

ANALYSIS OF GENETIC VARIATION IN *PEDICULUS HUMANUS* AND *POPULUS*
TRICHOCARPA

BY

BRETT PHILIP OLDS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Professor Ken N. Paige, Chair, Co-Director of Research
Professor Barry R. Pittendrigh, Co-Director of Research
Associate Professor Matt Hudson
Associate Professor Ripan S. Malhi

Abstract

With the advent of sequencing technologies that are both affordable and readily available, biologists are now able to address questions that were previously intractable. New species are having their genomes mapped at an increasing rate, including non-model organisms. Two such organisms are the human body louse, *Pediculus humanus corporis* and the black cottonwood, *Populus trichocarpa*. While unrelated, these two organisms each represent a study system with questions that challenge our current understanding of each organism.

Body and head lice, while closely related, are thought to be separate species with their most important difference being that only body lice vector disease to humans. The first question is: Are body lice (*Pediculus humanus coporis*) and head lice (*Pediculus humanus capitis*) the same species? A total of 10,771 body louse and 10,770 head louse transcripts were predicted from a combined assembly of Roche 454 and Illumina sequenced cDNAs from whole body tissues collected at all life stages and during pesticide exposure and bacterial infection treatments. Illumina reads mapped to the 10,775 draft body louse gene models from the whole genome assembly predicted nine presence/absence differences, but PCR confirmation resulted in a single gene difference. One novel microRNA was predicted in both lice species and 99% of the 544 transcripts from *Candidatus riesia* indicate that they share the same endosymbiont. Overall, few differences exist, which supports the hypothesis that these two organisms are ecotypes of the same species.

A second question is: Are there gene expression differences between these two organisms that cause the body louse to vector disease to humans while head lice do not? We utilized an RNAseq analysis on 7-day old head and body lice fed blood infected with *Bartonella quintana*,

the bacterium that causes trench fever, and control individuals to elicit gene expression differences. Eight immunoresponse genes came out significant, many associated with the Toll pathway; Fibrinogen-related protein (PHUM500950), Spaetzle (PHUM595260), Defensin 1 (PHUM365700) and 2 (PHUM595870), Serpin (PHUM311330), Cactus (PHUM345810), Scavenger receptor A (ScavA; PHUM066640) and Apolipoprotein D (PHUM427700). Increased expression of Fibrinogen-related protein and Spaetzle, both related to the Toll pathway, in treated body lice supports the hypothesis that body lice are fighting infection from *B. quintana*. But conflicting results in Defensin 1 and 2 based upon validation method suggest another mechanism in head lice alternative to the Toll pathway might be involved. Additionally, Scavenger receptor A was higher in both control and treated head lice, suggesting higher phagocytotic activity in head lice to curb infection.

In 1981, Whitham and Slobodchikoff hypothesized that long-lived plant species, like black cottonwood, that propagate both asexually and sexually, might develop, through accumulation of somatic mutations, as genetic mosaics, providing the potential to respond to or even outrun their insect herbivores. The initial question to assess this hypothesis is: What is the level of somatic variation that exists with an individual tree and its clonally derived offspring? We show that on the average there are 4,840 unique amino acid changes in 2,569 genes within each tissue sampled within an individual parental tree and its clonally derived offspring. In addition, genetic variation within a parental tree, and between a parental tree and its clonally derived offspring, show about half as much genetic variation among tissues as exists between genetically unrelated trees within a population. These results support the long-standing hypothesis that some tree species exist as mosaics of genetic variability and can evolve (i.e., genetic differentiation between parent and clonally derived offspring).

Acknowledgements

I would like to thank both of my advisors, Ken Paige and Barry Pittendrigh for their acceptance of me into their respective labs after my first year of graduate school. Each has provided me a great environment to learn and grow as a scientist. I consider both of you a mentor and friend and look forward to future endeavors with each of you. I would also like to thank each of my committee members, Matt Hudson and Ripan Malhi for valuable feedback during this process and for expanding my horizons into bioinformatics and population genetics.

I owe my successful journey to a lot of people throughout the years I have spent here in Illinois. First and foremost, I want to thank the graduate students, both past and present, of SIB and GEEB that have provided friendship, stimulating science conversation, training and stress relief through the years. This would not have been an enjoyable experience without you. I expect to see many of you at future conferences and have our paths cross as we become the next generation of scientists. Specifically, Cassandra Allsup, Nick Naeger, Rhiannon Peery, Jessica and Adam Brandt, Daniel Welsh and Sara Paver, thank you for your close friendship and support, I cherish each of you. I would be remiss if I did not mention Lisa Smith, Penny Broga and Susan Balfe. Each one of these terrific ladies has been instrumental in my success from financial needs to administrative and lab support while, at the same time, providing friendship. And finally, I have to thank my friends and family from around the country. Your love and support is always comforting and felt from a distance. As I move on to the next stage in life, I take with me friendship, memories and support from all of the amazing people that have contributed to my successful journey, thank you!

Table of Contents

Chapter 1: General Introduction.....	1
Chapter 2: Comparison of the transcriptional profiles of head and body lice.....	7
Chapter 3: Differentially gene expression in human head and body lice when challenged with <i>Bartonella quintana</i> , a disease causing agent.....	22
Chapter 4: Within plant genetic variation: An assessment of the somatic mutation hypothesis	37
Figures and Tables.....	45
Bibliography.....	64
Appendix A	77

Chapter 1: General Introduction

Variation exists at all levels of biological organization, between and within species, populations and individuals. From Charles Darwin and Alfred Russel Wallace (1858: evolution by means of natural selection) to Gregor Mendel (1865: mode of inheritance) to Thomas Hunt Morgan (1911: traits linked to specific chromosomes) to James D. Watson, Francis Crick and Rosalind Franklin (1953: structure of DNA), the field of genetics has slowly started to characterize the mechanisms underlining the variation observed between species, populations and individuals. Three fields of research, evolutionary biology, molecular biology and population genetics, culminated in the modern evolutionary synthesis during the 1930's and 1940's. It was at this juncture that evidence from these three fields supported the concept that genetic differences in the DNA code was responsible for observable phenotypic variation and could pass from generation to generation and flow between populations.

Over the last 50 years of the 20th century, numerous advances in techniques and empirical evidence have led to the ultimate development of the field of genomics. The driving force behind this movement was the desire to understand the link between individual genes, their function and how they changed to produce genetic variation observed phenotypically. During this time, scientists understood that DNA was the hereditary material, but were unclear of the mechanisms leading to replication or translation of this molecule. Furthermore, we lacked the knowledge of how the cell reads the genetic code. In 1961, Francois Jacob and Jacques Monod demonstrated messenger RNA to be the secondary molecule that moved between the nucleus, the location of DNA, and the ribosome where proteins are produced. During the late 1960's and early 1970's, the discovery of restriction enzymes (Meselson and Yaun, 1968; Arber and Linn, 1969) and reverse transcriptase (Temin and Mizutani, 1970) were instrumental in the beginnings

of the age of molecular cloning, allowing for individual fragments of DNA to be isolated and manipulated. The genetic code was cracked in 1971 by Marshall Nirenberg and colleagues who showed that a set of three nucleotides code for one of the 20 amino acids.

Scientists now had the genetic code, could isolate fragments of DNA and understood how DNA was copied, transcribed into RNA and translated into proteins. While we could measure variation in populations through electrophoretic techniques (Hubby and Lewontin, 1966), we still lacked an efficient way of sequencing or replicating DNA to understand what the variation was or how it affected gene function, in other words, linking phenotype to genotype. It was in 1975 that Gilbert and Sanger developed one of the first widely accepted methods for the sequencing specific regions of the DNA code, the Sanger method, still in use today, but has been largely replaced by next generation techniques. Kreitman and colleagues (1983) demonstrated single nucleotide polymorphisms in the *Adh* gene of 11 species of *Drosophila* linked to phenotypic and gene frequency differences in natural populations, while also suggesting that amino acid replacement is mostly deleterious to the organism. Kary Mullis in 1983 is credited with the discovery of the polymerase chain reaction (Rabinow, 1996). A two pronged attack was underway, the identification of the entire sequence from an organism and the link between those genes and function. The 1990's led to the first organisms from all 3 domains of life to be sequenced; the bacteria, *Haemophilus influenza* (Fleischmann *et al.*, 1995), the eukaryote, *Saccharomyces cerevisiae* (Goffeau *et al.* 1996) and the archaeon, *Methanococcus jannaschii* (Bult *et al.* 1996).

By 2003, the human genome was sequenced using the Sanger technique, costing \$2.7 billion and taking 13 years, (IHGSC, 2001). As of 2013, close to 10,000 species' genomes have been sequenced, primarily bacteria and viruses (NCBI, 2013). Of those, 792 eukaryotes have

been sequenced. The rate at which genomes are being completed is increasing at an exponential rate (Genbank 2013). This is mainly due to the decreasing costs of sequencing (NIH 2013). For example, in 2001, the average cost per genome was around \$100 million, but last year that number had decreased to just \$10,000. The largest drop in cost happened between 2007 and 2008 when sequencing centers largely switched from Sanger to next-generation sequencing (NIH 2013).

Next-generation or high throughput sequencing technology has exploded in recent years, due to the advent of Roche 454 (Indianapolis, IN, USA) and Solexa, now known as Illumina (San Diego, CA, USA). Originally Sanger sequencing used a method called chain-termination to sequence DNA fragments. While highly accurate, Sanger sequencing is limited to amplifying fragments of up to 1,000 base pairs (bp), but is slow, as it only sequences a single fragment at a time rather than running multiple fragments in parallel like next generation sequencing, and relatively expensive. Shotgun sequencing, of which Roche 454 and Illumina are primary players, has allowed fragments much larger than 1,000 bp, up to an entire genome, to be sequenced by splicing these DNA fragments into numerous shorter fragments. Each of these two techniques use a proprietary surface (Roche 454: well plate; Illumina: 7 lane slide) to attach DNA fragments to primers in colonies after which nucleotides are washed over these colonies to replicate each of these short DNA fragments. Both techniques use light to read the attachment of nucleotides as the DNA is replicated. Roche 454 results in up to 5 million reads of between 400-700 bp at a cost of \$10 per million bp, while Illumina produces up to 3 billion reads between 20-200 bp at a cost of up to \$0.15 per million bp. Each of these techniques produces an enormous amount of data, requiring high capacity computers to reassemble the reads into useable DNA sequences.

Moore's Law suggests that the number of transistors on integrated circuits doubles every two years. As a result, computer processing power has increased and the cost has decreased, resulting in computers that, now, have the same computing power as clusters of computers five years ago. In fact computers are now powerful enough that the bottleneck is disk space. Large next generation sequencing projects produce terabytes of data that need to be handled and stored. This has created a niche for bioinformaticians that can handle large datasets and a need for software capable of interpreting the reads into biologically significant data.

As with computer hardware and sequencing technology, software programs have quickly evolved to fill this niche. There are hundreds of programs on the internet that each has a specific purpose; *de novo* assembly, reference mapping, alignments, protein analysis, blast search, single nucleotide polymorphism (SNP) analysis and many others. Some of these tools are free to use while others require purchase. Recent software packages such as Genious and CLC Genomic workbench offer many of these tools in one canned software package.

With the advent of these new technologies biologists are now able to address questions that were previously intractable. New species are having their genomes mapped every year. Projects are underway to sequence multiple individuals from a single species such as the 1000 Human Genomes Project launched in 2008 (TGPC, 2012) or the 5000 Insect Genome Project (i5k; Robinson *et al.*, 2011), launched in 2011. These data, publicly available, allow investigators to determine SNP variation between individuals of a single species. As more organisms are sequenced, phylogenetic trees are being confirmed or resolved based on genomic data, rather than markers or morphological traits. Behavioral traits can now be matched with gene expression differences specific to the brain. As gene annotations improve, specific gene targets can be manipulated for desired outcomes such as GMO crops. Uncovering methylation of the DNA

sequence has suggested a Lamarckian way of passing on traits from parent to offspring. And finally, non-model organisms can now have their genomes mapped quickly and affordably.

Once an organism's genome is mapped, it becomes a platform for linking the biology of the organism to the genetic underpinnings of particular characteristics. Even if nothing genetic is previously known about an organism, with a genome in hand, additional sequencing of other individuals of that same species now allows for comparison of the genetic variation that exists between different species or within a single species. Here we use two organisms, the human body louse, *Pediculus humanus corporis*, and the black cottonwood tree, *Populus trichocarpa* to investigate genetic variation at a species and individual level, respectively.

In 2010, Kirkness and colleagues sequenced the 108 million base pairs (Mbp) genome of the body louse, consisting of 10,775 genes (Kirkness *et al.* 2010). A close relative, the head louse, *Pediculus humans capitis*, has not yet had its genome mapped, but previous work suggests these two organisms diverged from one another approximately 70,000 years ago with the advent of clothing in humans (Kittler *et al.*, 2003, 2004). What makes these two organisms an interesting study system is that they differ in one unique way, body lice can vector disease to humans whereas head lice cannot. While we see head lice commonly in children, body lice are rarely seen unless environmental conditions degrade, as with homeless people or after natural disasters. This leaves us with two important questions: 1) Are body lice (*Pediculus humanus coporis*) and head lice (*Pediculus humanus capitis*) the same species? 2) Are there gene expression differences between these two organisms that cause the body louse to vector disease to humans while head lice do not? Here we address these two questions using a combination of Roche 454 and Illumina sequence data along with an RNAseq analysis to elicit whole gene and gene expression differences between head and body lice. While the body/head louse is an ideal

system for studying the genetic variation between species, it is not the ideal system for investigating genetic variation within an individual.

The black cottonwood tree, *Populus trichocarpa* was the first woody plant to have its genome sequenced. Completed in 2007, the genome consisted of ~485 Mbp and 45,000 genes were annotated (Tuskan *et al.*, 2006). Consequently, *Populus trichocarpa* has now become the model organism for woody tree species. A commercially viable species, black cottonwood has two distinct characteristics which make it ideal for testing the somatic mutation hypothesis; it is relatively long lived (up to 200 years) and it can reproduce both sexually and asexually. Being a long-lived organism, trees encounter an increasing number of short-lived pests (rusts and insect herbivores). As a result, the generation and perpetuation of somatic mutations, and thus new variation within the individual, would be highly advantageous (Whitham and Slobodchikoff, 1981; Gill *et al.*, 1995). So our final question is, 3) what is the level of somatic variation that exists with an individual tree and its clonally derived offspring? We address this question with Illumina data and a SNP analysis of multiple tissues from 11 parent-offspring pairs of black cottonwood trees to describe the existing genetic variation. Consequently, what can this illuminate about the 32-year old somatic mutation hypothesis set forth by Whitham and Slobodchikoff (1981)?

Chapter 2: Comparison of the transcriptional profiles of head and body lice¹

Abstract

Head and body lice are both blood-feeding parasites of humans although only the body louse is a potent disease vector. In spite of numerous morphological and life history differences, head and body lice have recently been hypothesized to be ecotypes of the same species. We took a comparative genomics approach to measure nucleotide diversity by comparing expressed sequence tag data sets from head and body lice. A total of 10,771 body louse and 10,770 head louse transcripts were predicted from a combined assembly of Roche 454 and Illumina sequenced cDNAs from whole body tissues collected at all life stages and during pesticide exposure and bacterial infection treatments. Illumina reads mapped to the 10,775 draft body louse gene models from the whole genome assembly predicted nine presence/absence differences, but PCR confirmation resulted in a single gene difference. Read per million base pair estimates indicated that 14 genes showed significant differential expression between head and body lice under our treatment conditions. One novel microRNA was predicted in both lice species and 99% of the 544 transcripts from *Candidatus riesia* indicate that they share the same endosymbiont. Overall, few differences exist, which supports the hypothesis that these two organisms are ecotypes of the same species.

¹ This chapter appeared in its entirety in the journal, *Insect Molecular Biology* and is referred to later in this dissertation as Olds *et al.* 2012. Olds B.P., B.S. Coates, L.D. Steele, W. Sun, T.A. Agunbiade, K.S. Yoon, J.P. Strycarz, S.H. Lee, K.N. Paige, L.M. Clark and B.R. Pittendrigh. 2012. Comparison of the transcriptional profiles of head and body lice. *Insect Molecular Biology* **21**(2): 257-268.

Introduction

The body louse (*Pediculus humanus humanus* Linnaeus) was first described in 1758 and was differentiated from the head louse (*Pediculus humanus capitis* de Geer) in 1767, but their status as distinct species remains in debate (Durden & Musser, 1994; Khudobin, 1995). Although head and body lice do not interbreed in the wild (Busvine, 1948; Schaefer, 1978), fertile *F1* hybrids have been reported under laboratory conditions (Mullen & Durden, 2009). Busvine (1978) further indicated that head and body lice are specialized for the infestation of specific regions of human hosts, and do not migrate even in cases of dual infestation. Microsatellite DNA evidence from double infested humans suggests that these two organisms are distinct species that have differentiated through spatial isolation on the same host (Leo *et al.*, 2005). In contrast, a comparison of 10 haplotypes deduced from polymorphisms in the cytochrome oxidase I (COI) gene, suggested that these two organisms are conspecifics (Leo *et al.*, 2002). A phylogenetic reconstruction using polymorphism data from random intergenic spacer region sequence suggested that head lice populations might be a reservoir from which body lice emerge during instances of low hygiene (Li *et al.*, 2010). This evidence indicated that head and body lice are variants of a single species, which respond differently to environmental conditions, allowing elicitation of different region specificities during infestation. Interestingly, these phylogenetic data also estimated that head and body lice differentiated from other species of lice about 5.6 million years ago. This split in the lice lineage roughly coincides with the human and chimpanzee, and suggests that host-pathogen co-evolution may have given rise to unique louse lineages (Reed *et al.*, 2004; Patterson *et al.*, 2006). Head and body lice are thought to have further diverged into three mitochondrial clades over the last 2 million years, and contemporary populations are comprised of (1) lice found in Africa, (2) lice found in North and Central

America, and (3) lice found only in Ethiopia and Nepal (Pennisi, 2004; Reed *et al.*, 2004; Leo & Barker, 2005; Light *et al.*, 2008; Raoult *et al.*, 2008). This most recent differentiation of head and body lice may reflect the commonplace use of clothing by human cultures (Kittler *et al.*, 2003, 2004). Such differences between biotypes or species have classically been investigated through single nucleotide polymorphism and microsatellite analyses comparing populations; however, with recent advances in functional genomics we can investigate the transcriptional differences between closely related groups of organisms. Such comparisons can be carried out in a highly efficient manner with the existence of well-defined transcriptomes in both groups.

