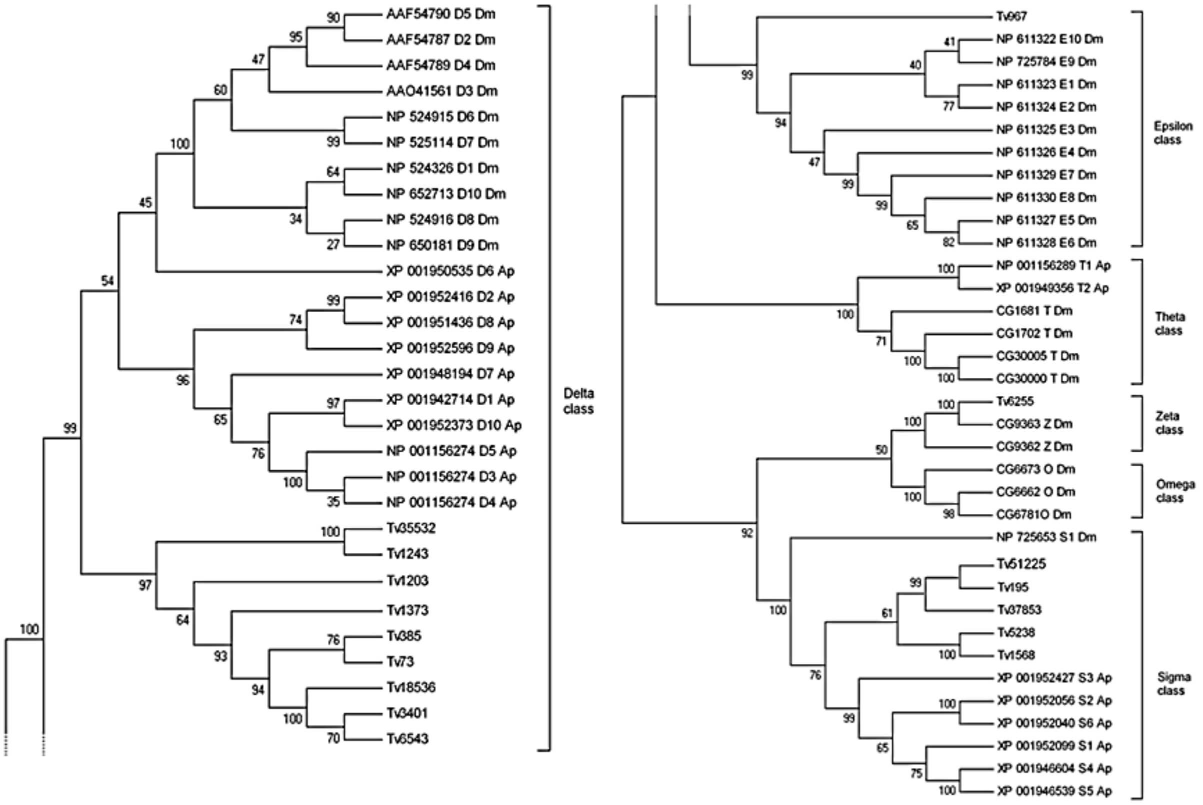
Table 2.3 Number of annotated insect glutathione-S-transferases in the order Hemiptera and their distribution across different clades in *Trialeurodes vaporariorum*, *Myzus persicae* and *Acyrthosiphon pisum*

CLASS	HEMIPTERA *Acyrthosiphon pisum*	 **Myzus persicae*	 * ***Trialeurodes vaporariorum***
Cytosolic GSTs			
Delta	10	8	**9**
Epsilon	0	0	**1**
Omega	0	0	**0**
Sigma	6	8	**5**
Theta	2	2	**0**
Zeta	0	0	**1**
Microsomal GSTs	2	2	**1**
GST Total	20	21	**17**

(* numbers based on transcriptome data)

Data taken from the this study and (Ramsey *et al.*, 2010)

Figure 2.3 Unrooted Neighbour joining tree showing the phylogenetic analysis of predicted cytosolic glutathione-S-transferases (GSTs) proteins of *Trialeurodes vaporariorum* (Tv) in relation to GST proteins from *Drosophila melanogaster* (Dm) and *Acyrthosiphon pisum* (Ap) (accession numbers given). Bootstrap values next to the nodes represent the percentage of 1000 replicate trees that preserved the corresponding clade. Positions containing alignment gaps and missing data were eliminated with pairwise deletion.



2.3.2.3 Putative carboxyl/cholinesterase transcripts

A total of 78 contigs were identified with the CCE protein domain in the *T. vaporariorum* transcriptome. 27 unique contigs were manually curated of which14 sequences were found to be full length. The overall number of putative CCEs identified was within the range of 22-29 found in other sequenced insects of the same order (*A. pisum*-29 and *M. persicae*-22). Classification of the putative *T. vaporariorum* CCEs was based on the closest BLAST hits in the NCBI nr database and by phylogenetic analysis. Carboxyl/cholinesterases (CCEs) can be classified into thirteen clades and three distinct classes (Ramsey *et al.*, 2007). Nine of these clades were represented in the *T. vaporariorum* transcriptome (Figure 2.4, Table 2.4). CCE clades absent in *T. vaporariorum* were clade B (alpha esterase), clade D (integument esterase) and clade F and G (Lepidopteran juvenile esterase). The majority of the CCEs belonged to the dietary/detoxification class (12), which includes esterases involved in the detoxification of insecticides. Compared to the other sequenced Hemipterans, *T. vaporariorum* showed a potential expansion in this class (Table 2.4). On the other hand, there was an apparent contraction in the hormone/pheromone processing class with reduced number of members in clade E as compared to the other sequenced Hemipterans. However, it was difficult to acertain that genes in this clade have been lost from the genome or was just absent in this dataset. The neuro/development class had nine members in this class as compared to seven in *A. pisum* and five in *M. persicae*. With the exception of clade J (acetylcholinesterase), all the other clades of this class are mainly non–catalytic and are involved in cell interactions (Ramsey *et al.*, 2010).

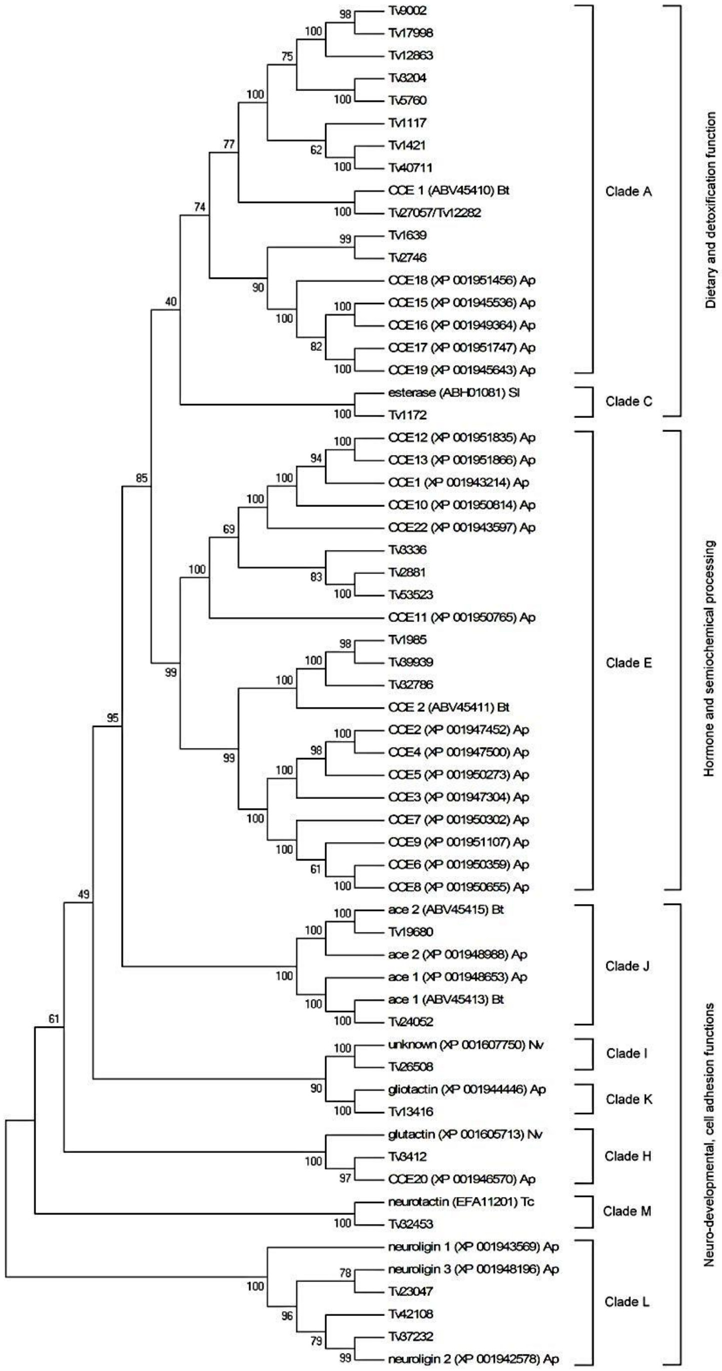
Table 2.4 Number of annotated insect carboxyl/choliesterases in the order Hemiptera and their distribution across different clades in *Trialeurodes vaporariorum*, *Myzus persicae* and *Acyrthosiphon pisum*

CLASS	HEMIPTERA		
	Acyrthosiphon pisum	**Myzus persicae*	* ***Trialeurodes vaporariorum***
Dietary class			
A clade	5	5	**11**
B clade	0	0	**0**
C clade	0	0	**1**
Hormone/semiochemical processing			
D clade	0	0	**0**
E clade	18	12	**6**
F clade	0	0	**0**
G clade	0	0	**0**
Neuro/developmental			
H clade	1	0	**1**
I clade	0	1	**1**
J clade	2	3	**2**
K clade	1	1	**1**
L clade	3	0	**3**
M clade	0	0	**1**
CCE Total	29	22	**27**

(* numbers based on transcriptome data)

Data taken from the this study and (Ramsey *et al.*, 2010)

Figure 2.4 Unrooted Neighbour joining tree showing the phylogenetic analysis of predicted carboxyl/cholinesterases (CCEs) proteins of *Trialeurodes vaporariorum* (Tv) in relation to CCE proteins from Acyrthosiphon pisum (Ap), *Bemisia tabaci* (Bt), *Nasonia vitripennis* (Nv), *Spodoptera littoralis* (Si) and *Tribolium castaneum* (Tc) (accession numbers given). Bootstrap values next to the nodes represent the percentage of 1000 replicate trees that preserved the corresponding clade. Positions containing alignment gaps and missing data were eliminated with pairwise deletion.



2.4 DISCUSSION

Herbivore insects can be broadly classified as generalist and specialist, where specialists feed on a small number of closely related plants species and generalists consume a variety of plants (Bernays and Graham, 1988). *Trialeurodes vaporariorum* falls under the category of generalist as it forages on a diverse range of plant families. The generalists are exposed to a wider range of environmental chemicals (Ramsey *et al.*, 2010). In order to protect themselves from the harmful xenobiotics, insects can either avoid such environments or develop a mechanism for disarming these toxic compounds. *T. vaporariorum* and other insect species, rely heavily on three enzyme superfamilies (P450s, CCEs and GSTs) for metabolising xenobiotics such as plant secondary metabolites and insecticides. Annotation of these three enzyme superfamilies in the sequenced insect genomes shows a large variation (Oakeshott *et al.*, 2010). It has been hypothesised that organisms that encounter diverse xenobiotic substances in their environment, tend to have a significantly higher number of P450s, GSTs and CCEs (Ranson *et al.*, 2002).

Representatives of all the three enzyme superfamilies were found in the transcriptome of *T. vaporariorum*. The data suggests that *T. vaporariorum* has modest detoxification capabilities. Of the three superfamilies, putative P450s in particular are lower than the number found in other sequenced Hemipterans. In contrast, the number of CCEs and GSTs were quite comparable to other sequenced insects of the same order. Comparative analysis of *T. vaporariorum* dataset with other insects of the same order revealed several putative candidates that could be involved in insecticide resistance. Furthermore, it highlighted the general and species specific trends amongst sequenced Hemipterans, namely *T. vaporariorum*, *A. pisum* and *M. persicae*.

There is evidence for P450 evolution through tandem duplication. Duplication of P450s in *M. persicae* has been associated with insecticide resistance (Puinean *et al.*, 2010). Phylogenetic analysis of the *T. vaporariorum* dataset revealed species specific duplication events. One of the most apparent examples of this is the set of three sequences (CYP4G59, CYP4G60 and CYP4G61) in the CYP4 clade (Figure 4.2).

Over expression of *CYP6M1* and *CYP6CY3* contribute to neonicotinoid insecticide resistance in Hemipterans, *B. tabaci* and *M. persicae* (Puinean *et al.,* 2010, Karunker *et al.,* 2008). In *T. vaporariorum* the closet homolog to *CYP6CM1 was CYP6CM2* and *CYP6CM3* with 68% and 67% similarity, respectively. And *CYP6DP1 and CYP6DZ1* had 60% and 59% similarity to *CYP6CY3*. Several members of the CYP3 and CYP4 clades, particularly CYP6, CYP9 and CYP4 families, have been linked with insecticide resistance. Therefore, *CYP6CM2, CYP6CM3, CYP6DP1, CYP6DZ1* and other members of CYP3 and CYP4 clades identified in this study could be implicated in imparting insecticide resistance to *T. vaporariorum*. However, their potential role in insecticide resistance would require further investigation.

The number of GSTs identified in the *T. vaporariorum* dataset was within the range of GSTs identified in other insects of the same order. The insect specific GST classes, delta and epsilon are often associated with detoxification. On the other hand, the sigma class of GSTs is associated with oxidative stress response. Therefore it is not surprising that in Hemipterans, highest number of GSTs belong to the delta and sigma class. The epsilon and zeta classes had one member each in *T. vaporariorum* but these classes were missing in *A. pisum* and *M. persicae*. Absence of these two classes in *A. pisum* and *M. persicae* suggests that these GST classes maybe be absent from all aphids in general. The omega class was found to be absent in all the sequenced Hemipterans (*T. vaporariorum*, *A. pisum* and *M. persicae*) which could be true for all members of this order. The theta class had two members each in *A. pisum* and *M. persicae* but it was absent in *T. vaporariorum*. However, it is difficult to acertain whether the lack of a particular GST class in *T. vaporariorum* is due to its absence in the current transcriptomic dataset or these genes have been lost from the genome.

In case of the CCEs, the esterases involved in detoxification of xenobiotics belong to the dietary class. The main difference in the number of CCEs between sequenced Hemipterans was found to be associated with clade A of this class and Clade E of hormone/semiochemical processing class. *T. vaporariorum* had twice as many CCEs in clade A (11) as compared to *A. pisum* and *M. persicae* which had 5 members each. Phylogenetic analysis suggests a potential expansion of clade A in *T. vaporariorum* (Figure 4.4). One of the putative members of clade A (contig 12282) had high homology to COE1 gene (accession number ABV45410) of *B. tabaci* which is found to be over expressed in strains resistant to organophosphates (Alon *et al.*, 2008). Another member of the same clade (contig 12863) was identified in only TV6 library which was constructed using imidacloprid resistant strain. This could therefore have a potential role in the neonicotinoid resistance of this strain. Another clade associated with insecticide resistance is clade J (acetylcholinesterase), members of this clade are targets for organophosphate and carbamate insecticides (Grafton-Cardwell *et al.*, 2004). Two putative acetylcholinesterases were identified in this dataset, which could also have a potential role in insecticide resistance.

2.5 CONCLUSION

Trialeurodes vaporariorum is a major pest of ornamental and edible crops that has developed resistance against several insecticides. Until recently, research involving the characterisation of molecular mechanism underlying insecticide resistance was hampered due to lack of genomic information. 454 pyrosequencing of the transcriptome generated 54,748 unique contigs which has dramatically improved the sequence information available for this species. Analysis of the *T. vaporariorum* transcriptome has revealed several gene candidates for the biosynthesis and detoxification of xenobiotics. Potential detoxification genes represented in this dataset included 17 GSTs, 57 P450s and 27 CCEs. The result suggests that *T. vaporariorum* has a wide range of xenobiotic metabolising enzymes in the same classes as other sequenced Hemipterans, but with distinct changes in the population of each clade/class. This suggests that these three superfamilies expand independent of the clade/class and depend on the substrate specificity required to cope with a new xenobiotic stress. Comparison with the other members of this order shows that the dataset represent nearly complete collection of such genes from the three superfamilies. The elucidation of the function of the various classes and clades of these superfamilies can provide an important insight into the ability of *T. vaporariorum* to respond to environmental changes. The results also prove that the 454 pyrosequencing of ESTs is a good technique for gene discovery in the absence of the reference genomic resource.

CHAPTER 3

Annotation of the cytochrome P450 gene superfamily in *Heliconius melpomene*

3.1 INTRODUCTION

The cytochrome P450s (CYP) are a diverse class of enzymes found ubiquitously in nature (Alzahrani, 2009). In insects they are found in almost all tissues and are involved in many important tasks, including steps in the insect hormone synthesis pathways and the metabolism of foreign compounds (Scott, 2008). Insect genomes typically contain in the range of 50 to 150 *CYP* genes (Feyereisen, 1999). The insects have twice as many P450s as mammals but only a third of the number found in plants (Feyereisen, 2006). P450s have been characterised into four major clades, each containing some CYP families from vertebrate species: CYP2, CYP3, CYP4 and the mitochondrial clade.

The advent of next generation sequencing technology has made it possible to sequence genomes rapidly. This has given the opportunity to examine the genomics of large number of non-model organisms. Ever growing numbers of genome sequences have made it possible to identify more and more P450s and a variety of novel characteristics have been revealed. With the dawn of whole genome sequencing and the establishment of expressed sequenced tag (EST) libraries it has become possible to identify every *CYP* gene in a genome and thus make valid comparisons between species. Further, the recognition of different CYP clades, coupled with functional analyses of the enzymes within these clades, has allowed insights to be made into the evolution of *CYP* genes and connections to be made with insect life histories. For example, the honeybee *Apis mellifera* has a total *CYP* complement of just 46 genes, and has a very different distribution of *CYP* genes across the different clades than seen in other insects (Consortium, 2006). For example, *A. mellifera* is particularly depauperate in clade 4 CYPs, having only four sequences. Because many clade 4 enzymes are implicated in environmental response and the detoxification of xenobiotics, this finding offers one explanation for the susceptibility of honeybees to pesticides (Claudianos *et al.*, 2006).

The CYP2 and mitochondrial clades contain conserved genes involved in essential physiological functions, such as the *Halloween* genes that mediate synthesis of the moulting hormone 20-hydroxyecdysone (Rewitz *et al.*, 2006), as well as rapidly evolving P450s that have broader substrate specificities and are implicated in the metabolism of xenobiotics. The most conserved *CYP* genes are in clade 2, whose members encode enzymes with essential physiological roles such as *Phantom* (*CYP18A1*), a *Halloween* gene that encodes a 25-hydroxylase which functions in the metabolic pathway leading to synthesis of the insect moulting hormone 20-hydroxyecdysone (Rewitz *et al.*, 2007, Feyereisen, 2006). The number of clade 2 *CYP* genes, however, varies little between insects.

Genes in the CYP3 clan are numerous and are often found in gene clusters that in some cases appear to be evolutionarily ancient and in others represent remarkably recent duplication events. While the CYP3 and CYP4 clades include genes that encode enzymes with important physiological and developmental roles, their best-characterized members are associated with resistance to pesticides. For example, *CYP6G1* has been implicated in resistance to DDT in *Drosophila* and overexpression of *CYP6Z1* is associated with pyrtheroid resistance in *Anopheles* (Daborn *et al.*, 2002, Nikou *et al.*, 2003). The largest P450 gene set so far documented is in the mosquito *Aedes aegypti* (Nene *et al.*, 2007). In this species there appears to be a large expansion of *CYP9* genes, belonging to clade 3, which has resulted from multiple gene duplication events that have occurred since the *Aedes* lineage diverged from that of the malaria vector *Anopheles gambiae* (Waterhouse *et al.*, 2008). *CYP9* genes can be differentially induced in the moth *Manduca sexta* by supplementing diets with plant allelochemicals and xenobiotics (Stevens *et al.*, 2000), and are constitutively overexpressed in some insecticide-resistant strains of another Lepidopteran, *Helicoverpa armigera* (Yang *et al.*, 2006).

Related CYP enzymes function in the metabolism of plant defensive compounds encountered in the diets of herbivorous insects, and the evolution of CYPs has allowed some species to adapt to host plants that are unavailable to others (Wen *et al.*, 2006). *H. melpomene* are monophagous as larvae and individuals of this species are found almost exclusively on *Passiflora menispermifolia* (Naisbit, 2001). Consequently, we may expect that the *CYP* gene set involved in the metabolism of dietary toxins will be narrow and specialised. A recent analysis of *CYP* gene counts in a generalist and a specialist species of aphid has supported this hypothesis, with the gene superfamily apparently being at least 40% larger in the polyphagous *Myzus persicae* and the greatest expansions being in clade 3 (Nene *et al.*, 2007). However, *H. melpomene* are not digestively restricted and grow equally well on several different species of *Passiflora* (Smiley, 1978). This suggests that the choice of a single host plant is not associated with a lack of detoxification genes and an inability to digest the defensive chemicals of different *Passiflora* but likely results from ecological factors such as differential ant predation or competition from other *Heliconius* species.

The genome of the *H. melpomene* butterfly was sequenced recently (The *Heliconius* Genome Consortium, 2012). The current study was part of the analysis carried out for the *H. melpomene* genome consortium. The main aim of this project was to identify all the CYP sequences found in the genome and transcriptome of *H. melpomene*. The annotated *CYP* genes were classified into different clades and their number compared to the validated cytochrome P450s found in other insect genomes. The results of the *H. melpomene* genome wide analysis of the cytochrome P450s are reported in this chapter.

3.2 METHODOLOGY

3.2.1 454 transcriptome assembly

For the assembly, all available *Heliconius melpomene* transcriptome data for wing disc and midgut (accession numbers SRX005618, SRX005617 and SRX058058) was assembled with the 6,010 publicly available *H. melpomene* ESTs on NCBI. The assembly was performed using Roche Newbler software version 2.5p1 (454 Life Sciences, Branford, Conneticut, USA) from the command line with the -cdna flag, and SMART IIa adaptor was trimmed from the raw 454 data using the -vt option. Assembly was conducted on a 64-bit UNIX server (running on Debian OS) with 96GB RAM.

3.2.2 Identification of CYPs in the genome

Putative P450 sequences from *H. melpomene* were curated from the assembled 454 transcriptomic sequences and genomic scaffolds, using *Bombyx mori* P450 protein sequences as a reference. The reference protein sequences for the gene superfamily were obtained from NCBI protein database. All the reference sequences were checked for P450 conserved domain using interproscan (Zdobnov and Apweiler 2001) prior to using them for BLAST searches (Altschul *et al.*, 1990). The genome sequences were obtained from the *Heliconius* genome project website (http://www.butterflygenome.org/). TBLASTn searches were conducted using the *H. melpomene* genome scaffolds as the query sequence and *B. mori* P450 sequences as reference. Having identified the location of each C*YP* gene in the *H. melpomene* scaffolds, putative gene sequences were extracted and used in BLAST search to find the closest match in the NCBI database. The best- match alignment was used to inform manual annotation of each gene in Artemis (Rutherford *et al.*, 2000).

3.2.3 Identification of CYPs in the transcriptome

454 transcriptomic sequences, encoding P450s, were identified using previously annotated *B. mori* P450 protein sequences. Each of the *B. mori* reference sequence was used as input for BLAST search against the assembled *H. melpomene* transcriptome. TBLASTn searches were

conducted using standalone BLAST. To ensure exhaustive search, each contig was further BLAST searched to obtain the reads corresponding to the particular contig. All the reads obtained were reassembled using geneious software (Drummond *et al.*, 2009). The consensus sequence was checked for frame shifts using NCBI BLASTx. Each contig was manually corrected for any misassembly. The final corrected consensus sequence was translated using EXPASY translate tool (Gasteiger *et al.*, 2003). The longest open reading frame was selected to carry out BLASTp searches to confirm the presence of the P450 gene superfamily and to check if the sequence has complete or partial protein domain.

3.2.4 Validation of gene models

The MAKER pipeline (Cantarel *et al.*, 2008) generated gene models (obtained from the *Heliconius* genome project website - http://www.butterflygenome.org/), corresponding to *CYP* genes, were validated using the manually annotated P450 sequences. All the curated sequences from the genome and transcriptome were imported into the Apollo genome annotation and viewer tool (Lewis *et al.*, 2002) to help in the validation and refinement of the gene models. The MAKER predicted gene models were manually verified and edited to correct errors in it. The intron splice-sites were determined using the canonical GT/AG rule and using the matching EST sequences when available**.** The peptide sequences of the validated gene models where send to Dr David Nelson (cytochrome P450 consortium) for further analysis and assigning names.

3.2.5 Phylogenetic analysis

To get a better insight into the evolution and duplication of the different P450 gene families, phylogenetic analysis was carried out. The curated *H. melpomene* sequences along with the reference protein sequences from other insect genome were aligned. MUSCLE software (Edgar, 2004) was used to perform multiple sequence alignment (MSA). This alignment was manually refined, where required. For the construction of the phylogenetic tree, the Neighbor-Joining (NJ) method (JTT matrix with different rates among sites, gamma parameter = 1.0, bootstrap test = 1000 replicates) was applied to the MSA using MEGA 5 (Tamura *et al*,. 2007).

3.3 RESULTS

3.3.1 454 transcriptome assembly

The 454 transcriptome assembly had 13,422 isogroups, 36,363 isotigs (with an N50 of 2370bp) and 44,327 contigs (with an N50 of 1370bp), where 'isogroups' are defined as transcripts, 'isotigs' are splice variants and 'contigs' are separate exons. The current assembly of the *Heliconius melpomene* genome (draft genome assembly Version 1 as of 5 July 2012) encodes 100 predicted P450s including one pseudogene (Appendix B TableS3.1). The majority of these sequences belonged to CYP3 clade (43), followed by CYP4 (39) and nine genes each in CYP2 and mitochondrial clades. Distribution of these genes in different clades spread across fifty three scaffolds. Thirty seven of the predicted genes had EST support available (Appendix B TableS3.2). The EST collection consisted of *Heliconius melpomene* transcriptome data for wing disc, midgut and publicly available *H. melpomene* ESTs on NCBI. The overall *CYP* gene count in the genome of *H. melpomene* was consistent with the other fully sequences genomes. However, there were variations in the distribution of the *CYP* genes into different gene families and subfamilies within each clade. The difference can be explained largely by the CYP3 and CYP4 clade. The frequency distribution of *H. melpomene CYP* genes into the four clades closely matched those in *Bombyx mori* and the Dipterans *Drosophila melanogaster* and *Anopheles gambiae* (Table 3.1).

Table 3.1 Number of validated cytochrome P450s in the insect genomes and their distribution across the P450 clades.

Order	Diptera			Hymenoptera		Hemiptera	Coleoptera	Lepidoptera	
CLADE	*Drosophila melanogaster*	*Anopheles gambiae*	*Aedes aegypti*	*Apis mellifera*	*Nasonia vitripennis*	*Acyrthosiphon pisum*	*Tribolium castaneum*	*Bombyx mori*	***Heliconius melpomene***
CYP2	6	10	11	8	7	10	8	7	**9**
CYP3	36	42	84	28	48	33	70	30	**43**
CYP4	32	45	59	4	30	32	44	36	**39**
Mito	11	9	10	6	7	8	9	11	**9**
P450 Total	85	106	164	46	92	83	131	84	**100**

*Data taken from Adams *et al.* (2000), Holt *et al.* (2002), Bin *et al.* (2005), Claudianos *et al.* (2006), Strode *et al.* (2008), T.G.S. Consortium (2008), Ramsey *et al.* (2010), Oakeshott *et al.* (2010) and *Junwen et al.* (2011).

3.3.2 Distribution of *CYP* genes in *H. melpomene*

3.3.2.1 CYP2 and the mitochondrial clade

The distribution of the *CYP* genes in the CYP2 or mitochondrial clades, was consistent with that found in other insect genomes (Table 3.1). The gene count of these two clades is within the range of 6-11 found for most of these clades in the other species. *H. melpomene* has nine members of CYP2 clades which were further distributed into gene families - CYP15, CYP18, CYP303, CYP306 and CYP307 – which had one gene each whereas CYP304 and CYP305 gene families had two genes each. Lowest numbers of *CYP2* genes are found in *Drosophila* (6) and highest in *Aedes aegyptii* (11). The mitochondrial clade also had nine members – eight genes and a single pseudogene. Largest family was CYP333 containing three genes and one pseudogene (Figure 3.1). The largest number of genes in the mitochondrial clade are found in *Drosophila* (11) and lowest number found in *Apis mellifera* (6).

