SHORT TERM EVOLUTION IN THE IMMUNE RESPONSE OF *DROSOPHILA MELANOGASTER*: INSIGHTS FROM STUDIES OF POPULATION GENETICS AND THE EPIDEMIOLOGY OF BACTERIAL INFECTION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Punita Juneja

January 2011

SHORT TERM EVOLUTION IN THE IMMUNE RESPONSE OF *DROSOPHILA MELANOGASTER*: INSIGHTS FROM STUDIES OF POPULATION GENETICS AND THE EPIDEMIOLOGY OF BACTERIAL INFECTION

Punita Juneja, Ph.D.

Cornell University 2011

Studies of natural populations reveal that tremendous phenotypic variation in immune function exists within species. Selection on extant variation drives the short term evolution of the immune response, potentially resulting in the temporary maintenance of genetic variation in populations or in the fluctuation of allele frequencies. Immune response genes also frequently show evidence of elevated rates of adaptive evolution between species. I used two approaches to study how genetic variation within a population is related to long term evolutionary patterns. From an in-depth study of the pathogen recognition molecule Eater, I find evidence for a recent partial selective sweep in a single population of *Drosophila melanogaster*. The putatively selected allele has a significantly higher level of gene expression, suggesting that gene regulation rather than protein structure is the target of selection. In a broader study of over 200 immune genes using target enrichment and high-throughput sequencing, I find that genes with the highest rates of adaptive evolution between species have low levels of variation within a population. This suggests that selective sweeps, which reduce variation, occur in rapidly evolving genes. Genes that recognize infection and transduce signal within the immune response have low levels of variation consistent with selective sweeps, supporting the idea that these two aspects of the immune system are subject to elevated pathogen pressures.

Our ability to understand the selective pressures that shape the antibacterial immune response is limited by our lack of knowledge about the epidemiology of disease in natural populations. I have performed a survey of natural bacterial pathogens in wild populations of *D. melanogaster* in Ithaca, New York, with the aim of understanding the rates, distributions, and identities of bacterial infections in the wild. I find that 0.3% to 2% of wild flies are infected with a diverse array of opportunistic pathogens. The identification and subsequent characterization of natural pathogens will lead to a better understanding of the selective pressures that drive the evolution of the insect immune response. A complete understanding of the evolution of resistance to infection requires consideration of the short term evolutionary dynamics measured through population genetics and phenotypic study of individuals and their pathogens within populations.

BIOGRAPHICAL SKETCH

Punita Juneja graduated from University of California, Berkeley, in 2003 with Bachelor of Arts in Molecular and Cell Biology. As an undergraduate, she performed research in the lab of Dr. Laurence Tecott at the University of California, San Francisco, under the guidance of Dr. Evan Goulding. Her research goal was to help develop and implement methods for quantitating mouse feeding, drinking, and locomotory behavior. It was there that she gained an appreciation for the importance of computing and bioinformatics in biological research.

As an undergraduate, Punita studied abroad in Costa Rica as part of the the University of California Education Abroad Program under the guidance of Dr. Frank Joyce. There she developed a fascination with entomology and also embarked on a decade of vegetarianism. After returning to Berkeley, she took a research position with her entomology professor, Dr. Alexander Purcell, and later with his former graduate student, Dr. Rodrigo Almeida. In both labs, she studied insects that vector plant pathogens.

At Cornell University, Punita did a rotation in the lab of Dr. Ann Hajek, where she helped develop a method to quantitate entomopathogenic fungal resting spores in soil. She joined the lab of Dr. Brian Lazzaro in 2005, where she studied the evolutionary genetics of the immune response of *Drosophila melanogaster*. In the course of her graduate career, she greatly improved her skills in statistics, population genetics, bioinformatics, microbiology, and molecular biology. In her spare time, she enjoyed the many things Ithaca has to offer, including but not limited to triathlons, cross-country skiing, indoor rock climbing, beer brewing, dance parties, drinking coffee at Gimme, and swimming at Six Mile Creek.