In addition to genetic differences between head and body lice, distinct feeding habits, morphologies and disease vectoring capacities have been observed. Head lice aggregate and feed on the head region of humans, where females attach eggs to the base of hair shafts. In contrast, body lice feed upon body regions, secure eggs to clothing (Light *et al.*, 2008) and tend to be slightly larger in size compared to head lice (Light *et al.*, 2008; Bonilla *et al.*, 2009). Although head lice are common amongst school-aged children, symptoms including itching and loss of sleep pose relatively few health risks (Light *et al.*, 2008; Toloza *et al.*, 2008). In contrast, body lice are competent and significant vectors of human diseases (Light *et al.*, 2008), and can transmit the bacteria *Bartonella quintana*, *Borrelia recurrentis* and *Rickettsia prowazekii* to human hosts during feeding (Light *et al.*, 2008; Bonilla *et al.*, 2009). *B. quintana* is the causative agent of trench fever, which afflicted an estimated 1 million humans during World War I and remains problematic in regions of the world where human living conditions are compromised (Sasaki *et al.*, 2006; Bonilla *et al.*, 2009; Mullen & Durden, 2009). The bacterium *B. recurrentis* can cause relapsing fever that, if untreated, results in 30–70% mortality rates in humans, and remains problematic in parts of the world where living conditions are poor and access to medical

care is limited (Rhee & Johnson, 2009). Epidemic typhus caused by *R. prowazekii* infections was responsible for more than 3 million deaths during World War I when it was rampant in crowded prison conditions. Body lice are seldom observed in the USA, but remain a public health concern because of infestations of homeless individuals or persons affected by natural disasters when hygiene conditions have declined (Sasaki *et al.*, 2006).

Understanding louse biology and host interactions, as well as genome function, are important factors for the development of novel tactics to reduce vector competency and disease transmission in human populations. The current body louse gene models are primarily based upon *ab initio* predictions from the draft body louse genome (Kirkness *et al.*, 2010), which predicted 163 novel protein coding genes that show no known orthologues within other species. Before now, no direct evidence has been provided from expressed sequence tag (EST) data to assist in gene prediction schema. Gene prediction algorithms have limited capabilities for the detection of nested genes and accuracy in defining N- and C-terminal coding regions (Salamov & Solovyev, 2000), and often require transcriptome ‘training sets’ to improve predictions (Munch & Krogh, 2006). EST and digital gene expression data remain resources for novel gene discovery (Nesbitt *et al.*, 2010), functional genome annotation (Srinivasan *et al.*, 2005), and for comparative genomics (Mitrevva *et al.*, 2004). Herein, EST data sets for head and body lice were generated by next generation sequencing (NGS). Transcript data were used to (1) validate *ab initio* gene predictions for the body louse (Kirkness *et al.*, 2010), (2) identify novel genes within the body louse, (3) characterize genes from the head louse, and (4) compare sequence and relative gene expression-levels between head and body lice. The EST and gene expression data are valuable for on-going genome annotation efforts, and provide key comparative genomics

information that addresses on-going questions of recent speciation or functional variance between head and body lice.

Experimental Procedures

Samples, cDNA preparation and pyrosequencing

Two body louse (BL) and three head louse (HL) strains were used. A colony of human body louse (Frisco-BL) was originally collected from nine homeless individuals in San Francisco by Dr Jane Koehler in December 2008 (unpublished data) and a head louse collection (SF-HL) was obtained in south Florida (Yoon *et al.*, 2004). Frisco-BL and SF-HL are maintained on an *in vitro* rearing system with human blood. Additional head lice were obtained from Bristol, UK (BR-HL) and Korean (KR-HL) colonies. The Culpepper strain of body lice was maintained on rabbits. Frisco-BL and BR-HL colonies were dipped into 1% solutions in ethanol of each of the following pesticides; permethrin, malathion, ivermectin, spinosad, fipronil and imidacloprid for 10 m, 30 m and 1 h time intervals. Lice from the Frisco-BL and BR-HL colonies were fed human blood infected with *Escherichia coli* and *Staphylococcus aureus*. All other strains of lice were not stressed with either pesticides or bacteria.

Body lice (Frisco-BL and Culpepper) and head lice (SF-HL, BR-HL and KR-HL) colony samples from pesticide and infected blood treatments were pooled separately with approximately 20 mg of each life stages (egg, first, second, third instar and adult) from the same colony, and flash frozen in liquid N₂. Total RNA was extracted from each pooled sample using the Qiagen RNA extraction kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. The RNA samples were submitted to the Keck Center at the University of Illinois at Urbana-Champaign, normalized using the procedure described by Lambert *et al.* (2010), sequenced on a full plate on the Roche 454 sequencer (Roche, Indianapolis, IN, USA) and data were outputted in

.sff format. The remaining portion of the same RNA samples were subsequently sequenced on a single flow cell of the Illumina Genome Analyzer 2 (GA2) (San Diego, CA, USA) using 50 pyrosequencing cycles and data obtained as .fastq files.

Sequence analysis and read mapping

Raw sequence read data from head and body louse cDNA pools were analyzed using the CLC Genomics Workbench 4x (Cambridge, MA, USA). Preprocessing involved trimming of nucleotides _50 from each 454 read, and for quality scores of <20 (both 454 and Illumina GA2 reads). The processed read data from head and body lice were mapped (aligned) independently to (1) the body louse gene models, (2) the body louse genome sequence assembly and (3) the genome of the endosymbiont *C. riesia*. The methods for these processed samples are described below.

The body louse gene models v. 1.2 were downloaded from Vectorbase.org ([http://phumanus.vectorbase.org/GetData/ Downloads/](http://phumanus.vectorbase.org/GetData/Downloads/)) in the file 'phumanus.TRANSCRIPTS-PhumU1.2.fa.gz', unpacked and uploaded into the CLC Genomics Workbench 4x. Processed head and body louse read data were independently mapped to the body louse gene models using the parameters: mismatch cost = 2, insertion and deletion cost = 3 and similarity $\geq 80\%$. Subsequent consensus transcriptome sequences were generated using a majority rule and were outputted in FASTA format for both head and body louse data sets. Unmapped read data from head and body louse cDNA pools were subsequently aligned against the body louse whole genome sequence supercontigs (Kirkness *et al.*, 2010; file: 'phumanus.SUPERCONTIG-USDA.PhumU1.fa.gz') using parameters and output as described previously. 'Consensus transcriptomes' for head and body louse were created by combining sequences generated by mapping to body louse gene models and supercontigs. These analyses allowed for the assignment

of putative orthology between head and body louse genes, as well as for the potential discovery of novel expressed genes not discovered by *ab initio* gene predictors.

Functional annotations of head and body louse ‘reference transcriptomes’ were retrieved from the annotations previously assigned to body louse gene models v. 1.2 (Kirkness *et al.*, 2010). Novel genes, defined as reads that mapped only to the *P. humanus* supercontigs, were used as queries of the NCBI nr/nt and dbEST, and nr databases using blastn and blastx, respectively (word size = 11, match cost = 1, mismatch cost = -3, gap cost = 5 and extension cost = 2). Any blast hit with an *e*-value smaller than $E-10$ and mapping to an organism other than head or body lice was investigated for coverage and similarity. Pairwise alignments were generated amongst putative orthologous transcripts from head and body lice reference transcriptome sequences and the body louse gene models using a local (bl2seq from BLAST) alignment and default parameters (word size = 11, match cost = 1, mismatch cost = -3 and gap cost = 5). The percent amino acid sequence identity was estimated for each comparison (orthologue), and 1 minus this value plotted.

Processed read data that did not map to either body louse gene models v. 1.2 or supercontigs were mapped against the 575 kb body louse endosymbiont *C. riesia* genome sequence (GenBank accession CP001085; RefSeq NC-014109.1; Kirkness *et al.*, 2010), using parameters and output procedures described previously.

PCR validation of putative species-specific genes

Of the total 10,775 potential genes described by Kirkness *et al.* (2010), four were not observed from body louse EST data and five were not present within the head louse ESTs. These differences were validated by extracting total mRNA individually from whole adult head and body louse using the Qiagen RNA extraction kit (Qiagen) according to the manufacturer’s

instructions, followed by synthesizing first strand cDNAs with the Superscript III Reverse Transcriptase (Invitrogen, Life Technologies, Carlsbad, CA, USA) with oligo(dT) (50 uM) as described by the manufacturer. Template cDNA was quantified on a NanoDrop 2000 (Thermo Scientific, Waltham, MA, USA) and diluted to 20 ng/ml with nuclease free water. Gene specific oligonucleotide primers for all nine putative differentially expressed genes were designed using PRIMER3 (Rozen & Skaletsky, 2000) using default parameters and synthesized at Integrated DNA Technologies (Coralville, IA, USA). PCR reactions were performed in 50 µl reaction volumes that contained 10 µl of 10x reaction buffer (Promega Corporation, Madison, WI, USA), 10 µl of 1.5 mM MgCl₂, 8 µl of 10 mM deoxynucleotide triphosphate, 0.25 µl (5 U/µl) of *Taq* polymerase (Promega Corporation), 1 µl of 10 pmol reverse and forward primers and 1 µl of 20 ng/ml of cDNA template. Identical reactions using 20 ng/ml of genomic DNA were used in positive control PCR reactions from both head and body lice. PCR reactions were carried out in a PTC-200 thermal cycler (MJ Research, Inc. Waltham, MA, USA), which included a 94 °C denaturation for 3 min, followed by 31 cycles of 94 °C for 1 min, 55 °C for 1 min and 72 °C for 1 min, then a final extension step of 72 °C for 10 min. The PCR products were checked on a 2% agarose gel impregnated with ethidium bromide, and visualized under UV illumination.

Reads per kilobase of exon model

Trimmed reads obtained from the Illumina GA2 output were used in a RNA-Seq analysis from CLC Genomics Workbench, modeled after Mortazavi *et al.* (2008) except that spliced ‘reference transcriptome sequence’ was used as a scaffold. Read mapping parameters on CLC Genomic Workbench included paired distance of 180–250, minimum reads of 10 and maximum mismatches of 2, and the data were treated as prokaryotic because of the omission of introns. Values for each transcript were reported as RPKM values (Mortazavi *et al.*, 2008). In order to

evade the effects of sampling error of rare events, any data that contained more than 25 mapped reads to a single gene within either the head or body louse were omitted from the data set. Read data from the head louse were normalized based upon the ratio of the RPKM value for the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) transcript level in the body louse (ratio = 1.129323). Correlation between head and body louse RPKM values was determined by linear regression. The fold-increase (+ value) or decrease (- value) in RPKM from head louse data compared to that in body louse was calculated, and the distribution of values were used to determine 99.0 and 99.9% CIs. Significance of any differences in relative transcript abundance was determined by use of a pseudo *P*-value estimated at the 99.9% CI, and administered as an arbitrary significance threshold ($P = 0.001$).

Results

Samples, cDNA preparation and pyrosequencing

Raw Roche 454 data generated 418 234 (body louse) and 346 902 (head louse) reads with an average length of 331 and 318 base pairs (bp), respectively. After trimming based on quality (Phred < 20) and removing reads fewer than 50 bp in length, there were 372 983 and 304 606 reads with an average length of 303 and 287 bp, respectively. In contrast, raw Illumina data consisted of 90.1 (body louse) and 94.1 (head louse) million paired reads with an average length of 84 and 83 bp, respectively. After trimming, there was a loss of ~200 reads from each sample but average length decreased to 76 and 75 bp, respectively. There were also 2 (body louse) and 1.8 (head louse) million single reads that decreased in average length from 52 and 51 bp to 39 and 38 bp, respectively.

Sequence analysis and read mapping

Mapping of Illumina read data against the 10,775 body louse gene models resulted in matches to 10,771 and 10,770 transcripts, respectively, from body and head lice (Table 2.1), and showed that four transcripts were not found in body lice and five transcripts were not found in head lice (nine presence/absence differences; Table 2.1). A total of 217 transcripts, combined from head and body louse data sets, were predicted to have $\geq 3x$ coverage. Within this subset of 217 genes, 131 (60.4%) were the same in both head and body lice (Table 2.2). Only 66 transcripts, combined from head and body louse data sets, were identified with 1x coverage and 38 of these were the same in both head and body lice.

Further analysis only considered the 7401 consensus sequences with lengths that spanned $\geq 70\%$ of the reference. Gapped pairwise alignments of consensus body and head louse EST sequences respectively showed means of 99.5 ± 0.6 and $99.1 \pm 0.8\%$ similarities with the body louse gene models v. 1.2 (Figures 2.1, 2.2). These mean estimates decreased to 99.6 ± 0.4 and $99.2 \pm 0.6\%$ for analogous ungapped pairwise alignments. The percent identity ranged from 87.2 to 100% for gapped alignment and 92.1 to 100% for ungapped alignment in both species (Figure 2.1). There was a weak positive correlation between the variance in head or body lice EST consensus sequences and the percent that the transcript covered the reference gene model sequence (Figure 2.2). Grouping of genes based upon ‘conserved’ (predicted genes with homology in another species), ‘hypothetical’ (predicted genes lacking homology and function), ‘predicted’ (predicted genes with sequence similarity in other species but function is unknown) and ‘other’ (all other genes with a known functional role) genes, showed that the ‘hypothetical’ group had higher intraspecific variation compared to the other three groups. The 93 genes annotated as immune-responsive in the body louse genome were represented within both head

and body lice consensus EST sequences (coverage range 3–12 000x). A comparison of the percent identity from the body louse gene model for corresponding body and head lice transcripts (orthologues) indicated that three genes (PHUM56303, 018550 and 248870) are $\leq 90\%$ from the gene models with gapped alignment. Additionally, PHUM56303 and 248870 show $\geq 10\%$ divergence between head louse and the consensus EST sequences, whereas PHUM018550 is $\geq 10\%$ divergent from consensus in both head and body lice (Figure 2.1).

A new microRNA (miRNA) was predicted in both the head and body louse ‘reference transcriptome’ sequences. The putative 62 bp head and body lice miRNA showed 100% sequence similarity between species, and when aligned with the latter 62 of the 71 bp from the *Tribolium castaneum* miR276 (Figure 2.3A). Despite omission of the initial nine nucleotides, the predicted head and body louse miRNA formed a secondary structure that is analogous to that of miR276 (Figure 2.3B, 2.3C). Additionally, we found that 538 and 539 of the 544 *Candidatus riesia* transcripts are present within body and head louse ‘consensus transcriptomes’, respectively (Table 2.1). These data indicate that lice share this obligate endosymbiont, which can be detected using metagenomic analysis.

PCR validation of putative species-specific genes

Using a bioinformatics approach, a total of nine genes were unique to one EST data set or the other. Specifically, four and five genes were missing from the body and head louse, respectively. Primers for all nine putative species-specific genes were obtained and used to validate presence/ absence in each head or body louse RNA sample. Five of the nine PCR primer pairs (Table 2.3) resulted in PCR amplification for body lice, whereas four of nine produced a fragment for head lice. Overall, only five out of nine primer pairs worked to generate a fragment when body louse genomic DNA was used as a template; however no fragments were generated

within negative controls (Table 2.4). PHUM091050, PHUM621980, PHUM623680 and PHUM624080 were not able to amplify products in any sample and could represent falsely predicted transcripts from the original body louse genome project. These results also indicate that the hypothetical protein encoded by gene PHUM540560 was only present in body lice (Table 2.4).

Reads per kilobase of exon model

Normalization of head louse reads per kilobase of exon model (RPKM) values based upon the ratio at the GAPDH gene (1.129323) resulted in approximate equilibration of expression data from a housekeeping gene. Read mapping resulted in the estimation of highly correlated gene expression data from head and body lice (Figure 2.4A; $R^2 = 0.9663$). The distribution of fold-differences in normalized read count data showed a normal tendency around the ratio of 1.0 (no expression difference between species; range -6.25 to 6.40). Application of a significance threshold of $P = 0.001$ estimated from the 99.9% confidence interval (CI; Figure 2.4B) resulted in the setting of critical gene expression ratios at $3.741 \geq X \geq -3.741$. This threshold suggests that 14 genes showed a statistically significant difference in transcript level between species, and corresponded to nine head louse genes that were more highly expressed (+ expression) and five genes that showed lower (–) estimated expression values (Table 2.5).

Discussion

Head and body lice have been considered separate species for centuries and only recently has that idea been studied with the latest molecular tools. According to the numerous definitions of a species (ie biological, ecological, morphological, evolutionary and phylogenetic), head and body lice could be considered separate species based upon a few key traits; head and body lice occupy separate niches, head lice are smaller than body lice, head lice feed constantly and body

lice feed twice a day, migrating from the clothing to the skin and back again, and body lice can vector disease whereas head lice cannot (Mayr, 1942; Busvine, 1948, 1978; Schaefer, 1978; de Queiroz, 2005). Yet, a few characteristics led to the questioning of this hypothesis. First, there is anecdotal evidence that the two species can interbreed in captivity (Mullen & Durden, 2009). Second, it has been hypothesized that because body lice are seldom seen in the wild except when environmental conditions degrade, head lice might be a reservoir from which body lice emerge (Li *et al.*, 2010). Third, head lice have been shown to develop phenotypic characteristics of body lice in the lab (Alpatov & Nastukova, 1955; Levene & Dobzhansky, 1959). Numerous insect genomes have been assembled using NGS technology, which is increasingly affordable. These draft genomes are typically mapped but not well annotated, requiring additional refinement by genetic linkage or physical mapping in order for complete scaffolds to be produced such as those from model species (i.e. *Drosophila melanogaster*, *Mus musculus*, *Arabidopsis thaliana*). For those researchers interested in the description of new species, a basic draft genome allows for the rapid identification of gene and transcripts (ISGC 2008, Werren *et al.*, 2010), which provides data for comparative genomic and phylogenetic reconstruction methods. NGS data were used to investigate the species concept for head and body lice.