3.3.2.2 CYP3 and CYP4 clades

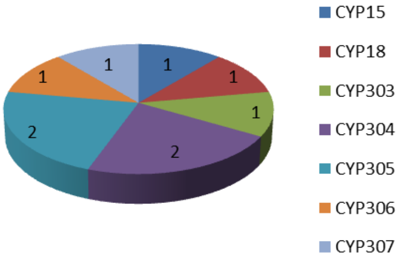
The variation in gene number can be explained largely by the CYP3 and CYP4 clades. The CYP3 clade has been found to be most variable among insect genomes. The majority of the *Heliconius* P450 genes belonged to the

CYP3 clade (43). The highest number belonged to CYP6 family (23) closely followed by CYP9 (7). In the CYP6 family the CYP6AB subfamily had the largest number of genes (9). The largest P450 gene set so far documented is in the mosquito *Aedes aegyptii* (Nene *et al.*, 2007). In this species there appears to be a large expansion of *CYP9* genes, belonging to clade 3, which has resulted from multiple gene duplication events that have occurred since the *Aedes* lineage diverged from that of the malarial vector *A. gambiae* (Waterhouse *et al.*, 2008). The lowest numbers of *CYP3* genes are found in *Apis mellifera* (28). Largest member of CYP4 clade are found in the order Diptera – *Aedes aegypti* (59) and *Anopheles gambiae* (45) and the lowest number found in the order Hymenoptera – *Apis mellifera* with only four members found. In *H. melpomene* CYP4 clade, the largest family was CYP4 (15) followed by CYP340 (9).

Figure 3.1 Distribution of *Heliconius melpomene* cytochrome P450s into the four clades – CYP2, CYP3, CYP4 and Mitochondrial clade. The four clades are further divided into families, subfamilies and isoforms, including pseudogenes. The table represents the number of *CYP* genes in each subfamily. The pie chart shows the distribution of *CYP* genes into different families in each clade.

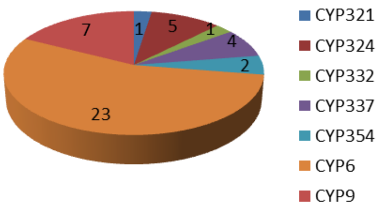
CYP2

Family	Subfamily	gene
CYP15	C	1
CYP18	A	1
CYP303	A	1
CYP304	F	2
CYP305	B	2
CYP306	A	1
CYP307	A	1



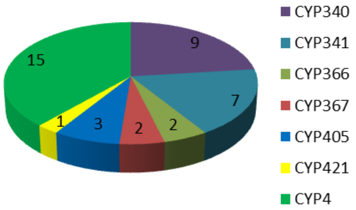
CYP3

Family	Subfamily	gene
CYP321	C	1
CYP324	A	5
CYP332	A	1
CYP337	C	4
CYP354	A	2
CYP6	AB	**9**
	AE	**6**
	AN	5
	CT	1
	FA	2
CYP9	A	6
	G	1



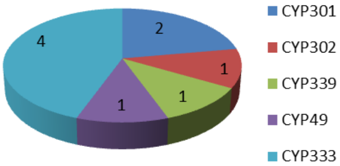
CYP4

Family	Subfamily	gene
CYP340	N	1
	P	1
	Q	1
	R	**6**
CYP341	A	3
	E	3
	F	1
CYP366	B	2
CYP367	A	1
	B	1
CYP405	A	3
CYP421	A	1
CYP4	CG	**9**
	G	2
	L	1
	M	2
	S	1



Mito

Family	Subfamily	gene
CYP301	A	1
	B	1
CYP302	A	1
CYP339	A	1
CYP49	A	1
CYP333	A	1
	B	2
	un (pseudogene)	1



3.3.3 Putative candidates for cyanogenesis

Figure 3.2 represents a bootstrap tree that contains representative full length P450s from *Drosophila melanogaster*, *Anopheles gambiae*, *Apis mellifera*, *Tribolium castaneum*, *Zygaena filipendulae*, *Helicoverpa armigera*, *Manduca sexta*, and *Bombyx mori* along with 100 P450s from *Heliconius melpomene*. The four insect CYP clades are coloured - CYP3 (red), CYP4 (green), CYP2 (pink) and the mitochondrial clade (blue). All *H. melpomene* sequences are shown as bold and the P450s found in clusters on the scaffolds are highlighted. Phylogenetic analysis of protein sequences identified *H. melpomene* CYP3 and CYP4 sequences that have potential orthologies with enzymes with roles in biosynthesis of cyanogenic defense compounds in *Z. filipendulae* (marked as asterisk in Figure 3.2). CYP332A3 of *Z. filipendulae* shows close orthology with CYP332A1 of *H. melpomene* with bootstrap support of 100% (Figure 3.3a). CYP332A clusters in CYP3 insect clade which contained P450s generally involved in xenobiotic metabolism. CYP332A4 and CYP332A5 in *M. sexta* are highly expressed in the midgut and putative common function is more likely to be in the detoxification of xenobiotics. CYP332A3 of *Z. filipendulae* and CYP332A1 of *H. melpomene* clustered with other CYP332A found only in Lepidoptera species (Jensen *et al.*, 2011) however, they formed a separate node, which indicates it has been recruited from the orthologous CYP332A into the cyanogenic pathways. The other P450 involved in cyanogenesis belongs to CYP405A subfamily of CYP4 clade. In *Z. filipendulae* it is represented by CYP405A2. Three member of CYP405A subfamily (CYP405A4, CYP405A5 and CYP405A6) were found in *H. melpomene* (Figure 3.3b). These three CYP405As clustered with the *Z. filipendulae* CYP405As with bootstrap support of more than 90%. CYP332A1, CYP405A4, CYP405A5 and CYP405A6 thus represent the best putative candidates for cyanogenesis in *H. melpomene.*

Figure 3.2 Unrooted Neighbour joining tree showing the phylogenetic analysis of cytochrome P450s (CYP) proteins of *Heliconius melpomene* (shown in bold) in relation to CYP proteins from *Drosophila melanogaster* (Dm), *Anopheles gambiae* (Ag), *Apis mellifera* (Am), *Tribolium castaneum* (Tc), *Zygaena filipendulae* (Zf), *Helicoverpa armigera* (Ha), *Manduca sexta* (Ms) and *Bombyx mori* (Bm). Distance bootstrap values of >70% (1000 replicates) are indicated at the corresponding nodes. The four insect CYP clades are coloured distinctively - CYP3 (red), CYP4 (green), CYP2 (pink) and mitochondrial clade (blue). *H. melpomene* P450s that are found in clusters on the scaffolds have been highlighted. The position of putative candidates for cyanogenesis, are marked with an asterisk.

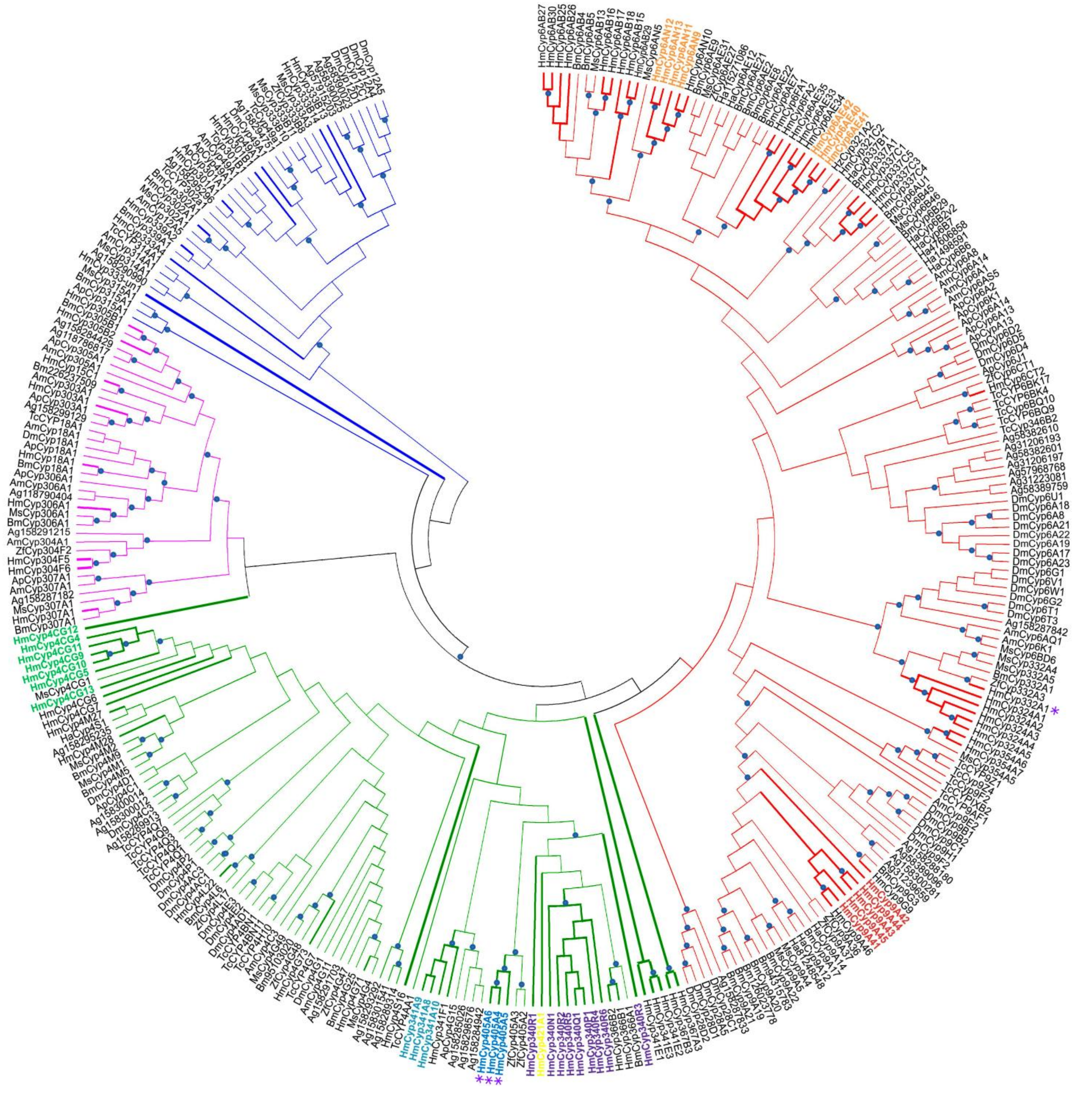
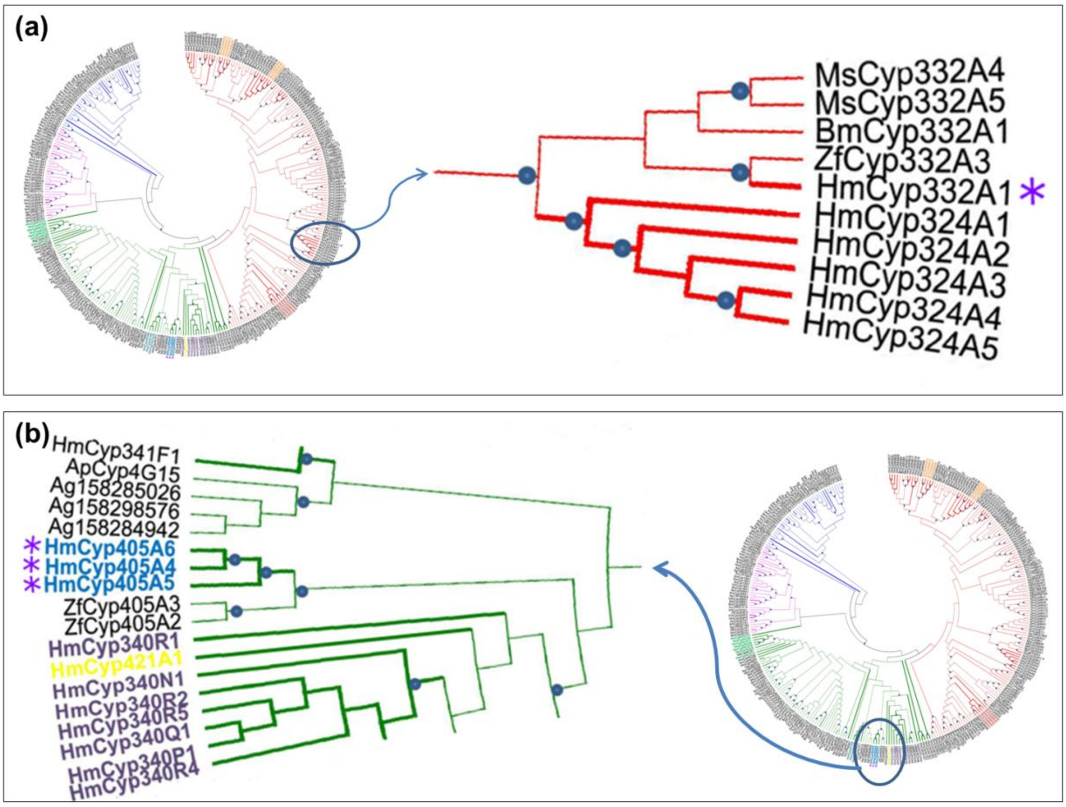


Figure 3.3 Sub tree showing the putative cytochrome P450 candidates for cyanogenesis in *Heliconius melpomene* (shown in bold), (a) putative candidate from CYP332A subfamily and (b) putative candidates from Cyp405A subfamily. The position of putative candidates is marked with an asterisk.

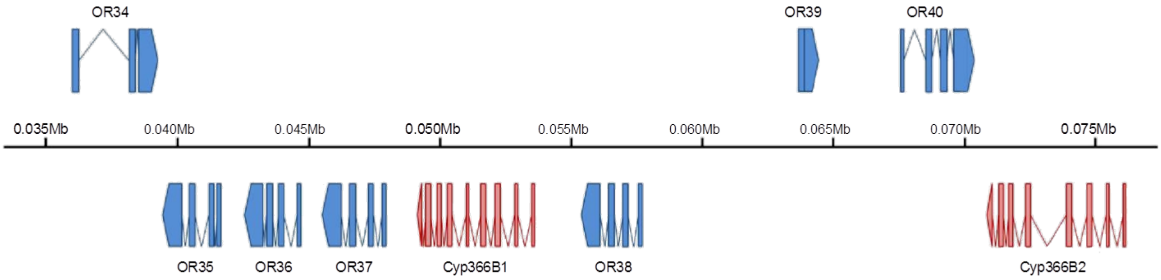


3.3.4 OR P450 gene cluster

Interestingly, two *CYP* genes were also found amongst a cluster of olfactory receptor (OR) genes (Figure 3.4). The *CYP* genes situated alongside the *OR* genes of *H. melpomene* (CYP3336B1 and CYP3336B2) were predicted to encode clade 4 enzymes. Due to the proximity of these genes to *OR* genes, the potential role of the proteins encoded by these genes was also examined. Using a BLAST (Altschul *et al.*, 1990) search it was found that the best match to a characterised protein was 30% amino acid sequence identity to CYP4AW1 of the Scarab beetle *Phyllopertha diversa*. This CYP is specifically expressed in the antenna of *P. diversa*, where it is believed to act as a pheromone-degrading enzyme (Maibeche-

Coisne *et al.*, 2004). The finding of two *CYP* genes with predicted products showing amino acid sequence similarity to this pheromone-degrading enzyme, amongst a cluster of *OR* genes, was intriguing. The possibility that the products of these genes operate in the same system for the detection and metabolism of related compounds is an exciting one and is worthy of further experimental investigation.

Figure 3.4 Diagrammatic representation of the *Heliconius melpomene* scaffold containing clusters of olfactory receptor genes (in red) along with P450 genes (blue). In case of each gene the exon structure is depicted



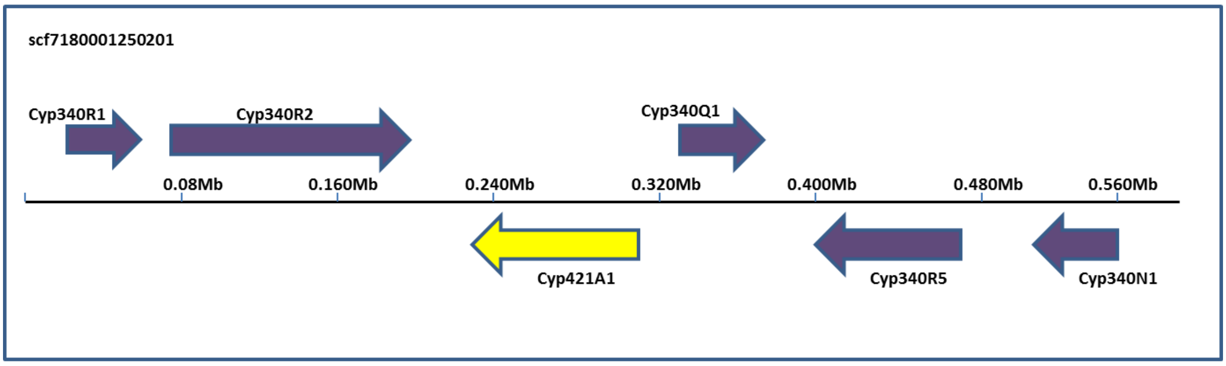
3.3.5 Gene clusters in *Heliconius melpomene*

It has been seen that uneven distribution of P450s in the genome occurs due to clustered organization. In *H. melpomene* 36 P450s were found in nine clusters of three or more (Table 3.2). All the gene clusters belonged to CYP3 or CYP4 clades. Six genes from CYP4 clade formed the largest gene cluster. Three genes were in the forward orientation and the other three genes were in the reverse orientation (Figure 3.5). It consisted of five genes belonging to CYP340 family and one gene from CYP421 family. The presence of large P450 gene clusters, sometimes containing members of different P450 families, has been observed in other insects and plants e.g. *Drosophila melanogaster* and *Arabidopsis thaliana* (Tijet *et al.*, 2000).

Table 3.2 Cytochrome P450 gene clusters in *Heliconius melpomene* genome

Gene	Orientation	HMEL Stable ID	Scaffold number	Gene number/ cluster
CYP6AE40	reverse	-	scf7180001249711	3
CYP6AE41	reverse	-	scf7180001249711	
CYP6AE42	reverse	-	scf7180001249711	
CYP6AN9	forward	HMEL007566	scf7180001249436	4
CYP6AN11	forward	HMEL007566	scf7180001249436	
CYP6AN12	reverse	HMEL007567	scf7180001249436	
CYP6AN13	forward	HMEL007568	scf7180001249436	
CYP9A45	reverse	HMEL013741	scf7180001250505	5
CYP9A41	forward	HMEL013737	scf7180001250505	
CYP9A42	forward	HMEL013742	scf7180001250505	
CYP9A43	reverse	HMEL013739	scf7180001250505	
CYP9A44	reverse	HMEL013738	scf7180001250505	
CYP405A5	reverse	HMEL016673	scf7180001250781	3
CYP405A6	reverse	HMEL016674	scf7180001250781	
CYP405A4	reverse	HMEL016675	scf7180001250781	
CYP4CG5	reverse	HMEL003583	scf7180001245460	4
CYP4CG9	reverse	HMEL003581	scf7180001245460	
CYP4CG10	reverse	HMEL003582	scf7180001245460	
CYP4CG13	reverse	HMEL003584	scf7180001245460	
CYP4CG4	reverse	HMEL008770	scf7180001249777	3
CYP4CG11	reverse	HMEL008768	scf7180001249777	
CYP4CG12	reverse	HMEL008772	scf7180001249777	
CYP340R6	reverse	HMEL013957	scf7180001250535	4
CYP340P1	reverse	HMEL013954	scf7180001250535	
CYP340R3	reverse	HMEL013955	scf7180001250535	
CYP340R4	reverse	HMEL013956	scf7180001250535	
CYP340R2	forward	HMEL010931	scf7180001250201	6
CYP340R5	reverse	HMEL010936	scf7180001250201	
CYP340N1	forward	HMEL010939	scf7180001250201	
CYP340R1	forward	HMEL010930	scf7180001250201	
CYP421A1	reverse	HMEL010934	scf7180001250201	
CYP340Q1	forward	HMEL010935	scf7180001250201	
CYP341A8	forward	HMEL017148	scf7180001250807	3
CYP341A9	forward	HMEL017149	scf7180001250807	
CYP341A10	forward	HMEL017150	scf7180001250807	

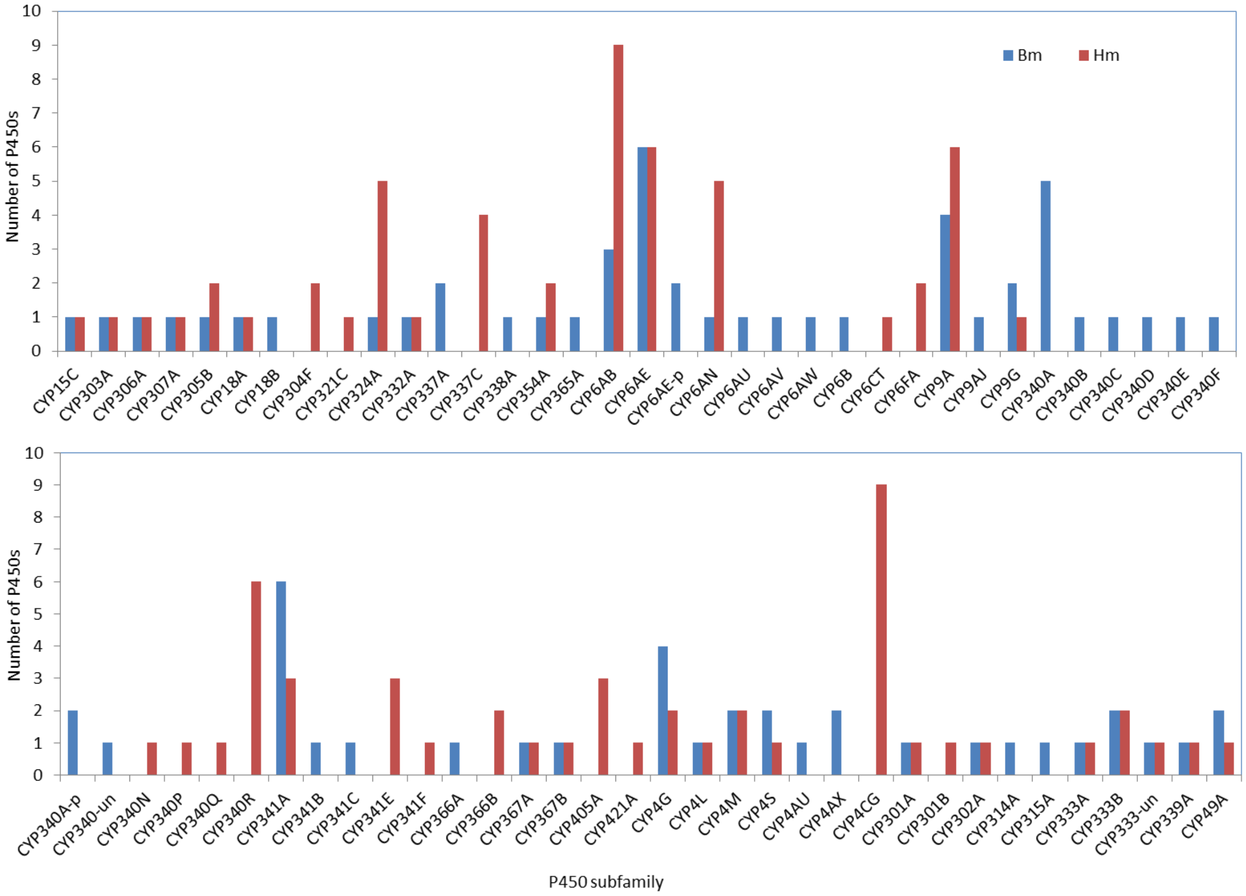
Figure 3.5 Largest cytochrome P450 gene cluster (belonging to CYP4 clade) in *Heliconius melpomene*. The scaffold number is presented on the left hand side of the figure and the approximate position of each gene on the scaffold is depicted. The arrow represents the orientation of the gene on the scaffold



3.3.6 Comparison of *Bombyx mori* and *Heliconius melpomene*

The *B. mori* genome contains 78 genes and 6 pseudogenes. On the other hand in case of *H. melpomene*, 99 genes and only a single pseudogene was found. In *B. mori* 38 P450s are found in eight clusters of 3 or more and 34 singletons. Whereas in *H. melpomene* 36 P450s were found in nine clusters of three or more and 27 singletons (Figure 3.6). However unlike *B. mori* no cluster was found with genes from mixed clades. *B. mori* is the only insect in which clusters from mixed clades (microsomal and mitochondrial) have been reported so far (Junwen *et al.*, 2011). The single pseudogene found in *H. melpomene* was CYP333-un which was also found in *B. mori*, suggesting that this was a true pseudogene rather than an artefact of sequencing or assembling error. Figure 3.6 clearly shows that there are differences in the distribution of genes in different families/subfamilies of P450s in *B. mori* and *H. melpomene* with subfamilies showing genome specific gene expansions.

Figure 3.6 Distribution of P450 subfamilies in *Bombyx mori* and *Heliconius melpomene*



3.4 DISCUSSION

Analysis of the available insect *CYP* genes indicate that they fall into four major clades – CYP2, CYP3, CYP4 and mitochondrial clades. Each insect genome contains about 100 or so P450 genes, all evolved from a common ancestor and each coding for a different P450 enzyme (Feyereisen, 1999). Given the incomplete nature of the genome and the limited transcriptomic data examined it can not be determined that all of the P450 genes in the genome have been sampled here. However the number of P450s in *H. melpomene* was comparable to the number found in other insect genomes. This seemed to follow the general trend across different insect orders with majority of the enzymes belonging to the CYP3 clade followed by the CYP4 clade. CYP2 and mitochondrial clades generally forms 1:1 orthologues for all insects.

The number of P450s in insect genomes currently ranges from 48 in *Apis mellifera* to 164 in *Aedes aegyptii.* In a given species, the majority of P450s are primarily for the detoxification of xenobiotics (Scott, 2008). As more and more insect genomes are being sequenced, it is apparent that P450s have a much broader role in insects. Analysis of other insect genomes suggests that *Heliconius melpomene* appears to have a fairly typical number of P450s. Although the total number of P450s is similar to other insect genomes, it differs in the distribution into various families and subfamilies within each clade. Large scale gene duplication events have been assumed to correspond to major evolutionary changes (De Bodt *et al.*, 2005). Acquisition of new biological functions has been associated with proliferation of gene families followed by functional diversification of paralogues (Ohno, 1970).

Comparing gene families responsible for xenobiotic metabolism among different insects can point to species specific genes. On comparing the distribution of P450s in *B. mori* and *H. melpomene*, two striking differences can be seen in CYP3 and CYP4 clades. The CYP6 family, of CYP3 clade, in *H. melpomene* has undergone gene expansion especially in the CYP6AB subfamily. It has nine members in this subfamily as compared to only three members found in *B. mori.* Another gene subfamily with a large number of genes is CYP4CG of CYP4 clade. It has nine members in *H. melpomene* but no members found in *B. mori* (Figure 3.6). Members of CYP405A subfamily are also missing from *B. mori* genome. It includes one of the two P450s involved in cyanogenesis. *H. melpomene* contains three members of this subfamily which show close orthology to *Z. filipendulae*. Resistance of *Zygaena* species to hydrogen cyanides (HCN) has been well known since the beginning of the twentieth century (Zagrobelny, 2004). The biosynthetic pathway of cynaogenic glucosides is very simple and the entire pathway is encoded by three genes: two cytochrome P450s and a UDP-glycosyltransferase (UGT). All three genes (*CYP332A3, CYP405A2 and UGT33A1*) have been recently identified in *Z. filipendulae* (Jensen *et al.*, 2011). CYP332A3 of *Z. filipendula,* shows close orthology with CYP332A1 of *H. melpomen*e with bootstrap support of 100%. CYP405A2 of *Z. filipendulae* show close orthology to *H. melpomene* CYP405A4, CYP405A5 and CYP405A6, however, at this stage we cannot predict which of the three is most likely to be involved in cyanogenesis. (Figure 3.3).

Co-localisation of cytochrome P450 genes with olfactory receptor (OR) genes was also revealed during annotation of P450 encoding genes in the *H. melpomene* genome. Two *CYP* genes (CYP366B1 and CYP366B2) were found amongst a cluster of *OR* genes. ORs are seven-transmembrane domain G protein-coupled receptors expressed in the antennae and maxillary palps of insects. These receptors bind to environmental chemicals and transform the chemical signal into activation of neurons, allowing the detection and perception of odours that may be important in finding a suitable mate or food source (Sanchez-Gracia *et al.*, 2009). In the genome of *H. melpomene*, 70 *OR* genes have been identified. Several of these exist in small genomic clusters, and appear from phylogenetic analyses to be *Heliconius* specific and therefore the product of recent gene duplication events. The most striking of these is the cluster identified on scf7180001250804 on chromosome 5, which includes the genes

HmelOR35, *36*, *38-40* and the more divergent *HmelOR37* (The *Heliconius* Genome Consortium, 2012).