To my family for their unconditional support and love.

ACKNOWLEDGEMENTS

I would like to thank my major advisor, Dr. Brian Lazzaro, for the support he has given me during my time in his lab. He has taught me how to be research scientist by allowing me independence when it came to choosing the direction of my projects. At the same time, he has taught me not to stray too far from my goals and has ensured that my years here were productive. I also appreciate his financial support, which allowed me to spend the majority of my time doing research. I would also like to thank my committee members, Drs. Andy Clark and Ann Hajek, for being willing to meet with me on a regular basis, for their invaluable feedback on manuscripts, and for the advice they have offered me throughout my graduate career.

My fellow labmates, especially Madeline Galac, Sarah Short, and Jacob Crawford, have made coming to work a pleasure. They have taught me a great deal about scientific discourse and have been sounding boards for many questions of experimental design and analysis. I will especially miss them on our official lab holiday, October 31. I would also like to thank Miguel Rosado for his assistance and enthusiasm in the lab. Members of the Entomology department and the Cornell community have greatly enhanced my education, especially Drs. Jeff Scott, Laura Harrington, Carlos Bustamante and Chip Aquadro. Financial support from the Entomology department via the Rawlins Fund and from the graduate school has also greatly assisted my research.

I have made many friendships in Ithaca that will transcend our few short years in graduate school. My friendships with "The Brewing Conspiracy" (Ben and Gretchen Heavner, Annie Rowe, Mike Booth, Heather Fullerton, Ben Logsdon, Arend van der Zande, Cresten Mansfeldt, Cloelle Giddings, and Maddy the dog) and the "Girls Night Gang" (Melissa Hardstone, Lauren Cator, Made-

line Galac, Sarah Short, Sarah Jandricic, Susan Villarreal, Annie Rowe and Michelle Helinski) have been steady constants in my tenure here. A special thanks to Michelle Helinski who faithfully dragged me out of the lab to go running, rain or shine, light or dark, snow or not. Lastly, Annie Rowe, Asti Bhatt, Shamoni Maheshwari, Cresten Mansfeldt, Madeline Galac, and Rick Pampuro have provided me with endless hours of support and entertainment, and without them, I would not have survived graduate school.

Finally, I am thankful for my family, including my grandparents, aunts, uncles, and cousins, who instilled in me the value of hard work and education. My siblings, Poonam and Vikram Juneja, are two of the kindest and hardest working individuals I know, which is a tribute to our parents, Indu and Vinod Juneja. My parents have made me who I am today and for that I thank them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

POPULATION GENETICS OF INSECT IMMUNE RESPONSES*

1.1 Introduction

The immune system that we can observe and measure today is but a snapshot of a dynamic and evolving process, a moment in an ongoing genetic battle between hosts and their pathogens. Indications of this conflict are etched in the genome as signatures of adaptive evolution in the host immune system. These evolutionary signatures can also be read experimentally to give insight into the nature of host-pathogen interactions. This chapter will examine the evolutionary genetics of insect immune systems over both short and long timescales. In several instances, comparisons and contrasts will be drawn between species with distinct ecologies in order to elucidate commonalities and idiosyncrasies of insect immune evolution.

Adaptive evolution can manifest in evolutionarily favored amino acid substitutions within genes as well as in genomic diversification of gene families. Both processes can be measured by comparing homologous genes and gene families across related species. Adaptive amino acid evolution is generally detected as a significantly elevated rate of amino acid substitution relative to an expectation based on the evolutionary rate at genetically silent positions (McDonald and Kreitman, 1991; Yang et al., 2000; Anisimova and Liberles, 2007) (Figure 1.1). Adaptive gene family expansion can be inferred from an increased rate of duplication relative to that of other gene families in the genome (Hahn

* Presented with minor modifications from the originally published book chapter "Juneja, P., Lazzaro, B. P., 2009. Population genetics of insect immune responses. In: *Insect Infection and Immunity: Evolution, Ecology and Mechanisms*, edited by Stuart Reynolds and Jens Rolff."