Our data also indicate that the gene predictions provided by Kirkness *et al.* (2010) are highly complete. The EST sequence and expression data generated in this study are also, to our knowledge, the first genomic comparison of head lice to body lice. EST data demonstrate that transcript sequences and expression levels are highly similar between these two organisms. Our data also suggest that the majority of body and head lice transcripts show 0.1– 1.3% nucleotide diversity and only a small number of genes have a level of differentiation >5–15% different. The levels of interspecific (between the body louse gene models and head louse consensus EST

sequences) and intraspecific nucleotide variation (between the body louse gene models and body louse consensus EST sequences) were not significantly different. The intraspecific genetic variation amongst humans is as low as 0.1% and greater than 1% depending on the method used for detection (Pang *et al.*, 2010). Boyce *et al.* (1994) showed a within-species nucleotide diversity of 0.1–1.3% present in restriction-site polymorphisms of the mitochondrial genome of the bark weevil, *Pissodes strobi*. Additionally, Boyce *et al.* (1994) demonstrated a 2–16% difference in nucleotide diversity in a closely related species of the family Curculionidae. The genetic diversity between humans and great ape species has been estimated to be ~2% depending on whether insertions/deletions (indels) are included (Wetterbom *et al.*, 2006). Additionally, we observed a single gene difference between body and head lice (presence/absence in expression only). Unfortunately the function of this gene is unknown and will need to be investigated further. Our data suggest that the hypothesis that at least some strains of head and body lice may be ecotypes of the same species may be valid (Li *et al.*, 2010). These conclusions are in contrast to previous studies using neutral molecular markers that supported the idea that head and body lice are separate species (Leo & Barker, 2005; Leo *et al.*, 2005); however our data, by using total RNA, ignore mutational differences found in noncoding regions used by many of the previous studies.

MiRNAs, a class of nonprotein coding small RNAs, are single stranded RNAs of about 18–24 nucleotides in length and originate from hairpin-shaped or stem-loop structures of approximately 70 nucleotides (Hori *et al.*, 2010; Shruti *et al.*, 2011). Our initial BLAST search, through the National Center for Biotechnology Information (NCBI) database, identified a novel miRNA, mir-276, from the EST reads in both body lice and head lice. According to the <http://www.mirbase.org> database, mir-276 has been identified in at least 12 of the *Drosophila* species

and over 11 other organisms, including *Apis mellifera*, *Bombyx mori*, *Daphnia pulex* and *Nasonia vitripennis* (Griffiths-Jones, 2004; Griffiths-Jones *et al.*, 2006, 2008; Kozomara & Griffiths-Jones, 2011). Although the specific function(s) of most miRNAs are still essentially unknown, two separate studies on the honey bee brain have implicated mir-276 as potentially playing roles in larval development and age dependent behavioral changes in *A. mellifera* (Behura & Whitfield, 2010; Hori *et al.*, 2010). Although the molecular or biochemical role of this miRNA was not investigated within head and body lice, miRNAs have been shown to have roles in developmental regulation, viral and metabolic diseases and post-transcriptional regulation (Singh *et al.*, 2008; Hori *et al.*, 2010). These will be interesting avenues for future louse genomics research.

The EST and gene expression data from the present study validated the gene models predicted by Kirkness *et al.* (2010), and provided insight into the level of gene sequence diversity between body and head lice. Comparative genomic methods for species description remain to be fully developed, but our comparison of closely related species/ecotypes provided a unique opportunity to analyze genomic data in a phylogenetic context. As new species' genomes are sequenced, the genetic species concept might gain traction, leading to a more universal definition of a species, possibly identifying species in the process of speciation. This study provides valuable genomic data for sequence comparison, with subsequent analyses suggesting that head and body lice might be conspecific species or subspecies. Furthermore, the reference consensus EST sequences for the head louse provides potential candidate genes with differential expression with the body louse and may lead to the identification of differences that allow the vectoring of human diseases.

Chapter 3: Differential gene expression in human head and body lice when challenged with *Bartonella quintana*, a disease causing agent

Abstract

Head and body lice are thought to be ecotypes of the same species, but differ in one very important aspect; body lice vector disease to humans and head lice do not. Immunoresponse genes are suspected in the disease vector competency differences between the two organisms. We utilized an RNAseq analysis on 7-day old head and body lice fed blood infected with *Bartonella quintana*, the bacterium that causes trench fever, and control individuals to elicit gene expression differences. Eight immunoresponse genes came out significant, many associated with the Toll pathway; Fibrinogen-related protein (PHUM500950), Spaetzle (PHUM595260), Defensin 1 (PHUM365700) and 2 (PHUM595870), Serpin (PHUM311330), Cactus (PHUM345810), Scavenger receptor A (ScavA; PHUM066640) and Apolipoprotein D (PHUM427700). Increased expression of Fibrinogen-related protein and Spaetzle, both related to the Toll pathway, in treated body lice supports the hypothesis that body lice are fighting infection from *B. quintana*. But conflicting results in Defensin 1 and 2 based upon validation method suggest another mechanism in head lice alternative to the Toll pathway might be involved. Additionally, Scavenger receptor A was higher in both control and treated head lice, suggesting higher phagocytotic activity in head lice to curb infection. These eight genes represent targets for future RNAi knockdown experiments to confirm the disease vector competency differences in head and body lice.

Introduction

Head (*Pediculus humanus capitis*) and body lice (*Pediculus humanus corporis*) are thought to be closely related species or even ecotypes of the same species (Leo *et al.*, 2002; Li *et al.*, 2010; Olds *et al.*, 2012). Although head and body lice do not interbreed in the wild (Busvine, 1948; Schaefer, 1978), fertile F₁ hybrids have been reported in laboratory conditions (Mullen and Durden, 2009). However, they differ in some well characterized life-history patterns/traits. Head lice aggregate and feed on the head region of humans, where females attach eggs to the base of hair shafts. In contrast, body lice feed upon body regions, secure eggs to clothing (Light *et al.*, 2008), and are larger in size as compared to head lice (Bonilla *et al.*, 2009; Light *et al.*, 2008).

One of the most important practical differences between head and body lice is that head lice do not transmit human disease, whereas body lice are vectors of the following bacterial diseases to humans: trench fever, caused by *Bartonella quintana*, relapsing fever caused by *Borrelia recurrentis*, and epidemic typhus caused by *Rickettsia prowazekii* (Bonilla *et al.*, 2009; Light *et al.*, 2008). Whether or not an insect can be a competent disease vector is most likely influenced by the immune-response of the host organism. To date, several studies have investigated immune system gene comparisons in mosquito (Christophides *et al.*, 2002), honeybee (Evans *et al.*, 2006), flour beetle (Zou *et al.*, 2007) and silkworm (Tanaka *et al.*, 2008), highlighting both similarities in gene families but also striking differences. For example, both head and body lice lack the immunoresponse genes, immune deficiency (IMD), gram-negative bacteria-binding (GNBP) and Fas-associated (apoptosis) protein with death domain (FADD), but show differential immune responses. Kim *et al.* (2011) demonstrated a similar pattern of phagocytotic activity in head and body lice when challenged with *Staphylococcus aureus*, a

gram-positive bacteria, but a different pattern of phagocytotic activity in head and body lice when challenged with *Escherichia coli*, a gram-negative bacteria, suggesting differential gene expression between the head and body louse.

With the completion of the body louse genome being annotated along with relatively complete body and head louse transcriptomes (Kirkness *et al.*, 2010; Olds *et al.*, 2012), we are now in a position to begin to ascertain transcriptome differences giving us insights into why body lice vector disease and head lice do not. Understanding how body and head lice differentially respond to *B. quintana* is an important part of elucidating why one is a competent vector and the other is not. A first step towards this aforementioned goal is a transcriptome analysis comparing the transcriptional profiles of head and body lice exposed or not exposed to *B. quintana* in their blood meals. Utilizing an Illumina platform and subsequent RNAseq analysis, we measured differential gene expression in both head and body lice when challenged with a disease causing bacteria, *Bartonella quintana*.

Experimental Procedures

Bacterial strain and culture conditions

B. quintana were obtained from Dr. Jane Koehler (University of California-San Francisco) and maintained in a biosafety level 2 (BSL2) laboratory, at the University of Massachusetts-Amherst. *B. quintana* were cultured on chocolate agar plates following Dr. Koehler's protocols (37 °C, culture plates in CO₂ jars) for 7 to 10 days and harvested by adding 4 ml of PBS (0.1 M, pH 7.2). Colony forming unit (CFU) per microliter were estimated using the dilutions of bacterial suspension in PBS (10 µl, 10⁴- to 10⁷-fold dilutions) plated on the chocolate agar plates.

Experimental infection of body and head lice with B. quintana

A field population of human body lice (Frisco-BL) originally collected from 9 homeless individuals in San Francisco (December, 2008) by Dr. Jane Koehler's research group and head lice (SF-HL) collected from South Florida (Yoon *et al.*, 2004), both maintained on the *in vitro* rearing system with human blood, were infected with *B. quintana*. Harvested *B. quintana* suspension (ca. 50 µl) in PBS was mixed with 1 ml human blood to prepare 2.0×10^3 CFU/µl blood (Seki *et al.*, 2007). Adult female lice (body and head lice, respectively) were fed on the *B. quintana* infected blood using the *in vitro* rearing system (Yoon *et al.*, 2006) until fully satiated. Fed lice were transferred into a clean rearing unit with non-infective blood and maintained for 7 days. For control experiments, lice were fed on non-infective blood for 7 days. Louse samples (30 lice/experiment) were stored in a liquid N₂ tank. Collected samples from 3 replicated experiments (30 infected body lice (BLT), 30 infected head lice (HLT), 30 uninfected body lice (BLC), and 30 uninfected head lice (HLC)) for a total of 12 pooled samples were packaged in dry-ice and shipped to Dr. Barry Pittendrigh's laboratory (University of Illinois at Urbana-Champaign) for RNAseq analysis.

RNA Extraction

Total RNA was extracted from each pooled sample using the Qiagen RNA extraction kit (Qiagen, Valencia CA) according to manufacturer instructions. Samples were submitted to the Keck Center at the University of Illinois at Urbana-Champaign and sequenced on an Illumina Genome Analyzer 2 (GA2) using 50 pyrosequencing cycles. Two samples, each library tagged independently, per lane were sequenced for a total of six lanes of Illumina and data obtained as .fastq files.

Raw data processing and RNAseq analysis

Raw sequence read data from head and body louse cDNA pools were analyzed using the CLC Genomics Workbench 4x. Pre-processing involved trimming of nucleotides with quality scores of < 20 from Illumina GA2 reads. The processed read data from head and body louse were mapped (aligned) independently to 1) the body louse gene models, 2) the genome of the endosymbiont *Candidatus riesia* and 3) the genome of 24 bacterial species. The body louse gene models version 1.2 were downloaded from Vectorbase.org, the 25 bacterial species, including *Candidatus riesia* were downloaded from NCBI and uploaded into the CLC Genomics Workbench 4x. Processed head and body louse read data from each of the 12 samples (4 treatments X 3 replicates) were independently mapped to the body louse gene models and bacterial reference genomes using the parameters: mismatch cost = 2, insertion and deletion cost = 3, and similarity $\geq 80\%$.

RNAseq analysis was modeled after Mortazavi *et al.* (2008), using RPKM values, except that in each case, the “reference genome” included only transcripts as to allow direct comparison of individual transcripts from each treatment. RNAseq parameters on CLC Genomic Workbench included paired distance of 180-250 bps, a minimum of 10 reads, maximum mismatches of 2, minimum length fraction of 0.9 and minimum similarity fraction of 0.8. The data were treated as prokaryotic due to omission of annotation with the transcript files we used for each organism. Values for each transcript were reported as both reads per kilobase of exon model per million mapped reads (RPKM; Mortzavi *et al.*, 2008) and unique reads mapped.

For each transcript, RPKM values were square-root transformed due to a Poisson distribution of count data. Any RPKM value less than 1 was counted as a 0. Consequently the

square-root transformed mean RPKM value was subtracted for each of the (4) 2-way comparisons (HLC-HLT, BLC-BLT, HLC-BLC and HLT-BLT) and then divided by the standard error mean difference, resulting in a list of p-values for each transcript. This list was then ordered from lowest to highest and using a cumulative p-value, all transcripts less than 0.05 were considered significant.

A second analysis, assuming a negative binomial distribution, was conducted on unique reads mapped and analyzed with the R (R 2008) and DEseq (Anders and Huber, 2010) software packages. Each transcript was analyzed for (4) 2-way comparisons (HLC-HLT, BLC-BLT, HLC-BLC and HLT-BLT). DEseq uses a geometric mean to adjust for uneven distribution of reads across replicates and treatments to allow for direct comparison of resultant values to calculate fold change and p-value. Consequently log2 fold change and an adjusted p-value are calculated. To control for false discovery rate (FDR), the adjusted p-value was calculated with a Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The resulting list of transcripts was considered significantly differentially expressed based upon the 2-way comparison performed.

Finally, a qRT-PCR was run on 17 immunoresponse genes. Multiple primer sets were constructed for each of the 17 target genes to determine a single best primer set for qPCR (Supplemental Table 1). From our 12 samples, cDNA was created using Promega GOScript reverse transcriptase system. Elongation factor 1- α (EF1 α) was used as control between samples. cDNA samples were diluted based upon the lowest concentration of EF1 α so that concentrations were the same across all samples. Equalized samples were prepared for qPCR using the Promega GOTaq qPCR master mix and then analyzed on the A&B Bio Systems Step One Plus real time PCR machine. Each gene was run in triplicate.

Results

Sequence analysis and read mapping

Mapping of Illumina read data against the 10,992 body louse gene models resulted in an average of 283x coverage from all 3 replicates. RNAseq analysis resulted in an average of 68,367,908 total reads with an average of 47,067,912 reads uniquely mapped.

Analysis of differential gene expression

Using RPKM values, 4, 3, 3272, and 3402 genes were found significantly differentially expressed for the (4) pairwise comparisons, HLC-HLT, BLC-BLT, HLC-BLC and HLT-BLT, respectively (Table 3.1). Alternatively, using only uniquely mapped reads, 23, 19, 552, and 954 genes were significantly differentially expressed, respectively. Only 127 genes matched between RPKM and unique reads mapped for the HLC-BLC comparison and 188 genes matched for the HLT-BLT. Due to the negative binomial analysis being more conservative than the Poisson analysis, log₂ fold change values for genes found in both the HLC-BLC and HLT-BLT pairwise comparisons (N=437) were plotted against one another (Figure 3.1). Most genes showed a similar fold change pattern regardless if they were control or treated lice. Only five genes showed a different fold change pattern when compared between control lice and treated lice. Two genes are hypothetical proteins, but PHUM427700 (Apolipoprotein-D), PHUM595870 (Defensin 2) and PHUM365700 (Defensin 1) are all immunoresponse genes.

Specifically, we were interested in all immunoresponse genes, of which there are 93 in the head and body lice with an additional 12 genes associated with the immune response. Using a combination of both RPKM value results and unique reads mapped and adding in a qRT-PCR confirmation, a few genes showed overlapping results between two or more of the analysis

techniques. With a Poisson distribution (RPKM values), 36 of the 93 and 5 of 12 genes showed differential expression. Conversely, with a negative binomial distribution (unique reads mapped), 10 of 93 and 1 of 12 genes were significant (Table 3.2). With such disparity between analyses using RPKM or unique reads mapped, qRT-PCR was conducted on 17 immunoresponse genes that were either common to both analyses or unique to one or the other. Overall, the qRT-PCR more closely matched the results from negative binomial analysis (Table 3.2).

When comparing within ecotype (head or body lice), no genes were shown to be significant between control versus treated lice by all three analysis methods (i.e. Poisson, negative binomial or qRT-PCR; Figure 3.2). Although, PHUM427700 (Apolipoprotein D) was found to be significantly higher in treated body lice as compared to control lice in the negative binomial analysis and confirmed by qRT-PCR. PHUM595870 (Defensin 2) showed up as significant in both negative binomial (higher in treated body lice versus control body lice) and qRT-PCR (higher in control body lice versus treated body lice) but concluded the opposite result.

When comparing between treatment (control versus treated lice), three genes were significant between both Poisson and negative binomial analyses and confirmed by qRT-PCR; PHUM500950 (Fibrinogen-like protein) and PHUM596260 (Spaetzle) were significantly higher in body lice for both control and treated lice, while PHUM365700 (Defensin 1) was only higher in control head lice as compared to control body lice. A few other genes (PHUM066640, PHUM311330, PHUM427700) were confirmed to be significantly differentially expressed by qRT-PCR and negative binomial analyses but not Poisson analysis (Figure 3.3).

Discussion

Head and body lice share the same genetic background, including all 93 immunoresponse genes. Both organisms lack the recognition gene, GGBP and both IMD and FADD from the

IMD pathway. But, there are no genes that are present in one and not the other organism (Olds *et al.*, 2011), which leaves gene expression as one possible culprit for disease vector competency difference between head and body lice.

Overall, of the hundreds of genes found differentially expressed common to both the Poisson and negative binomial distributions, a gene ontology (GO) analysis revealed no enrichment categories of significance. A few metabolic enrichment categories were observed with low enrichment scores, but none of these explained the differences in vector competency. So we limited our search to genes specifically associated with the immune response in lice. We identified seven immunoresponse genes that showed significant differential expression in at least one RNAseq analysis method (RPKM; Poisson distribution or unique reads; negative binomial distribution) and confirmed with qRT-PCR when comparing between ecotypes.

Two genes, Fibrinogen-related protein (PHUM500950) and Spaetzle (PHUM595260), were confirmed by all 3 analysis methods in both between treatment comparisons resulting in higher transcript levels in both control and treated body lice. The Fibrinogen-related protein, part of the cellular response, is a common bacterial recognition protein conserved in invertebrates and mammals, and its role is to directly tag microbial sugar moieties for phagocytic degradation in blood serum or hemolymph (Faik *et al.*, 2011). A BLAST search of the protein sequence resulted in a 65% identity match with Fibrinogen-related-proteins from *Drosophilla mojavensis*, *Drosophila virilis* and *Drosophila ananassae*, all of which included the important C-terminal domain, which is responsible for microbial sensing (Faik *et al.*, 2011). In *D. melanogaster*, the louse Fibrinogen-related protein has 65% amino acid identity to a scabrous protein, which is a Fibrinogen-related-protein implicated in a variety of processes, which also contains the Fibrinogen C-terminal domain.

The transcript level of Fibrinogen in BLC was greater when compared to HLC (2.00-fold) and this level increased when BLT was compared to HLT (2.77-fold). There was also a *B. quintana* infection effect in body lice, as BLT resulted in 1.34-fold more transcript than detected in BLC. By the 8 day post-infection time point, head lice had killed off or contained the invading *B. quintana* whereas it was still proliferating and spreading in the body lice. Because this protein targets bacteria for endocytosis by phagocytes in the hemolymph, it seems likely that body lice were responding significantly to infection by increasing the transcript level of the Fibrinogen-related-protein, which would increase the cellular response by enhancing phagocytosis.

Spaetzle is an extracellular protein that is the ligand for the Toll receptor. In order to become active, extracellular recognition factors initiate a serine protease cascade(s), which proteolytically cleaves Spaetzle into its active form (Mizguguchi *et al.*, 1998). Once active, it binds to the Toll receptor, which leads to a cascade of intracellular events eventually leading to the transcription of Toll effector genes such as antimicrobial peptides (AMPs).