It is difficult to accurately predict the function of the *OR* genes at this locus. Mouse models suggest that a single OR may respond to multiple odorants, and one odour may activate several ORs (Malnic *et al.*, 1999). Consistent with an ability to recognise diverse ligands, wide variation in primary peptide sequence is found between ORs, and the difficulty of crystallising membrane proteins has meant it is challenging to obtain structural and functional information (Floriano *et al.*, 2000). As a result, the molecular mechanisms underlying OR activity and the response to chemical stimulants remain unclear, and the specificity of an OR may not be readily predicted from the gene sequence that encodes it; whether the chromosome 5 *OR* genes are associated with the response to pheromones or plant odorants awaits further study. In *D. melanogaster*, the *OR* genes are roughly evenly spread throughout the genome (Robertson *et al.*, 2003). Although there are two loci at which an *OR* gene is positioned very close to a *CYP* gene (Or49b is adjacent to Cyp9h1 and Or85e is ~8 kb from Cyp313b1), there are no directly comparable clusters in this species.

3.5 CONCLUSION

Insects have evolved to tolerate plant allelochemicals and insecticides by diverse strategies such as the amplification P450s superfamily. Analysis of the *H. melpomene* genome has revealed several candidates for the biosynthesis and detoxification of xenobiotics. Potential members of this superfamily represented in our dataset included 100 P450s. The result suggests that *H. melpomene* has a wide range of xenobiotic metabolising enzymes in the same clade as other insects, but with distinct changes in the population of each clade/family. This suggests that this superfamily expand independent of the clade/family and depend on the substrate specificity required to cope with a new xenobiotic stress. On comparison with the other insect genomes, it shows that this dataset represent nearly complete collection of P450 genes. The elucidation of the function of the various families and subfamilies of each clade can provide an important insight into the ability of *H. melpomene* to respond to environmental changes. In summary, the annotation of the *CYP* genes in *H. melpomene* finds a total complement that is consistent with other insect genomes. There are no striking expansions or deletions in any of the CYP clades. However, within clades there are differences in gene count between different families, and the phylogenetic analysis (Figure 3.2) reveals where Lepidopteran-specific and *Heliconius*-specific expansions have occurred. Together with the identification of P450s with EST support this study can help to determine the key genes, such as the genes involved in *Heliconius* adaptation to cyanogenic host plants.

CHAPTER 4

Analysis of the *Lucilia sericata* transcriptome: xenobiotic metabolising enzyme superfamilies and SNP detection

4.1 INTRODUCTION

The blowflies *Lucilia sericata* and *Lucilia cuprina* are responsible for causing myiasis in sheep and other animals (Stevens and Wall, 1996). They are facultative ectoparasites that have become the primary pest of domesticated sheep around the world (Hartley *et al.*, 2006). Although similar in morphology and ecology, the different populations of these sibling species are known to vary in their importance as pests in different parts of the world (Stevens and Wall, 1996). The blowfly *Lucilia sericata* is also one of the primary species of blowflies used in maggot therapy (Tourle *et al.*, 2009). Further, *L. sericata* larvae are a potential source of novel antimicrobials, as they produce compounds with antibiotic activities which can influence the formation of biofilms in bacteria (Cazandera *et al.*, 2009). This species, along with other similar species, are used in forensic studies to estimate post mortem intervals (Tarone *et al.*, 2007, Gallagher *et al.*, 2010). In short, *L. sericata* has been associated with numerous areas of research in the field of medical, forensic, ecological, evolutionary and veterinary sciences (Paul *et al.*, 2009, Anderson, 2000, Denno and Cothran, 1976, Mellenthin *et al.*, 2006, Fischer *et al.*, 2004) and is therefore both an important and interesting pest.

Genomic information on *L. sericata* is still sparse, despite its association with numerous areas of research (Sze *et al.*, 2012). This is because the majority of research to date has been focused on its sibling species, *L. cuprina*, which is a model organism for insecticide resistance (Lee *et al.*, 2011). *L. cuprina* has evolved resistance to two forms of organophosphorous insecticides (OP), diazinon and malathion. Diazinon resistance being more prevalent than malathion resistance (Hartley *et al.*, 2006). In Australian *L. cuprina*, two single nucleotide polymorphisms in the *aE7* gene encoding esterase-3 are associated with OP resistance

(Newcomb *et al.*, 2005). However, very little information is available for OP resistance outside of Australasia. Limited information available for *L. sericata* also suggests the occurrence of these mutations in the New Zealand population but at a much lower frequency than within its sibling species *L. cuprina* (Hartley *et al.*, 2006).

In the past the high cost of sequencing could have been one of the main reasons for the limited sequence data available for *L. sericata*. Although genome sequencing costs have reduced drastically, for many non-model organisms whole genome sequencing is still impractical. Transcriptome sequencing, however, provides a cost effective alternative to get a quick access to the genetic information for any organism (Kumar and Blaxter, 2010). Concentrating sequencing efforts on transcribed regions allows analysis of only the expressed part of the genome, which cannot be easily predicted with the genome sequence alone. Sequencing of the normalised cDNA libraries decreases the bias towards sequencing highly expressed genes (Mundry *et al.*, 2012).

Recent advances in next generation sequencing technologies and bioinformatics analysis has made it possible to identify novel genes and gene functions in non-model systems. 454 pyrosequencing is an ideal technique for transcriptome sequencing as the sequence length is comparable to that achieved with traditional Sanger sequencing (Wall *et al.*, 2009). Although the sequence depth is lower than the other high throughput sequencing technologies, 454 generates comparatively longer reads. This makes it possible to generate a reliable assembly even in the absence of a sequenced reference genome, which is usually the case when dealing with non-model organisms (Rothberg and Leamon, 2008). In addition, the newer version of Newbler assembler available for assembling 454 reads, allows reliable *de novo* assembly of the cDNA sequences and grouping together of the presumptive gene isoforms (Nyberg *et al.*, 2012).

In the current study, *L. sericata* reads generated by 454 pyrosequencing were assembled to generate a reference transcriptome. The assembled data was utilised to identify gene families associated with detoxification. Insects mainly rely on three enzyme superfamilies - cytochrome P450s, glutathione-S-transferases and carboxyl/cholinesterases to detoxify potentially toxic xenobiotics in their diets or habitats. Insecticide resistance is generally attributed to the up-regulation of enzymes associated with xenobiotic metabolism and detoxification (Ranson *et al.*, 2002). On the other hand underrepresentation of these genes in an organism makes then more sensitive to insecticides, for example as in the Honey bee, *Apis mellifera* (Consortium, 2006). Therefore the main goal of this project was to identify putative members of these three gene superfamilies in the *L. sericata* transcriptome.

This newly generated genetic information was also used to identify single nucleotide polymorphisms (SNPs) in the UK population. The project utilised Illumina RNA-seq data generated for UK populations to carry out SNP discovery. The Illumina data also enabled a comprehensive assessment of the extent of E3 mutation in the UK. This study can serve as a basis for future SNP analysis of the *Lucilia sericata* population outside Australasia.

4.2 METHODOLOGY

4.2.1 Datasets used

For the reference transcriptome, *Lucilia sericata* samples from five different geographical locations (Australia, New Zealand, South Africa, UK and USA) were obtained to generate a mixed source 'reference' transcriptome. The cDNA library was prepared from pooled whole larvae and adults of five *L. sericata* populations. Enriched, full length cDNA was generated from 2µg RNA using SMART PCR cDNA synthesis kit (BD Clonetech). PrimeScript reverse transcription enzyme (Takara) was used to perform reverse transcription. In order to reduce over abundant transcripts, the double stranded cDNA was further normalised using the Trimmer cDNA normalisation kit (Evrogen). The normalised cDNA library was 454 sequenced at the Advanced Genomic facility at the University of Liverpool.

The Illumina single end RNA-seq data for the UK population, sequenced at University of Exeter, was further used for SNP detection.

4.2.2 454 transcriptome assembly

The assembly was performed using Roche Newbler software version 2.5p1 from the command line with the -cdna flag, and SMART adaptors were trimmed from the raw 454 data using the -vt option. Assembly was conducted on a 64-bit UNIX server (running on Debian OS) with 96GB RAM.

4.2.3 BLAST2GO analysis

The assembled contigs were analysed using automated Blast2GO software suite V2.6.2 (Conesa *et al.*, 2005). The analysis included BLAST searches, functional annotation with GO, InterPro terms (InterProScan), EC codes and metabolic pathways. BLASTx searches were performed remotely on NCBI server against non-redundant (nr) protein database using QBLAST. The program extracted the GO terms associated with blast hits and returned a list of GO annotations which are represented as hierarchical categories with increasing specificity. Next the GO terms were modulated using ANNEX, the annotation augmentation tool followed by GOSlim (consisting of a subset of the GO vocabulary encompassing key ontological terms and a mapping function between full GO and GOSlim). EC codes and KEGG metabolic pathway annotations were obtained by directly mapping the GO terms to their equivalent enzyme codes. InterPro searches were performed on the InterProEBI web server remotely via BLAST2GO.

4.2.4 Manual curation of sequences encoding xenobiotic metabolising enzyme superfamilies

Sequences encoding cytochrome P450s (P450s), glutathione-S-transferases (GSTs) and carboxyl/cholinesterases (CCEs) were identified using previously annotated *Drosophila melanogaster* P450, GSTs and CCEs protein sequences. Each of the full length *D. melanogaster* reference sequence was used as input for BLAST search against the assembled *L. sericata* transcriptome. TBLASTn searches were conducted using standalone BLAST. To ensure exhaustive search, each contig was further BLAST searched to obtain the reads corresponding to the particular contig. All the reads obtained were reassembled using geneious software (Drummond *et al.*, 2009). The consensus sequence was checked for frame shifts using NCBI BLASTx. Each contig was manually corrected for any misassembly, where possible. The final corrected consensus sequence was translated using EXPASY translate tool. Longest open reading frame was selected to carry out BLASTp searches to confirm the presence of the P450s, GSTs and CCEs superfamily and to check if the sequence has complete or partial protein domain. The curated cytochrome P450 sequences were sent to Dr David Nelson (cytochrome p450 consortium) for further analysis and to assign them names. In case of GSTs and CCEs classification was done on the basis of phylogenetic analysis (see next section).

4.2.5 Phylogenetic analysis

Consensus phylogenetic trees were constructed for each gene superfamily. All the putative full length and partial protein sequences were further analysed to classify them into individual gene families/clades. The curated *L. sericata* sequences for each superfamily along with the corresponding reference protein sequences from other insect genome were aligned. MUSCLE software (Edgar, 2004) was used to perform multiple sequence alignment. This alignment was manually refined, where required. For the construction of the phylogenetic tree, the Neighbor-Joining (NJ) method with bootstrap analysis of 1000 replicates was applied to the multiple sequence alignment using MEGA 5.0 (Tamura *et al.*, 2011).

4.2.6 Variant calling

RNA-seq reads (Illumina sequences from UK population) were first aligned to the reference transcriptome using Bowtie2 (Langmead and Salzberg, 2012). It is an ultra-fast and memory efficient tool for aligning sequence reads to the reference. As the RNA-seq reads were unpaired, Bowtie command with '–U' flag was used with all the other parameters set at default. Post processing of the mapped data and variant calling was done with SAMTools software package version 0.1.18 (Li *et al.*, 2009). The Bowtie2 output SAM file was converted to the binary version (BAM file) using SAMTools. The BAM file was sorted and indexed and SAMTools function 'mpileup' and 'bcftools' were used for variant calling. The output of this step served as the starting dataset for variant analysis. The initially identified variants were further filtered with vcfutils.pl using 'varfilter' command. Variants were called only at positions where the minimum mapping quality (-Q) and coverage (-d) were 25 and maximum read depth (-D) was set at 200.

4.3 RESULTS

4.3.1 454 transcriptome assembly

521,493 *L. sericata* reads were assembled into 32,928 contigs using Roche Newbler software version 2.5p1. The contigs were further assembled into isotigs and isogroups, where ‘isogroups’ are defined as transcripts, ‘isotigs’ are splice variants and ‘contigs’ are separate exons. The 454 transcriptome assembly had 26,348 isotigs with an average length of 1465 bp (N50 = 1972 bp) and average contig coverage of 2.8. These were grouped into 14,741 isogroups having an average contig coverage of 2.2 and average isotig coverage of 1.8. After removing contigs shorter than 100 bp, the trimmed assembly had a total of 29,115 contigs. The overall GC content of the contigs was 33.74%. Table 4.1 has the detailed assembly statistics for the 454 transcriptome assembly.

Table 4.1 Sequence assembly statistics for the Newbler assembly of *Lucilia sericata* transcriptome

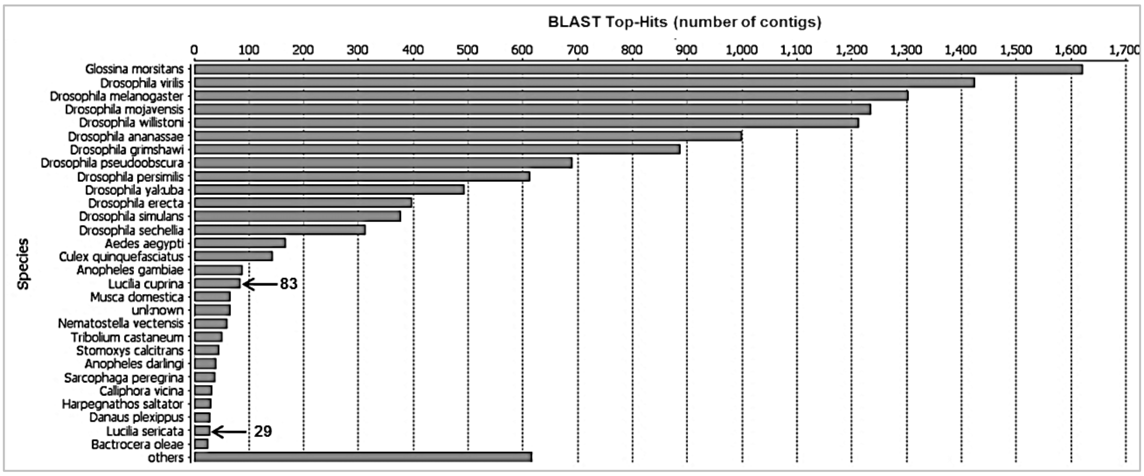
Reads	
Total number of reads	521493
Average read length	364.04
Number of bases	190905830
Isogroup Metrics	
Number of Isogroups	14741
Average contig coverage	2.2
Average isotig coverage	1.8
Isotig Metrics	
Number of Isotigs	26348
Average contig coverage	2.8
Number of bases	38613379
Average isotig size	1465
Largest isotig size	17221
N50 Isotig size	1972
Large Contig Metrics	
Number of contigs	15464

Number of bases	16252105
Average contig size	1050
N50 contig size	1192
Largest contig size	6600
All Contig Metrics	
Number of contigs	32926
Number of bases	20325906
Number of trimmed contigs	29115
Percent GC content	33.74%

4.3.2 Sequence analysis of the assembled transcriptome (BLAST2GO)

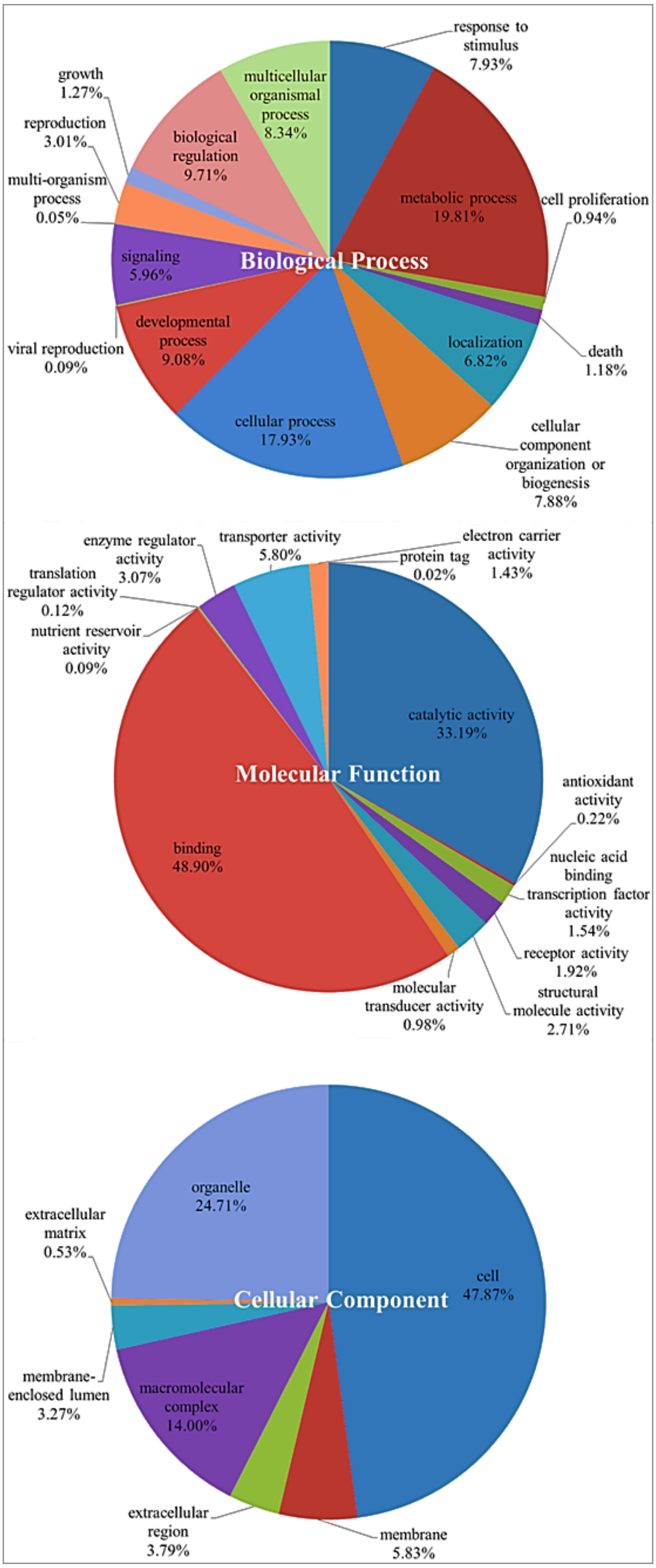
Out of 29,115 contigs, 45.21% (13163nsequences) returned an above cut-off BLAST hit to the NCBI non-redundant (nr) database (BLASTx 1e-3). The majority of the BLASTx hits for the *L. sericata* transcriptome sequences were against *Drosophila* (Figure 4.1). This is because *Drosophila* spp. are the best studied Dipterans with several sequenced genomes and thus make up the bulk of all the Dipteran sequences present in GenBank. Only a small percentage of sequences had top Blast hits against *Lucilia cuprina* and *Lucilia sericata* themselves (highlighted in Figure 4.1). This is due to the paucity of sequence information currently available for these species. Figure 4.1 shows the species distribution of the top Blast hits for *L. sericata* contigs.

Figure 4.1 Species distribution of the top blast hits for *Lucilia sericata* contigs



Gene ontology (GO) terms were assigned to predict the function of the sequences and to categorise them. A total of 8,961 of *L. sericata* contigs could be annotated with GO terms. These contigs were assigned to the three standard GO classifications - molecular functions, cellular components and biological processes. Within the classification of molecular functions, 7,174 sequences were identified, majority of these were predicted to have binding activities, of which 48.90% appeared under binding and 33.19 % appeared under catalytic activity. A total of 4606 sequences were annotated with cellular component classification, 47.87% belonged to cell and 24.71% belonged to the organelle category. Within classification of biological processes, of the total of 5,884 annotated sequences, 19.81% appeared in the metabolic process category followed by 19.93% in the cellular process category (Figure 4.2). The combined number of sequences across the different classifications is more than the sequences annotated with the GO terms as a single sequence may be described by several terms in the three classifications.

Figure 4.2 Gene ontology (GO) assignment for the *Lucilia sericata* contigs. The data presented represents the level 2 analysis, illustrating general functional categories.



The InterPro database was also used to classify the sequences on the basis of the putative function. In total, 16,743 contigs returned a hit against the InterPro database. The summary of the top twenty superfamilies and domains for the *L. sericata* contigs is shown in Table 4.2. Higher frequency of the cytochrome P450 and protein kinase domains were observed (as seen in Table 4.2). Cytchrome P450s are a diverse family of enzymes which have a crucial role in xenobiotic metabolism. However, *L. sericata* sequences also had overrepresented small GTPase superfamilies and Zinc finger domains which is not the case in other insect datasets available (Pauchet *et al.*, 2009, Hull *et al.*, 2013).

Table 4.2 Summary of the top twenty InterPro superfamilies and domains represented in the *Lucilia sericata* transciptome.

SUPERFAMILY			**DOMAIN**		
Interpro	Frequency	Description	Interpro	Frequency	Description
IPR001128	652	Cytochrome P450	IPR007087	739	Zinc finger, C2H2
IPR000175	212	Sodium:neurotransmitter symporter	IPR000719	688	Protein kinase, catalytic domain
IPR020849	206	Small GTPase superfamily, Ras type	IPR015880	567	Zinc finger, C2H2-like
IPR002198	175	Short-chain dehydrogenase/reductase SDR	IPR002557	426	Chitin binding domain
IPR011701	153	Major facilitator superfamily	IPR011009	368	Protein kinase-like domain
IPR003579	152	Small GTPase superfamily, Rab type	IPR002290	368	Serine/threonine-/ dual specificity protein kinase, catalytic domain
IPR002401	136	Cytochrome P450, E-class, group I	IPR017986	347	WD40-repeat-containing domain
IPR001806	134	Small GTPase superfamily	IPR000504	326	RNA recognition motif domain
IPR002041	121	Ran GTPase	IPR013087	326	Zinc finger C2H2-type/integrase DNA-binding domain
IPR003578	113	Small GTPase superfamily, Rho type	IPR015943	292	WD40/YVTN repeat-like-containing domain
IPR000718	108	Peptidase M13	IPR016040	268	NAD(P)-binding domain
IPR001548	106	Peptidase M2, peptidyl-dipeptidase A	IPR013783	260	Immunoglobulin-like fold
IPR002213	106	UDP-glucuronosyl/UDP-glucosyltransferase	IPR016024	258	Armadillo-type fold
IPR000618	98	Insect cuticle protein	IPR001650	247	Helicase, C-terminal
IPR005828	96	General substrate transporter	IPR020683	231	Ankyrin repeat-containing domain
IPR001314	90	Peptidase S1A, chymotrypsin-type	IPR016024	223	Armadillo-type fold
IPR001757	87	Cation-transporting P-type ATPase	IPR012336	220	Thioredoxin-like fold
IPR004119	80	Protein of unknown function DUF227	IPR012677	212	Nucleotide-binding, alpha-beta plait
IPR000734	79	Lipase	IPR020635	210	Tyrosine-protein kinase, catalytic domain
IPR000276	76	G protein-coupled receptor, rhodopsin-like	IPR011989	204	Armadillo-like helical

4.3.3 Manual curation of xenobiotic metabolising enzyme superfamilies

Representatives of all the three main xenobiotic metabolising enzyme superfamilies (P450s, GSTs and CCEs) were identified in the *L. sericata* transcriptome and curated manually.

4.3.3.1 Putative glutathione S-transferases (GSTs)

A total of 32 contigs were identified which were predicted to encode the GST protein domain in the *L. sericata* transcriptome. 28 unique contigs were manually curated of which 15 sequences were found to be full length. Analysis of the *L. sericata* transcriptome identified 28 putative cytosolic GSTs. This was similar to the number found in the genomes of mosquitoes *Aedes aegypti* and *Anopheles gambiae* but fewer than in the genome of *Drosophila melanogaster* (Table 4.3). No microsomal GSTs were identified from the *L. serciata* transcriptome. There are six classes of cytosolic GSTs – delta, epsilon, omega, theta, zeta and sigma (Low *et al.*, 2007). The distribution of these GSTs across the six classes differs substantially across different insect orders. Delta and epsilon classes are insect specific and are most strongly associated with detoxification functions (Claudianos *et al.*, 2006). Delta and epsilon classes are the largest classes in most of the insect orders (Oakeshott *et al.*, 2010). In case of the *L. sericata* transcriptome, the number of delta and epsilon GSTs was comparable to the number found in the genomes of other sequenced insects in the same order, namely 8-12 delta GSTs and 8-14 epsilon GSTs (Table 4.3). However seven theta GSTs were found in the transcriptome of *L. sericata*, about twice the number found in the genomes of *A. aegypti* (4), *D. melanogaster* (4) or *A. gambiae* (2). Local gene duplication is the most likely possibility for high number of GSTs in this particular class (Figure 4.3) but this would need to be confirmed by an examination of the genome. Similarly, for the zeta class, the number of GSTs (3) found in the blowfly transcriptome was higher as compared to the genomes of other insects of the same order. In contrast to the higher number of epsilon and theta GSTs, no EST support was found for the existence of sigma and omega classes in the blowfly (Table 4.3).

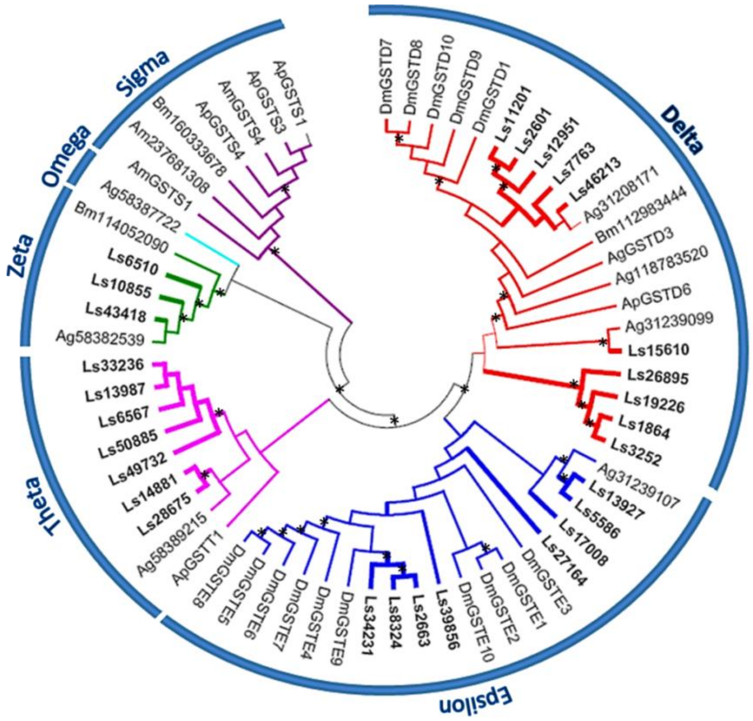
Table 4.3 Number of annotated insect glutathione-S-transferases (GSTs) and their distribution across different GST classes in the order Diptera

CLASS	DIPTERA			
	Drosophila melanogaster	*Anopheles gambiae*	*Aedes aegypti*	***Lucilia sericata***
Cytosolic GSTs				
Delta	11	12	8	**10**
Epsilon	14	8	8	**8**
Omega	5	1	1	**0**
Sigma	1	1	1	**0**
Theta	4	2	4	**7**
Zeta	2	1	1	**3**
Microsomal GSTs	0	3	3	**0**
GST Total	37	28	27	**28**

(* numbers based on transcriptome data)

Data taken from the this study and (Oakeshott *et al.*, 2010)

Figure 4.3 Un-rooted Neighbour joining tree showing the phylogenetic analysis of glutathione S-transferase (GSTs) proteins *of Lucilia sericata* (shown in bold) in relation to other GST proteins from *Drosophila melanogaster* (Dm), *Anopheles gambiae* (Ag), *Apis mellifera* (Am), *Acyrthosiphon pisum* (Ap) and *Bombyx mori* (Bm). Bootstrap values higher than 50% are marked with an asterisk (*) next to the nodes and represent the percentage of 1,000 replicate trees that preserved the corresponding clade. Positions containing alignment gaps and missing data were eliminated with pairwise deletion.



4.3.3.2 Putative cytochrome P450s (P450s)

Analysis of the *L. sericata* transcriptome identified a total of 131 P450 related contigs. These contigs were manually curated. All the curated sequences were sent to Dr David Nelson (Cytochrome P450 naming consortium) to assign them their correct names. Fifty six sequences were assigned complete names and the rest were placed in different gene 'bins' and assigned partial names. Most of the sequences with partial names were truncated and contained incomplete P450 domains. The presence of an incomplete domain could result from sequencing errors or misassemblies. For further analysis, only the sequences which were assigned complete names were used (Appendix C Table S4.1). The majority of these sequences belonged to the CYP3 clade (36), followed by the CYP4 clade (10), the CYP mitochondrial clade (8) and the least number belonging to the CYP 2 clade (2) (Table 4.4).

