

**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 French-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

<b>Title Page</b> .....	1
<b>Abstract</b> .....	2
<b>Acknowledgements</b> .....	4
<b>Contents</b> .....	5
<b>List of figures</b> .....	9
<b>List of tables</b> .....	11
<b>Abbreviations/Glossary</b> .....	13
<b>Chapter 1- Introduction: Next generation sequencing and xenobiotic metabolism</b> .....	15
1.1 Next generation sequencing (NGS) technologies.....	15
1.1.1 Pyrosequencing on the Roche 454.....	16
1.1.2 Illumina (Solexa).....	16
1.1.3 SOLiD (Sequence by Oligo Ligation Detection).....	17
1.1.4 Third generation NGS platforms.....	19
1.2 Applications of NGS technologies .....	19
1.3 Xenobiotic metabolising enzyme superfamilies .....	21
1.3.1 Glutathione-S-transferases (GSTs) .....	22
1.3.1.1 Classification and nomenclature .....	22
1.3.1.2 Role of insect GSTs .....	22
1.3.2 Cytochrome P450s (P450s) .....	23
1.3.2.1 Classification and nomenclature .....	24
1.3.2.2 Role of insect P450s .....	24
1.3.3 Carboxyl/Cholinesterases (CCEs) .....	25
1.3.3.1 Classification and nomenclature .....	25
1.3.3.2 Role of insect CCEs .....	26
1.4 Thesis overview .....	29
<b>Chapter 2 - Identifying putative xenobiotic metabolising enzyme superfamilies in the <i>Trialeurodes vaporariorum</i> transcriptome</b> .....	32
2.1 INTRODUCTION .....	32
2.2 METHODOLOGY.....	35
2.2.1 Libraries and sequencing .....	35
2.2.2 454 transcriptome assembly .....	35

2.2.3	Manual curation of sequences encoding xenobiotic metabolising enzyme superfamilies .....	35
2.2.4	Phylogenetic analysis .....	36
2.3	RESULTS .....	37
2.3.1	454 transcriptome assembly .....	37
2.3.2	Manual curation of xenobiotic metabolising enzyme superfamilies.....	38
2.3.2.1	Putative cytochrome P450 transcripts .....	38
2.3.2.2	Putative glutathione-S-transferase transcripts.....	41
2.3.2.3	Putative carboxyl/cholinesterase transcripts .....	43
2.4	DISCUSSION.....	47
2.5	CONCLUSION .....	49
<b>CHAPTER 3 - Annotation of the cytochrome P450 gene superfamily in <i>Heliconius melpomene</i></b> .....		
3.1	INTRODUCTION.....	52
3.2	METHODOLOGY .....	55
3.2.1	454 transcriptome assembly .....	55
3.2.2	Identification of CYPs in the genome .....	55
3.2.3	Identification of CYPs in the transcriptome .....	55
3.2.4	Validation of gene models .....	56
3.2.5	Phylogenetic analysis.....	57
3.3	RESULTS.....	58
3.3.1	454 transcriptome assembly .....	58
3.3.2	Distribution of <i>CYP</i> genes in <i>H. melpomene</i> .....	59
3.3.2.1	CYP2 and the mitochondrial clade.....	59
3.3.2.2	CYP3 and CYP4 clades .....	59
3.3.3	Putative candidates for cyanogenesis .....	62
3.3.4	OR P450 gene cluster .....	64
3.3.5	Gene clusters in <i>Heliconius melpomene</i> .....	65
3.3.6	Comparison of <i>Bombyx mori</i> and <i>Heliconius melpomene</i> ....	67
3.4	DISCUSSION .....	69
3.5	CONCLUSION .....	71
<b>CHAPTER 4 - Analysis of <i>Lucilia sericata</i> transcriptome: xenobiotic metabolising enzymes superfamilies and SNP detection</b> .....		
4.1	INTRODUCTION.....	74