The transcript levels of Spaetzle were 1.64-fold higher in BLC samples compared to HLC samples. At 8 days post-infection, BLT samples were 2.30-fold higher than HLT samples; supporting the contention that *B. quintana* must proliferate and invade the hemocoel to elicit an immune response, specifically by up-regulation of the Toll pathway. Also, BLT samples were 1.4-fold greater than BLC samples, which once again supports the hypothesis that *B. quintana* might have entered the hemocoel to elicit an immune response.

The Defensins are effector genes within the Toll pathway, which are translated into AMPs that are produced in almost all immune-related and epithelial cells. Defensins secreted by

epithelial cells in the alimentary track are one of the first humoral immune defense barriers against foreign pathogens (Kim *et al.*, 2012).

The connection between Spaetzle and Defensin 1 and 2 is unclear. Interestingly, Defensin 1 (PHUM365700) and 2 (PHUM595870) are the only 2 genes to show up significant in all four pairwise comparisons. Unfortunately, neither gene was confirmed by two analysis methods except for HLC-BLC in which Defensin 1 is 0.67-fold change higher in head lice. Otherwise, Defensin 1 and 2 conflict in their results, where the negative binomial and qRT-PCR analysis methods each suggested that head or body lice had the higher expression level. If Spaetzle is up-regulated in body lice, then Defensin genes would also be up-regulated, but we do not see this in our results. A potential hypothesis for the higher level of Defensin 1 and 2 in treated head lice, based upon qRT-PCR analysis, is that a separate mechanism, independent of the Toll pathway is responsible for the increased Defensin gene expression.

Serpin (PHUM311330), a serine protease inhibitor gene, is another gene that supports the idea of a separate mechanism independent of the Toll pathway, which is controlling the basal level constitutive overexpression of AMPs in uninfected head lice. This upstream regulator of the Toll pathway inhibits the cleavage of Spaetzle into its active form, ultimately down regulating the Toll pathway.

The transcript level of Serpin in HLC was 0.3-fold greater than in BLC samples and supported by both negative binomial and qRT-PCR analysis methods. This process would result in reduced activation of the Toll pathway in HLC versus BLC, although Serpin function in insects is not well studied and the limited research done in *D. melanogaster* has shown Serpins have a wide variety of functions (Reichart *et al.*, 2011). Nevertheless, a reduction in Toll

pathway activity should subsequently lead to a reduction in the production of AMPs. However, head lice maintain 0.67-fold higher Defensin 1 transcript under uninfected conditions, a finding that again supports the hypothesis that head lice control their basal transcription level of AMPs by a mechanism separate from the induction of the Toll pathway.

Cactus (PHUM345810) is an intracellular protein that is downstream of the Toll receptor that acts as a Toll pathway inhibitor. Specifically it is an inhibitor of transcription factor NF κ B, which up-regulates effector genes such as AMPs (Wu and Anderson, 1998). If Cactus is up-regulated, more will bind and inhibit the proteins Dif and Dorsal, which ultimately activate NF κ B and up-regulate AMPs, therefore Cactus up-regulation results in lower AMP transcript levels.

The transcript level of Cactus in BLC samples was 1.38-fold greater when compared to the transcript levels in HLC samples. Since Cactus is further downstream than Spaetzle, it may have a more direct inhibitory effect on the production of AMPs, and could be part of the reason for the 0.67-fold increase in the transcript level of Defensin 1 in HLC versus BLC samples, although it is still likely a different mechanism that may be used for the resulting higher constitutive basal expression of Defensin 1 in head lice. After 7 days post-infection, there was no statistical difference between the transcript levels of Cactus from BLT versus HLT samples, a finding that once again supports the idea that *B. quintana* upon reaching the hemocoel, is causing body lice to down-regulate the immune inhibitory effects of Cactus thus, increasing its immune response by increasing Toll pathway activation.

Scavenger receptor A (ScavA; PHUM066640) is one of many receptors expressed on the outside of macrophages involved in the recognition of microbes for phagocytic degradation (Goh

et al., 2010). Although there is limited research done in insects, human ScavA has an affinity for gram-negative bacteria and targets *E. coli* for phagocytosis (Peiser *et al.*, 2000).

The transcript level of ScavA was 0.66-fold greater in HLC versus BLC and was 0.75-fold greater in HLT versus BLT samples. It has been previously shown that head lice have more rapid and active phagocytes resulting in higher phagocytic activity than do body lice, specifically against its gram-negative endosymbiont *Reisia* during its migration from the stomach mycetome to the filarial mycetome (Perotti *et al.*, 2007). The overexpression of ScavA in head lice may be a possible reason for the heightened phagocyte over-activity in head lice as well as another possible factor in the head louse's ability to curb *B. quintana* proliferation.

Apolipoprotein D (PHUM427700) is a general stress reducing anti-oxidant, and is shown to be age-dependently up-regulated in the human brain, as well as in Alzheimer patients (Kalman *et al.*, 2000; Loerch, 2008). When over-expressed in *D. melanogaster*, it also increases the flies' lifespan, as well as functioning as a lipid antioxidant conferring resistance to oxidative stress (Walker *et al.*, 2006).

The transcript level of Apolipoprotein D was 0.69-fold greater in HLC versus BLC samples. Interestingly, a 2.19-fold increase in its transcript level was seen in BLT versus BLC samples. Because Apolipoprotein D functions as a general stress reducing agent, it is likely up-regulated due to the oxidative stress associated with *B. quintana* invasion into the hemocoel after proliferation in the gut. The head louse under uninfected conditions is thought to keep its immune and phagocytic response up-regulated to rapidly deal with *B. quintana* infection likely using more energy in the process and creating oxidative stress (John Clark, personal communication). The body louse under infected conditions and hemocoel invasion must up-

regulate its immune and phagocytic response causing oxidative stress and subsequent up-regulation of the stress-reducing Apolipoprotein D.

c-Jun NH₂ terminal kinase (JNK basket gene; PHUM128040) was not verified to be significantly different between any treatment using qRT-PCR techniques, however it was differentially expressed between ecotypes in both control and treated lice when analyzed using 2-way Poisson statistical analysis. This gene is integral in the pathway leading to regeneration of gut epithelial tissues (Buchon *et al.*, 2009). When a BLAST search was run comparing protein sequence, there was a 95% match to stress-activated protein kinase JNK-like from *Camponotus floridanus* and *Megachile rotundata*, and 88% match to *D. melanogaster*.

Based upon previous findings by Kim *et al.* (2012) that identified PGRP, Defensin 1 and Defensin 2 as potential genes involved in vector competence, we examined transcription levels of these genes following *B. quintana* infection in body louse midgut. None of these genes appeared to be significant in our evaluation of midgut transcription. The discrepancy in results could be due to the use of different bacteria for infection. Previous results by Kim *et al.* (2012) used *E. coli*-infected blood whereas our experiments used *B. quintana*. This may suggest that the nature of the bacterial species is integral to the louse's response and provides potential for microbial gut community level differences.

While there is still some question as to the exact mechanism for the disease vector competency between head and body lice, our results clearly demonstrate that a few immunoresponse genes are being differentially expressed and these may be potentially involved in the phenotypic difference. To further tease apart this complex interaction, an RNAi knockdown approach might be helpful to elicit gene function. Also, differences in phagocytic

response between head and body lice must be explored. The introduction of green fluorescent protein labeled *B. quintana* into the hemocoel of head and body lice coupled with a phagocytosis activity assay would allow us to compare head and body lice's cellular immune response. Ultimately this present study does provide the next step in deciphering this complex interaction.

Chapter 4: Within plant genetic variation: An assessment of the somatic mutation hypothesis

Abstract

The somatic mutation hypothesis proposes that individual plants accumulate spontaneous mutations and ultimately develop as genetic mosaics (Whitham, 1981; Whitham and Slobodchikoff, 1981; Gill *et al.*, 1995). The modular and layered nature of plant tissues provides the basis for mutations to occur and then be perpetuated through cell division and differentiation, potentially creating genetic heterogeneity within an individual plant. Such variation could subsequently spread through both sexual and asexual means. Although this hypothesis has been addressed to some degree from both theoretical (e.g., genetic models of somatic mutation; Antolin and Strobeck, 1985; Reusch and Boström, 2011) and empirical perspectives (e.g., neutral markers such as microsatellites; Lushai and Loxdale, 2002; Slatkin, 1984) the level of somatic variation within an individual plant has not been assessed from a gene level/whole genome perspective, and thus the role somatic mutation potentially plays as a mechanism for evolutionary change is unknown. Here we assess the magnitude of within plant genetic variation in the model system black cottonwood, *Populus trichocarpa*, and the prospects for evolutionary change in their clonally derived offspring as measured by changes in amino acids. We show that on the average there are 4,840 unique amino acid changes in 2,569 genes within each tissue sampled within an individual parental tree and its clonally derived offspring. In addition, genetic variation within an individual tree shows about half as much genetic variation among tissues as exists between genetically unrelated trees within a population. These results support the long-standing hypothesis that some tree species exist as mosaics of

genetic variability and can evolve (i.e., genetic differentiation between parent and clonally derived offspring).

Whitham and Slobodchikoff (1981) argued that somatic mutations may be an important source of heritable variation in plants through naturally occurring mechanisms of sexual and asexual reproduction. Somatic mutations would be particularly favored in long-lived, clonally propagating plants such as trees and could be of evolutionary importance by providing a way in which plants could keep up with, or potentially out-run, their pests and pathogens in a coevolutionary arms race or serve in fine-tuning an individual to changing environmental conditions. Over the past thirty-two years this idea has been bantered about with no clear resolution as to its ecological or evolutionary importance primarily because it is not known how much variation can be generated and perpetuated within an individual plant at the level of the gene.

Here we assess the magnitude of within-plant genetic variation in the model system black cottonwood, *Populus trichocarpa*, and the prospects for evolutionary change in their clonally derived offspring as measured by changes in amino acids. Tissue samples from eleven individual trees and their clonally derived offspring were collected from Washington State near Mount Rainier. Five tissues per individual tree, three from the parental tree (floral or leaf buds from the top of the tree, stem tissue from the middle of the tree and root material between the parent and offspring) and two tissues from their clonally derived offspring (stem tissue from the middle of the tree and floral or leaf bud tissue from the top of the tree), were resequenced on the Illumina platform.

Since we were interested in the potential for genetic change between parent and clonally derived offspring and the level of variation within an individual, we needed to ensure that trees

thought to be parent and offspring were in fact parent and offspring. We tried to establish absolute relatedness between parent and offspring through a shared root connection, however, verification in the field was difficult and not absolute. To resolve the relatedness issue, we counted the number of SNPs common to 2, 3, 4 and 5 tissues for each of the parent-offspring pairs. As a control we examined 20 replicates of common SNPs for each of 2, 3, 4 and 5 randomly selected unrelated tissues from across the population. As the number of tissues increased, when the tissues were related (i.e. from parent/offspring pair) the number of shared SNPs increased, whereas when the tissues were randomly selected from non-related trees they decreased to a probability of zero of being related once 4 and 5 tissues were included (Figure 4.1). Overall, we uncovered 7 parent-offspring pairs among the 9 trees in which we had pairs for comparison (two were missing offspring, see methods).

Approximately 77 million reads (6.8 billion base pairs) mapped to the black cottonwood genome, resulting in an average coverage of 16.4x (Table 4.1). SNP analysis of the 19 chromosomes revealed approximately 2.8 million SNPs per tissue. On average, 15,687 (range 7,407 to 37,627) unique SNPs were found within exonic regions of each tissue within an individual tree including its clonally derived offspring. Of these, there were 4,408 (range 2,292-8,031) unique amino acid changes in 2,366 genes (range 1,232- 4,323) in each tissue within an individual tree and its clonally derived offspring (Figure 4.2a). In addition, there were 432 (range 212-775) insertions/deletions that resulted in unique amino acid changes in 361 genes (range 177- 655) per tissue within an individual tree and its clonally derived offspring. Thus, on the average, there were 4,840 unique amino acid changes in 2,569 genes (with 158 overlapping for SNP and insertion/deletion data) per tissue within an individual plant, verifying that individual plants are variable at the gene level. Overall, the mutation rate of unique amino acid

changes across tissues within an individual tree and its clonally derived offspring averaged 4.74×10^{-8} (range 2.29×10^{-7} to 3.51×10^{-9}), calculated by dividing the average number of unique amino acid changes per tissue by the total genome size (378,545,895 bps) divided by the number of cell divisions from the ground to the tissue sampled (140.8 cell divisions/m; based on data from *Populus tremuloides*; Ally *et al.*, 2008). Comparatively, the overall rate of mutation from the SNP analysis averaged 2.83×10^{-5} (range 1.08×10^{-4} to 2.45×10^{-6}).

We also assessed whether the mutational variation derived in one tissue can spread to adjacent tissues within an individual tree and its clonally derived offspring. To address this we compared the number of shared mutations within 2 and 3 adjacent tissues. We show that on the average 1,636 (range 955 to 2,597) amino acid changes are shared between 2 adjacent tissues and 1,597 (range 1,374 to 1,901) are shared among 3 adjacent tissues (see Figure 4.2b and c). Thus, there is mutational spread within an individual tree from relatively localized mutations to those that can spread throughout much of the tree, creating a complex mosaic of genetic variability.

Given the level of variation observed we also assessed how genetic variation within an individual compared with the variation observed across unrelated trees within the population, in other words, is the level of variation generated within an individual as great as the variation observed across unrelated trees? To address this issue we compared pairwise differences in amino acid changes among all tissues within an individual tree, parental or offspring, to pairwise comparisons of tissues between a parent and its offspring and also to pairwise comparisons across randomly sampled trees within the population (Figure 4.3). Results showed that within plant genetic variation is approximately 45% and 41%, respectively, of that observed across unrelated trees in the population (Kruskal-Wallis: 418.7, $p < 0.0001$). The variation within an

individual parental tree or offspring was on the average greater than that between parent-offspring comparisons ($42,149 \pm 3,143$ amino acid changes versus $38,495 \pm 2,220$) but not significantly so (Kruskal-Wallis: -1.920 , $p > 0.05$).

The genes affected by somatic mutation were overrepresented by those involved in three categories across all trees and tissues: those involved in 1) apoptosis and the immune response, 2) metabolism and protein binding, and 3) cell and pollen recognition, pollination, pollen-pistil interactions and reproductive cellular processes. These overrepresented genes/gene pathways were ranked from highest (apoptosis/immune response) to lowest (cell and pollen recognition) by their enrichment scores (see Figure 4.4).

Results here support the long-standing hypothesis (Whitham, 1981; Whitham and Slobodchikoff, 1981) that some tree species can exist as mosaics of genetic variability and evolve. High levels of unique amino acid change were observed among tissues within all individual trees. Clonally derived offspring also acquired a unique suite of amino acid changes through time, differentiating them from their parents (over 8,000 unique amino acid changes based on two tissues within the offspring). In addition, these amino acid changes have the potential for being spread through sexual reproduction, given the high levels of amino acid change observed in reproductive bud tissue. There is also no distinction between the germ line and the soma in plants (Derman, 1960; Sutherland and Watkinson, 1986), thus a somatic mutation can easily be incorporated into the germ line. In addition, overrepresentation of amino acid change in genes involved in pollen recognition, pollen-pistil interactions and reproductive cellular processes in floral buds supports the notion that within plant genetic variation can be perpetuated through sexual reproduction. Thus, an individual can evolve as delineated by the

gene level changes shown here in particular, the changes observed from parent to clonally derived offspring.

Of course it will be necessary now to begin the task of tying within plant genetic variation to natural selection and fitness to address the potential adaptive significance of such variation. Whitham and Slobodchikoff (1981) argued that somatic mutations could be of evolutionary importance by providing a way in which long-lived plants could keep up with, or potentially out-run, their rapidly evolving pests and pathogens in a coevolutionary arms race. Our results are consistent with such a notion in that there is an overrepresentation of amino acid change in genes involved in plant defense to pathogen attack including the nucleotide-binding site R genes (Tuskan *et al.*, 2006). In addition, once genetic differences exist there is the potential for natural selection to modify gene frequencies within an individual through the process of differential growth (Sutherland and Watkinson, 1986). We show that there is differential spread of a subset of approximately 1,200-1,300 mutations across two to three tissues within an individual tree; whether these results reflect selective spread remains to be determined.

We also illustrate that approximately half of the genetic diversity within the tree population is generated within the individual plant via somatic mutation. This result potentially has important implications for the field of conservation biology in that it is unlikely that such plants will suffer the effects of inbreeding, typical of small populations, given the high levels of somatic mutation generated at the gene level and ultimately perpetuated through both sexual and asexual reproduction. In addition, results may address the observation that asexual populations can often be as genotypically polymorphic as sexual ones (Ellstrand and Roose, 1987). Such variation could also have unexpected consequences for ecological, evolutionary, and plant

molecular genetic studies that assume the maintenance of genetically identical clonally-derived lineages (e.g., through the propagation of branch cuttings as is frequently done in *Populus*).

Methods

DNA Extraction and Sequencing

Five tissues were collected from eleven mature black cottonwood trees of similar size, age and clonal spread taken from three locations within the Lewis and Yakima counties of Washington State and shipped on dry ice to the University of Illinois. Tissues were taken from the oldest, largest “parental” tree and one clonally derived offspring. Five tissues per individual tree - three from the parental tree (floral or leaf buds from the top of the tree, stem tissue from the middle of the tree and root material between the parent and offspring) and two tissues from their clonally derived offspring (stem tissue from the middle of the tree and floral and leaf bud tissue from the top of the tree) were resequenced on the Illumina platform, for a total of forty-eight tissue samples (seven tissues, one in each of three parents and four from two offspring, were not included in our samples and not resequenced given that they were either not collected (4 samples from two offspring), did not yield sufficient DNA (2 parental samples) or were lost (1 parental sample).

Total DNA from root samples, stem core tissue, and buds were extracted using CTAB extraction protocols or a Plant Tissue Kit from Qiagen. Extracted DNA was sent to the W. M. Keck Center for Comparative and Functional Genomics, on the campus of the University of Illinois, where the DNA was resequenced using the Illumina Genome Analyzer IIx with paired-end reads. The Illumina Genome Analyzer IIx utilizes an 8 lane flowcell and generates 8 to 30 gigabases per lane with read-lengths of 100 bp. Over the course of the study, improvements in sequencing chemistry increased the yield per lane from 8-10 gigabases to 20-30 gigabases per

lane. We sequenced 2 tissues per lane for a total of 24 lanes and an expected coverage of 10-15x across all 48 tissues.