Four distinct clades could be identified using phylogenetic analysis (Figure 4.4). The CYP3 clade includes the large CYP6 and CYP9 families in insects (Feyeresein, 2006). In *L. sericata,* the CYP 6 and CYP9 families are represented by twenty-four and two members respectively. The CYP3 and CYP4 clades were both abundant in all the Dipteran insects annotated so far, with the exception of *L. sericata* where the CYP4 clade contained only ten putative P450s (Table 4.4). Under-representation of the CYP4 clade in *L. sericata* could be due to the absence of members of this clade in the current transcriptomic dataset and additional genes may therefore await discovery.

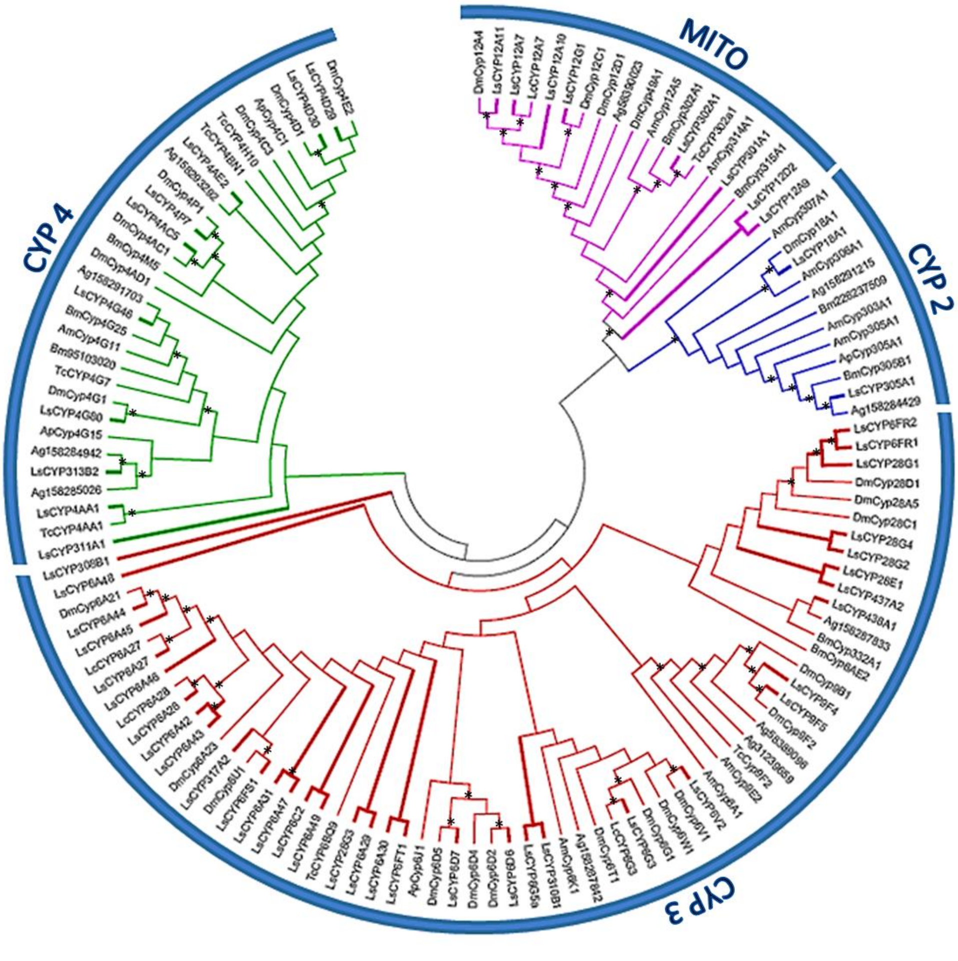
Table 4.4 Number of annotated insect cytochrome P450s and their distribution across different clades in the order Diptera

	DIPTERA			
CLADE	*Drosophila melanogaster*	*Anopheles gambiae*	*Aedes aegypti*	***Lucilia sericata***
CYP4	32	45	59	**10**
CYP3	36	42	84	**36**
CYP2	6	10	11	**2**
Mitochondrial	11	9	10	**8**
P450 Total	85	106	164	**56**

(* numbers based on transcriptome data)

Data taken from the this study and (Oakeshott *et al.*, 2010)

Figure 4.4 Un-rooted Neighbour joining tree showing the phylogenetic analysis of cytochrome P450s (P450s) proteins *of Lucilia sericata* (shown in bold) in relation to other P450 proteins from *Drosophila melanogaster* (Dm), *Anopheles gambiae* (Ag), *Apis mellifera* (Am), *Acyrthosiphon pisum* (Ap), *Bombyx mori* (Bm), *Tribolium castaneum* (Tc) and *Lucilia cuprina* (Lc). Bootstrap values higher than 50% are marked with an asterisk (*) next to the nodes and represent the percentage of 1,000 replicate trees that preserved the corresponding clade. Positions containing alignment gaps and missing data were eliminated with pairwise deletion.



4.3.3.3 Putative carboxyl/cholinesterases (CCEs)

Carboxyl/cholinesterases (CCEs) can be classified into thirteen clades and three distinct classes (Ramsey *et al.*, 2007). Phylogenetic analysis was carried out in order to classify the *L. sericata* CCEs into various clades (Figure 4.5). Nine of these clades were represented in the *L. sericata* transcriptome (Table 4.5, Figure 4.5). Clades with no *L. sericata* CCEs were clade A, clade C, clade G and clade I. Thirty four members of the CCE superfamily were found in the *L. sericata* transcriptome, compared to thirty five, fifty one and fifty four CCEs that have been identified in the genomes of *D. melanogaster, A. gambiae* and *A. aegytpi* respectively (Table 4.5). Thirty putative *L. sericata* CCEs were classified into the different CCE classes/clades. However four CCE sequences were too short to be reliably classified into any one of the clades.

The dietary/detoxification class consists of clade A-C which is involved in the detoxification of xenobiotics. The majority of CCEs in the *L. sericata* transcriptome belong to this class. Clade B of this class was represented by sixteen putative members which is equal to the number found in the *A. gambiae* genome and significantly higher than that of *D melanogaster* (13). *A. aegyptii* (22) has the highest number of CCEs belonging to clade B. In Australian *L. cuprina*, two single nucleotide polymorphisms in αE7 gene encoding the esterase 3 (E3) are associated with organophosphate resistance (Hartley *et al.*, 2006). Two *L. sericata* sequences (Ls11154, Ls44527) of the clade B were found to be highly similar to esterase 3 (E3) of *L. cuprina*.

L. sericata shows an apparent reduction in the diversity in CCEs involved in hormone/pheromone processing (clades D-G). Five members belonged to clade D as compared to three in *D. melanogaster*. Clade D is absent in *A. gambiae* and *A. aegyptii*. Members of clade E were in the range of 2-4 in other sequenced Dipterans. One putative member was found in clade F, which is low as compared to three members in *D. melanogaster*, six members each in *A. gambiae* and *A. aegyptii*. No members were found in clade G. Similarly neuro/development class (clades H-M) had a significantly lower number as compared to the other sequenced Dipterans (Table 4.5). As previously noted, however, the apparent lack of these genes might be an artefact of the transcriptome approach.

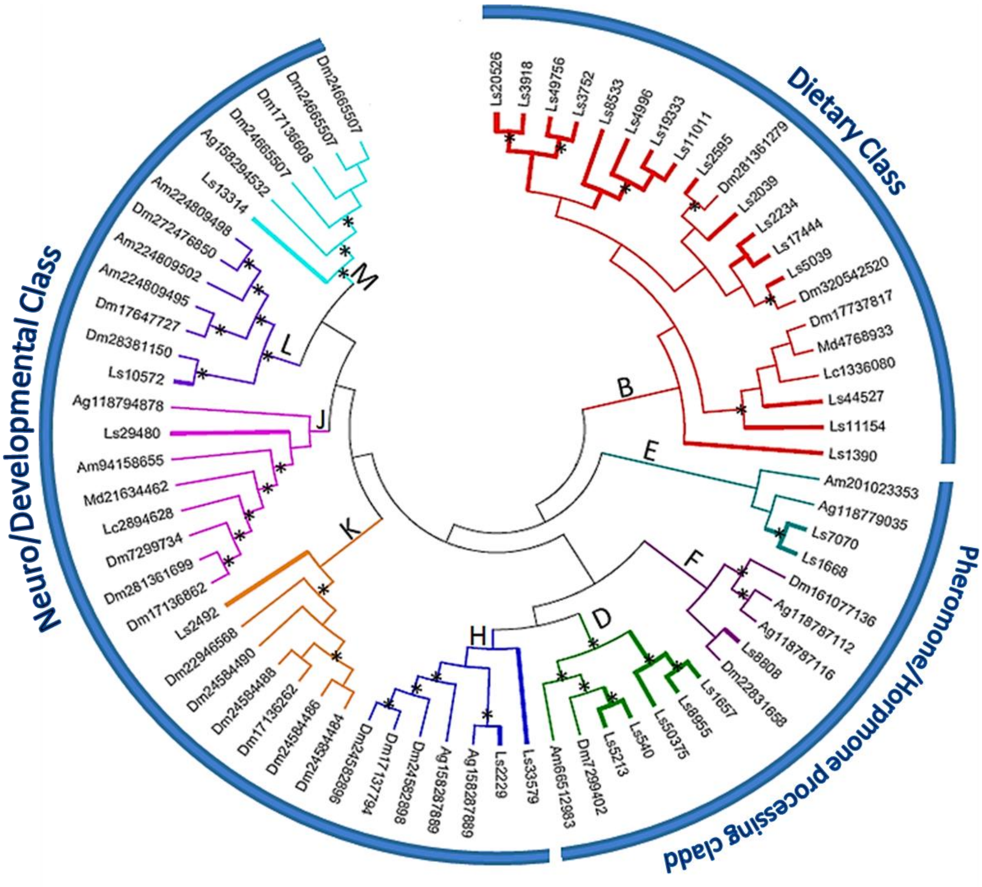
Table 4.5 Number of annotated insect carboxyl/cholinesterases (CCEs) and their distribution across different clades in the order Diptera

CLASS	DIPTERA			
	Drosophila melanogaster	*Anopheles gambiae*	*Aedes aegypti*	***Lucilia sericata***
Dietary class				
A clade	0	0	0	**0**
B clade	13	16	22	**16**
C clade	0	0	0	**0**
Hormone/semiochemical processing				
D clade	3	0	0	**5**
E clade	2	4	2	**2**
F clade	3	6	6	**1**
G clade	0	4	6	**0**
Neuro/developmental				
H clade	5	10	7	**2**
I clade	1	1	1	**0**
J clade	1	2	2	**1**
K clade	1	1	1	**1**
L clade	4	5	5	**1**
M clade	2	2	2	**1**
Unknown	-	-	-	**4**
CCE Total	35	51	54	**34**

(* numbers based on transcriptome data)

Data taken from the this study and (Oakeshott *et al.*, 2010)

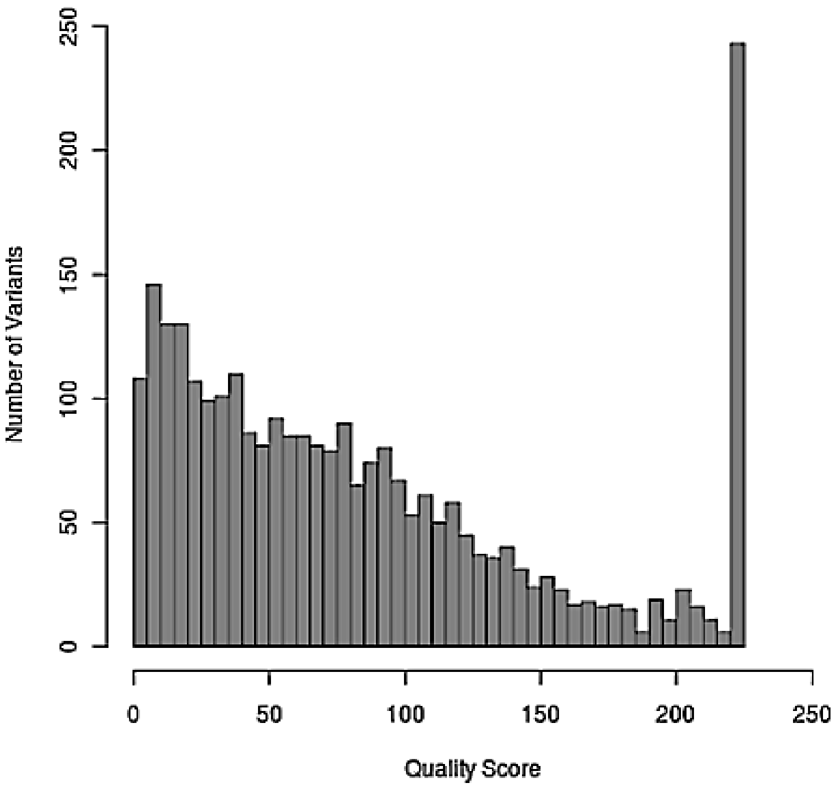
Figure 4.5 Un-rooted Neighbour joining tree showing the phylogenetic analysis of carboxyl/cholinesterases (CCEs) proteins *of Lucilia sericata* (shown in bold) in relation to other CCE proteins from *Drosophila melanogaster* (Dm), *Anopheles gambiae* (Ag), *Apis mellifera* (Am) and *Lucilia cuprina* (Lc). Bootstrap values higher than 50% are marked with an asterisk (*) next to the nodes and represent the percentage of 1000 replicate trees that preserved the corresponding clade. Positions containing alignment gaps and missing data were eliminated with pairwise deletion. The letters on the nodes represent the CCE clades as listed in Table 4.4



4.3.4 Variant calling

The assembled transcriptome was also used as a reference for SNP calling. Single end, RNA-seq reads were mapped to the reference transcriptome using Bowtie2. Only reliable, high quality SNPs were considered. In total, SAMtool detected 26,478 variant sites, when the Illumina sequences from the UK population where mapped to the reference transcriptome. Out of the total variant sites, 11,050 were SNPs and the rest were 'indels' (insertions or deletions) (Appendix C Figure S4.1). The initially detected SNPs were filtered using stringent parameter (coverage and mapping quality of 25 and maximum depth of 200) to select only reliable good quality SNPs. 2800 SNPs had mapping quality and coverage higher than the set filters (Figure 4.6).

Figure 4.6 Distribution of SNP quality scores in the UK population of *Lucilia sericata*. The graph represents only good quality SNPS after filtering using stringent parameter.



4.3.4.1 E3 mutation in the UK population

Single nucleotide polymorphisms in the *aE7* gene encoding esterase 3 (E3) at position 137 and 251 are associated with diazinon and malathion resistance respectively. Diazinon resistance is absent or rare outside of Australasia whereas malathion resistance is more widespread (Hartley *et al.*, 2006). In order to find out the extent of the E3 mutation in the UK population, single end RNA-seq reads were mapped against the manually curated CCE sequences with high similarity to *L. cuprina* E3 (Ls44527, Ls11154). Although Ls44527 had good coverage, it was truncated and hence contained only the diazinon region. A single Illumina read was found to have a SNP at the 137 position. As the SNP was at the end of the sequence, it could not be determined with confidence whether it is a very low frequency true SNP or simply a sequencing error (Appendix C Figure S4.2). However several other good quality SNPs were detected within this contig. On the other hand Ls11154 had very low coverage and no SNPs were detected in either position 137 or 251.

4.4 DISCUSSION

Lucilia sericata has been associated with numerous areas of research in the field of medical, forensic, ecological, evolutionary and veterinary sciences (Paul *et al.*, 2009, Anderson, 2000, Denno and Cothran, 1976, Mellenthin *et al.*, 2006, Fischer *et al.*, 2004). Despite its association with several areas of research there is still paucity of the sequence information available publically for this species. Transcriptome sequencing provides an alternative to get a quick access to the genetic information for any organism. It is cost effective and computationally less intensive than whole genome sequencing (Kumar and Blaxter, 2010). Transcriptome analysis using 454 pyrosequencing has now been reported for several non-model organisms (Pauchet *et al.*, 2009, Karatolos *et al.*, 2011, Hull *et al.*, 2013). A similar strategy was implemented in the current project to generate a reliable reference transcriptome to enhance the genomic resources available for this species. Although Sze *et al.* (Sze *et al.*, 2012) have recently attempted to assemble a *L. sericata* transcriptome; their work mainly focuses on the different strategies for assembling the data as compared to a thorough analysis of the assembled transcriptome itself. In contrast this study provides a detailed analysis of the assembled 454 transcriptome itself. The N50 values generated for this assembly were also higher than any of the assemblies generated by Sze *et al.,* using various parameters (Table 4.1).

Blast2GO analysis of the assembled sequences illustrate that the distribution of GO terms within the three categories was consistent with other insect transcriptomes (Figure 4.2). The InterProScan results further support this observation. Apart from the superfamilies/domains expected to be dominant in the insect sequences (such as cytochrome P450 and protein kinase), overrepresentation of small GTPase superfamilies and zinc finger domains were observed (Table 4.2). These superfamilies and domains are associated with various signalling pathways (Raymond *et al.*, 2001, Berrocal-Lobo *et al.*, 2010). The higher frequency of the mentioned superfamilies and domains in this dataset is intriguing and worthy of further investigation.

The transcriptomic data was further used to identify xenobiotic metabolising enzyme superfamilies (P450s, GSTs and CCEs). Representatives of all the three enzyme superfamilies were found in the transcriptome of *L. sericata*. The number of putative members is comparable to the other sequenced Dipterans, which further suggests the relative completeness of the assembled transcriptome. The data suggests that *L. sericata* has modest detoxification capabilities. Of the three superfamilies, putative P450s in particular were lower than the number found in other sequenced Dipterans (Table 4.4). However, additional P450s may await discovery due to the absence from this dataset. In contrast, the number of CCEs and GSTs were within the range of other sequenced insects of the same order (Table 4.3, Table 4.5). The data highlights the general and species specific trends amongst sequenced Dipterans, namely *D. melanogaster, A. gambiae*, *A. aegypti* and *L. sericata.*

Identification of sequences related to detoxification revealed putative candidates that could be involved in insecticide resistance. Several members of the CYP3 and CYP4 clades of P450s, particularly the CYP6, CYP9 and CYP4 families, have been linked with insecticide resistance. Similarly, the insect specific GST classes, delta and epsilon are often associated with detoxification. In case of the CCEs, the esterases involved in detoxification of xenobiotics belong to the dietary class. Another clade associated with insecticide resistance is clade J (acetylcholinesterase), members of this clade are targets for organophosphate and carbamate insecticides (Grafton-Cardwell *et al.*, 2004). Single putative acetylcholinesterase was identified in this dataset, which could also have a potential role in insecticide resistance. However, their potential role in insecticides resistance would require further investigation.

This newly generated genetic information was also used to identify single nucleotide polymorphisms (SNPs) in the UK population. The Illumina data further enabled to carry out a comprehensive assessment of the extent of E3 mutation in the UK population. In Australian *L. cuprina*, two single nucleotide polymorphisms in *aE7* gene encoding the esterase 3 (E3) are associated with two forms of organophosphate resistance. Single amino acid replacement (Gly137Asp) is associated with diazinon resistance and two replacements (Trp251Leu/Ser) are associated with malathion resistance (Claudianos *et al.*, 1999). Esterase based OP resistance is well documented for *L. cuprina*, especially in the populations from Australasia. However, limited information is available for OP resistance in *L. sericata* (Hartley *et al.*, 2006). In the UK population of *L. sericata*, a single example of diazinon associated SNP was found, but it could not be confirmed that this was not due to a sequencing error. The ability to compare sequencing data from different population around the world will substantially enhance our understanding of the spread of malathion and diazinon resistance outside Australasia. This study can therefore serve as the basis for future SNP analysis of different *L. sericata* populations from around the world.

Thus *L. sericata* transcriptome could also serve as a useful genomic resource in facilitating the future investigation in functional genomic, gene mapping and other genetics studies. This will in turn benefit medical, agricultural, ecology, evolution, and forensic research in this species. The large dataset generated could also be used in population level genomics and comparative genomic studies. Furthermore, the annotated sequences will facilitate the investigation of fundamental biology of *L. sericata* and provide a helpful comparison point for genomic studies in both sibling species *L. sericata* and *L. cuprina.*

4.5 CONCLUSION

L. sericata is a non-model organism that is associated with numerous areas of research but lacked genomic resources. Until recently, research involving the characterisation of molecular mechanism underlying insecticide resistance in *L. sericata* was hampered due to lack of genomic information. 454 pyrosequencing of the transcriptome generated 29,115 unique contigs which has dramatically improved the sequence information available for this species. Analysis of the *L. sericata* transcriptome has revealed several gene candidates for the biosynthesis and detoxification of xenobiotics. Potential detoxification genes represented in this dataset included 28 GSTs, 56 P450s and 34 CCEs. Comparison with the other members of this order shows that the dataset represent nearly complete collection of such genes from the three superfamilies. The elucidation of the function of the various classes and clades of these superfamilies can provide an important insight into the ability of *L. sericata* to respond to environmental changes. This newly generated genetic information further identified 2800 good quality SNPs in the UK population. This study can therefore serve as the basis for future SNP analysis of different *L. sericata* populations from around the world. The results also prove that transcriptome sequencing is a good technique to enhance the genomic resources available for non-model organisms.

CHAPTER 5

Removing microbial messages from 454 generated beetle transcriptomes

5.1 INTRODUCTION

In this chapter we analysed the midgut transcriptomes of five beetle species and their gut associated microorganisms. The datasets consisted of 454 transcriptomes which had been generated by others in previous projects in our laboratory (Pauchet *et al.*, 2009, Pauchet *et al.*, 2010). The the main objective of this project was to survey the midgut transcriptomes of five beetle species (*Gastrophysa viridula*, *Chyrsomela tremulae, Leptinotarsa decemlineata*, *Sitophilus oryzae* and *Callosobrunchus maculatus*) and to separate the microbial messages 'contaminating' them. It has been observed that when sequencing RNA from field collected organisms, microbial sequences often get sequenced along with those of the host. Dealing with these mixed sequences is therefore a major challenge when sequencing samples from wild populations, especially if the associated microbe is a eukaryotic endoparasite.

Rapid improvement in the efficiency and the speed of sequencing along with the falling cost has made it possible even for small research groups to sequence their favourite organisms of interest. This has shifted focus from a few laboratory based 'model' organisms to 'non-model' organisms and even those obtained directly from nature. Although next generation sequencing technologies allows us to directly sequence organisms obtained from the environment, analysing such data can have several technical challenges. The most important of these is to deal with sequences of the associated microorganisms that get sequenced along with the host RNA (Friedel *et al.*, 2005). Prior knowledge of all the sequences of microbes associated with the host is one potential way of dealing with this problem. It is therefore possible to physically separate the microbe and the host DNA/RNA by electronic subtraction. However this process is time consuming and laborious and does not guarantee the final sequenced host data set will be free from sequences of the microbial contaminant. It becomes even more difficult to obtain pure samples in case of endoparasites (Emmersen *et al.*, 2007).

As more and more species are being sequenced, such data can offer an opportunity to discover additional sequences derived from their associated microbes. For example, Salzberg and others (Salzberg *et al.*, 2005) discovered three new strains of the endosymbiont *Wolbachia pipientis* by surveying three different species of the fruit flies *Drosophila ananassae*, *D. simulans*, and *D. mojavensis*. To do this they simply electronically extracted and assembled all the sequences that matched with the genome of the previously sequenced *Drosophila melanogaster wolbachia* strain (wMel) by identifying raw reads that were similar to wMel.

Similarly, sequencing projects designed to target the Pea aphid (*Acyrthosiphon pisum*) also generated sequences of the primary symbiotic bacteria *Buchnera aphidicola* (Consortium, 2010). Screening of the aphid genome data for sequences of bacterial origin revealed large number of bacterial sequences, even after a pre-processing step to remove vector contamination. However by separating and reassembling the bacterial sequences along with gap closure by PCR resulted in a complete reconstruction of the genome of *Buchnera*. Sequences from aphid secondary symbionts, *Regiella insecticola* and *Hamiltonella defensa*, were also obtained. Similarly an improved understanding of the relationship between host and their associated microorganisms, has been facilitated by the development of genomic resources in several other aphid species (Huang *et al.*, 2010, Huerta-Cepas *et al.*, 2010). For example, analysis of the Soya bean aphid transcriptome resulted in the identification of potential transcripts belonging to both the aphid and its associated symbionts (Liu *et al.*, 2012).

However, when sequencing a new species, genomic information of all its associated microorganisms is not always available. To this end however, instead of considering the sequencing of mixed samples as a problem, it can also be regarded as a means of gaining insight into the simultaneous gene expression of both the host and the endoparasite as performed in new techniques such as dual-RNA seq (Westermann *et al.*, 2012). To achieve this efficiently, a reliable method is required which can classify the sequences according to their origin with a reasonable degree of accuracy without the need to know the genome/transcriptome sequences of both the host and endoparasite beforehand.

In conclusion, all the above mentioned examples took advantage of the existing genomic information, in order to computationally separate the sequences of the host and associated microbes. However, this richness of background genomic/transcriptomic information is not always available for all the organisms. Kumar and Blaxter (Kumar and Blaxter, 2011) used a strategy that could be used in this situation where the genomic sequences where previously unknown or were significantly divergent from the related species. The current project used similar approaches by testing bioinformatic pipelines to identify and separate insect sequences from those of their eukaryotic endoparasites.

5.2 METHODOLOGY

5.2.1 Beetle 454 derived transcriptomes

The five beetle datasets consisted of 454 derived transcriptomes from the larval midgut of *G. viridula*, *C. tremulae* and *L. decemlineata*, the adult midgut of *S. oryzae* and whole *C. maculatus* larvae. The full length, enriched, cDNAs were generated from 2μg RNA using SMART PCR cDNA synthesis kit (BD Clonetech). PrimeScript reverse transcription enzyme (Takara) was used to perform reverse transcription. In order to reduce over abundant transcripts, the double stranded cDNAs were further normalised using the Trimmer cDNA normalisation kit (Evrogen). To obtain the transcriptomes of the five beetle species, the normalised cDNA libraries were 454 sequenced at the Advanced Genomic facility at the University of Liverpool (Pauchet *et al.*, 2009, Pauchet *et al.*, 2010).

5.2.2 Preliminary assembly

C. tremulae reads obtained by 454 pyrosequencer were pre-processed (removal of Poly-A tails and SMART adaptors) and clustered using the MIRA v2.9.26x3 assembler with the *de novo*, normal, EST, 454 parameters. Specifying a minimum read length of 40 nt, a minimum sequence overlap of 40 nt and a minimum percentage overlap identity of 80%. Pre-processing and assembly of the remaining beetle datasets (*G. viridula*, *L. decemlineata*, *S. oryzae* and *C. maculatus*) was achieved using our in-house pipeline called 'est2assembly' (Papanicolaou *et al.*, 2009).

5.2.3 Estimation of taxonomic composition of the preliminary assembly

We used the following software and scripts in the analysis pipeline:

1) Standalone blast suite+ downloaded and installed from ftp://ftp.ncbi.nih.gov/blast/executables/blast+/LATEST
2) NCBI blast database downloaded from ftp://ftp.ncbi.nih.gov/blast/db/
3) NCBI taxonomy dumps downloaded and unpacked from ftp://ftp.ncbi.nih.gov/pub/taxonomy/gi_taxid_prot.dmp.gz; ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz

4) Perl script blast_taxonomy_report.pl (Kumar and Blaxter, 2011)

Contigs from each beetle preliminary assembly were queried against the NCBI nr database using BLASTx with default parameters. To estimate the taxonomic composition, blast_taxonomy_report.pl script was used. This perl script annotated each contig with the taxonomic order of the best BLAST hit. Graphs were generated to visualise the taxonomic distribution at phylum and order level.

5.2.4 Separation of contigs on the basis of most abundant taxonomic groups

In order to separate contigs of the most abundant taxonomic groups in the mixed transcriptomes:

1) Taxon specific blast databases were created for the most abundant groups. BLASTx searches with E-value cutoff of 1e-5, were carried out against the specific databases to separate the contigs belonging to different organisms. Using the BLAST similarity result, contigs from preliminary assembly were separated as belonging to one of the taxonomic groups, all or none. Contigs that gave hits with all the databases with similar bit scores were classified as all. Contigs with no blast hits or a bit score of less than 50 were classified as none.
2) The pre-existing Perl script blast_separate_taxa.pl developed by Kumar and Blaxter (Kumar and Blaxter, 2012) was used to separate the contigs based on the above classification. Further BLASTn searches were carried out to find out reads corresponding specific contigs.
3) A custom Perl script written in house was used to obtain reads corresponding to the contigs of interest

5.2.5 Taxon specific re-assembly

The separated reads were used to generate taxon specific assemblies using Roche Newbler software version 2.5p1 from the command line with the -cdna flag. Assembly was conducted on a 64-bit UNIX server (running on Debian OS) with 96GB RAM.

