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

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)