Data Analysis and Bioinformatics

Reads generated from Illumina sequencing were imported and analyzed using CLC Genomic Workbench 4.x. First reads were trimmed to eliminate bases with Phred quality scores less than 20. Next reads were mapped to the *Populus trichocarpa* reference genome downloaded from Phytozome consisting of 19 chromosomes and 2,499 additional unknown contigs. After reads were mapped, a SNP analysis was conducted. Results (approximately 150 million SNPs) from the SNP analysis were exported to a Postgres SQL database for further analysis. Using Ruby and PERL scripts tissues were compared for SNPs unique to and common among tissues and trees. Gene ontology (GO) analysis was conducted with DAVID, on the amino acid changes unique to each individual tissue and the *Arabidopsis thaliana* orthologous identifier.

Figures and Tables

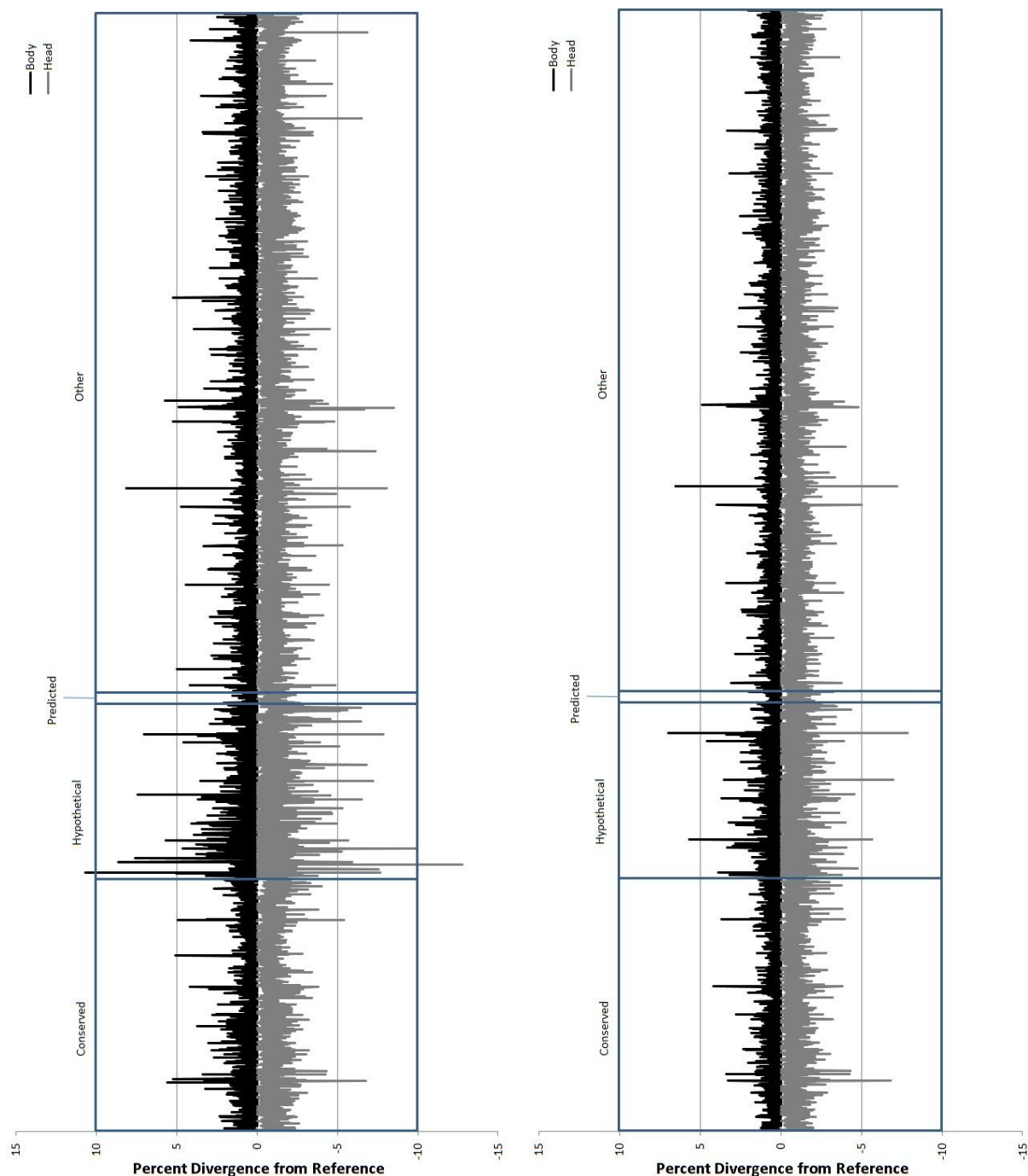


Figure 2.1a. Percent divergence (including gaps) of body and head louse consensus transcripts from the body louse gene model v. 1.2 sequences, grouped by gene category designated by Kirkness et al. 2010: “conserved”, “hypothetical” and “other”. **Figure 2.1b.** Percent divergence (ignoring gaps) of body and head louse consensus transcripts from the body louse gene model v. 1.2 sequences, grouped by gene category: “conserved”, “hypothetical” and “other”.

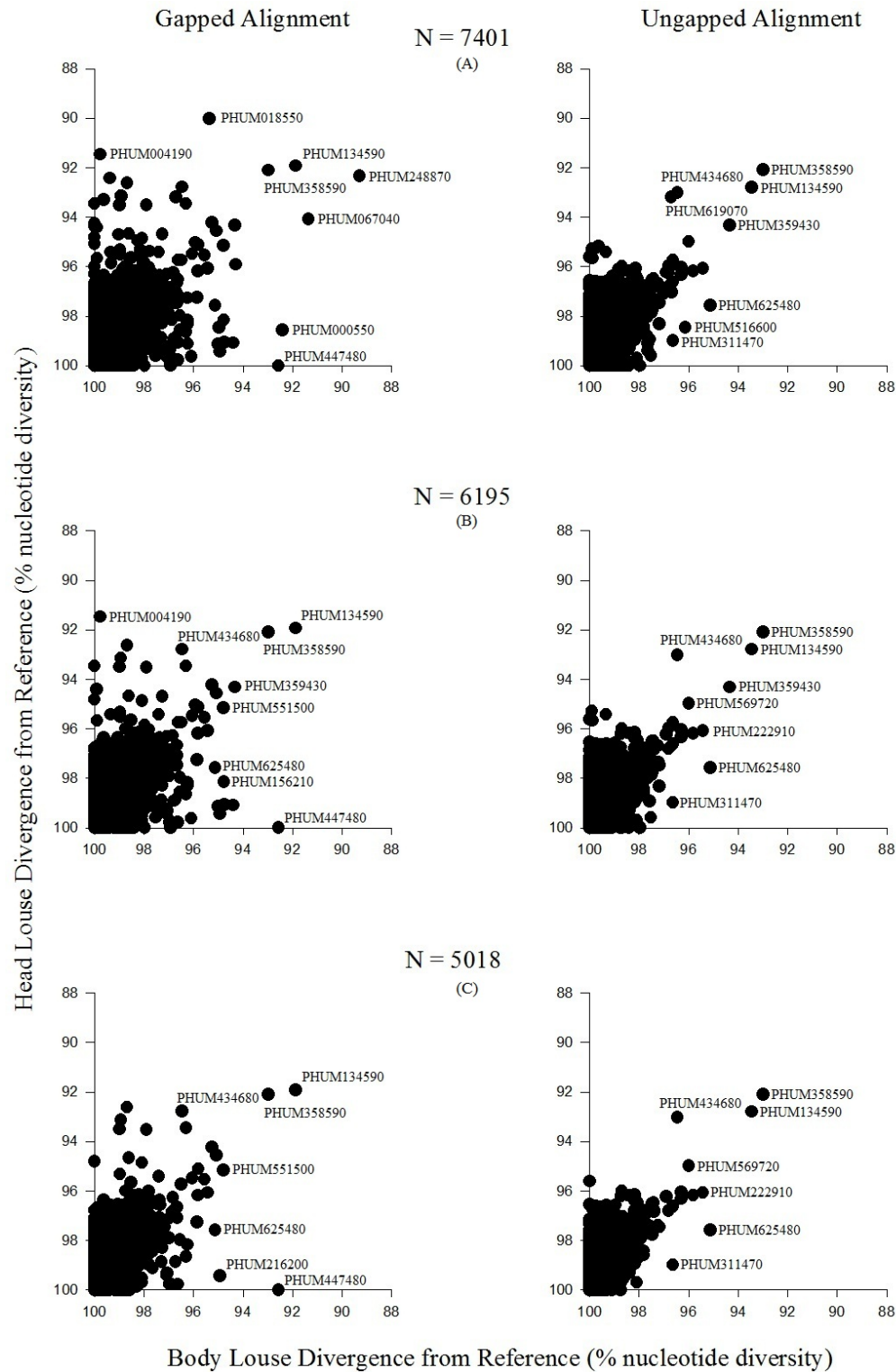


Figure 2.2. Percent identity of derived body and head louse proteins generated from consensus transcript sequences from the published gene models plotted on x-axis and y-axis respectively: (A) Genes with alignment > 70%, (B) Genes with alignment > 85% and (C) Genes with alignment > 95%.

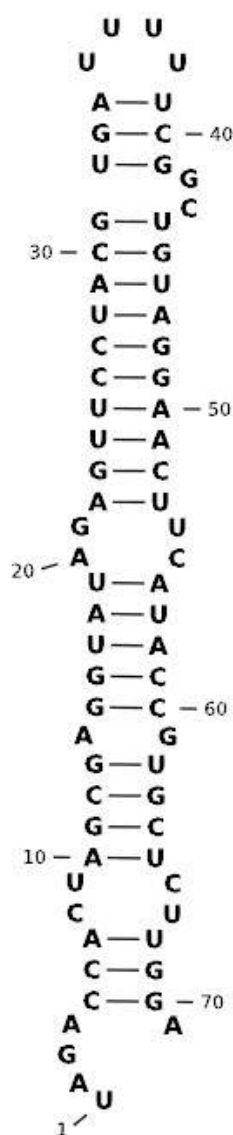
A)

```

Tribolium_mir276      UAGACCACUAGCGAGGUAUAGAGUCCUACGUGAUUUUUCGGCUGUAGGAACUUC AUACC 60
Body_Louse            -----AGCGAGGUAUAGAGUCCUACGUGAUUUUUCGGCUGUAGGAACUUC AUACC 51
                        *****
Tribolium_mir276      GUGCUCUUGGA 71
Body_Louse            GUGCUCUUGGA 62
                        *****

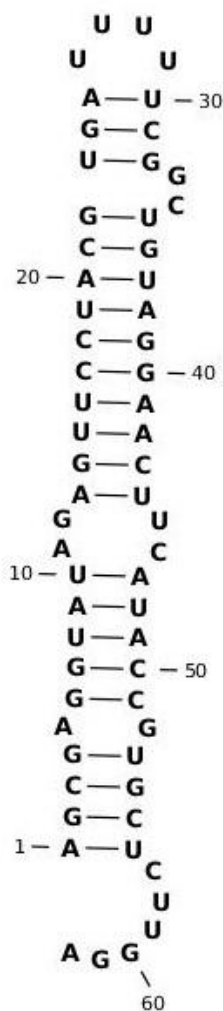
```

B) *T. castaneum* mir276



$\Delta G = -31.9 \text{ kcal/mol}$

C) Louse miRNA



$\Delta G = -26.7 \text{ kcal/mol}$

Figure 2.3. Prediction of a new miRNA from body and head louse ESTs: **(A)** Alignment with the *Tribolium castaneum* mir276 sequence. Putative secondary structures are shown for **(B)** the *T. castaneum* mir276 and **(C)** the body and head louse miRNA

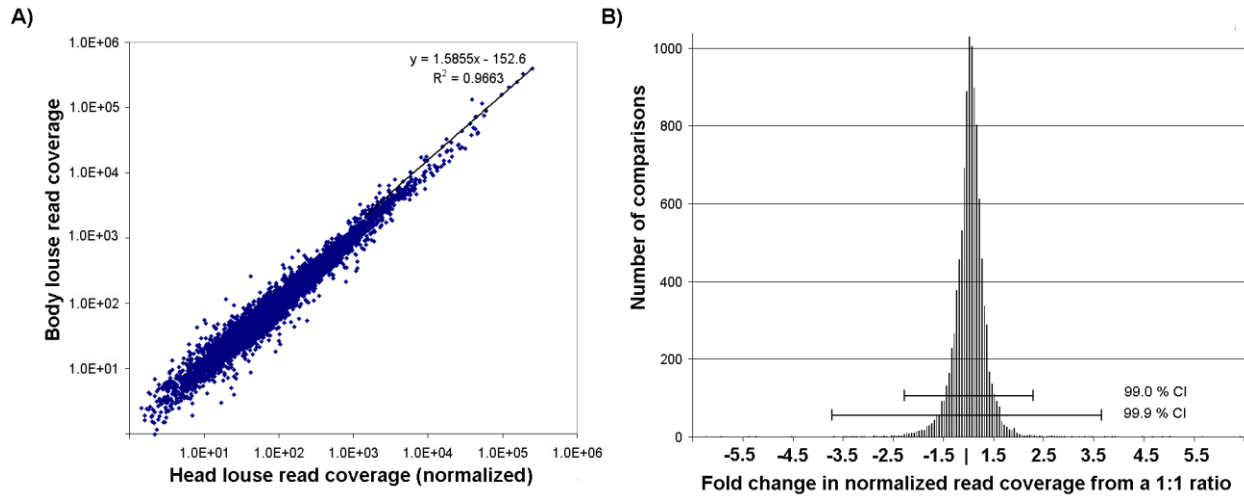


Figure 2.4. Analysis of normalized transcript coverage from head and body louse read data **(A)** Correlation of gene expression between body louse and normalized head louse transcript data based upon the read per million base pair (RPKM) estimates. **(B)** Distribution of fold change in the ratio of RPKM values between body and normalized head louse. This was used to set a significance threshold at the 99.9% confidence interval (CI; $\sim P = 0.001$), which corresponds to the 0.1% of genes showing the greatest divergence in RPKM $3.741 \geq X \leq -3.741$ fold difference between species.

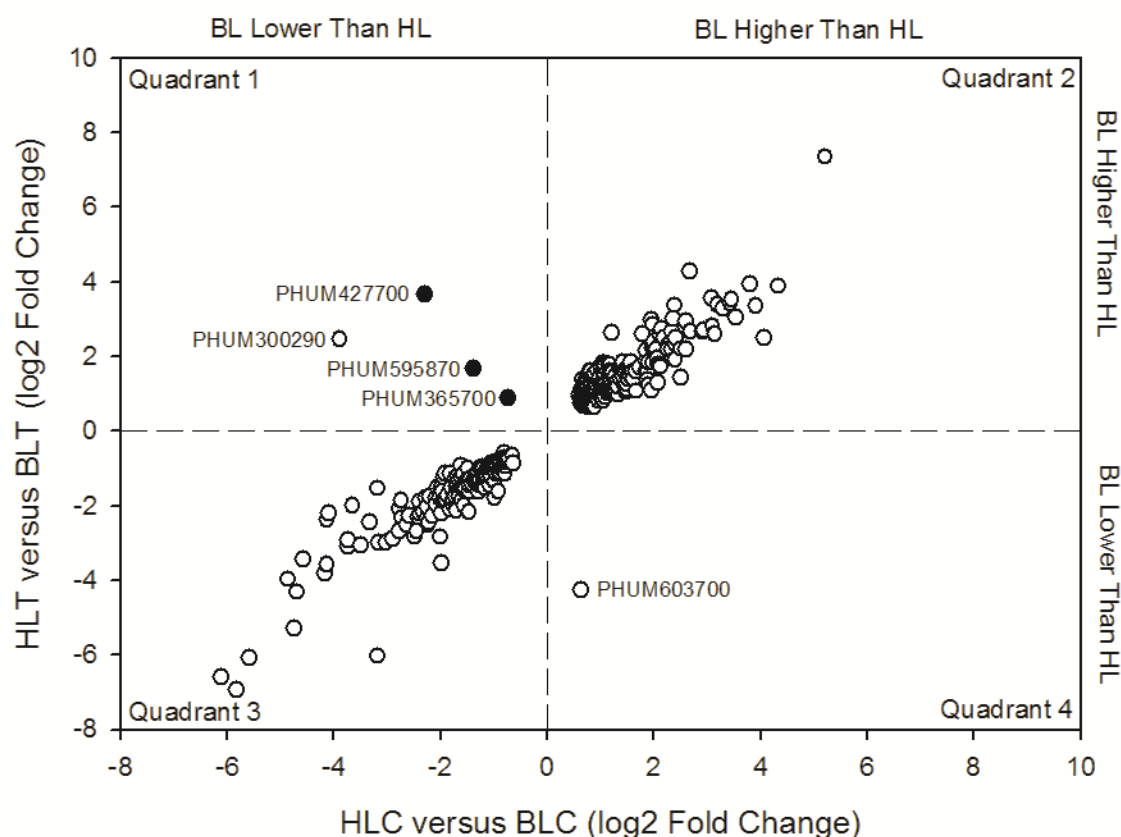


Figure 3.1. Log2 fold change values plotted for the 437 genes common to both HLC-BLC and HLT-BLT pairwise comparisons. Quadrant 1 is genes where control body lice had lower gene expression than control head lice but treated body lice had higher gene expression than treated head lice. Quadrant 2 is genes where control body lice had higher gene expression than control head lice but treated body lice had higher gene expression than treated head lice. Quadrant 3 is genes where control body lice had lower gene expression than control head lice but treated body lice had lower gene expression than treated head lice. Quadrant 4 is genes where control body lice had higher gene expression than control head lice but treated body lice had lower gene expression than treated head lice.