5.2.6 BLAST2GO analysis

The assembled contigs were analysed using automated Blast2GO software suite V2.6.2 (Conesa *et al.*, 2005). This analysis included BLAST searches, functional annotation with GO, InterPro terms (InterProScan), EC codes and metabolic pathways. BLASTx searches were performed remotely on NCBI server against non-redundant (nr) protein database using QBLAST. The program extracted the GO terms associated with blast hits and returned a list of GO annotations which are represented as hierarchical categories with increasing specificity. Next the GO terms were modulated using ANNEX, the annotation augmentation tool followed by GOSlim (consisting of a subset of the GO vocabulary encompassing key ontological terms and a mapping function between full GO and GOSlim). EC codes and KEGG metabolic pathway annotations were obtained by directly mapping the GO terms to their equivalent enzyme codes. InterPro searches were performed on the InterProEBI web server remotely via BLAST2GO.

5.2.7 Codon usage and GC content

The overall GC content and the codon usage was also calculated for each assembly using EMBOSS tool ‘cusp’ (Rice *et al.*, 2000).

5.3 RESULTS

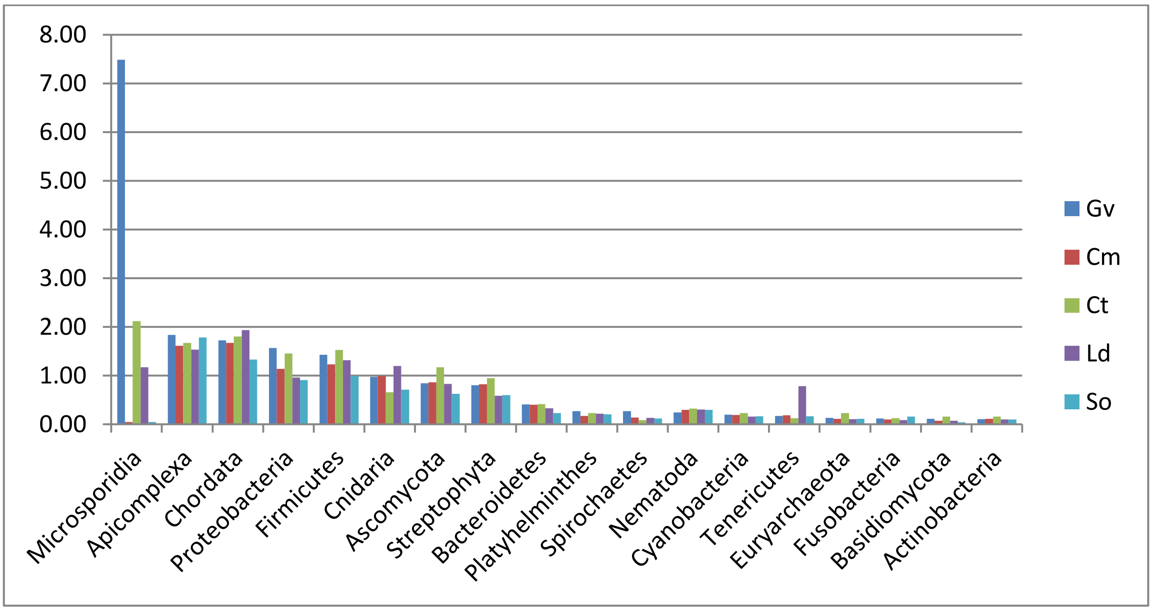
5.3.1 BLAST annotation and estimation of taxonomic composition

The transcriptome assemblies of the five beetle species were screened to estimate the taxonomic composition by conducting BLASTx searches against NCBI nr database and annotating each contig with the taxonomic information. In the five beetle datasets, the potential level of associated microbes/contaminants varied from approximately 16%-27% (Table 5.1). Green dock beetle had the highest percentage of contigs having homology to species other than that of arthropods. With the exception of the Green dock beetle dataset, the other four beetle datasets had less than 2% of contigs in each non arthropoda phylum/division (Figure 5.1). In the four beetle datasets (Cowpea weevil, Poplar leaf beetle, Colorado potato beetle and Rice weevil) low percentage of contigs were found to belong to different genera of microorganisms (less than 2%) so they were not analysed further. However in case of the Green dock beetle a high percentage of contigs were found to belong specifically to microsporidia. Approximately 7% of the total contigs showed homology to microsporidia (Figure 5.1). Hence, the Green dock beetle dataset was used as a test-bed in which to trial the different bioinformatic pipelines to accurately classify and separate insect sequences from those of their eukaryotic endoparasites.

Table 5.1 Summary of Blast results and taxon analysis for the five beetle 454 transcriptome datasets

Number of contigs	***Gastrophysa viridula***	***Callosobruchus maculatus***	***Chrysomela tremulae***	***Leptinotarsa decemlineata***	***Sitophilus oryzae***
Total Contigs	20817	32584	10969	21692	22989
BLASTX hits	19056	29915	10356	20593	21617
No hits	1761	2669	613	1099	1372
Arthropoda tophits (percent)	13828 (72.57)	24312 (81.27)	8240 (79.57)	16263 (78.97)	18137 (83.90)
Tophits with other phylums (percent)	5228 (27.43)	5603 (18.73)	2116 (20.43)	4330 (21.03)	3480 (16.10)

Figure 5.1 Phylum level distribution of microbes/contaminants in the five beetle datasets. X axis shows the most abundant phylums/division found in the five datasets and Y-axis shows the percentage of contigs belonging to each phylum/division



Gv – *Gastrophysa viridula* (Green dock beetle)

Cm- *Callosobruchus maculatus* (Cowpea weevil)

Ct- *Chrysomela tremulae* (Poplar leaf beetle)

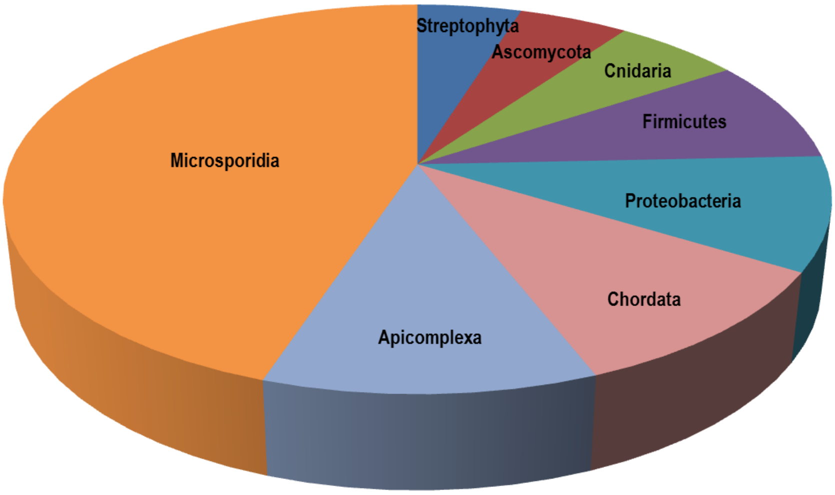
Ld- *Leptinotarsa decemlineata* (Colorado potato beetle)

So- *Sitophilus oryzae* (Rice weevil)

5.3.2 Classification of contigs and separation of reads

Analysis of BLAST and annotation results for the 20817 Green dock beetle contigs revealed that a total of 1284 contigs (7.4% of the green dock contigs with BLASTx hits) had BLASTx hits to microsporidia. Figure 5.2 shows the distribution of the abundant group of microbes/contaminants found in the Green dock beetle dataset. Microsporidia formed the major group of associated microbes in this datasets. Microsporidia and arthropoda specific databases were used as filters to create subsets of contigs belonging to the two phylum/division of interest. For this subset of NCBI database specific to the two taxa were downloaded and BLASTx searches were carried out using all green dock beetle contigs as query against each database with the E value cutoff of 1e-5. Out of the total 20817 contigs, 15778 contigs returned Blast hits against the arthropoda database whereas 4777 contigs returned blast hits against the microporidial database. Contigs which gave hits against both the databases were reclassified or discarded on the basis of the bit score. If the bit score was more than 50 against one of the database it was reclassified otherwise the contig was discarded. Following the classification step, 14772 contigs were classified as arthropoda and 1279 contigs were classified as microsporidia. Custom perl scripts were used to separate the classified contigs into two groups. BLASTn searches and custom Perl scripts were used to extract pre-processed reads corresponding to each contig.

Figure 5.2 Pie chart showing the distribution of the most abundant microbes/contaminants found in the *Gastrophysa viridula* 454 midgut transcriptome dataset



5.3.3 Taxon specific re-assembly

The 745,794 putative Green dock beetle 454 reads obtained for arthropoda specific contigs were reassembled using Roche Newbler software version 2.5p1. Of these, 667,487 reads were assembled into 15,276 contigs. These contigs were then assembled into 12,368 'isotigs' having an average length of 1,676 bp (N50 = 2111 bp) and average contig coverage of 2.4. These contigs were further grouped into 7,633 isogroups with average contig coverage of 2 and average isotig coverage of 1.6. All the contigs shorter than 100 bp were removed. The resulting final 12,787 contigs had an average length of 970.83 bp and average read coverage of 19.83. The overall GC content of all the contigs was 38.09%.

Similarly 105,425 putative microsporidial reads were reassembled using Roche Newbler software version 2.5p1. Of these, 103,670 reads were assembled into 1,638 contigs. These contigs were further assembled into 1495 isotigs with an average length of 1508 bp (N50 = 1956 bp) and average contig coverage of 1.4. The contigs were grouped into 1,294 isogroups having an average contig coverage of 1.3 and average isotig coverage of 1.2. After removing contigs shorter than 100 bp, the trimmed assembly had a total of 1,464 contigs with an average length of 1306.33

bp and average read coverage of 19.45. The overall GC content of the contigs was 34.05%. Detailed assembly statistics for the two assemblies are given in table 5.2

Table 5.2 Sequence assembly statistics for the Newbler reassembly of separated Green dock beetle midgut (*Gastrophysa viridula*) and microsporidial sequences

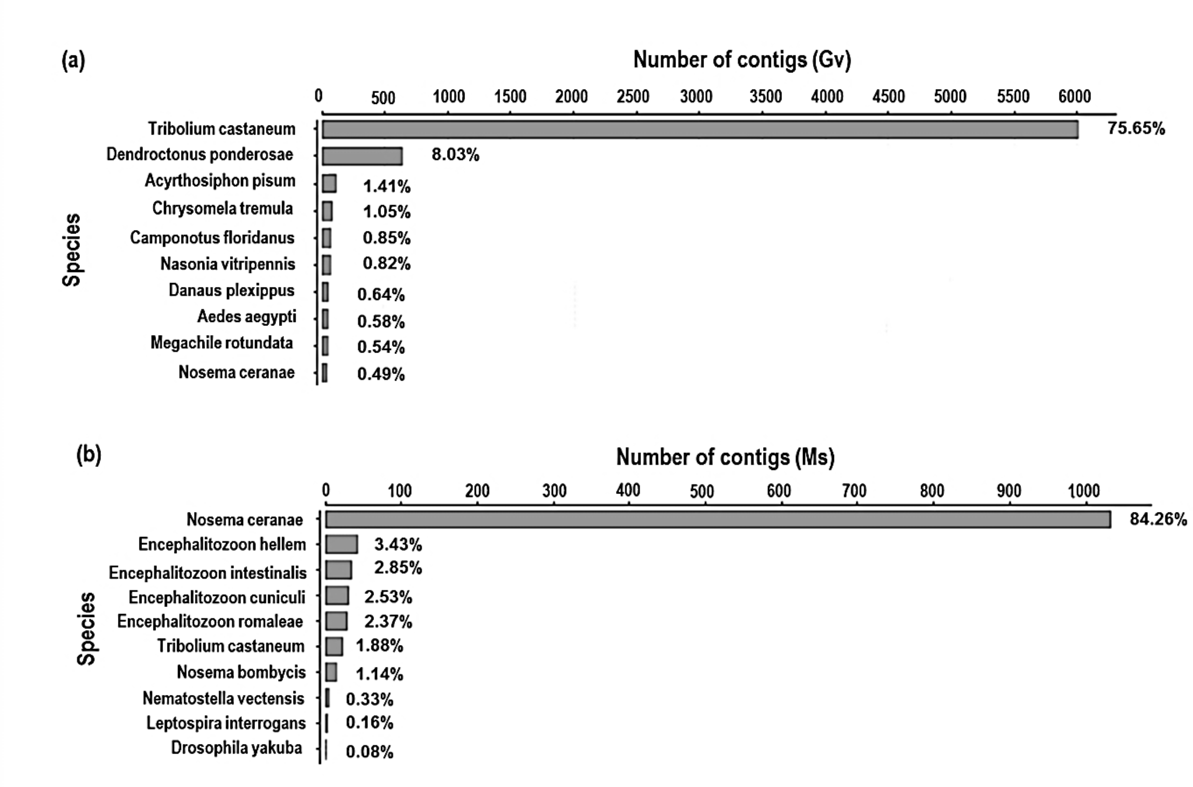
	G. viridula	***Microsporidia***
Reads		
Total number of reads	745794	105425
Average read length	364.41	368.05
Number of bases	278497163	40052880
Number of reads assembled	667487	103670
Isogroup Metrics		
Number of Isogroups	7633	1294
Average contig coverage	2	1.3
Largest contig coverage	305	12
Number with one contig	5308	1109
Average isotig coverage	1.6	1.2
Largest isotig coverage	81	9
Number with one isotig	5408	1151
Isotig Metrics		
Number of Isotigs	12368	1495
Average contig coverage	2.4	1.4
Largest contig coverage	15	7
Number with one contig	5753	1223
Number of bases	20735431	2255376
Average isotig size	1676	1508
Largest isotig size	7744	6395
N50 Isotig size	2111	1956
Large Contig Metrics		
Number of contigs	8098	1216
Number of bases	11156733	1823773
Average contig size	1377	1499
N50 contig size	1661	1832

Largest contig size	5934	6395
Q40 plus bases	10784323 (96.66%)	1782421 (97.73%)
Q39 minus bases	372410 (3.34%)	41352 (2.27%)
All Contig Metrics		
Number of contigs	15276	1638
Number of bases	12516583	1916044
Number of trimmed contigs	12787	1464
Average contig length	970.83	1306.33
Average contig coverage	19.83	19.45
Percent GC content	38.09	34.05

5.3.4 Sequence analysis of the taxon specific reassemblies (BLAST2GO)

The majority of the Blast hits for the Green dock beetle stringent assembly were against the Red flour beetle, *Tribolium casteneum* (75.65%). This may simply be because *T. casteneum* is the only beetle genome that has been fully sequenced to date and thus makes up the bulk of all the beetle sequences present in GenBank. A small percentage of shorter sequences had Blast hits against *Nosema ceranae* (0.49%) but most of these were against hypothetical proteins. Figure 5.3(a) shows the species distribution of the top ten Blast hits for the Green dock beetle contigs. Similarly, for the microsporidial assembly, 96.58% of sequences had significant blast hits against the microsporidial division of fungi. Figure 5.3(b) shows the species distribution of the top ten blast hits for the microsporidial contigs. 84.26% of the sequences had a top Blast hit against the microsporidium *Nosema ceranae*. Approximately 3% of the sequences had top Blast hits against taxa other than microspordia.

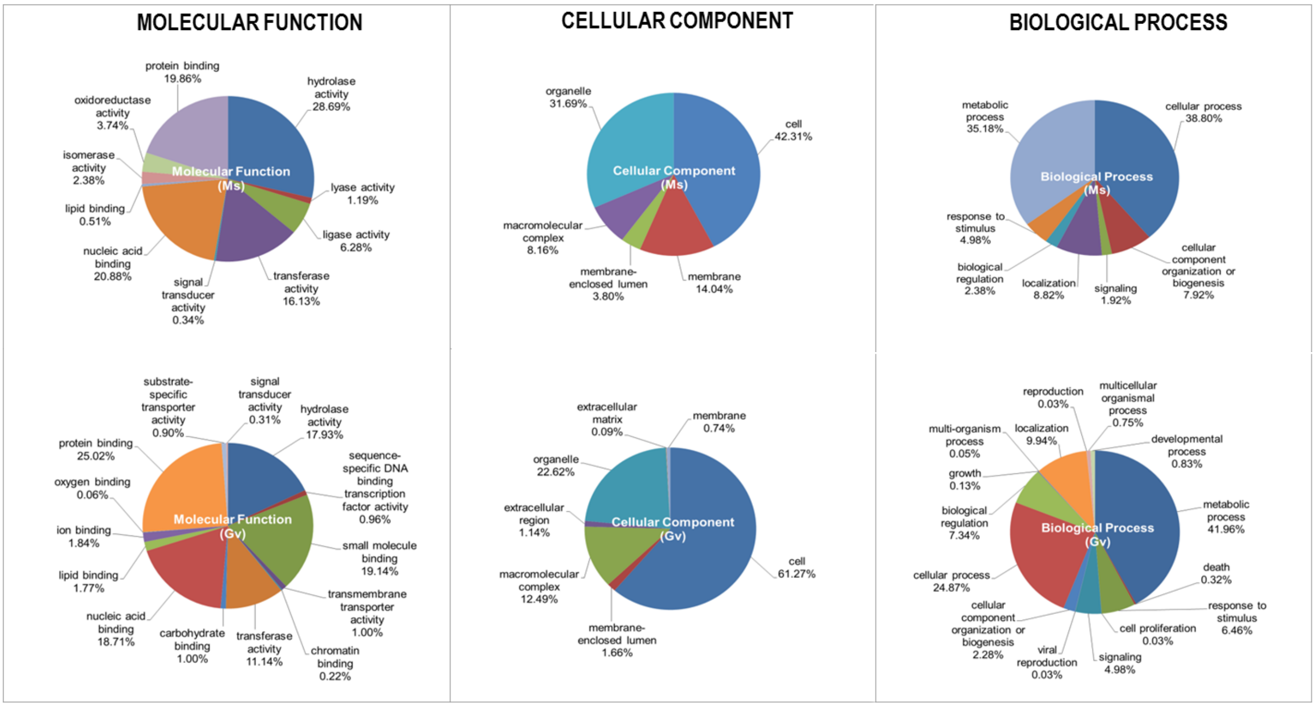
Figure 5.3 Species distribution of the top ten blast hits (a) Green dock beetle contigs (Gv), (b) microsporidial contigs (Ms)



Gene ontology (GO) terms were assigned to predict the function of the sequences and to categorise them. A total of 7,933 of Green dock beetle contigs had blast hits, of which 3,782 (47.67%) could be annotated with GO terms. These contigs were assigned to the three standard GO classifications, namely molecular function, cellular components and biological processes. Within the classification of molecular processes, 2,276 sequences were identified, the majority of these were predicted to have binding activities, of which 25.02% appeared under protein binding and 19.14 % appeared under small molecular binding. A total of 1,390 sequences were annotated with the cellular component classification, with 61.22% belonging to the ‘cell’ and 22.62% belonging to the ‘organelle’ category. Within the classification of ‘biological processes’, of the total 1,546 annotated sequences, 41.96% appeared in the ‘metabolic process’ category and 24.87% in the ‘cellular process’ category (Figure 5.4). The combined number of sequences across the different classifications is more than the sequences annotated with the GO terms as a single sequence can be described by several terms in the three classifications.

However, 1,226 of the microsporidial contigs had Blast hits, of which 824 (67%) were annotated with GO terms. A total of 339 sequences were annotated with the molecular processes classification, 28.69 % belonged to hydrolase activity and 20.88% belonged to nucleic acid binding category. Of the total of 224 sequences annotated with cellular component classification, 42.31% appeared in the cell and 31.60% belonged to the organelle category. Within the classification of biological processes, 343 sequences were annotated, 38.80% appeared in the cellular process category followed by 35.18% in the metabolic process category (Figure 5.4).

Figure 5.4 Gene ontology (GO) assignment for the *Gastrophysa viridula* contigs (Gv) and microsporidial contigs (Ms). The data presented represents the level 2 analysis, illustrating general functional categories



InterPro database was also used to classify the sequences on the basis of the putative function. The summary of the top ten superfamilies and domains for the Green dock beetle and microsporidial sequences is shown in Table 5.3. Insect gut has a major role in digestion and xenobiotic metabolism therefore higher frequency of the cytochrome P450s and protein kinase is expected (as seen in Table 5.3). However Green dock beetle sequences had over-represented small GTPase superfamilies and Zinc finger domains which is not the case in other available insect midgut datasets (Pauchet *et al.*, 2009, Pauchet *et al.*, 2010). The most frequent InterPro superfamily/domain in the microsporidial sequences were proteasome/chaperonin superfamilies and WD 40 repeat domains which are used in protein-protein interactions.

Table 5.3 (a) Summary of the top ten InterPro superfamilies and domains represented in the *Gastrophysa viridula* larval midgut transcriptome (Gv)

SUPERFAMILY (Gv)			**DOMAIN (Gv)**		
Interpro	Frequency	Description	Interpro	Frequency	Description
IPR001806	41	Small GTPase superfamily	IPR007087	173	Zinc finger, C2H2
IPR003579	34	Small GTPase superfamily, Rab type	IPR015880	151	Zinc finger, C2H2-like
IPR020849	34	Small GTPase superfamily, Ras type	IPR013087	139	Zinc finger C2H2-type/integrase DNA-binding domain
IPR001128	33	Cytochrome P450	IPR011009	105	Protein kinase-like domain
IPR005828	32	General substrate transporter	IPR015943	97	WD40/YVTN repeat-like-containing domain
IPR002198	31	Short-chain dehydrogenase/reductase SDR	IPR016040	95	NAD(P)-binding domain
IPR011701	30	Major facilitator superfamily	IPR017986	89	WD40-repeat-containing domain
IPR003578	29	Small GTPase superfamily, Rho type	IPR000719	86	Protein kinase, catalytic domain
IPR002347	27	Glucose/ribitol dehydrogenase	IPR013083	83	Zinc finger, RING/FYVE/PHD-type
IPR013128	21	Peptidase C1A, papain	IPR016024	73	Armadillo-type fold

(b) Summary of the top ten InterPro superfamilies and domains represented in the separated microsporidial sequences (Ms).

SUPERFAMILY (Ms)			**DOMAIN (Ms)**		
InterPro	Frequency	Description	InterPro	Frequency	Description
IPR001353	9	Proteasome, subunit alpha/beta	IPR015943	31	WD40/YVTN repeat-like-containing domain
IPR002423	8	Chaperonin Cpn60/TCP-1	IPR017986	29	WD40-repeat-containing domain
IPR001208	7	Mini-chromosome maintenance, DNA-dependent ATPase	IPR011009	26	Protein kinase-like domain
IPR003579	7	Small GTPase superfamily, Rab type	IPR000719	24	Protein kinase, catalytic domain
IPR015712	6	DNA-directed RNA polymerase, subunit 2	IPR003593	20	AAA+ ATPase domain
IPR001806	6	Small GTPase superfamily	IPR014001	20	Helicase, superfamily 1/2, ATP-binding domain
IPR003578	6	Small GTPase superfamily, Rho type	IPR001650	19	Helicase, C-terminal
IPR020849	6	Small GTPase superfamily, Ras type	IPR016024	17	Armadillo-type fold
IPR017998	6	Chaperone, tailless complex polypeptide 1	IPR002290	16	Serine/threonine- / dual-specificity protein kinase, catalytic domain
IPR023332	5	Proteasome A-type subunit	IPR014729	14	Rossmann-like alpha/beta/alpha sandwich fold

5.3.5 GC content and codon usage

The overall GC content and the codon usage of the separated Green dock beetle and the microsporidial assemblies were compared with the other four beetle datasets (Cowpea weevil, Poplar leaf beetle, Colorado potato beetle and Rice weevil). The overall GC content of majority of the beetle dataset was in the range of 37-38% with the exception of rice weevil (*Sitophilus oryzae*) which had a GC content of 35.69%. The microsporidial assembly had a lower GC content (34.08%) than the beetle assemblies (Table 5.4). Similarly the codon bias results showed differences in the frequency of codon usage for three amino acids, between the beetle and the microsporidial datasets. (Table 5.5, Appendix D TableS5.1)

TABLE 5.4 Overall %GC content for each of the five beetle transcriptomes and microporidial sequences

Organism	%GC content
Callosobruchus maculatus	38.38
Chrysomela tremulae	37.92
Leptinotarsa decemlineata	37.76
Sitophilus oryzae	35.69
Gastrophysa viridula	38.12
Microsporidia	34.05

TABLE 5.5 Codon usage frequencies (percent) for three amino acids – isoleucine, serine and valine in *Callosobruchus maculatus* (Cm), *Chrysomela tremulae* (Ct), *Leptinotarsa decemlineata* (Ld), *Sitophilus oryzae* (So), *Gastrophysa viridula* (Gv) and microsporidia (Ms)

Codon	Amino acid	Cm	Ct	Ld	So	Gv	Ms
ATA	Isoleucine	35.8	33.9	33.5	36.6	35.5	**44.7**
ATC	Isoleucine	24.4	24	24.1	20.7	24	19.4
ATT	Isoleucine	**39.8**	**42.1**	**42.4**	**42.7**	**40.5**	35.8
AGC	Serine	14	12	11.4	12.3	10.8	10.6
AGT	Serine	18.5	17.7	17.8	19.3	18.1	**25.3**
TCA	Serine	**22.8**	**25.3**	**24.3**	21	**26.1**	24
TCC	Serine	13.9	14.6	14.9	14.2	13.3	10.5
TCG	Serine	9.4	9	10.2	10.6	10.7	8.6
TCT	Serine	21.4	21.4	21.4	**22.6**	21.1	20.9
GTA	Valine	27.3	25.5	25.4	28	26.9	**37.9**
GTC	Valine	18.7	18	18.6	17.8	18.3	15.9
GTG	Valine	21.3	22.4	20.5	18.6	22.2	21.2
GTT	Valine	**32.7**	**34.1**	**35.5**	**35.6**	**32.6**	24

5.4 DISCUSSION

The main aim of this chapter was to survey the five beetle transcriptomes and to classify and separate sequences of microbes/contaminants, if present in the sampled tissue. It was found that out of the five datasets only the Green dock beetle dataset had a significant proportion of microsporidial sequences present. Hence this dataset was analysed further using a set of in house derived analysis pipelines. As the sequence dataset was from a non-model organism, it was essential to use a strategy which could classify the sequences according to their taxa with a reasonable degree of confidence and without the prior need to know the sequence of the host or the microbe.

Separation of mixed sequences can be achieved by different methods ranging from homology based approach, machine learning and statistical methods. The canonical approach uses homology searches to resolve the origin of the species. Whereas, other approaches, separate sequences based on GC content, codon usage, etc (such as GC method (Huitema *et al.*, 2003), problastic approach (Maor *et al.*, 2003), likelihood method (Hraber and Weller, 2001) and Support Vector Machines (Rudd and Tetko, 2005, Friedel *et al.*, 2005). These methods have different caveats and the choice of method depends on the type of dataset. If the genome sequence of the organism (or closely related species) is available, the sequences can be directly mapped to the genomes. But for non-model organism this is not the case. Methods that rely on GC content or codon usage cannot be used if both the organisms have similar codon usage. Similarly, biased representation of the taxa in existing databases decreases the reliability of homology approaches such as BLAST. This problem has been tackled by using advanced methods where restricted databases were used for homology search. Hsiang and Goodwin (Hsiang and Goodwin, 2003) used subset of database consisting of a single plant and fungal genome, each closely related to the infected plant and fungal pathogen. Mixed sequences can also be classified using metagenomic methods (Mitra *et al.*, 2011, Brady and Salzberg, 2009). But most of these methods are often standardised for classifying prokaryotes and cannot reliably be used for eukaryotes. Kumar and Blaxter (Kumar and Blaxter, 2011) used a method similar to that used in metagenomics to separate mixed sequences from a nematode and its bacterial endosymbiont. They utilized the information of the sequence similarity search, sequence coverage and GC content of each sequence.

Separating mixed sequences becomes even more challenging if both the organisms are eukaryotic and have insufficient sequencing data available for the same or related species (Maor *et al.*, 2003). The accuracy also depends on the depth of coverage which is very low in case of data obtained from 454 pyrosequencing. The separation of the Green dock beetle dataset was difficult as the mixed sequences belonged to eukaryotic organisms. The sequences had low coverage and were found to have similar GC content, hence the sequence coverage and DNA composition was not considered while separating the mixed sequences. Instead the advanced homology based approach was implemented where two restricted databases consisting of arthropoda and microsporidia were used.