4.2	METHODOLOGY .....	77
4.2.1	Datasets used .....	77
4.2.2	454 transcriptome assembly .....	77
4.2.3	BLAST2GO analysis .....	77
4.2.4	Manual curation of sequences encoding xenobiotic metabolizing enzyme superfamilies .....	78
4.2.5	Phylogenetic analysis .....	78
4.2.6	Variant calling .....	79
4.3	RESULTS .....	80
4.3.1	454 transcriptome assembly .....	80
4.3.2	Sequence analysis of the assembled transcriptome (BLAST2GO) .....	81
4.3.3	Manual curation of xenobiotic metabolising enzyme superfamilies .....	86
4.3.3.1	Putative glutathione S-transferases (GSTs) .....	86
4.3.3.2	Putative cytochrome P450s (P450s) .....	88
4.3.3.3	Putative carboxyl/cholinesterases (CCEs) .....	90
4.3.4	Variant calling .....	93
4.3.4.1	E3 mutation in the UK population .....	95
4.4	DISCUSSION .....	96
4.5	CONCLUSION .....	99
<b>CHAPTER 5 – Removing microbial messages from 454 generated beetle transcriptomes .....</b>		<b>101</b>
5.1	INTRODUCTION .....	101
5.2	METHODOLOGY .....	104
5.2.1	Beetle 454 derived transcriptomes .....	104
5.2.2	Preliminary assembly .....	104
5.2.3	Estimation of taxonomic composition of the preliminary assembly .....	104
5.2.4	Separation of contigs on the basis of most abundant taxonomic groups .....	105
5.2.5	Taxon specific re-assembly .....	105
5.2.6	BLAST2GO analysis .....	106
5.2.7	Codon usage and GC content .....	106
5.3	RESULTS .....	107

5.3.1 BLAST annotation and estimation of taxonomic composition .....	107
5.3.2 Classification of contigs and separation of reads .....	109
5.3.3 Taxon specific reassembly.....	110
5.3.4 Sequence analysis of the taxon specific reassemblies (BLAST2GO) .....	112
5.3.5 GC content and codon usage .....	117
5.4 DISCUSSION .....	119
5.5 CONCLUSION .....	122
<b>Chapter 6 - An overview of the impact of next generation sequencing on the annotation of gene families involved in xenobiotic metabolism</b> .....	125
6.1 NGS and Xenobiotic metabolising gene families in insects .....	125
6.2 Advantages of utilising NGS technologies in non-model species research .....	128
6.3 Bioinformatics challenges and limitations of NGS data in non-model organism research .....	128
6.4 Future Perspectives.....	131
<b>Appendices</b> .....	134
<b>Bibliography</b> .....	147



## LIST OF FIGURES

Figure 2.1	Characteristics of assembled <i>Trialeurodes vaporariorum</i> 454 contigs and BLASTx alignments against <i>Drosophila melanogaster</i> .....	38
Figure 2.2	Neighbour joining tree showing the phylogenetic analysis of P450s of <i>Trialeurodes vaporariorum</i> and other insects .....	40
Figure 2.3	Neighbour joining tree showing the phylogenetic analysis of GSTs of <i>Trialeurodes vaporariorum</i> and other insects .....	43
Figure 2.4	Neighbour joining tree showing the phylogenetic analysis of CCEs of <i>Trialeurodes vaporariorum</i> and other insects .....	46
Figure 3.1	Distribution of <i>Heliconius melpomene</i> cytochrome P450s into clades .....	61
Figure 3.2	Neighbour joining tree showing the phylogenetic analysis of P450s of <i>Heliconius melpomene</i> and other insects.....	63
Figure 3.3	Sub tree showing the putative cytochrome P450 candidates for cyanogenesis in <i>Heliconius melpomene</i> .....	64
Figure 3.4	Diagrammatic representation of the <i>Heliconius melpomene</i> scaffold containing clusters of olfactory receptor genes along with P450 genes.....	65
Figure 3.5	Largest cytochrome P450 gene cluster in <i>Heliconius melpomene</i> .....	67
Figure 3.6	Distribution of P450 subfamilies in <i>Bombyx mori</i> and <i>Heliconius melpomene</i> .....	68
Figure 4.1	Species distribution of the top blast hits for <i>Lucilia sericata</i> contigs.....	82
Figure 4.2	Gene ontology (GO) assignment for the <i>Lucilia sericata</i> contigs .....	83
Figure 4.3	Neighbour joining tree showing the phylogenetic analysis of GSTs of <i>Lucilia sericata</i> and other insects.....	88
Figure 4.4	Neighbour joining tree showing the phylogenetic analysis of P450s of <i>Lucilia sericata</i> and other insects .....	90