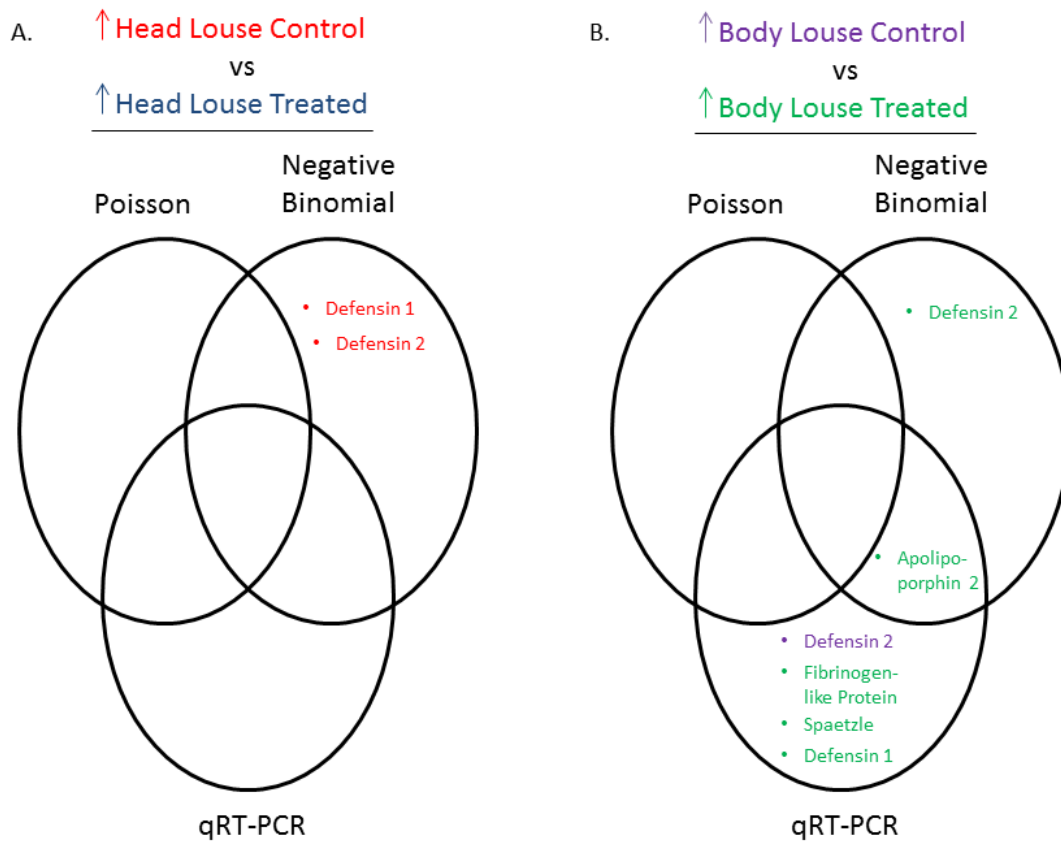


Figure 3.2. Venn diagrams comparing statistically significant immune-related genes from transcriptome analysis when head louse control samples are compared to head louse treated samples (A) and when body louse control samples are compared to body louse treated samples (B). Gene transcription level in whole head and body lice under uninfected conditions and 7 days post *B. quintana*-infection were assessed using two statistical methods, 2-way Poisson and negative binomial distributions, and verified by qRT-PCR. Above are statistically significant immune-related genes ($P < 0.05$).

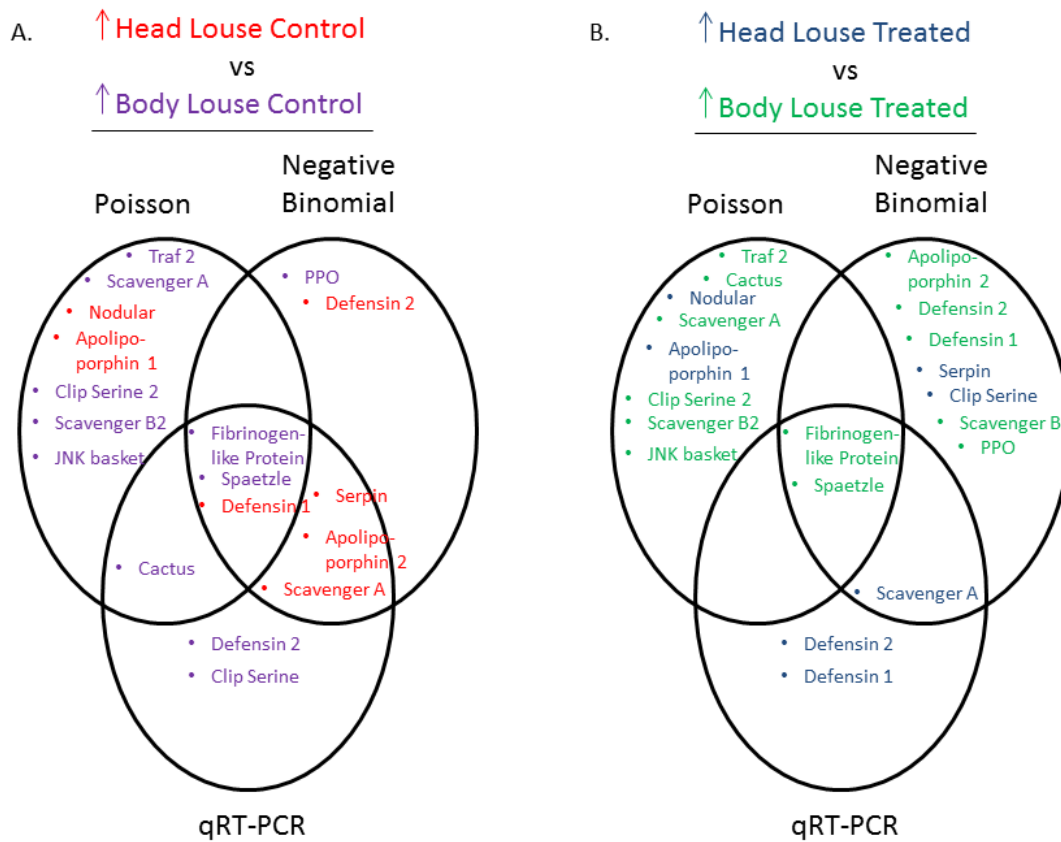


Figure 3.3. Venn diagrams comparing statistically significant immune-related genes from transcriptome analysis when head louse control samples are compared to body louse control samples (A) and when head louse treated samples are compared to body treated samples (B). Gene transcription level in whole head and body lice under uninfected conditions and 7 days post *B. quintana*-infection were assessed using two statistical methods, 2-way Poisson and negative binomial distributions, and verified by qRT-PCR. Above are statistically significant immune-related genes ($P < 0.05$).

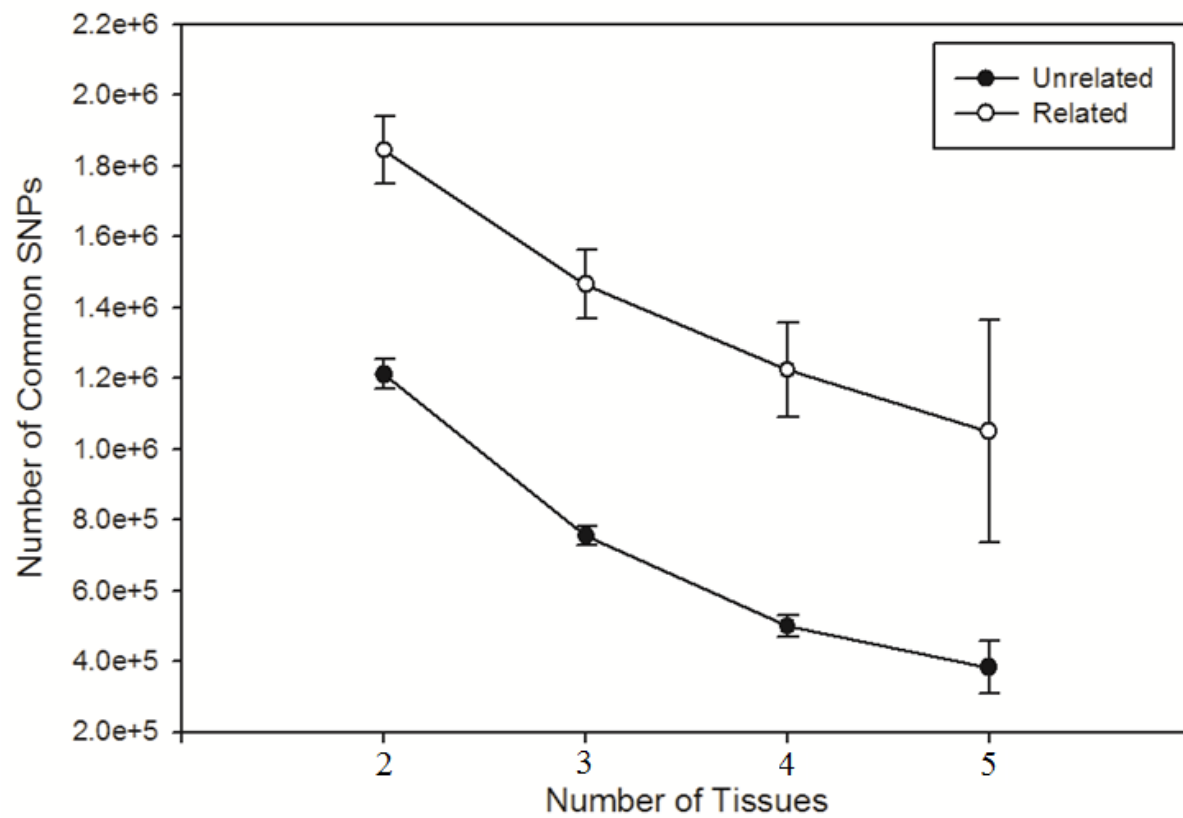


Figure 4.1. Validation of parent-offspring relatedness using the number of SNPs common to 2, 3, 4 and 5 tissues from related tissues as compared to the number of SNPs common to 2, 3, 4 and 5 tissues from unrelated trees.

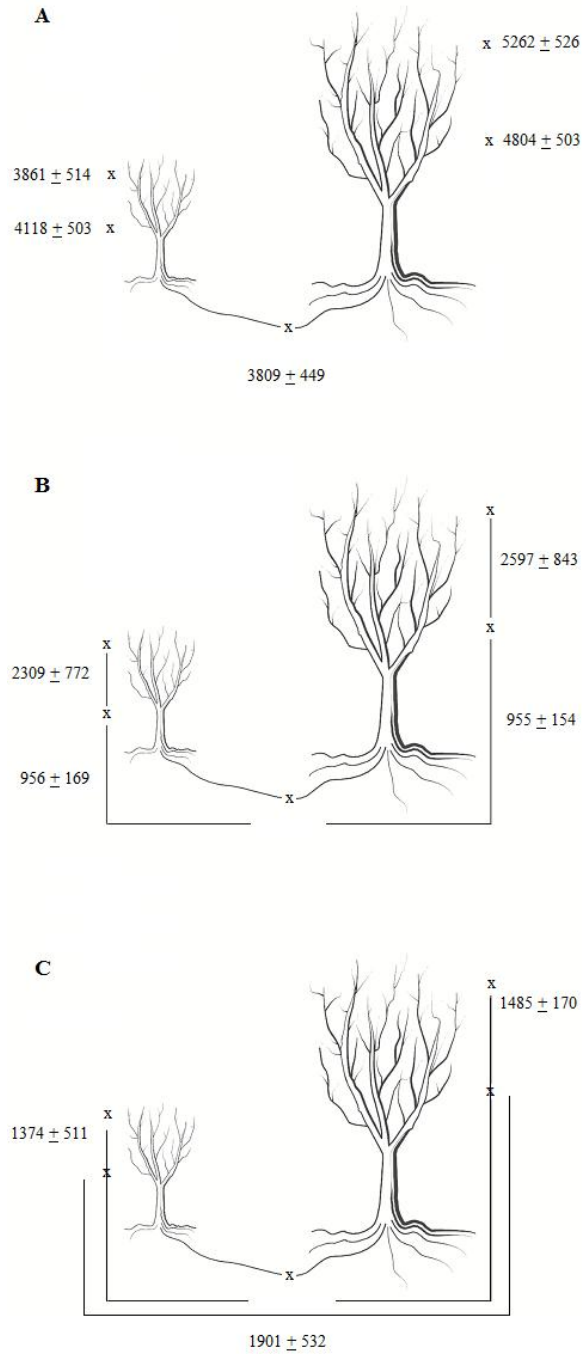


Figure 4.2. (a) Amino acid changes unique to single tissue in the same parent-offspring pair. (b) Amino acid changes common to any two adjacent tissues but unique from other tissues in the same parent-offspring pair. (c) Amino acid changes common to any three adjacent tissues but unique from other tissues in the same parent-offspring pair. Background image was created by Susan Balfe (2013).

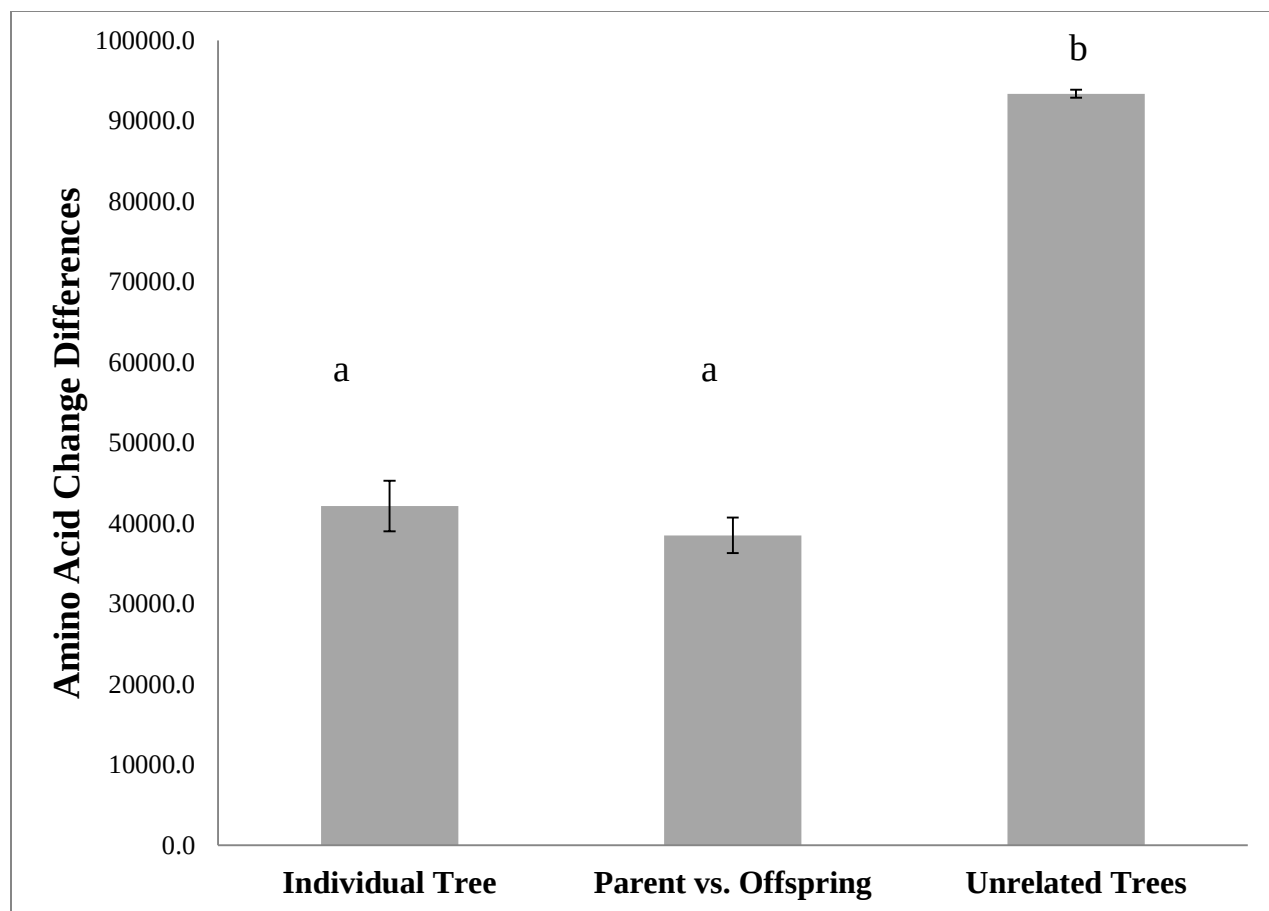


Figure 4.3. Amino acid change differences between two tissues from within an individual tree, between parent and offspring or between completely unrelated trees.

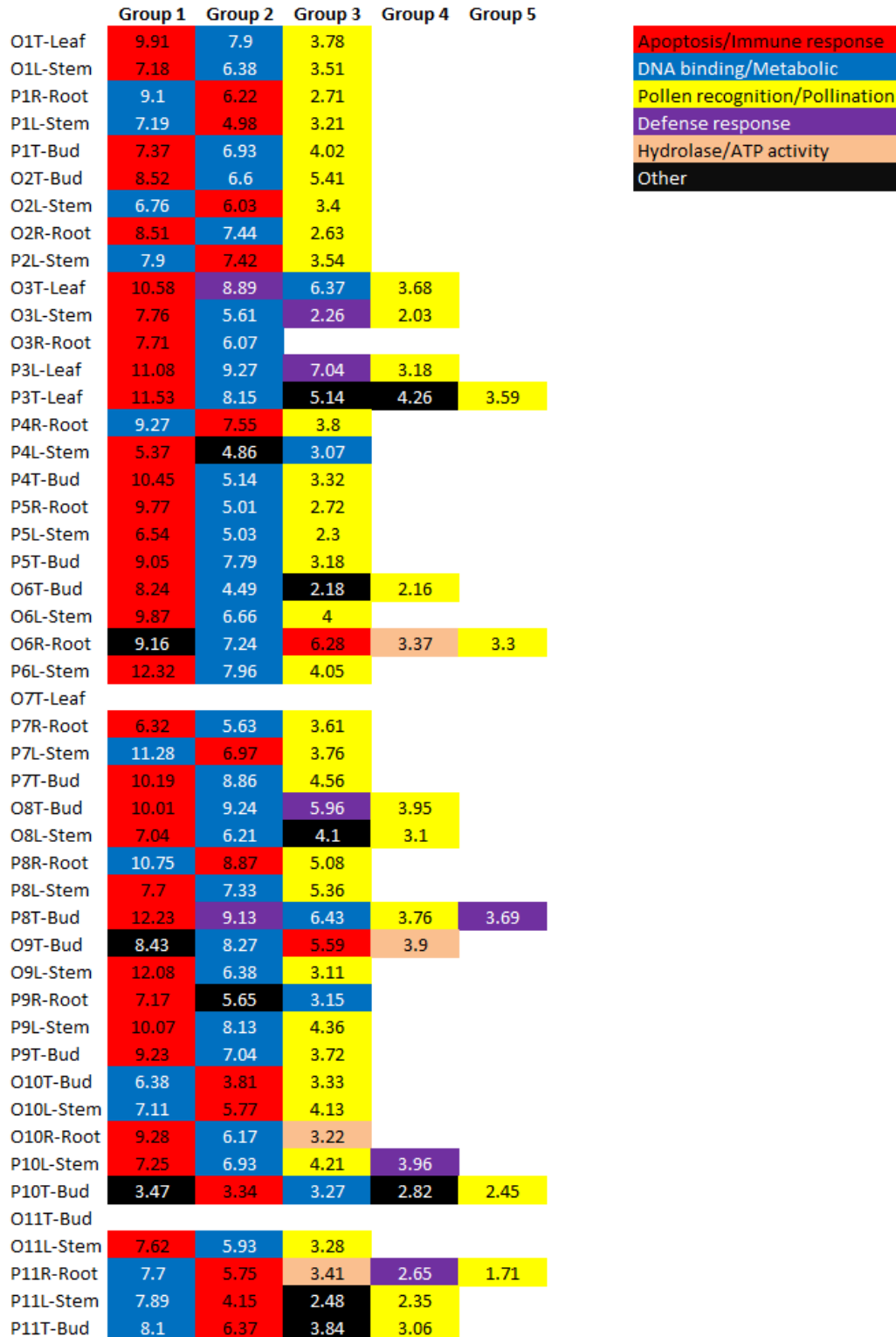


Figure 4.4. Results from gene ontology (GO) analysis of the genes containing a unique amino acid change. Numbers within each color corresponds to the enrichment score obtained for that GO category.