Blast annotation and taxonomic composition results clearly show that out of the five beetle datasets, Green dock beetle had the highest percentage of contamination. The majority of the non arthropoda sequences had homology to microsporidia (Figure 5.2). Pauchet *et al.* (Pauchet *et al.*, 2010) found similar results when they surveyed the same EST datasets for homologs of the ribosomal protein genes from *Tribolium.* For the four beetle EST datasets (Cowpea weevil, Poplar leaf beetle, Colorado potato beetle and Rice weevil), only a single set of transcipts were found which indicated that no contamination was present. In contrast, for the Green dock beetle EST dataset they obtained two distinct hits. According to their analysis the second set of transcripts came from another eukaryotic organism, presumably microsporidial contaminant. Approximately 6% of all green dock beetle contigs matched one of the two microsporidian genomes, *Nosema ceranae* or *Encephalitozoon cuniculi.*

Analysis of the separated sequences, suggest that the Green dock beetle was infected with microsporidia, when sequenced. The InterProScan results further support this observation. Apart from the superfamilies/domains expected to be dominant in the green dock beetle gut sequences (such as cytochrome P450 and protein kinase), overrepresentation of small GTPase superfamilies and Zinc finger domains were observed (Table 5.3). These superfamilies and domains are associated with various signalling pathways (Raymond *et al.*, 2001, Berrocal-Lobo *et al.*, 2010). Insects have an innate immune response system which helps it to effectively combat a pathogen attack. Different signalling pathways are used to activate this innate immune response, depending on the type of pathogen present (Tsakas and Marmaras, 2010). Several studies have shown that the midgut plays a critical role in the immune response as well as digestion and xenobiotic metabolism (Freitak *et al.*, 2009, Xu *et al.*, 2012). The higher frequency of the mentioned superfamilies and domains in this beetle dataset could thus represent a stress response to the microsporidial infection.

Similarly, higher frequency of proteasomes, chaperonins and WD 40 repeats in the microsporidial sequences could be associated with the variety of stress encountered by this organism within the host. Strong representation of the proteasomes involved in protein degradation and proteins of chaperonins family was needed to ensure reliable protein folding in another microsporidia, *T. hominis* (Heinz *et al.*, 2012). Proteasomes are multi subunit proteins in eukaryotes that are believed to be involved in selective intracellular proteolysis, antigen presentation and cell cycle regulation. The WD40 repeats are found in a number of eukaryotic proteins involved in signal transduction, RNA processing, gene regulation, cell division, cytoskeleton assembly and protein degradation (Reddy *et al.*, 2008, Smith *et al.*, 1999). Microsporidia have a highly reduced genome and like many other intracellular pathogens they tend to lose many protein and biochemical pathways as they rely on the host to provide the required substrate (Peyretaillade *et al.*, 2011). The proteins which are retained are highly derived and prone to misfolding, therefore the large representation of the above mentioned superfamilies/domains may be required to maintain the protein functionality. Hence suggesting that, these superfamilies/domains play an important role in cellular defence system and provide it resistance against host.

Microsporidia constitute a diverse group of obligate intracellular parasites that can infect many eukaryotes including humans and insects (Dussaubat *et al.*, 2012). These pathogens belong to a group of highly reduced fungi. There are over 1200 species of microsporidia and relatively small number of genomes are available till date (Katinka *et al.*, 2001). Genome mapping tools, homology searches and percent GC content were used in order to determine the species of microsporidia sequences obtained from the green dock beetle dataset. Although BLASTx results had maximum hits against *N. ceranea* sequences, less than 3 % of the sequences could be mapped to the *N. ceranea* genome. Similar results were obtained for the BLASTn results. The percent GC content of the microsporidial sequences too was different from the GC content of the *N. ceranea* genome. Diversity in the percent GC content within the same genus has also been observed in sequenced genomes of other microsporidial genus (Appendix D TableS5.2). All this information suggested that the microsporidial sequences in the current study do not belong to *N. ceranea* but a distantly related species of the same genus. However due to the lack of sequenced microsporidial genomes, the exact species could not be determined.

5.5 CONCLUSION

In summary, the survey of the five beetle transcriptome datasets revealed that Green dock beetle contained a significant number of sequences belonging to microsporidia. Overrepresentation of superfamilies and domains implicated in the immune response suggested that the Green dock beetle midgut was infected with this obligate intracellular organism at the time of sequencing. Further investigation of the contigs containing these motifs can give a better insight into the interaction of this beetle host with the microporidial pathogen. Therefore, instead of considering mixed samples as a problem, it can be regarded as means of gaining insight into simultaneous gene expression of both the host and the associated microorganism. The choice of method used for separating mixed sequences depends on the organisms and the sequence information available. Due to the affordability of the next generation sequencing technology, the wealth of sequencing information is expanding rapidly. Hence, the advanced homology method can be used to classify the sequences according to their origin with reasonable degree of accuracy without the need to know the genome/ transcriptome sequences beforehand.

CHAPTER 6

An overview of the impact of next generation sequencing on the annotation of gene families involved in xenobiotic metabolism

Four different NGS datasets from the non-model organisms *T. vaporariorum*, *H. melpomene*, *L. sericata* and *G. viridula* were analysed in this thesis. Molecular research in these insects was hindered in the past by limited sequence information. In the work presented here, transcriptome data generated by 454 pyrosequencing was used as a starting point to study the genomics of these ecologically and economically important insects. Transcriptome data was used to generate a reliable reference for more detailed functional characterisation (Chapters 2, 4 and 5). In the case of the butterfly, *H. melpomene*, transcriptome data was used to identify and validate the cytochrome P450 gene models, establish intronic-exonic regions and to identify alternatively spliced regions (Chapter 3). Furthermore, Illumina generated RNA-seq data was used for SNP characterisation in *L. sericata* (Chapter 4).

6.1 NGS and Xenobiotic metabolising gene families in insects

The projects discussed in this thesis describe the transcriptomics of non-model insect species, including the assembly and annotation of the associated ESTs, and their potential to develop functional genomics tools for these species. The main focus of the study was to annotate the three gene superfamilies (P450s, GSTs and CCEs) involved in detoxification and to carry out comparative genomics to identify potential genes associated with resistance. These gene superfamilies have been the subject of analysis for every sequenced insect genome, mainly because they include resistance gene candidates. The extent and diversity of these three gene families shows large variation in the annotated insect species *(Adams et al.*, 2000, Holt *et al.*, 2002, Nene *et al.*, 2007, Yu *et al.*, 2009, Junwen *et al.*, 2011) and this diversity will be further discussed in this closing chapter.

Insecticide resistance across many species has been attributed to the up-regulation of enzymes associated with the xenobiotic detoxification and metabolism (Feyereisen, 1995), whereas the under representation of these enzymes in species like the Honeybee have been linked to their unusual sensitivity to insecticides (Claudianos *et al.*, 2006). Major pests like Blowfly and Whitefly have proven remarkably adapt in evolving resistance to insecticides due to mutations or over expression of genes belonging to these gene families (Oakeshott *et al.*, 2003, Newcomb *et al.*, 2005, Wang *et al.*, 2010).

Across different insect orders, specific trends can be observed within different clades/classes of these gene families. The major differences are in the clades/classes involved in xenobiotic metabolism. In GSTs the delta and epsilon classes are often involved in insecticide resistance. Similarly, members of CYP3 and CYP4 clades of P450s are associated with xenobiotic metabolism. In case of CCE enzymes implicated in imparting resistance in most cases belong to the dietary/detoxification class. Identification and further characterisation of these genes can improve our understanding of the rapid evolution and adaptation of insects to changes in their environment.

The data generated in the current study helped to identify several putative candidates involved in xenobiotic metabolism. Potential detoxification transcripts represented in *T. vaporariorum* dataset included 17 GSTs, 57 P450s and 27 CCEs. Comparative analysis of *T. vaporariorum* dataset with other insects of the same order revealed several putative candidates that could be involved in insecticide resistance. 10 putative GSTs were found in the Delta-Epsilon class. In case of P450s - CYP6CM2, CYP6CM3, CYP6DP1, CYP6DZ1 and other members of CYP3 and CYP4 clades identified in this study could be implicated in imparting insecticide resistance to *T. vaporariorum.* It had twice as many CCEs in dietary/detoxification class as compared to other sequenced insects of the same order (*A. pisum* and *M. persicae*). One of the putative members of this class (contig 12282) had high homology to *COE1* gene (accession number ABV45410) of *B. tabaci* which is found to be over expressed in strains resistant to organophosphates (Alon *et al.*, 2008). Another member of the same clade (contig 12863) was identified in only TV6 library which was constructed using imidacloprid resistant strain. This could therefore have a potential role in the neonicotinoid resistance of this strain.

Similarly, representatives of all the three enzyme superfamilies were found in the transcriptome of *L. sericata.* Potential detoxification transcripts represented in this dataset included 28 GSTs, 56 P450s and 34 CCEs. Several putative candidates in the clades/classes implicated in xenobiotic metabolism were identified. Two putative *L. sericata* sequences (Ls11154, Ls44527) of the dietary/detoxification class of CCEs were found to be highly similar to *L. cuprina aE7* gene which is associated with organophosphate resistance (Hartley *et al.*, 2006).

Analysis of the *H. melpomene* genome revealed several candidates for the biosynthesis and detoxification of xenobiotics. The number of P450s in insect genomes currently ranges from 48 in *Apis mellifera* to 164 in *Aedes aegyptii* (Consortium, 2006, Nene *et al.*, 2007). 100 P450s were identified in *H. melpomene* genome. Comparison with *B. mori* revealed several *H. melpomene* and Lepidopteran specific gene expansions.

With an increasing number and diversity of insect genomes becoming available, the diversity of these three enzyme superfamilies can be better appreciated and their evolution in insects can be further understood. The availability of sequencing information from a wide range of insects may reveal more functions for these large superfamilies. This in turn can improve our understanding of their importance in the adaptation of insects to new ecological niches.

6.2 Advantages of utilising NGS technologies in non-model species research

Next generation sequencing technologies (NGS) have a great potential to help address numerous gene and genome level questions in molecular biology, by providing a rapid and cost effective means of generating sequencing resources for almost any organism. NGS provides a means to engage in the genomics of non-model species that lack other sequence resources and have no prior history of functional genetics. Thus shifting the research focus from few laboratory based model organisms towards natural populations of wider ecological and evolutionary relevance (Ekblom and Galindo, 2011).

The choice of NGS platform used for sequencing depends on the question being addressed and readily available genomic resources for the organism of interest or its related species. While Roche 454 pyrosequencing is better for *de novo* sequencing due to long sequence reads, Illumina is preferred for gene expression studies and SNP characterisation as it provides deep coverage. Most published works on non-model organisms have utilised Roche 454 pyrosequencing for assembly and annotation, due to its longer reads (400-700bp) (Kumar and Blaxter, 2010). As it is not feasible to invest heavily in genome resources for every species or natural populations, transcriptome sequencing projects for non-model organism are a better alternative, as they cost less and are computationally more tractable (Kumar, 2010). Furthermore, concentrating the sequencing effort on the expressed part of the genome allows analysis of the expressed region which cannot be predicted from the genome sequence alone (Mundry *et al.*, 2012).

6.3 Bioinformatic challenges and limitations of NGS data in non-model organism research

Rapid progress in the NGS technologies is exponentially increasing the sequence throughput and large scale studies in genomics/transcriptomics of non-model organisms are becoming a reality (Liu *et al.*, 2012). With the advancement in the NGS technologies, the focus is now shifting towards solving the bioinformatics challenges in order to make sense of large datasets. Although the actual cost of sequencing has substantially decreased, before planning the methodology, it is important to consider other expenses, skills and infrastructure required for sample preparation and data analysis. The storage and analysis of the large volume of sequencing data is very challenging (Mundry *et al.*, 2012). Computing power can also be an issue for the data processing of these large datasets. Computing systems with large memories and multiple cores are required for the parallel processing of the sheer volume of NGS data. Alternatively, cloud computing resources can be rented as a service and the end users do not require the knowledge for the maintenance, physical location or configuration of the system that delivers the service. It is not only an attractive framework for parallel computing for data analysis, but it also allows for massive data storage (Martin and Wang, 2011, Thudi *et al.*, 2012). Therefore, cloud computing is a better alternative especially for labs that occasionally generate NGS dataset, as it provides computational facilities and storage services on a need to use basis.

In addition to the infrastructure, another important consideration is the type of NGS data generated. If the whole genome sequencing is not the end goal, transcriptome or expressed sequence tags (ESTs) can be generated instead, which costs less but still yields sufficient information to develop genomic resources for any organism (Bouck and Vision, 2007). However, use of EST data for study of gene families has certain limitations. Transcripts from genes expressed at a low level may not be sequenced at all. Hence, absence of a particular transcript from a given dataset is not a strong evidence for the absence of the corresponding gene from the genome; it may simply remain to be discovered. When using EST data for gene discovery, it gives no information about the genomic position, gene order or location of structures such as introns. Without the positional information accurate identification of species specific genes and gene family expansions can be problematic. And homologous relationships between chromosomal regions cannot be conclusively established (Bouck and Vision, 2007). Although this can be deduced by comparing with genomic sequence from closely related species, for non-model organisms this is usually not the case.

Furthermore, it is difficult to distinguish alleles and alternative spliced variants from paralogs during assembly. Lack of a reference genome makes it difficult to validate the assembly and to evaluate the assembled contigs (Wang *et al.*, 2004). ESTs are subject to sampling bias (i.e. underrepresentation of rare transcripts) this can be overcome to a certain extent by sampling RNA from multiple tissue types and from different developmental stages. In order to further reduce redundancies, cDNA libraries can be standardised by using normalisation procedure. This can enrich the diversity of transcripts captured in the EST collection (Nagaraj *et al.*, 2007). However this is not required when carrying out gene expression analysis.

Another most widely used application of NGS in non-model organisms is rapid and cost effective SNP characterisation. SNPs can be obtained from either genome or transcriptome of any organism. Variant calling tools identify same base call discrepancies occurring in multiple sequences as SNPs, assuming that redundant reads represent actual SNPs rather than sequencing error. In most cases this can be dealt by using stringent filtering parameters to select only reliable good quality SNPs. One obvious problem with SNP detection from NGS data is that sequencing errors show a similar signature to low frequency SNP alleles. It is even more problematic to confirm SNPs with low coverage or towards the end of the reads where the sequence quality is comparatively lower.

As NGS is becoming affordable, the type of organisms being sequenced has shifted from few laboratory-based organisms to wild populations. These natural populations can often be contaminated and obtaining pure samples by separating mixed sequences can be problematic if there is no prior knowledge of the contaminant. Sequence separation becomes even more challenging if both the organisms are eukaryotic and have insufficient sequencing data available for the same or related species. Various strategies can be applied to separate mixed sequences, however they all have different caveats. It is essential to select the method on the basis of the type of organisms and the sequence information available. However instead of considering mixed sequences as a problem, it can be regarded as means of gaining insight into host-parasite interaction and studying pathogen-infected host tissues (Kumar and Blaxter, 2011).

With the decrease in sequencing costs and increases in accuracy, sequence length and coverage, the future bottle neck is most likely be the data analysis rather than generating sequence data. Assembling the data is computationally intensive and parameter settings need to be determined depending on the properties of the dataset. However, advances in high performance computing will greatly reduce the time required for analysis of large datasets. Despite the above mentioned caveats, NGS technologies and bioinformatics can open up avenues to develop functional genomics resources for diverse species of interest to ecologists and evolutionary biologists.

6.4 FUTURE PERSPECTIVES

The work presented in this thesis has helped in developing genomic resources for four non-model insect species (*T. vaporariorum*, *H. melopmene*, *L. sericata* and *G. viridula*), which in turn can help in their future research. In chapter 2, annotation of *T. vaporariorum* transcriptome for gene families involved in xenobiotic metabolism revealed a number of transcripts that have potential role in conferring insecticide resistance. Elucidation of the exact function of these gene candidates and their potential role in imparting insecticides resistance can provide an important insight into the ability of *T. vaporariorum* to respond to environmental changes.

The main aim of Chapter 3 was to identify all the *CYP* genes in the genome and transcriptome of *H. melpomene*. Annotation of P450 encoding genes in the *H. melpomene* genome revealed co-localisation of cytochrome P450 genes with olfactory receptor (*OR*) genes. The possibility that the products of these genes operate in the same system for the detection and metabolism of related compounds is an exciting one and is worthy of further experimental investigation. We also found four putative candidates with close orthology to *CYP* genes involved in cyanogenesis in *Z. filipendulae.* To confirm the role of these genes in *Heliconius* adaptation to cyanogenic host plants would require more experimentation.

The SNPs generated in Chapter 4 need to be validated by alternative methods before being used in further studies as true genetic markers. The transcriptomic reference and validated SNPs can therefore serve as the basis for future SNP analysis and enhance our understanding of the spread of malathion and diazinon resistance in different *L. sericata* populations from around the world. In addition, the annotated sequences will facilitate the investigation of fundamental biology of *L. sericata* and can be further compared to its sibling species *L. cuprina.*

Finally, the data in Chapter 5 can help in understanding host-parasite interactions. The Green dock beetle midgut data was found to be contaminated with microsporidia, suggesting that the insect was infected when sequenced. Overrepresentation of superfamilies and domains implicated in the immune response can give a better insight into the interaction of this beetle host with the microsporidial parasite. Furthermore, investigating contigs containing these motifs can help understand the innate response of the host to infection. Similarly thorough examination of the separated microsporidial data can help to understand how endoparasites like microsporidia can cope with variety of stress encountered inside the host. Thus providing the basis for further exploration and understanding of the fundamental mechanisms of stress response.

APPENDIX A

Table S2.1: List of putative *Trialeurodes vaporariorum* cytochrome P450s with complete names

CLADE	**CYP**	**CLADE**	**CYP**
CYP2	CYP18A1	CYP4	CYP4AV6
CYP2	CYP304G1	CYP4	CYP4C63
CYP2	CYP306A1	CYP4	CYP4CR1
CYP3	CYP6CM2	CYP4	CYP4CS1
CYP3	CYP6CM3	CYP4	CYP4CT1
CYP3	CYP6CM4	CYP4	CYP4G59
CYP3	CYP6DB2	CYP4	CYP4G60
CYP3	CYP6DP1	CYP4	CYP4G61
CYP3	CYP6DP2	CYP4	CYP380D1
CYP3	CYP6DQ1	CYP4	CYP380E1
CYP3	CYP6DR1	CYP4	CYP403A1
CYP3	CYP6DS1	CYP4	CYP403A2
CYP3	CYP6DS2	Mito	CYP301A1
CYP3	CYP6DS3	Mito	CYP301B1
CYP3	CYP6DT1	Mito	CYP302A1
CYP3	CYP6DT2	Mito	CYP314A1
CYP3	CYP6DT3	Mito	CYP315A1
CYP3	CYP6DT4	Mito	CYP353C1
CYP3	CYP6DT5	Mito	CYP404A1
CYP3	CYP6DT6		
CYP3	CYP6DT7		
CYP3	CYP6DU1		
CYP3	CYP6DV1		
CYP3	CYP6DV2		
CYP3	CYP6DV3		
CYP3	CYP6DW1		
CYP3	CYP6DX1		
CYP3	CYP6DX2		
CYP3	CYP6DY1		
CYP3	CYP6DZ1		
CYP3	CYP6DZ2		
CYP3	CYP6EA1		
CYP3	CYP401A1		
CYP3	CYP402A1		
CYP3	CYP402B1		

APPENDIX B

Table S3.1: Number of cytochrome P450 genes in *Heliconius melpomene*

CLADE	orientation	HMEL stable id	Scaffold	start of gene	end of gene
CYP2					
CYP15C1	forward	HMEL006305	scf7180001249124	36557	46518
CYP18A1	forward	HMEL005066	scf7180001247540	51408	55164
CYP306A1	reverse	HMEL005067	scf7180001247540	58791	52970
CYP303A1	reverse	HMEL007289	scf7180001249384	275986	264001
CYP304F6	reverse	HMEL015610	scf7180001250686	124281	114056
CYP304F5	forward	HMEL015784	scf7180001250700	236595	241698
CYP305B2	forward	HMEL010376	scf7180001250079	12208	24960
CYP305B1	forward	HMEL010377	scf7180001250079	35227	51919
CYP307A1	reverse	HMEL011518	scf7180001250246	191834	184993
CYP3					
CYP6AB18	reverse	HMEL011177	scf7180001250217	219238	216935
CYP6AB15	reverse	HMEL011177	scf7180001250217	212312	207936
CYP6AB29	reverse	HMEL011177	scf7180001250217	206504	203038
CYP6AB16	reverse	HMEL008256	scf7180001249662	12767	15027
CYP6AB17	reverse	HMEL008256	scf7180001249662	4935	6821
CYP6AB25	forward	HMEL003453	scf7180001245305	4308	7079
CYP6AB26	reverse	HMEL003682	scf7180001245570	187498	189657
CYP6AB27	reverse	HMEL012288	scf7180001250361	521377	522537
CYP6AB30	reverse	HMEL012288	scf7180001250361	511264	512659
CYP6AE33	reverse	HMEL010595	scf7180001250135	42305	44172
CYP6AE34	reverse	HMEL010595	scf7180001250135	45815	47967
CYP6AE35	forward	HMEL003566	scf7180001245448	281303	283277
CYP6AE40	reverse	HMEL008479	scf7180001249711	366070	364035
CYP6AE41	reverse	HMEL008479	scf7180001249711	369660	367633
CYP6AE42	reverse	HMEL008479	scf7180001249711	373241	371335
CYP6AN9	forward	HMEL007566	scf7180001249436	27545	29512
CYP6AN11	forward	HMEL007566	scf7180001249436	32787	34973
CYP6AN12	reverse	HMEL007567	scf7180001249436	37134	39677
CYP6AN13	forward	HMEL007568	scf7180001249436	40747	42513
CYP6AN10	forward	HMEL013102	scf7180001250438	161377	165974
CYP6CT2	forward	HMEL006595	scf7180001249232	2550	7523
CYP6FA1	forward	HMEL017153	scf7180001250808	10835	12839
CYP6FA2	reverse	HMEL017154	scf7180001250808	16092	17733
CYP9A45	reverse	HMEL013741	scf7180001250505	196327	203153
CYP9A41	forward	HMEL013737	scf7180001250505	159394	167933
CYP9A42	forward	HMEL013742	scf7180001250505	205630	212225

CYP9A43	reverse	HMEL013739	scf7180001250505	177227	191055
CYP9A44	reverse	HMEL013738	scf7180001250505	168863	176210
CYP9A46	forward	HMEL012008	scf7180001250330	4486	10545
CYP9G9	reverse	HMEL006125	scf7180001249092	1162	10532
CYP324A1	forward	HMEL017463	scf7180001250824	968626	973286
CYP324A3	reverse	HMEL017464	scf7180001250824	992015	999228
CYP324A4	reverse	HMEL017464	scf7180001250824	989411	986114
CYP324A2	reverse	HMEL005373	scf7180001247934	6490	10581
CYP324A5	reverse	HMEL005373	scf7180001247934	1	4237
CYP332A1	forward	HMEL013081	scf7180001250434	200752	205929
CYP337C1	forward	HMEL004530	scf7180001246721	343856	346033
CYP321C2	reverse	HMEL004447	scf7180001246642	19990	21489
CYP337C3	reverse	HMEL004449	scf7180001246642	90665	93667
CYP337C4	reverse	HMEL004449	scf7180001246642	94514	97719
CYP337C5	forward	HMEL009495	scf7180001249880	312093	313946
CYP354A6	forward	HMEL012954	scf7180001250418	546296	552720
CYP354A7	forward	HMEL012955	scf7180001250418	557646	560846
CYP4					
CYP405A5	reverse	HMEL016673	scf7180001250781	742844	746355
CYP405A6	reverse	HMEL016674	scf7180001250781	749004	754629
CYP405A4	reverse	HMEL016675	scf7180001250781	759942	764980
CYP4G71	reverse	HMEL016705	scf7180001250781	1337669	1347334
CYP4G73	reverse	HMEL006588	scf7180001249228	44520	50967
CYP4L22	forward	HMEL009876	scf7180001249962	66	7455
CYP4M27	forward	HMEL005932	scf7180001249024	200744	207450
CYP4M28	forward	HMEL005933	scf7180001249024	212577	216975
CYP4S16	forward	HMEL005553	scf7180001248286	127536	135130
CYP4CG6	reverse	HMEL010693	scf7180001250150	11219	19965
CYP4CG7	forward	HMEL002471	scf7180001242087	65878	71023
CYP4CG5	reverse	HMEL003583	scf7180001245460	17161	24633
CYP4CG9	reverse	HMEL003581	scf7180001245460	216	6396
CYP4CG10	reverse	HMEL003582	scf7180001245460	10610	15956
CYP4CG13	reverse	HMEL003584	scf7180001245460	27438	31702
CYP4CG4	reverse	HMEL008770	scf7180001249777	276044	282180
CYP4CG11	reverse	HMEL008768	scf7180001249777	269095	275154
CYP4CG12	reverse	HMEL008772	scf7180001249777	285346	286728
CYP340R6	reverse	HMEL013957	scf7180001250535	39781	50562
CYP340P1	reverse	HMEL013954	scf7180001250535	387	3811
CYP340R3	reverse	HMEL013955	scf7180001250535	11714	12931
CYP340R4	reverse	HMEL013956	scf7180001250535	24632	30083

CYP340R2	forward	HMEL010931	scf7180001250201	7526	19989
CYP340R5	reverse	HMEL010936	scf7180001250201	39882	47608
CYP340N1	forward	HMEL010939	scf7180001250201	50719	55126
CYP340R1	forward	HMEL010930	scf7180001250201	2218	4819
CYP421A1	reverse	HMEL010934	scf7180001250201	22139	30520
CYP340Q1	forward	HMEL010935	scf7180001250201	33999	38269
CYP341E1	reverse	HMEL013801	scf7180001250519	36696	39668
CYP341E2	forward	HMEL013797	scf7180001250519	33	2170
CYP341E3	forward	HMEL005374	scf7180001247939	5383	9558
CYP341A8	forward	HMEL017148	scf7180001250807	923477	926856
CYP341A9	forward	HMEL017149	scf7180001250807	929771	936399
CYP341A10	forward	HMEL017150	scf7180001250807	945945	948771
CYP341F1	reverse	HMEL006318	scf7180001249127	53909	58711
CYP366B1	reverse	HMEL017070	scf7180001250804	49153	53603
CYP366B2	reverse	HMEL017074	scf7180001250804	70866	76335
CYP367A3	reverse	HMEL002116	scf7180001240868	91265	98439
CYP367B3	reverse	HMEL002115	scf7180001240868	81527	86401
Mito					
CYP333B13	Forward	HMEL002625	scf7180001242575	58256	61189
CYP333B14	Forward	-	scf7180001242575	64347	66683
CYP301A1	Forward	HMEL003058	scf7180001243663	37559	49276
CYP301B1	Reverse	HMEL007856	scf7180001249521	36409	45025
CYP49A1	Reverse	HMEL007941	scf7180001249546	38041	42251
CYP333-un1	Forward	HMEL012587	scf7180001250375	2856	5962
CYP333A4	Forward	HMEL012676	scf7180001250380	387826	391496
CYP302A1	Forward	HMEL016930	scf7180001250793	212441	214527
CYP339A2	forward	HMEL017316	scf7180001250816	148411	152568

Table S3.2: List of *Heliconius melpomene CYP* genes having EST support

Clade	*CYP* Genes
CYP2	CYP18A1
CYP2	CYP303A1
CYP2	CYP304F6
CYP2	CYP304F5
CYP2	CYP306A1
CYP3	CYP6AB18
CYP3	CYP6AB15
CYP3	CYP6AB16
CYP3	CYP6AB17
CYP3	CYP6AE33
CYP3	CYP6AE34
CYP3	CYP6AE35
CYP3	CYP6AE40
CYP3	CYP6CT2
CYP3	CYP9A45
CYP3	CYP9A41
CYP3	CYP9A42
CYP3	CYP9A43
CYP3	CYP9A44
CYP3	CYP9G9
CYP3	CYP324A1
CYP3	CYP324A3
CYP3	CYP332A1
CYP3	CYP337C1
CYP3	CYP354A6
CYP4	CYP4G71
CYP4	CYP4L22
CYP4	CYP4M27
CYP4	CYP4S16
CYP4	CYP4CG4
CYP4	CYP4CG5
CYP4	CYP4CG6
CYP4	CYP4CG7
CYP4	CYP405A4
MITO	CYP301B1
MITO	CYP333A4
MITO	CYP333B13

APPENDIX C

Table S4.1: List of putative *Lucilia sericata* cytochrome P450s with complete names

CLADE	**CYP**	**CLADE**	**CYP**
CYP2	CYP305A1	**CYP4**	CYP311A1
CYP2	CYP18A1	**CYP4**	CYP313B2
CYP3	CYP28E1	**CYP4**	CYP4AA1
CYP3	CYP28G1	**CYP4**	CYP4AC5
CYP3	CYP28G2	**CYP4**	CYP4AD1
CYP3	CYP28G3	**CYP4**	CYP4AE2
CYP3	CYP28G4	**CYP4**	CYP4D29
CYP3	CYP308B1	**CYP4**	CYP4D30
CYP3	CYP310B1	**CYP4**	CYP4G46
CYP3	CYP317A2	**CYP4**	CYP4G80
CYP3	CYP437A2	**CYP4**	CYP4P7
CYP3	CYP438A1	**Mito**	CYP301A1
CYP3	CYP6A27	**Mito**	CYP12A11
CYP3	CYP6A28	**Mito**	CYP12A7
CYP3	CYP6A29	**Mito**	CYP12G1
CYP3	CYP6A30	**Mito**	CYP12A9
CYP3	CYP6A31	**Mito**	CYP302A1
CYP3	CYP6A42	**Mito**	CYP12A10
CYP3	CYP6A43	**Mito**	CYP12D2
CYP3	CYP6A44		
CYP3	CYP6A45		
CYP3	CYP6A46		
CYP3	CYP6A47		
CYP3	CYP6A48		
CYP3	CYP6A49		
CYP3	CYP6C2		
CYP3	CYP6D6		
CYP3	CYP6D7		
CYP3	CYP6FR1		
CYP3	CYP6FR2		
CYP3	CYP6FS1		
CYP3	CYP6FT1		
CYP3	CYP6G3		
CYP3	CYP6G5		
CYP3	CYP6V2		
CYP3	CYP9F4		
CYP3	CYP9F5		

Figure S4.1 Distribution of SNP quality scores in the UK population of *Lucilia sericata*. The graph represents default quality SNPS before filtering using stringent parameter.