Figure 4.5	Neighbour joining tree showing the phylogenetic analysis of CCEs of <i>Lucilia sericata</i> and other insects .....	93
Figure 4.6	Distribution of SNP quality scores in the UK population of <i>Lucilia sericata</i> .....	94
Figure 5.1	Phylum level distribution of microbes/contaminants in the five beetle datasets.....	108
Figure 5.2	Distribution of the most abundant microbes/contaminants found in the <i>Gastrophysa viridula</i> 454 midgut transcriptome dataset .....	110
Figure 5.3	Species distribution of the top ten blast hits (a) Green dock beetle contigs, (b) microsporidial contigs .....	113
Figure 5.4	Gene ontology (GO) assignment for the <i>Gastrophysa viridula</i> contigs and microsporidial contigs .....	115

## LIST OF TABLES

Table 1.1	Comparison of the three leading NGS sequencers with Sanger sequencer .....	18
Table 1.2	Applications and characteristics of Roche 454, Illumina and Solid sequencing platforms .....	20
Table 1.3	Number of annotated GSTs, P450s and CCEs in the insect genomes .....	27
Table 2.1	Summary statistics for <i>Trialeurodes vaporariorum</i> transcriptome assembly .....	37
Table 2.2	Number of annotated P450s in the order Hemiptera and their distribution across different clades .....	39
Table 2.3	Number of annotated GSTs in the order Hemiptera and their distribution across different clades .....	42
Table 2.4	Number of annotated CCEs in the order Hemiptera and their distribution across different clades. ....	45
Table 3.1	Number of validated P450s in the insect genomes and their distribution across the P450 clades.....	59
Table 3.2	Cytochrome P450 gene clusters in <i>Heliconius melpomene</i> genome.....	66
Table 4.1	Sequence assembly statistics for the Newbler assembly of <i>Lucilia sericata</i> transcriptome .....	80
Table 4.2	Summary of the top twenty InterPro superfamilies and domains represented in the <i>Lucilia sericata</i> transcriptome ....	85
Table 4.3	Number of annotated insect GSTs and their distribution across different GST classes in the order Diptera .....	87
Table 4.4	Number of annotated insect cytochrome P450s and their distribution across different clades in the order Diptera .....	89
Table 4.5	Number of annotated insect CCEs and their distribution across different clades in the order Diptera .....	92

Table 5.1	Summary of Blast results and taxon analysis for the five beetle 454 transcriptome datasets.....	108
Table 5.2	Sequence assembly statistics for the newbler reassembly of separated Green dock beetle midgut and microsporidial sequences.....	111
Table 5.3	Summary of the top ten InterPro superfamilies and domains represented in the <i>Gastrophysa viridula</i> larval midgut transcriptome and the separated microsporidial sequences	116
TABLE 5.4	Overall %GC content for each of the five beetle transcriptomes and microsporidial sequences.....	118
TABLE 5.5	Codon usage frequencies (percent) for three amino acids in the five beetle transcriptomes and separated microsporidial sequences.....	118

## **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
<i>De novo</i> 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)