A) Mapping to <i>Pediculus humanus corporis</i>		
	Body Louse	Head Louse
Number of mapped reads	60,413,275	59,347,119
Exp. number of transcripts*	10,775	10,775
Observed transcript number	10,771	10,770
Transcripts > 10x coverage	10,167	10,124
Transcripts <10x coverage	604	646
Transcripts <3x coverage	164	175
Transcripts <1x coverage	52	43

B) Mapping to <i>Candidatus rieszia</i>		
Number of mapped reads	40,054	50,330
Exp. number of transcripts**	544	544
Observed transcript number	538	539
Transcripts > 10x coverage	35	37
Transcripts <10x coverage	503	502
Transcripts <3x coverage	420	399
Transcripts <1x coverage	194	194

* *Pediculus humanus corporis* gene model version 1.2 (Kirkness et al. 2010)

** GenBank accession CP001085; RefSeq NC-014109.1

Table 2.1. Summary of reads mapped to reference body louse genome, expected and observed number of transcripts and coverage of transcripts found in body and head lice and their endosymbiont.

Body (n = 37)	Both (n = 131)
Ammonium: transporter	Adenylosuccinate: lyase
Class: A	Beta-3: adrenergic receptor
Conserved (7)	Beta-galactosidase
Glucose: dehydrogenase precursor	Conserved (14)
Gustatory: receptor PhGr6	Cuticle: protein 18.6, isoform B
Histone: H3	Formate: acetyltransferase
Hypothetical (21)	Glutamate: receptor
Micronuclear: linker histone polyprotein	G-protein: coupled receptor
Pancreas: transcription factor 1 subunit alpha	Growth-arrest-specific: protein
Predicted	GTP-binding: nuclear protein RAN1
Restnal: homeobox protein Rx-A	Hepatocyte: nuclear factor 3-gamma
	Homeobox: protein arx
	Homeobox: protein OTX2
	Homeobox: protein ceh-19
	Hypothetical (89)
	Inhibitory: POU protein
	Kek1
	Ladybird: homeobox corepressor
	Leucine-rich: repeat-containing protein
	Lymnokinin: receptor
	Meiotic: recombination protein SPO11
	Odorant: receptor PhOr3
	Organic: cation transporter
	Orphan: nuclear receptor NR2E1
	Orphan: nuclear receptor NR6A1
	Paired: mesoderm homeobox protein 2B
	Protein: MSP1
	Synaptic: vesicular amine transporter
	Thyrotroph: embryonic factor
	Ultrabithorax
Head (n = 49)	
Class: A (2)	
Conserved (5)	
DNA: methyltransferase	
Dynein: intermediate chain 3, ciliary	
Histone: H4	
Histone: H3	
Homeobox: protein Nkx-3.1	
Homeobox: protein Nkx-2.5	
Hypothetical (25)	
Neurotrimin: precursor	
Odorant: receptor PhOr2	
Odorant: receptor PhOr7	
Proboscipedia	
Protein: krueppel	
Rho/RAC: guanine nucleotide exchange factor	
Senseless	
Sialic: acid synthase	
Transient: receptor potential protein	
Tripsin-4: precursor	
Ubiquitin: ligase protein FANCL	

Table 2.2. Description of gene function/annotation (Kirkness et al. 2010) with less than 3x coverage in either the body or head lice transcript list along with genes that are less than 3x coverage in both lists.

Not found in	Gene ID	Gene Description	Primer	Nucleotide sequence (5' - 3')
Body Louse	PHUM199460	Hypothetical	199460_F	TGGAATCCAAGGTTGTTGAA
			199460_R	ACTTTTGCCAAACCTCATTCG
	PHUM279880	Hypothetical	279880_F	CAGCCGATAAAATTCGGTTT
			279880_R	CGGTGTCGGATATTTTCATCA
	PHUM594680	Hypothetical	594680_F	ACAATCGAGCAAGCAAACCT
			594680_R	CGTGAAATACGAAATTGATTGG
	PHUM623680	Adenylosuccinate	623680_F	TGACTGACTCCACGGTGCT
			623680_R	TTCAGCTCGTCAACCATAGTG
Head Louse	PHUM091050	Hypothetical	091050_F	AAGGAAACGGAAATGCTTTG
			091050_R	TTCTGCTTCCACTTTGTCCA
	PHUM193910	Hypothetical	193910_F	CTACCCTACGGATTGGGATG
			193910_R	CTGCTCCGATTTCAAATGCT
	PHUM540560	Hypothetical	540560_F	GTCATCCCGTTCGTTACGTC
			540560_R	TTCTTTTCCGATCGTTTTCG
	PHUM621980	Hypothetical	621980_F	CAGACCTGATCCTGCTCTCC
			621980_R	AGCATCGCTCTGGAGGATTT
	PHUM624080	Beta-galactosidase	624080_F	TGGAGTGAATACCACGACGA
			624080_R	ACATCCAGAGGCACTTCACC

Table 2.3. Primers used for confirmation of 9 genes based on EST differences observed in body and head lice.

	Gene								
	PHUM091050	PHUM193910	PHUM199460	PHUM279880	PHUM540560	PHUM594680	PHUM621980	PHUM623680	PHUM624080
Sequencing/Genomic Analysis									
Head Louse Mapping	-	-	+	+	-	+	-	+	-
Body Louse Mapping	+	+	-	-	+	-	+	-	+
PCR Analysis									
Head Louse RNA	-	+	+	+	-	+	-	-	-
Body Louse RNA	-	+	+	+	+	+	-	-	-
Positive Control (Body Louse DNA)	-	+	+	+	+	+	-	-	-
Negative Control (water)	-	-	-	-	-	-	-	-	-

Table 2.4. Sequencing and genomic analysis revealed 9 transcripts present in either the head or body louse reads mapped to the body louse reference genome, but not the other. PCR confirmation results of those same 9 transcripts to validate whether each transcript truly exists within each organism. Presence (+) and absence (-) are reported for each gene, RNA samples and positive and negative controls.

Gene	Difference	Annotation
PHUM547150	3.828	Conserved hypothetical protein
PHUM403540	3.853	Hypothetical protein
PHUM531970	4.428	Enzymatic polyprotein, putative
PHUM605200	4.449	Cuticle protein, putative
PHUM098910	4.619	Cytochrome P-450
PHUM409960	4.824	Triadin, putative
PHUM448780	4.951	Glucose dehydrogenase precursor, putative
PHUM153450	5.62	Trypsin-4 precursor, putative
PHUM233900	6.358	Parathyroid hormone receptor, putative
PHUM266290	-6.252	Agglutinin isolectin 2 precursor
PHUM278770	-5.970	Bardet-biedl syndrome 4, bbs4, putative
PHUM227900	-5.402	Predicted protein
PHUM482590	-5.251	Hypothetical protein
PHUM004160	-4.578	Histone H2B.3

Table 2.5. Putative differentially expressed genes identified from normalized head louse read data compared to that of body louse. Differences reported as a fold increase (+) or decrease (-) in read counts.

	Poisson				Negative Binomial			
	HLC-HLT	BLC-BLT	HLC-BLC	HLT-BLT	HLC-HLT	BLC-BLT	HLC-BLC	HLT-BLT
Total Significant Genes	4	3	3272	3402	23	19	546	945
Higher in treated	1	0	-	-	23	10	-	-
Higher in Control	3	3	-	-	0	9	-	-
Higher in BL	-	-	3054	3188	-	-	268	251
Higher in HL	-	-	218	214	-	-	278	694
Unique to comparison	4	3	87	217	21	17	109	517
Higher in treated	1	0	-	-	21	9	-	-
Higher in Control	3	3	-	-	0	8	-	-
Higher in BL	-	-	78	212	-	-	86	72
Higher in HL	-	-	9	5	-	-	23	436

Table 3.1. Statistically significant genes found to be differentially expressed in either the RPKM or unique reads mapped analyses of each pairwise comparison.

Gene Description	Vectorbase ID	Poisson				Negative Binomial				qPCR			
		(RKPM difference)				(Log2 Fold Change)				(2 ^{ΔΔC_t})			
	PHUM	HLC-HLT	BLC-BLT	HLC-BLC	HLT-BLT	HLC-HLT	BLC-BLT	HLC-BLC	HLT-BLT	HLC-HLT	BLC-BLT	HLC-BLC	HLT-BLT
^b Fibrinogen-like Protein	500950	-	-	^a -9.07	-8.21	-	-	1.49	1.26	-	1.34	2	2.77
Scavenger A	066640	-	-	-10.5	-9.73	-	-	-0.77	-0.93	-	-	0.66	0.75
Spaetzle	596260	-	-	-8.25	-9.7	-	-	0.83	1.05	-	1.4	1.64	2.3
Defensin 1	365700	-	-	8.5	-	-0.69	-	-0.74	0.9	-	2.23	0.67	0.19
^c Defensin 2	595870	-	-	-	-	-1.47	1.58	-1.38	1.67	-	0.36	1.94	0.45
Apolipo-porphin 2	427700	-	-	-	-	-	5.04	-2.29	3.66	-	2.19	0.69	-
Serpin	311330	-	-	-	-	-	-	-2.06	-1.51	-	-	0.3	-
Clip Serine	571420	-	-	-	-	-	-	-	-2.25	-	-	2.34	-
^d Cactus	345810	-	-	-10.95	-11.26	-	-	-	-	-	-	1.38	-
Scavenger B	351630	-	-	-	-	-	-	-	0.55	-	-	-	-
PPO	448900	-	-	-	-	-	-	0.68	0.89	-	-	-	-
Traf 2	129280	-	-	-37.36	-36.86	-	-	-	-	-	-	-	-
Nodular	249370	-	-	42.63	43.92	-	-	-	-	-	-	-	-
Apolipo-porphin 1	154960	-	-	25.07	24.19	-	-	-	-	-	-	-	-
Scavenger B2	569610	-	-	-11.98	-12.43	-	-	-	-	-	-	-	-
Clip serine 2	451100	-	-	-8.95	-9.26	-	-	-	-	-	-	-	-
JNK basket	128040	-	-	-20.67	-19.88	-	-	-	-	-	-	-	-

^aValue – Indicates over-expression in highlighted treatment. *All values listed are statistically significant (p<0.05)

^bGene Name- Indicates differential gene expression detected by both statistical methods and verified by qPCR

^cGene Name- Indicates differential gene expression detected by negative binomial statistics and verified by qPCR

^dGene name- Indicates differential gene expression detected by Poisson distribution statistics and verified by qPCR.

Table 3.2. Statistically relevant immune-related genes in head and body lice following infection by *B. quintana* and transcriptome analysis is determined by 2-way Poisson, negative binomial and verified using qRT-PCR techniques.

	<u>Averages</u>	<u>Min</u>	<u>Max</u>
Total Genome (bp)	378,545,895		
Total Transcriptome (bp)	64,868,036		
Total Reads	121,161,142	43,244,301	206,213,748
Total BP	10,517,085,909	3,686,574,180	17,847,402,180
Total Reads Mapped	76,877,917	13,034,666	137,358,394
Total BP Mapped	6,856,248,165	1,189,392,068	11,925,414,749
Average Coverage	16.4	2.9	28.6
Total SNPs	2,765,352	367,835	3,385,414
Heterozygous sites	1,461,015	157,036	1,857,572
Within exonic regions	517,921	82,550	606,219
Results in an amino acid (AA) change	88,047	13,594	106,554
Exonic and Heterozygous site	281,716	32,258	354,920
Exonic, Heterozygous site and AA change	48,311	5,877	61,379
SNPs unique to tissue	188,406	61,918	1,207,069
Within exonic regions	21,483	7,407	216,732
Results in an amino acid (AA) change	5,529	2,292	40,315

Table 4.1. Summary of sequencing coverage and results of single nucleotide polymorphisms (SNP) analysis. Averages and ranges provided from entire 48 tissue sample set.

Gene	Directionality	Primer Sequence
EF1a	F	CAAGAAGCTGTGCCAGGAGA
	R	GCACAATGACCTGAGCAAGG
PHUM066640	F	ACGGATGGCATTGGAACGAGAA
	R	AATGCATGTGACAGGAATGCCA
PHUM128040	F	TTGCTAAATGCGTTCACGCCTCA
	R	GACATACGCTCGTGATCCAAATCCA
PHUM129280	F	AGCATGAAATCCGGTGACAAGAGA
	R	ATCGTGAGGCGTCTGCATTATGT
PHUM154960	F	TCGGTTGTACAAGGTGAACCTCAA
	R	AGCGTGCAATGAATGCAAACCTGT
PHUM249370	F	AGTGCATGTGAAGATATGGTGCCT
	R	TTCGGTTCCTGTTATTGTGACCGT
PHUM311330	F	AGCAAGTTTCTTTCACGCCCGAT
	R	TTCAGTTGCTTTGTTTCGCGTCGT
PHUM345810	F	TGGTTTGGAGCTGACATAAATGCTCG
	R	CCATTGCTACCGATTGGACACTCAT
PHUM351630	F	AAGGGTCAACAAGCGAACTGTTTCC
	R	TCCAAATCACTGATCACAGTTGTGCC
PHUM365700	F	GTTGATTTGAATCCTGTCGCCGGT
	R	CAATGAGCGGCACAAGCAGAAATGA
PHUM427700	F	ATCGACGACGAGAGTCGAACAACAA
	R	GATGAACAACCGGAAATTGCGTGGA
PHUM448900	F	ACCGGCTAGAGAAGGTTTGGTCAAT
	R	AATGGTTCGCGACTCAAACGTA
PHUM451100	F	AAAGGCACAGACCGGAAGTTGTT
	R	AACCAAACCTGATGGATCTGGCA
PHUM500950	F	ACGAAGCATTGCACAGATTGACGA
	R	AAACCCTCGTATTCGGCAACCCATTT
PHUM569610	F	AAGAAGTTGAAACGAGCGGAGGAGT
	R	TGGTCCCGCAGGACAATAACAAT
PHUM571420	F	ATCGCCATCGACGTGCGCATATAAA
	R	TGGGTACGCAACACGTATAAAGCCA
PHUM595870	F	TGGAGGAAGATTCAGGAGAGCAACT
	R	AGCAACGTCCACCTTTGTAACCT
PHUM596260	F	ATGAAATCGTCGGCGTACCTGAAA
	R	TCGAACGTACGTCATCATTTCCCT

Supplemental Table 3.1. Sequence of primers used in qRT-PCR analysis.

Bibliography

- Ally, D., Ritland K., and Otto, S.P. 2008. Can clone size serve as a proxy for clone age? An exploration using microsatellite divergence in *Populus tremuloides*. *Mol Ecol* **17**: 4897-4911.
- Alpatov, V.V. and Nastukova, O.A. 1955. Transformation of the head form of *Pediculus humanus* into the body form under changed conditions of existence. *Bulletin of the Moscow Natural History Research Society* **60**(4): 79-92.
- Anders, S. and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Biology* **11**(10): R106.
- Antolin, M.F. and Strobeck, C. 1985. The population genetics of somatic mutation in plants. *Am Nat* **126**: 52-62.
- Arber, W. and Linn, S. 1969. DNA modification and restriction. *Annu. Rev. Biochem.* **38**: 467–500.
- Behura, S.K. and Whitfield, C.W. 2010. Correlated expression patterns of microRNA genes with age-dependent behavioural changes in honeybee. *Insect Mol Biol* **19**(4): 431-439.
- Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**: 289–300.
- Bonilla, D.L., Kabeya, H., Henn, J., Kramer, V.L. and Kosoy, M.Y. 2009. *Bartonella quintana* in body lice and head lice from homeless persons, San Francisco, California, USA. *Emerg Infect Dis* **15**(6): 912-915.

- Boyce, T.M., Zwick, M.E. and Aquadro, C.F. 1994. Mitochondrial DNA in the bark weevils: Phylogeny and evolution in the *Pissodes strobi* species group (Coleoptera: Curculionidae). *Mol Biol Evol* **11**(2): 183-194.
- Buchon, N., Broderick, N.A., Chakrabarti, S., and Lemaitre, B. 2009. Invasive and indigenous microbiota impact intestinal stem cell activity through multiple pathways in *Drosophila*. *Genes Dev* **23**(19): 2333-2344.
- Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., Fitzgerald, L.M., Clayton, R.A., Gocayne, J.D., *et. al.* 1996. Complete genome sequence of the Methanogenic Archaeon, *Methanococcus jannaschii*. *Science* **273**(5278): 1058-1073.
- Busvine, J.R. 1948. The 'head' and 'body' races of *Pediculus humanus* L. *Parasitology* **39**(1): 1–16.
- Busvine, J.R. 1978. Evidence from double infestations for the specific status of human head and body lice (Anoplura). *Syst Entomol* **3**: 1–1.
- Christophides, G.K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P.T., Collins, F.H., Danielli, A., Dimopoulos, G., Hetru, C., Hoa, N.T., Hoffmann, J.A., Kanzok, S.M., Letunic, I., Levashina, E.A., Loukeris, T.G., Lycett, G., Meister, S., Michel, K., Moita, L.F., Muller, H.M., Osta, M.A., Paskewitz, S.M., Reichhart, J.M., Rzhetsky, A., Troxler, L., Vernick, K.D., Vlachou, D., Volz, J., von Mering, C., Xu, J., Zheng, L., Bork, P. and Kafatos, F.C. 2002. Immunity-related genes and gene families in *Anopheles gambiae*. *Science* **298**: 159e165.