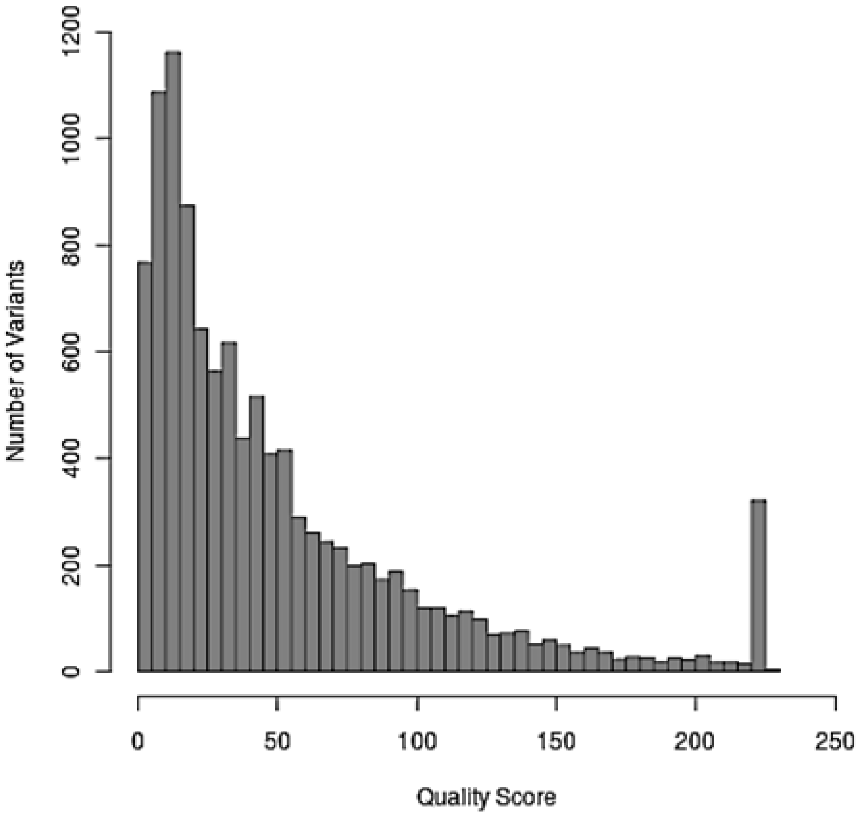
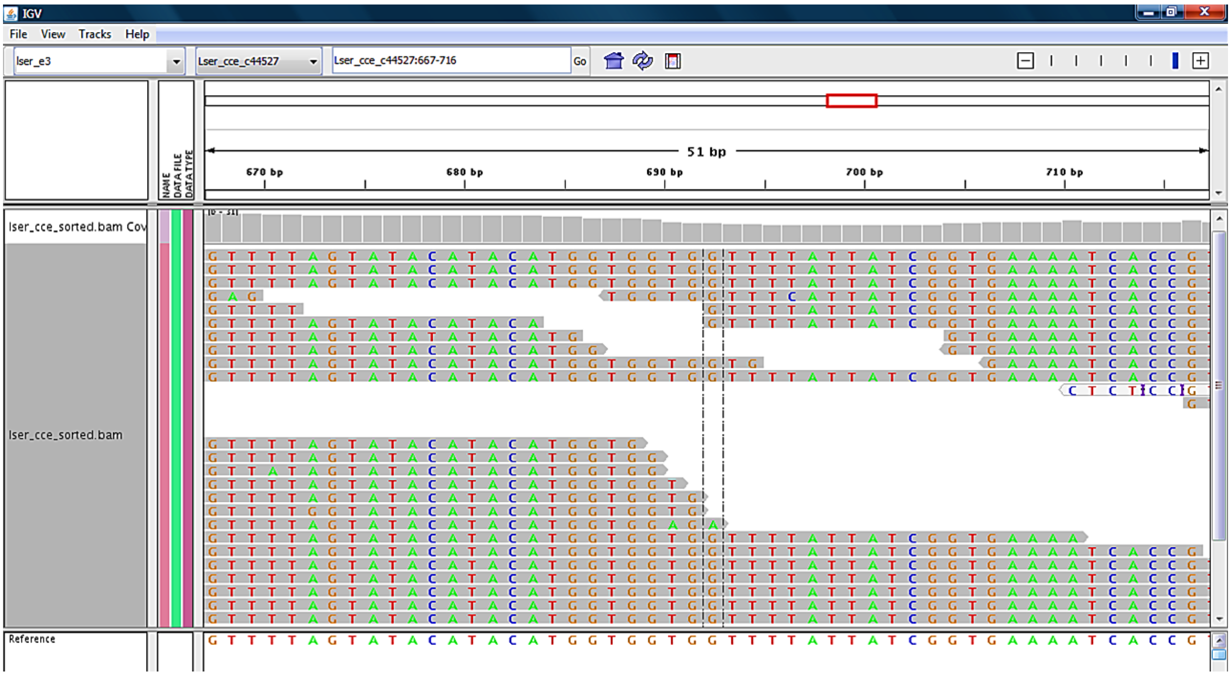


Figure S4.2 Screenshot of the IGV viewer showing diazinon region in the Ls44527 contig. The lser_cce_sorted.bam region is the mapped Illumina reads from the UK population. A single Illumina read has a SNP in the position associated with diazinon resistance.



lser_e3

Lser_cce_c44527

Lser_cce_c44527:667-716

APPENDIX D

TABLE S5.1 Codon usage frequencies (percent) for *Callosobruchus maculatus* (Cm), *Chrysomela tremulae* (Ct), *Leptinotarsa decemlineata* (Ld), *Sitophilus oryzae* (So), *Gastrophysa viridula* (Gv) and microsporidia (Ms)

codon	**AA**	**Cm**	**Ct**	**Ld**	**So**	**Gv**	**Ms**
GCA	A	0.346	0.352	0.348	0.322	0.361	0.44
GCC	A	0.201	0.219	0.211	0.215	0.204	0.14
GCG	A	0.132	0.116	0.127	0.145	0.128	0.123
GCT	A	0.321	0.312	0.314	0.318	0.306	0.292
TGC	C	0.392	0.38	0.376	0.351	0.362	0.307
TGT	C	0.608	0.62	0.624	0.649	0.638	0.693
GAC	D	0.381	0.346	0.366	0.373	0.347	0.388
GAT	D	0.619	0.654	0.634	0.627	0.653	0.612
GAA	E	0.66	0.68	0.678	0.689	0.682	0.709
GAG	E	0.34	0.32	0.322	0.311	0.318	0.291
TTC	F	0.358	0.386	0.387	0.315	0.41	0.341
TTT	F	0.642	0.614	0.613	0.685	0.59	0.659
GGA	G	0.337	0.356	0.359	0.341	0.364	0.374
GGC	G	0.203	0.185	0.184	0.197	0.185	0.164
GGG	G	0.167	0.186	0.177	0.17	0.169	0.127
GGT	G	0.292	0.274	0.28	0.292	0.282	0.335
CAC	H	0.387	0.372	0.378	0.377	0.366	0.37
CAT	H	0.613	0.628	0.622	0.623	0.634	0.63
ATA	I	0.358	0.339	0.335	0.366	0.355	**0.447**
ATC	I	0.244	0.24	0.241	0.207	0.24	0.194
ATT	I	**0.398**	**0.421**	**0.424**	**0.427**	**0.405**	0.358
AAA	K	0.665	0.674	0.683	0.712	0.676	0.643
AAG	K	0.335	0.326	0.317	0.288	0.324	0.357
CTA	L	0.13	0.12	0.115	0.127	0.127	0.151
CTC	L	0.107	0.117	0.121	0.093	0.12	0.071
CTG	L	0.15	0.146	0.147	0.125	0.148	0.104
CTT	L	0.187	0.184	0.19	0.182	0.179	0.194
TTA	L	0.207	0.195	0.196	0.27	0.188	0.245
TTG	L	0.219	0.238	0.231	0.203	0.239	0.236
ATG	M	1	1	1	1	1	1
AAC	N	0.398	0.362	0.382	0.367	0.356	0.35
AAT	N	0.602	0.638	0.618	0.633	0.644	0.65
CCA	P	0.389	0.405	0.39	0.379	0.401	0.422
CCC	P	0.167	0.186	0.178	0.171	0.167	0.124
CCG	P	0.13	0.117	0.135	0.15	0.131	0.13
CCT	P	0.314	0.291	0.297	0.3	0.301	0.324
CAA	Q	0.591	0.616	0.615	0.62	0.619	0.699
CAG	Q	0.409	0.384	0.385	0.38	0.381	0.301
AGA	R	0.342	0.372	0.346	0.339	0.359	0.387
AGG	R	0.208	0.211	0.198	0.186	0.19	0.164
CGA	R	0.146	0.151	0.165	0.158	0.175	0.149
CGC	R	0.087	0.071	0.073	0.084	0.073	0.07
CGG	R	0.084	0.086	0.09	0.093	0.081	0.067
CGT	R	0.134	0.109	0.128	0.14	0.122	0.163
AGC	S	0.14	0.12	0.114	0.123	0.108	0.106
AGT	S	0.185	0.177	0.178	0.193	0.181	**0.253**
TCA	S	**0.228**	**0.253**	**0.243**	0.21	**0.261**	0.24

TCC	S	0.139	0.146	0.149	0.142	0.133	0.105
TCG	S	0.094	0.09	0.102	0.106	0.107	0.086
TCT	S	0.214	0.214	0.214	**0.226**	0.211	0.209
ACA	T	0.367	0.37	0.36	0.361	0.381	0.451
ACC	T	0.2	0.212	0.198	0.192	0.184	0.119
ACG	T	0.142	0.124	0.134	0.149	0.133	0.126
ACT	T	0.291	0.294	0.308	0.298	0.302	0.304
GTA	V	0.273	0.255	0.254	0.28	0.269	**0.379**
GTC	V	0.187	0.18	0.186	0.178	0.183	0.159
GTG	V	0.213	0.224	0.205	0.186	0.222	0.212
GTT	V	**0.327**	**0.341**	**0.355**	**0.356**	**0.326**	0.24
TGG	W	1	1	1	1	1	1
TAC	Y	0.378	0.341	0.356	0.348	0.346	0.404
TAT	Y	0.622	0.659	0.644	0.652	0.654	0.596
TAA	*	0.384	0.355	0.37	0.479	0.333	0.408
TAG	*	0.243	0.221	0.216	0.226	0.226	0.256
TGA	*	0.373	0.424	0.413	0.295	0.441	0.337

TABLE S5.2 Percent GC content of scaffolds for the sequenced microsporidial genomes

Organism	**%GC**
N. parisii ERTm1	34.42
Nematocida sp1 ERTm2	38.34
N. parisii ERTm3	34.46
Nematocida sp1 ERTm6	38.3
V. culicis floridensis	39.75
V. corneae ATCC 50505	36.47
E. aedis	22.47
E. cuniculi EcI	46.9
E. cuniculi EcII	46.85
E. cuniculi EcIII	46.83
E. cuniculi GB-M1	47.31
E. intestinalis ATCC 50506	41.53
N. ceranae BRL01	25.27
A. algerae PRA109	23.3
A. algerae PRA339	23.41

Source: http://www.broadinstitute.org/annotation/genome/microsporidia_comparative/GenomeStats.html

APPENDIX E
LIST OF PUBLICATIONS

Related to the thesis

- **Chauhan, R.**, Jones, R., Wilkinson, P., Pauchet, Y., ffrench-Constant, R. H. (2013). Cytochrome P450 encoding genes from the *Heliconius* genome as candidates for cyanogenesis. (Insect Molecular Biology)

- The Heliconius Genome Consortium, Dasmahapatra, K. K., Walters, *et al.*, **Chauhan, R**., *et al.* (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487, 94-98. doi:10.1038/nature11041

- Karatolos, N., Pauchet, Y., Wilkinson, P., **Chauhan, R**., Denholm, I., Gorman, K., Nelson, D. R., Bass, C., ffrench-Constant, R. H., Williamson, M. S. (2011). Pyrosequencing the transcriptome of the greenhouse whitefly, *Trialeurodes vaporariorum*, reveals multiple transcripts encoding insecticide targets and detoxifying enzymes. *BMC Genomics*, 12, 56. doi:10.1186/1471-2164-12-56

Not related to the thesis

- Ferreira, P.G., Patalano, S., **Chauhan, R**., ffrench-Constant, R., Gabaldon, T., Guigo, R., Sumner, S. (2013). Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes. *Genome Biology*, 14, R20. doi:10.1186/gb-2013-14-2-r20

- Pauchet, Y., Wilkinson, P., **Chauhan, R**., ffrench-Constant, R. H. (2010). Diversity of beetle genes encoding novel plant cell wall degrading enzymes. PLoS One, 5, e15635. doi:10.1371/journal.pone.0015635

BIBLIOGRAPHY

ADAMS, M. D., CELNIKER, S. E., HOLT, R. A., EVANS, C. A., GOCAYNE, J. D. & AMANATIDES, P. G. E. A. (2000) The genome sequence of *Drosophila melanogaster*. *Science,* 287, 2185-2195.

ALDRIDGE, W. (1953) Serum esterases. 1. Two types of esterases (A and B) hydrolysing p-nitrophenyl acetate, propionate and butyrate, and a method for their determination. *Biochem J.,* 53, 110-117.

ALON, M., ALON, F., NAUEN, R. & MORIN, S. (2008) Organophosphates' 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 Biochem Mol Biol,* 38, 940-949.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990) Basic local alignment search tool. *J Mol Biol,* 215, 403-410.

ALZAHRANI, A. M. (2009) Insects Cytochrome P450 enzymes: Evolution, Function and Method of Analysis. *Global J Mol Sci,* 4, 167-179.

ANDERSEN, J., UTERMOHLEN, J. & FEYEREISEN, R. (1994) Expression of house fly CYPA1 and NADPH-cytochrome P450 reductase in *Escherichia coli* and reconstitution of an insecticide-metabolising P450 system. *Biochemistry,* 33, 2171-2177.

ANDERSON, G. S. (2000) Minimum and maximum development rates of some forensically important Calliphoridae (Diptera). *J Forensic Sci,* 45, 824–832.

ANSORGE, W. J. (2009) Next-generation DNA sequencing techniques. *New Biotechnol,* 25, 195-203.

BERNAYS, E. & GRAHAM, M. (1988) On the evolution of host specificity in phytophagous arthropods. *Ecology,* 69, 886–892.

BERROCAL-LOBO, M., STONE, S., YANG, X., ANTICO, J., CALLIS, J., RAMONELL, K. & SOMERVILLE, S. (2010) ATL9, a RING zinc finger protein with E3 ubiquitin ligase activity implicated in chitin- and NADPH oxidase-mediated defense responses. *PLoS One,* 5, e14426.

BIN, L., XIA, Q., LU, C., ZHOU, Z. & XIANG, Z. (2005) Analysis of cytochrome P450 genes in silkworm genome (*Bombyx mori*). *Sci China C Life Sci,* 48, 414-418.

BOUCK, A. & VISION, T. (2007) The molecular ecologist's guide to expressed sequence tags. *Mol Ecol,* 16, 907-24.

BRADY, A. & SALZBERG, S. L. (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods,* 6, 673-676.

BRØDSGAARD, H. & ALBAJES, R. (Eds.) (1999) *Insect and mite pests,* Dordrecht, The Netherlands, Kluwer Academic Publishers.

BROWN, J. K. & CZOSNEK, H. (2002) Whitefly transmission of plant viruses. *Adv Bot Res,* 36, 65-100.

BYRNE, D., BELLOWS, T. & PARRELLA, M. (Eds.) (1990) Whiteflies in agricultural systems, Andover, Hants, UK, Intercept Ltd.

CAMPBELL, P., ROBIN, G. D. Q., COURT, L., DORRIAN, S., RUSSELL, R. & OAKESHOTT, J. (2003) Developmental expression and gene/enzyme identifications in the alpha esterase gene cluster of *Drosophila melanogaster*. *Insect Mol Biol,* 12, 459-471.

CANTAREL, B. L., KORF, I., ROBB, S. M. C., PARRA, G., ROSS, E., MOORE, B., HOLT, C., ALVARADO, A.S., YANDELL, M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 14,188–196.

CAZANDERA, G., VEENA, K. E. B. V., BERNARDSB, A. T. & JUKEMA, G. N. (2009) Do maggots have an influence on bacterial growth? A study on the susceptibility of strains of six different bacterial species to maggots of *Lucilia sericata* and their excretions/secretions. *J Tissue Viability,* 18, 80-87.

CHELVANAYAGAM, G., PARKER, M. W. & BOARD, P. G. (2001) Fly fishing for GSTs: a unified nomenclature for mammalian and insect glutathione transferases. *Chem Biol Interact,* 133, 256-260.

CHE-MENDOZA, A., PATRICIA, P. R. & AMÉRICO, R. D. (2009) Insecticide resistance and glutathione S-transferases in mosquitoes: A review. *African J Biotech,* 8, 1386-1397.

CHOUGULE, N. & BONNING, B. (2012) Toxins for transgenic resistance to hemipteran pests. *Toxins,* 4, 405-429.

CLARK, A. & SHAMAAN, N. (1984) Evidence that DDT dehydrochlorinase from the house fly is a glutathione S transferase. *Pest Biochem Physiol,* 22, 249–261.

CLAUDIANOS, C., RANSON, H., JOHNSON, R. M., BISWAS, S., SCHULER, M. A., BERENBAUM, M. R., FEYEREISEN, R. & OAKESHOTT, J. G. (2006) A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Mol Biol,* 15, 615-636.

CLAUDIANOS, C., RUSSELLA, R. J. & OAKESHOTT, J. G. (1999) The same amino acid substitution in orthologous esterases confers organophosphate resistance on the house fly and a blowfly. *Insect Biochem Mol Biol,* 29, 675–686.

CONESA, A., GÖTZ, S., GARCÍA-GÓMEZ, J. M., TEROL, J., TALÓN, M. & ROBLES, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics,* 21, 3674-3676.

CONSORTIUM, H. G. S. (2006) Insights into social insects from the genome of the honeybee *Apis mellifera. Nature,* 443, 931-949.

CONSORTIUM, T. G. S. (2008) The genome of the model beetle and pest *Tribolium castaneum. Nature,* 452, 949-955.

CONSORTIUM, T. H. G. (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature,* 487, 94-98.

CONSORTIUM, T. I. A. G. (2010) Genome Sequence of the Pea Aphid *Acyrthosiphon pisum. PLoS Biology,* 8, e1000313.

COX-FOSTER, D., CONLAN, S., HOLMES, E., PALACIOS, G., EVANS, J., MORAN, N., QUAN, P., BRIESE, T., HORNIG, M., GEISER, D., MARTINSON, V., VANENGELSDORP, D., KALKSTEIN, A., DRYSDALE, A., HUI, J., ZHAI, J., CUI, L., HUTCHISON, S., SIMONS, J., EGHOLM, M., PETTIS, J. & LIPKIN, W. (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. *Science,* 318, 283–287.

DABORN, P. J., YEN, J. L., BOGWITZ, M. R., LE GOFF, G., FEIL, E., JEFFERS, S., TIJET, N., PERRY, T., HECKEL, D., BATTERHAM., P., FEYEREISEN, R., WILSON, T. G. & FFRENCH-CONSTANT, R. H. (2002) A single p450 allele associated with insecticide resistance in *Drosophila. Science,* 297, 2253-2256.

DE BODT, S., MAERE, S. & VAN DE PEER, Y. (2005) Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution,* 20, 591–597.

DENNO, R. F. & COTHRAN, W. R. (1976) Competitive interactions and ecological strategies of Sarcophagid and Calliphorid flies inhabiting rabbit carrion. *Ann Entomol Soc Am.,* 69, 109–113.

DEVONSHIRE, A. & MOORES, G. (1982) A carboxylesterase with broad substrate specificity causes organophosphorus, carbamate and pyrethroid resistance in peach-potato aphids (*Myzus persicae*). *Pesti Biochem Phys,* 18, 235–246.

DEVONSHIRE, A. (1998) The evolution of insecticide resistance in the peach-potato aphid, *Myzus persicae. Philos Trans R Soc Lond B Biol Sci,* 353, 1677–1684.

DING, Y., ORTELLI, F., ROSSITER, L., HEMINGWAY, J. & RANSON, H. (2003) The *Anopheles gambiae* glutathione transferase supergene family: annotation, phylogeny and expression profiles. *BMC Genomics,* 4, 35.

DRUMMOND, A. J., ASHTON, B., CHEUNG, M., HELED, J., KEARSE, M., MOIR, R., STONES-HAVAS, S., THIERER, T. & WILSON, A. (2009) Geneious v4.8. *available from http://www.geneious.com.*

DUSSAUBAT, C., BRUNET, J.-L., HIGES, M., COLBOURNE, J. K., LOPEZ, J., CHOI, J.-H., MARTı´N-HERNA, R., BOTıAS, C., COUSIN, M., MCDONNELL, C., BONNET, M., BELZUNCES, L. P., MORITZ, R. F. A., CONTE, Y. L. & ALAUX, C. (2012) Gut Pathology and Responses to the Microsporidium *Nosema ceranae* in the Honey Bee *Apis mellifera. PloS One,* 7, e37017.

EDGAR, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res,* 32, 1792-1797.

EKBLOM, R. & GALINDO, J. (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity,* 107, 1-15.

EMMERSEN, J., RUDD, S., MEWES, H.-W. & TETKO, I. V. (2007) Separation of sequences from host-pathogen interface using triplet nucleotide frequencies. *Fungal Genet Biol: FG & B,* 44, 231-41.

ENAYATI, A., RANSON, H. & HEMINGWAY, J. (2005) Insect glutathione transferases and insecticide resistance. *Insect Molecular Biology,* 14, 3-8.

FARIA, M. & WRAIGHT, S. P. (2001) Biological control of *Bemisia tabaci* with fungi. *Crop Prot,* 20, 767-778.

FEYEREISEN, R. (1999) INSECT P450 ENZYMES. *Annu Rev Entomol,* 44, 507-533.

FEYEREISEN, R. (2006) Evolution of insect P450s. *Biochem Soc Trans,* 34, 1252-1255.

FISCHER, O. A., MATLOVA, L., DVORSKA, L., SVASTOVA, P., BARTL, J. & AL, R. T. W. E. (2004) Blowflies *Calliphora vicina* and *Lucilia sericata* as passive vectors of *Mycobacterium avium* subsp. avium, *M. a. paratuberculosis* and *M. a. hominissuis. Med Vet Entomol,* 18, 116–122.

FLORIANO, W. B., VAIDEHI, N., GODDARD, W. A., SINGER, M. S. & SHEPHERD, G. M. (2000) Molecular mechanisms underlying differential odor responses of a mouse olfactory receptor. *Proc Natl Acad Sci USA,* 97, 10712-10716.

FREITAK, D., HECKEL, D. G. & VOGEL, H. (2009) Bacterial feeding induces changes in immune-related gene expression and has trans-generational impacts in the cabbage looper (*Trichoplusia ni*). *Front Zool,* 6.

FRIEDEL, C. C., JAHN, K. H. V., SOMMER, S., RUDD, S., MEWES, H. W. & TETKO, I. V. (2005) Support vector machines for separation of mixed plant-pathogen EST collections based on codon usage. *Bioinformatics,* 21, 1383-1388.

GALLAGHER, M. B., SANDHU, S. & KIMSEY, R. (2010) Variation in developmental time for geographically distinct populations of the Common Green Bottle Fly, *Lucilia sericata* (Meigen). *J Forensic Sci,* 55, 438–442.

GASTEIGER, E., GATTIKER, A., HOOGLAND, C., IVANYI, I., APPEL, R. D., *et al.* (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucl Acids Res* 31, 3784–3788.

GILAD, Y., PRITCHARD, J. K. & THORNTON, K. (2009) Characterizing natural variation using next-generation sequencing technologies. *Trends Genet: TIG,* 25, 463-71.

GORMAN, K., DEVINE, G., BENNISON, J., COUSSONS, P., PUNCHARD, N. & DENHOLM, I. (2007) Report of resistance to the neonicotinoid insecticide imidacloprid in *Trialeurodes vaporariorum* (Hemiptera: Aleyrodidae). *Pest Manag,* 63, 555-558.

GRAFTON-CARDWELL, E., OUYANG, Y., STRIGGOW, R., CHRISTIANSEN, J. & BLACK, C. (2004) Role of esterase enzymes in monitoring for resistance of California red scale, *Aonidiella aurantii* (Homoptera: Diaspididae), to organophosphate and carbamate insecticides. *J Econ Entomol,* 97, 606-613.

GRANT, D., DIETZE, E. & HAMMOCK, B. (1991) Glutathione S-transferase isozymes in *Aedes aegypti*: purification, characterization, and isozyme-specific regulation. *Insect Biochem,* 21, 421-433.

HARTLEY, C. J., NEWCOMB, R. D., RUSSELL, R. J., YONG, C. G., STEVENS, J. R., YEATES, D. K., LA SALLE, J. & OAKESHOTT, J. G. (2006) Amplification of DNA from preserved specimens shows blowflies were preadapted for the rapid evolution of insecticide resistance. *Proc Natl Acad Sci U S A,* 103, 8757-62.

HEINZ, E., WILLIAMS, T. A., NAKJANG, S., NOEL, C. J., SWAN, D. C., GOLDBERG, A. V., HARRIS, S. R., WEINMAIER, T., MARKERT, S., BECHER, D., BERNHARDT, J., DAGAN, T., HACKER, C., LUCOCQ, J. M., SCHWEDER, T., RATTEI, T., HALL, N., HIRT, R. P. & EMBLEY, T. M.

(2012) The Genome of the Obligate Intracellular Parasite *Trachipleistophora hominis*: New Insights into Microsporidian Genome Dynamics and Reductive Evolution. *PLoS Pathogens,* 8, e1002979.

HEMINGWAY, J. & KARUNARATNE, S. (1998) Mosquito carboxylesterases: a review of the molecular biology and biochemistry of a major insecticide resistance mechanism. *Med Vet Entomol,* 12, 1-12.

HEMINGWAY, J., HAWKES, N., MCCARROLL, L. & RANSON, H. (2004) The molecular basis of insecticide resistance in mosquitoes. *Insect Biochem Mol Biol,* 34, 653-665.

HENNEBERRY, T. J., JECH, L. F., HENDRIX, D. L. & STEELE, T. (2000) *Bemisia argentifolii* (Homoptera: Aleyrodidae) honeydew and honeydew sugar relationships to sticky cotton. *Southwest Entomol,* 25, 1-14.

HEYMANN, E. (1980). Carboxylesterases and amidases. In: JAKOBY, W. B. (Ed.) *Enzymatic Basis of Detoxification.* Academic Press, New York, Vol II, 291–316.

HOLT, R. A., SUBRAMANIAN, G. M., HALPERN, A., SUTTON, G. G., CHARLAB, R. & NUSSKERN, D. R. E. A. (2002) The genome sequence of the Malaria Mosquito *Anopheles gambiae. Science,* 298, 129-149.

HRABER, P. T. & WELLER, J. W. (2001) On the species of origin: diagnosing the source of symbiotic transcripts. *Genome Biology,* RESEARCH0037.

HSIANG, T. & GOODWIN, P. H. (2003) Distinguishing plant and fungal sequences in ESTs from infected plant tissues. *J Microbiol Methods*, 54339– 54351.

HUANG, T. Y., COOK, C. E., DAVIS, G. K., SHIGENOBU, S., CHEN, R. P. Y. & CHANG, C. C. (2010) Anterior development in the parthenogenetic and viviparous form of the pea aphid *Acyrthosiphon pisum*: hunchback and orthodenticle expression. *Insect Mol Biol,* 19, 75-85.