- Derman, H. 1960. Nature of plant sports. *Am Hort Mag* **39**: 123-174.
- Durden, L.A. and Musser, G.G. 1994. The sucking lice (Insecta: Anoplura) of the world: A taxonomic checklist with records of mammalian hosts and geographic distributions. *B Am Mus Nat Hist* **218**: 1-90.
- Ellstrand, N.C. and Roose, M.L. 1987. Patterns of genotypic diversity in clonal plant species. *Am J Bot* **74**: 123-131.
- Evans, J.D., Aronstein, K., Chen, Y.P., Hetru, C., Imler, J.L., Jiang, H., Kanost, M., Thompson, G.J., Zou, Z. and Hultmark, D. 2006. Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Mol Biol* **15**: 645e656.
- Faik, I., Oyedji, S., Idris, Z., Messias-Reason, I., Lell, B., Kremsner, P. and Kun, J. 2011. Ficolin-2 levels and genetic polymorphisms of FCN2 in malaria. *Hum Immunol* **72**: 74-79.
- Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., *et. al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd. *Science* **269**(5223): 496-512.
- GenBank. 2013. <http://www.ncbi.nlm.nih.gov/genbank/>. Accessed 10 January 2013.
- Gill, D.E., Chao, L., Perkins, S.L. and Wolf, J.B. 1995. Genetic mosaicism in plants and clonal animals. *Ann Rev Ecol Syst* **26**: 423-444.
- Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y.,

- Philippsen, P., Tettelin, H. and Oliver, S.G. 1996. Life with 6000 genes. *Science* **274** (5287): 546, 563–567.
- Goh, J., Tan, Y., Dodds, A., Reid, K. and Lu, J. 2010. The class A macrophage scavenger receptor type 1 (SR-A1) recognizes complement iC3b and mediates NF κ B activation. *Prot Cell* **1**(2): 174-187.
- Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res* **32**(Database Issue): D109-D111.
- Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. and Enright, A.J. 2006. mirBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**(Database Issue): D140-144.
- Griffiths-Jones, S., Saini, H.K., van Dongen, S. and Enright, A.J. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**(Database Issue): D154-D158.
- Hori, S., Kaneko, K., Saito, T.H., Takeuchi, H. and Kubo, T. 2010. Expression of two microRNAs, *ame*-mir-276 and -1000, in the adult honeybee (*Apis mellifera*) brain. *Apidologie*: 1-14.
- Hubby, J.L. and Lewontin, R.C. 1966. A molecular approach to the study of genetic heterozygosity in natural populations. I. the number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics* **54**(2): 577-594.
- (IHGSC) International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

- (ISGC) The International Silkworm Genome Consortium. 2008. The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem Molec* **38**(12): 1036-1045.
- Jacob, F. and Monod, J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**(3): 318-356.
- Kalman, J., McConathy, W., Araoz, C., Kasa, P. and Lacko, A. 2000. Apolipoprotein D in the aging brain and Alzheimers dementia. *Neurol Res* **22**: 330-336.
- Khudobin, V.V. 1995 The adaptive potentials of human head and clothes lice when parasitizing on man. *Medskaya Parasitology* **1**: 23-25.
- Kim, J.H., Min, J.S., Kang, J.S., Kwon, D.H., Yoon, K.S., Strycharz, J., Koh, Y.H., Pittendrigh, B.R., Clark, J.M., and Lee, S.H. 2011. Comparison of the humoral and cellular immune responses between body and head lice following bacterial challenge. *Insect Biochem Mol Biol* **41**(5): 332-339.
- Kim, J.H., Yoon, K.S., Previte, D.J., Pittendrigh, B.R., Clark, J.M., and Lee, S.H. 2012. Comparison of immune responses in the alimentary track tissues between body and head lice following *Escherichia coli* oral challenge. *J Asia Pac Entomol* (In Press).
- Kirkness, E.F., Haas, B.J., Sun, W., Braig, H.R., Perotti, M.A., et al. (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc Natl Acad Sci USA* **107**: 12168–12173.
- Kittler, R., Kayser, M. and Stoneking, M. 2003. Molecular evolution of *Pediculus humanus* and the origin of clothing. *Curr Biol* **13**(16): 1414-1417.

- Kittler, R., Kayser, M. and Stoneking, M. 2004. Erratum: Molecular evolution of *Pediculus humanus* and the origin of clothing. *Curr Biol* **14**(24): 2309.
- Kreitman, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase gene region of *Drosophila melanogaster*. *Nature* **304**: 412-417.
- Kozomara, A. and Griffiths-Jones, S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**(Database Issue): D152-D157.
- Lambert, J.D., Chan, X.Y., Spiecker, B. and Sweet, H.C. 2010. Characterizing the embryonic transcriptome of the snail *Ilyanassa*. *Integr Comp Biol* **50**: 768–777.
- Leo, N.P. and Barker, S.C. 2005. Unravelling the evolution of head lice and body lice of humans. *Parasitol Res* **98**: 44-47.
- Leo, N.P., Campbell, N.J.H., Yang, X., Mumcuoglu K. and Barker, S.C. 2002. Evidence from mitochondrial DNA that head and body lice of humans (Phthiraptera: Pediculidae) are conspecific. *J Med Entomol* **39**(4): 663-666.
- Leo, N.P., Hughes, J.M., Yang, X., Poudel, S.K.S., Brogdon W.G. and Barker, S.C. 2005. The head and body lice of humans are genetically distinct (Insecta: Phthiraptera, Pediculidae): Evidence from double infestations. *Heredity* **95**: 34-40.
- Levene, H. and Dobzhansky, T. 1959. Possible genetic difference between the head louse and the body louse. *Am Nat* **93**: 347.
- Li, W., Ortiz, G., Fournier, P., Gimenez, G., Reed, D.L., Pittendrigh B.R. and Raoult, D. 2010. Genotyping of human lice suggests multiple emergences of body lice from local head louse populations. *PLoS Neglect Trop D* **4**(3): e641. Doi:10.1371/journal.pntd.0000641.

- Light, J.E., Toups M.A. and Reed, D.L. 2008. What's in a name: The taxonomic status of human head and body lice. *Mol Phylogenet Evol* **47**(3): 1203-1216.
- Loerch, P. 2008. Evolution of the aging brain transcriptome and synaptic regulation. *PLoS One* **3**:e3329.
- Lushai, G. and Loxdale, H.D. 2002. The biological improbability of a clone. *Genet Res* **79**: 1-9.
- Mayr, E. 1942. *Systematics and the Origin of Species*. Columbia Univ. Press, New York.
- Meselson, M. and Yuan, R. 1968. DNA restriction enzyme from *E. coli*. *Nature* **217**(5134): 1110–1114.
- Mitreva, M., McCarter, J.P., Martin, J., Dante, M., Wylie, T., Chiapelli, B., Pape, D., Clifton, S.W., Nutman, T.B. and Waterston, R.H. 2004. Comparative genomics of gene expression in the parasitic and free-living nematodes *Strongyloides stercoralis* and *Caenorhabditis elegans*. *Genome Res* **14**(2): 209-220.
- Mizuguchi, K., Parker, J., Blundell, T., Gay, N. 1998. Getting knotted: A model for the structure and activation of Spaetzle. *Trends Biochem Sci* **23**: 239-242.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**(7): 621-628.
- Mullen, G. and Durden, L.A. eds., 2009. *Medical and Veterinary Entomology*. Academic Press.
- Munch, K. and Krogh, A. 2006. Automatic generation of gene finders for eukaryotic species. *BMC Bioinformatics* **7**: 263-275.

(NCBI) National Center for Biotechnology Information. 2013. <http://www.ncbi.nlm.nih.gov/>.

Accessed 10 January 2013.

(NIH) National Institute of Health. 2013. <http://www.nih.gov/>. Accessed 10 January 2013.

Nesbitt, M.J., Moerman, D.G. and Chen, N. 2010. Identifying novel genes in *C. elegans* using SAGE tags. *BMC Mol Biol* **11**: 96-105.

Olds B.P., Coates, B.S., Steele, L.D., Sun, W., Agunbiade, T.A., Yoon, K.S., Strycarz, J.P., Lee, S.H., Paige, K.N., Clark, J.M. and Pittendrigh, B.R. 2012. Comparison of the transcriptional profiles of head and body lice. *Insect Molecular Biology* **21**(2): 257-268.

Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J., Rafiq, M.A., Conrad, D.F., Park, H., Hurles, M.E., Lee, C., Venter, J.C., Kirkness, E.F., Levy, S., Feuk, L. and Scherer, S.W. 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* **11**: R52 doi:10.1186/gb-2010-11-5-r52.

Patterson, N., Richter, D.J., Gnerre, G., Lander E.S. and Reich, D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**: 1103-1108.

Peiser, L., Gough, P., Kodama, T. and Gordon, S. 2000. Macrophage class A scavenger receptor-mediated phagocytosis of *Escherichia coli*: role of cell heterogeneity, microbial strain and culture conditions *in vitro*. *Infect Immun* **68**: 1953-1963.

Pennisi, E. 2004. Louse DNA suggests close contact between early humans. *Science* **308**(5694): 210.

Perotti, M., Allen, J., Reed, D., Braig, H. 2007. Host-symbiont interaction of the primary endosymbiont of human head and body lice. *FASEB J* **21**: 1058-1066.

- de Queiroz, K. 2005. Ernst Mayr and the modern concept of species. *Proc Natl Acad Sci USA* **102**: 6600–6607.
- R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Rabinow, P. 1996. *Making PCR: A Story of Biotechnology*. Chicago: University of Chicago Press.
- Raoult, D., Reed, D., Dittmar, K., Kirchman, J.J., Rolain, J.M., Guillen S. and Light, J.E. 2008. Molecular identification of lice from Pre-Columbian mummies. *J Infect Dis* **197**(4): 535-543.
- Reed, D.L., Smith, V.S., Hammond, S.L., Rogers A.R. and Clayton, D.H. 2004. Genetic analysis of lice supports direct contact between modern and archaic humans. *PLoS Biol* **2**(11): e340. doi:10.1371/journal.pbio.0020340.
- Reichhart, J., Gubb, D. and Leclerc, V. 2011. The *Drosophila* Serpins: Multiple functions in immunity and morphogenesis. *Methods in Enzymol* **11**: 205-225.
- Reusch, T.B.H. and Boström, C. 2011. Widespread genetic mosaicism in the marine angiosperm *Zostera marina* is correlated with clonal reproduction. *Evol Ecol* **25**: 899-913.
- Rhee, K.Y. and Johnson Jr., W.D. 2009. *Borrelia* species (relapsing fever). In: G.L. Mandell, J.E. Bennett and R. Dolan, eds. 2009. *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*. Saunders Elsevier. volume 2, section F.

- Robinson, G.G., Hackett, K.J., Purcell-Miramontes, M., Brown, S.J., Evans, J.D., Goldsmith, M.R., Lawson, D., Okamuro, J., Robertson, H.M. and Schneider, D.J. 2011. Creating a buzz about insect genomes. *Science* **331**: 1386.
- Rozen, S. and Skaletsky, H.J. 2000. Primer3 on the WWW for general users and for biologist programmers. In: S. Krawetz and S. Misener, eds. 2000. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press. pp. 365-386.
- Salamov, A.A. and Solovyev, V.V. 2000. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res* **10**: 516-522.
- Sasaki, T., Poudel, S.K.S., Isawa, H., Hayashi, T., Seki, N., Tomita, T., Sawabe, K. and Kobayashi, M. 2006. First molecular evidence of *Bartonella quintana* in *Pediculus humanus capitis* (Phthiraptera: Pediculidae), collected from Nepalese children. *J Med Entomol* **43**: 110–112.
- Schaefer, C.W. 1978. Ecological separation of the human head and body lice (Anoplura: Pediculidae). *T Roy Soc Trop Med H* **72**: 669-670.
- Seki, N., S. Kasai, N. Saito, O. Komagata, M. Mihara, T. Sasaki, T. Tomita, T. Sasaki, and M. Kobayashi. 2007. Quantitative analysis of proliferation and excretion of *Bartonella quintana* in body lice, *Pediculus humanus* L. *Am J Trop Med Hyg* **77**: 562-566.
- Shruti, K., Shrey, K. and Vibha, R. 2011. Micro RNAs: Tiny sequences with enormous potential. *Biochem Bioph Res Co* **407**: 445-449.
- Singh, S.K., Bhadra, M.P., Girschick, H.J. and Bhadra, U. 2008. MicroRNAs – micro in size but macro in function. *FEBS J* **275**: 4929-4944.

- Slatkin, M. 1984. Somatic mutations as an evolutionary force. In: Greenwood, P.J. Harvey, P.H. and Slatkin, M. eds. *Evolution in Honour of John Maynard Smith*. Cambridge University Press pp. 19-29.
- Srinivasan, B.S., Caberoy, N.B., Suen, G., Taylor, R.G., Shah, R., Tengra, F., Goldman, B.S. Garza, A.G. and Welch, R.D. 2005. Functional genome annotation through phylogenomic mapping. *Nat Biotechnol* **23**(6): 691-698.
- Sutherland, W.J. and Watkinson, A.R. 1986. Do plants evolve differently? *Nature* **320**: 305.
- Tanaka, H., Ishibashi, J., Fujita, K., Nakajima, Y., Sagisaka, A., Tomimoto, K., Suzuki, N., Yoshiyama, M., Kaneko, Y., Iwasaki, T., Sunagawa, T., Yamaji, K., Asaoka, A., Mita, K. and Yamakawa, M. 2008. A genome-wide analysis of genes and gene families involved in innate immunity of *Bombyx mori*. *Insect Biochem Mol Biol* **38**: 1087e1110.
- Temin, H.M. and Mizutani, S. 1970. Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus. *Nature* **226** (5252): 1211–1213.
- (TGPC) 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56-65.
- Toloza, A.C., Vassena, C. and Picollo, M.I. 2008. Ovicidal and adulticidal effects of monoterpenoids against permethrin-resistant human head lice, *Pediculus humanus capitis*. *Med Vet Entomol.* **22**: 335-339.
- Tuskan, G.A., DiFazio, S., Jansson, S., Bohlman, J., Grigoriev, I., Hellsten, U., *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. and Gray). *Science* **313**: 1596-1604.

- Walker, D., Muffat, J., Rundel, C. and Benzer, S. 2006. Overexpression of a *Drosophilla* homolog of apolipoprotein D leads to increased stress resistance and extended lifespan. *Curr Biol* **16**: 674-679.
- Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., Gadeau, J., Colbourne, J.K., *et al.* 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* **327**: 343-347.
- Wetterbom, A. 2010. Genome and transcriptome comparisons between human and chimpanzee. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine*. Uppsala: Acta Universitatis Upsaliensis.
- Whitham, T.G. 1981. Individual trees as heterogenous environments: adaptation to herbivory or epigenetic noise? In: Denno, R.F. and Dingle, H. eds. *Insect Life History Patterns: Habitat and Geographic Variation*. New York, Springer-Verlag. pp. 9-27.
- Whitham, T.G. and Slobodchikoff, C.N. 1981. Evolution by individuals, plant-herbivore interactions, and mosaics of genetic variability: the adaptive significance of somatic mutations in plants. *Oecologia* **49**: 287-292.
- Yoon, K. S., Strycharz, J.P., Takano-Lee, M., Edman, J.D. and Clark, J.M. 2006. An improved *in vitro* rearing system for the human head louse allows the determination of resistance to formulated pediculicides. *Pestic Biochem Physiol* **86**: 195-202.
- Yoon, K.S., Gao, J.R., Lee, S.H., Coles, G.C., Meinking, T.L., Takano-Lee, M., Edman, J.D. and Clark, J.M. 2004. Resistance and cross-resistance to insecticides in human head lice from Florida and California. *Pestic Biochem Phys* **80**: 192-201.

Zou, Z., Evans, J.D., Lu, Z., Zhao, P., Williams, M., Sumathipala, N., Hetru, C., Hultmark, D.
and Jiang, H. 2007. Comparative genomic analysis of the *Tribolium* immune system.
Genome Biol **8**: R177.

Appendix A

Methods and Materials

DNA Extraction and Sequencing

Initially, five tissues were collected from eleven mature black cottonwood trees of similar size, age and clonal spread taken from three locations within the Lewis and Yakima counties of Washington State (Figure 1) and shipped on dry ice to the University of Illinois. Tissues were taken from the oldest, largest “parental” tree and one of the clonally derived offspring furthest out from the parental tree (roots can be traced from the parent to the offspring with minimal digging to establish connections). Stem tissue was taken from a randomly derived branch approximately mid-way up the trunk of the parental tree and mid-way up the trunk of the clonally derived offspring (the outer cortex of the L-II layer will be dissected away from the L-I and L-III layers for DNA extraction and sequencing). Floral tissue was taken from the top of the plants in all cases to maximize age related mutational variation. Root tissue was taken from the clonally derived root runner closest to the offspring.

Total DNA from root samples, stem core tissue, and floral buds were extracted using standard CTAB extraction protocols or Plant Tissue Kit from Qiagen. Extracted DNA was sent to the W. M. Keck Center for Comparative and Functional Genomics, on the campus of the University of Illinois, where the DNA was sequenced (resequenced) using the Illumina Genome Analyzer IIx with paired-end reads. The Illumina Genome Analyzer IIx utilizes an 8 lane flowcell and generates 8 to 30 gigabases per lane with read-lengths of 100 bp. Over the course of the study, improvements in sequencing chemistry increased the yield per lane from 8-10 gigabases to 20-30 gigabases per lane. We sequenced 2 tissues per lane for a total of 24 lanes and an expected coverage of 10-15x across all 48 tissues.

Data Analysis and Bioinformatics

Reads generated from Illumina sequencing were imported and analyzed using CLC Genomic Workbench 4.x. The *Populus trichocarpa* reference genome was downloaded from Phytozome.net consisting of 19 chromosomes and 2,499 additional unknown contigs. Raw reads were trimmed to eliminate bases with Phred quality scores less than 20. Trimmed reads were mapped to the 19 chromosomes with the following parameters; similarity of 0.8, length fraction of 0.5, insertion and deletion costs of 3, mismatch cost of 2, min/max distance of 180-250. If any read mapped to multiple places in the genome, it were ignored. After reads were mapped, a SNP (single nucleotide polymorphisms) analysis was conducted with the following parameters; a windows length of 11, minimum central quality of 20, minimum average quality of 15, minimum coverage of 6 and minimum variant frequency of 35%. Similarly, a DIP (deletions and insertion polymorphisms) analysis was conducted with the same parameters. Results (approximately 150 million) from the SNP and DIP analyses were exported to Postgres SQL database for further analysis. Using Ruby and PERL scripts, tissues were compared for polymorphisms unique to and common among tissues and trees. Statistics on tissue comparisons were analyzed in Systat 13. Gene ontology (GO) analysis was conducted with the web software, DAVID (<http://david.abcc.ncifcrf.gov/>), on the amino acid changes unique to each individual tissue and the *Arabidopsis thaliana* orthologous identifier.