HUERTA-CEPAS, J., MARCET-HOUBEN, M., PIGNATELLI, M., MOYA, A. & GABALDÓN, T. (2010) The pea aphid phylome: a complete cataogue of evolutionary histories and arthropod orthology and paralogy relationships for *Acyrthosiphon pisum* genes. *Insect Mol Biol,* 19, 13-21.

HUITEMA, E., TORTO, T. A., STYER, A. & KAMOUN, S. (2003) Combined ESTs from plant-microbe interactions: using GC counting to determine the species of origin. *Methods Mol Biol,* 236, 79–84.

HULL, J. J., GEIB, S. M., FABRICK, J. A. & BRENT, C. S. (2013) Sequencing and *De Novo* Assembly of the Western Tarnished Plant Bug (*Lygus hesperus*) Transcriptome. *PLoS One,* 8, e55105.

ILLUMINA - http://www.illumina.com/systems/sequencing.ilmn.

JANET HEMINGWAY, NICOLA J HAWKES, LYNN MCCARROLL & RANSON, H. (2004) The molecular basis of insecticide resistance in mosquitoes. *Insect Biochem Mol Biol,* 34, 653-665.

JENSEN, N. B., ZAGROBELNY, M., HJERNØ, K. & OLSEN, C.-E. (2011) Convergent evolution in biosynthesis of cyanogenic defence compounds in plants and insects. *Nat Commun,* 2, 273.

JONES, D. R. (2003) Plant Viruses Transmitted by Whiteflies. *Eur J Plant Pathol,* 109, 195-219.

JONES, R. T., BAKKER, S. E., STONE, D., SHUTTLEWORTH, S. N., BOUNDY, S., MCCART, C., DABORN, P., FFRENCH-CONSTANT, R. H. & VAN DEN ELSEN, J. M. (2010) Homology modelling of *Drosophila* cytochrome P450 enzymes associated with insecticide resistance. *Pest Manag Sci,* 66, 1106-1115.

JUNWEN, A., YONG, Z., JUN, D., QUANYOU, Y., GAOJUN, Z., FEI, W. & ZHONG-HUAI, X. (2011) Genome-wide analysis of cytochrome P450 monooxygenase genes in the silkworm, *Bombyx mori. Gene,* 480, 42-50.

KARATOLOS, N., DENHOLM, I., WILLIAMSON, M., NAUEN, R. & GORMAN, K. (2010) Incidence and characterisation of resistance to neonicotinoid insecticides and pymetrozine in the greenhouse whitefly, *Trialeurodes vaporariorum* Westwood (Hemiptera: Aleyrodidae). *Pest Manag Sci,* 66, 1304-1307.

KARATOLOS, N., PAUCHET, Y., WILKINSON, P., CHAUHAN, R., DENHOLM, I., GORMAN, K., NELSON, D. R., BASS, C., FFRENCH-CONSTANT, R. H. & WILLIAMSON, M. S. (2011) Pyrosequencing the transcriptome of the greenhouse whitefly, *Trialeurodes vaporariorum* reveals multiple transcripts encoding insecticide targets and detoxifying enzymes. *BMC Genomics,* 12, 56.

KARUNKER, I., BENTING, J., LUEKE, B., PONGE, T., NAUEN, R., RODITAKIS, E., VONTAS, J., GORMAN, K., DENHOLM, I. & 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 Mol Biol,* 38, 634-644.

KASAI, S., WEERASHINGHE, I. S., SHONO, T. & YAMAKAWA, M. (2000) Molecular cloning, nucleotide sequence and gene expression of a cytochrome P450 (CYP6F1) from the pyrethroid-resistant mosquito, *Culex quinquefasciatus* Say. *Insect Biochem Mol Biol,* 30, 163-171.

KATINKA, M., DUPRAT, S., CORNILLOT, E., METENIER, G., THOMARAT, F., PRENSIER, G., BARBE, V., PEYRETAILLADE, E., BROTTIER, P., WINCKER, P. & AL, E. (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi. Nature,* 414, 450-453.

KOSTAROPOULOS, I., PAPADOPOULOS, A., METAXAKIS, A., BOUKOUVALA, E. & PAPADOPOULOU-MOURKIDOU, E. (2001) Glutathione S-transferase in the defence against pyrethroids in insects. *Insect Biochem Mol Biol,* 31, 313-319.

KUMAR, S. & BLAXTER, M. L. (2010) Comparing *de novo* assemblers for 454 transcriptome data. *BMC Genomics,* 11, 571.

KUMAR, S. & BLAXTER, M. L. (2011) Simultaneous genome sequencing of symbionts and their hosts. *Symbiosis,* 55, 119-126.

LANDER, J., PARSONS, J., RIFE, C., GILLILAND, G. & ARMSTRONG, R. (2004) Parallel evolutionary pathways for glutathione transferases: structure and mechanism of the mitochondrial class Kappa enzyme rGSTK1–1. *Biochemistry,* 43, 352–261.

LANGMEAD, B. & SALZBERG, S. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods,* 9, 357-359.

LEE, S. F., CHEN, Z., MCGRATH, A., GOOD, R. T. & BATTERHAM, P. (2011) Identification, analysis, and linkage mapping of expressed sequence tags from the Australian sheep blowfly. *BMC Genomics,* 12, 406.

LESHKOWITZ, D., GAZIT, S., REUVENI, E., GHANIM, M., CZOSNEK, H., MCKENZIE, C., SHATTERS, R. L. & BROWN, J. K. (2006) Whitefly (*Bemisia tabaci*) genome project: analysis of sequenced clones from egg, instar, and adult (viruliferous and non-viruliferous) cDNA libraries. *BMC Genomics,* 7.

LEWIS, S. E., SEARLE, S. M. J., HARRIS, N., GIBSON, M., IYER, V., RICTER, J., WIEL, C., BAYRAKTAROGLU, L., BIRNEY, E., CROSBY, M. A., KAMINKER, J. S., MATTHEWS, B., PROCHNIK, S. E., SMITH, C. D., TUPY, J. L., RUBIN, G. M., MISRA, S., MUNGALL, C. J. & CLAMP, M. E. (2002) Apollo: a sequence annoatation editor. *Genome Biology,* 3, research0082.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & SUBGROUP, G. P. D. P. (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics,* 25, 2078-2079.

LIU, L., LI, Y., LI, S., HU, N., HE, Y., PONG, R., LIN, D., LU, L. & LAW, M. (2012) Comparison of Next-Generation Sequencing Systems. *J Biomed Biotechnol,* 2012.

LIU, S., CHOUGULE, N. P., VIJAYENDRAN, D. & BONNING, B. C. (2012) Deep Sequencing of the Transcriptomes of Soybean Aphid and Associated Endosymbionts. *PLoS One,* 7, e45161.

LI, X., SCHULER, M. A. and BERENBAUM, M. R. (2007) Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. *Annu Rev Entomol*, 52, 231-253.

LOW, W. Y., NG, H. L., MORTON, C. J., PARKER, M. W., BATTERHAM, P. & ROBIN, C. (2007) Molecular Evolution of Glutathione S-Transferases in the Genus *Drosophila. Genetics,* 177, 1363-1375.

MAIBECHE-COISNE, M., NIKONOV, A. N., JACQUIN-JOLY, E. & LEAL, W. S. (2004) Pheromone anosmia in a scarab beetle induced by in vivo inhibition of a pheromone-degrading enzyme. *Proc Natl Acad Sci*, 101, 11459-11464.

MALNIC, B., HIRONO, J., SATO, T. & BUCK, L. B. (1999) Combinatorial receptor codes for odors. *Cell,* 96, 713-723.

MAO, W., RUPASINGHE, S., ZANGERL, A., BERENBAUM, M. & SCHULER, M. (2007) Allelic variation in the *Depressaria pastinacella* CYP6AB3 protein enhances metabolism of plant allelochemicals by altering a proximal surface residue and potential interactions with cytochrome P450 reductase. *J Biol Chem,* 282, 10544-52.

MAOR, R., KOSMAN, E., GOLOBINSKI, R., GOODWIN, P. & SHARON, A. (2003) PF-IND: probability algorithm and software for separation of plant and fungal sequences. *Current Genetics,* 43, 296-302.

MARDIS, E. R. (2008) Next-Generation DNA Sequencing Methods. *Annu Rev Genomics Hum Genet,* 9, 387-402.

MARTIN, J. & WANG, Z. (2011) Next-generation transcriptome assembly. *Nat Rev Genet,* 7, 671-682.

MARTIN, J. H., MIFSUD, D. & RAPISARDA, C. (2000) The whiteflies (Hemiptera: Aleyrodidae) of Europe and the Mediterranean Basin. *Bulletin Entomol Res,* 90, 407-448.

MELLENTHIN, K., FAHMY, K., ALI, R. A., HUNDING, A., ROCHA, S. D. & BAUMGARTNER, S. (2006) Wingless signaling in a large insect, the blowfly *Lucilia sericata*: a beautiful example of evolutionary developmental biology. *Dev Dyn,* 235, 347–360.

METZKER, M. L. (2010) Sequencing technologies [mdash] the next generation. *Nat Rev Genet,* 11, 31-46.

MISRA, J. R., HORNER, M. A., LAM, G. & THUMMEL, C. S. (2011) Transcriptional regulation of xenobiotic detoxification in *Drosophila*. *Genes Dev,* 25, 1796-1806.

MITRA, S., RUPEK, P., RICHTER, D. C., URICH, T., GILBERT, J. A., MEYER, F., WILKE, A. & HUSON, D. H. (2011) Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics,* 12, S21.

MONTELLA, I., SCHAMA, R. & VALLE, D. (2012) The classification of esterases: an important gene family involved in insecticide resistance--a review. *Mem Inst Oswaldo Cruz,* 107, 437-449.

MOREL, F., RAUCH, C., PETIT, E., PITON, A., THERET, N., COLES, B. & GUILLOUZO, A. (2004) Gene and protein characterization of the human glutathione S-transferase kappa and evidence for a peroxisomal localization. *J Biol Chem,* 279, 16246–16253.

MOROZOVA, O. & MARRA, M. A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics,* 92, 255-264.

MOROZOVA, O., HIRST, M. & MARRA, M. A. (2009) Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet,* 10, 135-51.

MUNDRY, M., BORNBERG-BAUER, E., SAMMETH, M. & FEULNER, P. (2012) Evaluating Characteristics of *De Novo* Assembly Software on 454 Transcriptome Data: A Simulation Approach. *PLoS One,* 7, e31410.

NAGARAJ, S. H., GASSER, R. B. & RANGANATHAN, S. (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinform,* 8, 6-21.

NAISBIT, R. E. (2001) Ecological divergence and speciation in *Heliconius cydno* and *H. melpomene*. *PHD Thesis, University College London.*

NEBERT, D. & GONZALEZ, F. (1987) P450 genes: Structure, evolution and regulation. *Annu Rev Biochem,* 56, 945-993.

NENE, V., WORTMAN, J. R., LAWSON, D., HAAS, B., KODIRA, C., TU, Z. J., LOFTUS, B., XI, Z., MEGY, K., GRABHERR, M., REN, Q., ZDOBNOV, E. M., LOBO, N. F., CAMPBELL, K. S., BROWN, S. E., BONALDO, M. F., ZHU, J., SINKINS, S. P., HOGENKAMP, D. G., AMEDEO, P., ARENSBURGER, P., ATKINSON, P. W., BIDWELL, S., BIEDLER, J., BIRNEY, E., BRUGGNER, R. V., COSTAS, J., COY, M. R., *et al.* (2007)

Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science,* 316, 1718-1723.

NEWCOMB, R. D., GLEESON, D. M., YONG, C. G., RUSSELL, R. J. & OAKESHOTT, J. G. (2005) Multiple mutations and gene duplications conferring organophosphorus insecticide resistance have been selected at the Rop-1 locus of the sheep blowfly, *Lucilia cuprina*. *J Mol Evol,* 60, 207-220.

NIKOU, D., RANSON, H. & HEMINGWAY, J. (2003) An adult-specific CYP6 P450 gene is overexpressed in a pyrethroid-resistant strain of the malaria vector, *Anopheles gambiae*. *Gene,* 318, 91-102.

NYBERG, K. G., CONTE, M. A., KOSTYUN, J. L., FORDE, A. & BELY, A. E. (2012) Transcriptome characterization via 454 pyrosequencing of the annelid *Pristina leidyi* an emerging model for studying the evolution. *BMC Genomics*, 13, 287.

OAKESHOTT, J. G., HOME, I., SUTHERLAND, T. D. & RUSSELL, R. J. (2003) The genomics of insecticide resistance. *Genome Biology,* 4, 202.

OAKESHOTT, J. G., JOHNSON, R. M., BERENBAUM, M. R., RANSON, H., CRISTINO, A. S. & CLAUDIANOS, C. (2010) Metabolic enzymes associated with xenobiotic and chemosensory responses in *Nasonia vitripennis*. *Insect Mol Biol,* 19, 147-163.

OAKESHOTT, J., CLAUDIANOS, C., CAMPBELL, P., NEWCOMB, R. & RJ, R. R. (Eds.) (2005) *Biochemical genetics and genomics of insect esterases,* Amsterdam, The Netherlands, Elsevier.

OHNO, S. (1970) Evolution by Gene Duplication. *Springer-Verlag,* New York.

OLIVEIRA, M. R. V., HENNEBERRY, T. J. & ANDERSON, P. (2001) History, current status, and collaborative research projects for *Bemisia tabaci*. *Crop Prot,* 20, 709-723.

ORTELLI, F., ROSSITER, L., VONTAS, J., RANSON, H. & HEMINGWAY, J. (2003) Heterologous expression of four glutathione transferase genes genetically linked to a major insecticide resistance locus, from the malaria vector *Anopheles gambiae*. *Biochem J,* 373, 957–963.

PAPANICOLAOU, A., STIERLI, R., FFRENCH-CONSTANT, R. H. & HECKEL, D. G. (2009) Next generation transcriptomes for next generation genomes using est2assembly. *BMC Bioinformatics,* 10, 447.

PAREEK, C. S., SMOCZYNSKI, R. & TRETYN, A. (2011) Sequencing technologies and genome sequencing. *J Appl Genet,* 52, 413-435.

PASZKIEWICZ, K. & STUDHOLME, D. J. (2010) *De novo* assembly of short sequence reads. *Brief Bioinform,* 44.

PAUCHET, Y., WILKINSON, P., CHAUHAN, R. & H, R. (2010) Diversity of Beetle Genes Encoding Novel Plant Cell Wall Degrading Enzymes. *PloS One,* 5, e15635.

PAUCHET, Y., WILKINSON, P., VAN MUNSTER, M., AUGUSTIN, S., PAURON, D. & FFRENCH-CONSTANT, R. H. (2009) Pyrosequencing of the midgut transcriptome of the poplar leaf beetle *Chrysomela tremulae* reveals new gene families in Coleoptera. *Insect Biochem Mol Biol,* 39, 403-413.

PAUCHET, Y., WILKINSON, P., VOGEL, H., NELSON, D. R., REYNOLDS, S. E., HECKEL, D. G. & FFRENCH-CONSTANT, R. H. (2009) Pyrosequencing the *Manduca sexta* larval midgut transcriptome: messages for digestion, detoxification and defence. *Insect Mol Biol,* 19, 61-75.

PAUL, A. G., AHMAD, N. W., LEE, H. L., ARIFF, A. M., SARANUM, M. & AL., A. S. N. E. (2009) Maggot debridement therapy with *Lucilia cuprina*: a comparison with conventional debridement in diabetic foot ulcers. *Int Wound J,* 6, 39–46.

PEIRIS, H. & HEMINGWAY, J. (1993) Characterisation and inheritance of elevated esterases in organophosphorus and carbamate insecticide resistant *Culex quinquefasciatus* (Diptera: Culicidae) from Sri Lanka. *Bull Entomol Res,* 83, 127–132.

PEYRETAILLADE, E., H, E. A., DIOGON, M., POLONAIS, P., PARISOT, N., BIRON, D., PEYRE, P. & DELBAC, F. (2011) Extreme reduction and compaction of microsporidian genomes. *Res Microbiol,* 162, 598-606.

PRAPANTHADARA, L., HEMINGWAY, J. & KETTERMAN, A. (1993) Partial purification and characterization of glutathione S-transferases involved in DDT resistance from the mosquito *Anopheles gambiae*. *Pest Biochem Physiol,* 47, 119–133.

PUINEAN, A., FOSTER, S., OLIPHANT, L., DENHOLM, I., FIELD, L., MILLAR, N., WILLIAMSON, M. & BASS, C. (2010) Amplification of a cytochrome P450 gene is associated with resistance to neonicotinoid insecticides in the aphid *Myzus persicae*. *PLoS Genetics,* 6.

RAMSEY, J. S., RIDER, D. S., WALSH, T. K., DE VOS, M., GORDON, K. H. J., PONNALA, L., MACMIL, S. L., ROE, B. A. & JANDER, G. (2010) Comparative analysis of detoxification enzymes in *Acyrthosiphon pisum* and *Myzus persicae*. *Insect Mol Biol,* 19, 155-164.

RAMSEY, J., WILSON, A., DE VOS, M., SUN, Q., TAMBORINDEGUY, C., WINFIELD, A., MALLOCH, G., SMITH, D., FENTON, B., GRAY, S. & JANDER, G. (2007) Genomic resources for *Myzus persicae*: EST sequencing, SNP identification, and microarray design. *BMC Genomics,* 8, 423.

RANSON, H. & HEMINGWAY, J. (Eds.) (2004) *Insect pharmacology and control: glutathione S-transferases,* Oxford, UK, Elsevier Ltd.

RANSON, H., CLAUDIANOS, C., ORTELLI, F., ABGRALL, C., HEMINGWAY, J. & SHARAKHOVA, M. (2002) Evolution of supergene families associated with insecticide resistance. *Science,* 298, 179-181.

RAYMOND, K., BERGERET, E., DAGHER, M., BRETON, R., GRIFFIN-SHEA, R. & FAUVARQUE, M. (2001) The Rac GTPase-activating protein RotundRacGAP interferes with Drac1 and Dcdc42 signalling in *Drosophila melanogaster. J Biol Chem,* 276, 35909-35916.

REDDY, D. M., ASPATWAR, A., DHOLAKIA, B. & GUPTA, V. (2008) Evolutionary analysis of WD40 super family proteins involved in spindle checkpoint and RNA export: Molecular evolution of spindle checkpoint. *Bioinformation,* 2, 461-468.

REWITZ, K. F., O'CONNOR, M. B. & GILBERT, L. I. (2007) Molecular evolution of the insect Halloween family of cytochrome P450s: Phylogeny, gene organization and functional conservation. *Insect Biochem Mol Biol,* 37, 741-753.

REWITZ, K. F., RYBCZYNSKI, R., WARREN, J. T. & GILBERT, L. I. (2006) The Halloween genes code for cytochrome P450 enzymes mediating synthesis of the insect moulting hormone. *Biochem Soc Trans,* 34, 1256-60.

RICE, P., LONGDEN, I. & BLEASBY, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet,* 16, 276-277.

ROBERTSON, H. M., WARR, C. G. & CARLSON, J. R. (2003) Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster. Proc Natl Acad Sci,* 100, 14537-14542.

ROCHE - http://454.com/products/gs-flx-system/index.asp.

ROTHBERG, J. M. & LEAMON, J. H. (2008) The development and impact of 454 sequencing. *Nature Biotechnol,* 26, 1117 - 1124.

RUDD, S. & TETKO, I. V. (2005) Eclair—a web service for unravelling species origin of sequences sampled from mixed host interfaces. *Nucl Acids Res,* 33, W724–W727.

RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M. & BARRELL, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics,* 16, 944-945.

SABOURAULT, C., GUZOV, V., KOENER, J., CLAUDIANOS, C., JR, F. P. & FEYEREISEN, R. (2001) Overproduction of a P450 that metabolizes diazinon is linked to a loss-of-function in the chromosome 2 ali-esterase (*MdalphaE7*) gene in resistant house flies. *Insect Mol Biol,* 10, 609-618.

SALINAS, A. & WONG, M. (1999) Glutathione S-transferases – A review. *Current Med Chem,* 6, 279-309.

SALZBERG, S. L., HOTOPP, J. C. D., DELCHER, A. L., POP, M., SMITH, D. R., EISEN, M. B. & NELSON, W. C. (2005) Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species. *Genome Biology,* 6, 402.

SANCHEZ-GRACIA, A., VIEIRA, F. G. & ROZAS, J. (2009) Molecular evolution of the major chemosensory gene families in insects. *Heredity,* 103, 208-216.

SATOH, T. & HOSOKAWA, M. (2006) Structure, function and regulation of carboxylesterases. *Chem Biol Interact,* 162, 195-211.

SCHADT, E. E., TURNER, S. & KASARSKIS, A. (2010) A window into third-generation sequencing. *Human Mol Genet,* 19, R227-R240.

SCOTT, J. (1999) Cytochromes P450 and insecticide resistance. *Insect Biochem Mol Biol,* 29, 757-777.

SCOTT, J. G. (2008) Insect cytochrome P450s: Thinking beyond detoxification. In: LIU, N. (Ed.) *Recent Advances in Insect Physiological, Toxicology and Molecular Biology.* Trivandrum, Research Signpost.

SHIOTSUKI, T. & KATO, Y. (1999) Induction of Carboxylesterase Isozymes in *Bombyx mori* by Infection of *E coli. Insect Biochem Mol Biol,* 29, 731-736.

SIMS, A. H., GENT, M. E., ROBSON, G. D., DUNN-COLEMAN, N. S. & OLIVER, S. G. (2004) Combining transcriptome data with genomic and cDNA sequence alignments to make confident functional assignments for *Aspergillus nidulans* genes. *Mycol Res,* 108, 853-857.

SMILEY, J. (1978) Plant chemistry and the evolution of host specificity: new evidence from *Heliconius* and Passiflora. *Science,* 201, 745-747.

SMITH, T., GAITATZES, C., SAXENA, K. & EJ, E. N. (1999) The WD repeat: a common architecture for diverse functions. *Trends Biochem Sci,* 24, 181-185.

SNYDER, M. & GLENDINNING, J. (1996) Causal connection between detoxification enzyme activity and consumption of a toxic plant compound. *J Comp Physiol A,* 179, 255-261.

STEVENS, J. & WALL, R. (1996) Species, subspecies and hybrid populations of the blowflies *Lucilia cuprina* and *Lucilia sericata* (Diptera: Calliphoridae). *Proc Royal Soc B,* 263, 1335-1341.

STEVENS, J. L., SNYDER, M. J., KOENER, J. F. & FEYEREISEN, R. (2000) Inducible P450s of the CYP9 family from larval *Manduca sexta* midgut. *Insect Biochem Mol Biol,* 30, 559-568.

STRODE, C., WONDJI, C. S., DAVID, J. P., HAWKES, N. J., LUMJUAN, N. & NELSON, D. R. (2008) Genomic analysis of detoxification genes in the mosquito *Aedes aegypti*. *Insect Mol Biol,* 38, 113-123.

SZE, S.-H., DUNHAM, J. P., CAREY, B., CHANG, P. L., LI, F., EDMAN, R. M., FJELDSTED, C., SCOTT, M. J., NUZHDIN, S. V. & TARONE, A. M. (2012) A de novo transcriptome assembly of *Lucilia sericata* (Diptera: Calliphoridae) with predicted alternative splices, single nucleotide polymorphisms and transcript expression estimates. *Insect Mol Biol,* 21, 205-221.

TAMURA, K., DUDLEY, J., NEI, M. & KUMAR, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol,* 24, 1596-1599.

TAMURA, K., PETERSON, D., PETERSON, N., STECHER, G., NEI, M. & KUMAR, S. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol,* 28, 2731-2739.

TARONE, A. M., JENNINGS, K. C. & FORAN, D. R. (2007) Aging blow fly eggs using gene expression: a feasibility study. *J Forensic Sci,* 52, 1350–1354.

TASKIN, V., KENCE, M. & GOCMEN, B. (2004) Determination of malathion and diazinon resistance by sequencing the *Md alpha E7* gene from Guatemala, Colombia, Manhattan, and Thailand housefly (*Musca domestica* L.) strains. *Genetika,* 40, 478-481.

THUDI, M., LI, Y., JACKSON, S., MAY, G. & VARSHNEY, R. (2012) Current state-of-art of sequencing technologies for plant genomics research. *Brief Funct Genomics,* 11, 3-11.

TIJET, N., HELVIG, C. & FEYEREISEN, R. (2000) The cytochrome P450 gene superfamily in *Drosophila melanogaster*: Annotation, intron-exon organization and phylogeny. *Gene,* 262, 189-198.

TOURLE, R., DOWNIE, D. & VILLET, M. (2009) Flies in the ointment: a morphological and molecular comparison of *Lucilia cuprina* and *Lucilia sericata* (Diptera: Calliphoridae) in South Africa. *Med Vet Entomol,* 23, 6-14.

TSAKAS, S. & MARMARAS, V. (2010) Insect immunity and its signalling: an overview. *ISJ,* 7, 228-238.

TSUBOTA, T. & SHIOTSUKI, T. (2010) Genomic analysis of carboxyl/cholinesterase genes in the silkworm *Bombyx mori*. *BMC Genomics,* 11, 377.

WALL, P., LEEBENS-MACK, J., CHANDERBALI, A., BARAKAT, A., WOLCOTT, E., LIANG, H., LANDHERR, L., TOMSHO, L., HU, Y., CARLSON, J., MA, H., SCHUSTER, S., SOLTIS, D., SOLTIS, P., ALTMAN, N. & DEPAMPHILIS, C. (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics,* 10, 347.

WANG, J., LINDSAY, B., LEEBENS-MACK, J., CUI, L., WALL, K., MILLER, W. & CW, C. D. (2004) EST clustering error evaluation and correction. *Bioinformatics,* 20, 2973-2984.

WANG, X.-W., LUAN, J.-B., LI, J.-M. & BAO, Y.-Y. (2010) *De novo* characterization and comparison of the whitefly transcriptomes reveals genes associated with development and insecticide resistance. *BMC Genomics,* 11, 400.

WARDLOW, L., LUDLAM, A. & BRADLEY, L. (1976) Pesticide resistance in the glasshouse whitefly (*Trialeurodes vaporariorum* (Westwood)). *Pestic Sci,* 7, 320–324.

WATERHOUSE, R. M., WYDER, S. & ZDOBNOV, E. M. (2008) The *Aedes aegypt*i genome: a comparative perspective. *Insect Mol Biol,* 17, 1-8.

WEN, Z., RUPASINGHE, S., NIU, G., BERENBAUM, M. R. & SCHULER, M. A. (2006) CYP6B1 and CYP6B3 of the black swallowtail (*Papilio polyxenes*): adaptive evolution through subfunctionalization. *Mol Biol Evol,* 23, 2434-2443.

WESTERMANN, A. J., GORSKI, S. A. and VOGEL, J. (2012) Dual RNA-seq of pathogen and host. *Nat Rev Microbiol,* 10, 618-630.

WILCE, M. C. & PARKER, M. W. (1994) Structure and function of glutathione S-transferases. *Biochem Biophys Acta,* 1205, 1–18.

WINTERMANTEL, W. M. (2004) Emergence of Greenhouse Whitefly (*Trialeurodes vaporariorum*) Transmitted Criniviruses as Threats to Vegetable and Fruit Production in North America. *APSnet Features.*

WISLER, G. & DUFFUS, J. (Eds.) (2001) *Transmission properties of whitefly-borne criniviruses and their impact on virus epidemiology,* San Diego, CA, Academic Press.

XU, C., LI, C. & KONG, A. (2005) Induction of phase I, II and III drug metabolism/transport by xenobiotics *Arch Pharm Res,* 28, 249–268.

XU, Q., LU, A., XIAO, G., YANG, B. & AL, J. Z. E. (2012) Transcriptional Profiling of Midgut Immunity Response and Degeneration in the Wandering Silkworm, *Bombyx mori. PloS One,* 7, e43769.

YANG, Y., CHEN, S., WU, S., YUE, L. & WU, Y. (2006) Constitutive overexpression of multiple cytochrome P450 genes associated with pyrethroid resistance in *Helicoverpa armigera. J Econ Entomol,* 99, 1784-1789.

YU, Q.-Y., LU, C., LI, W.-L., XIANG, Z.-H. & ZHANG, Z. (2009) Annotation and expression of carboxylesterases in the silkworm, *Bombyx mori. BMC Genomics,* 10, 553.

ZAGROBELNY, M. (2004) Cyanogenic glucosides and plant–insect interactions. *Phytochemistry,* 65, 293-306.

ZDOBNOV, E. M. & APWEILER, R. (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847-848.

ZHANG, Y. & JELTSCH, A. (2010) The Application of Next Generation Sequencing in DNA Methylation Analysis. *Genes,* 1, 85-101.

ZHU, Y., JOHNSON, T., MYERS, A. & KANOST, M. (2003) Identification by subtractive suppression hybridization of bacteria-induced genes expressed in *Manduca sexta* fat body. *Insect Biochem Mol Biol,* 33, 541-559.