Evidence for natural selection can be revealed through examination of rates of DNA sequence evolution (reviewed in Anisimova and Liberles (2007)). The null model for these studies is that genes evolve by neutral evolutionary processes. Neutral, selectively equivalent mutations arise by chance and, in the absence of natural selection, occasionally become fixed by random genetic drift. The rate with which this happens is the neutral substitution rate. Many mutations that arise within functional genes cause deleterious changes to protein structure or function. These mutations are constrained from rising to high frequency by negative, or purifying, selection and are assumed to rarely fix between species. In contrast, mutations that are advantageous, such as those that confer resistance to disease, may rapidly rise in frequency by positive, or directional, selection. Positive selection leads to a short term reduction in genetic diversity as the favored allele replaces existing variation in a population. A sufficiently high number of recurrent adaptive fixations may also increase long term divergence between species. Alternatively, multiple polymorphisms can be maintained in populations by balancing selection, which increases genetic diversity. In very rare cases, balanced polymorphism can occur when there is a heterozygote advantage, or overdominance, where heterozygote combination of two alleles has a higher fitness than homozygotes of either allele. More frequently, temporal or spatial variation in selection can maintain multiple alleles if each variant is advantageous in a different time or place.

Adaptive evolution can be detected by comparing DNA sequence of homologous genes from closely related species. This is generally achieved by comparing the rate of nonsynonymous, amino acid replacing, substitutions (d_N) to the rate of synonymous substitutions (d_S), which do not affect amino acid sequence. Synonymous substitutions are assumed to be invisible to selection and thus reflect neutral evolution. If all nonsynonymous mutations were also selectively neutral, d_N would equal d_S, and the ratio d_N/d_S would equal one. Positive selection on amino acid substitutions would result in an increase in the rate of nonsynonymous substitutions, or d_N greater than d_N. The ubiquity of purifying selection, however, means that the empirically observed rate of nonsynonymous substitutions over whole genes is much smaller than the rate of synonymous substitution, and d_N/d_S is almost always much less than one across entire genes. A more sophisticated implementation of this test, Phylogenetic Analysis by Maximum Likelihood (Yang et al., 2000), uses gene sequences from multiple species to test the hypothesis that d_N/d_S varies among codons in a gene, allowing localization of the target of selection to particular residues or gene regions. Another test for natural selection, the McDonald-Kreitman test (McDonald and Kreitman, 1991), uses information about polymorphism species and divergence between species. It tests the null hypothesis that the ratio of nonsynonymous and synonymous substitutions segregating within species is the same as the corresponding ratio between species. In this test, positive selection is detected as a proportional excess of nonsynonymous fixed differences between species. Selection favoring allelic diversification within species, in contrast, would lead to an excess of nonsynonymous polymorphisms. These tests, among others, allow inference of natural selection acting on specific genes and gene regions.

Figure 1.1: Detecting adaptive evolution in the genome

et al., 2005). The recent availability of whole-genome sequences from several insect species allows such comparisons to be made on a wide scale.

Innate immunity, which is shared by homology between vertebrates and insects, is hardwired within the genome and lacks the antibody production that characterizes the adaptive immune response of higher vertebrates. The insect innate immune system is capable of recognition and subsequent eradication of microbes and multi-cellular parasites through humoral and cellular defense mechanisms (reviewed in Lemaitre and Hoffmann (2007)). Humoral immunity is mediated by production of microbicidal peptides, enzymes, oxidative free radicals, and other compounds that are secreted directly into the insect hemolymph (blood). The humoral defense against microbial infection is genetically well understood in *D. melanogaster*. Invading microbes are detected by recognition molecules performing surveillance, signal is transduced through two primary signaling pathways, and defense is effected in part by abundantly produced antimicrobial peptides (AMPs). The two signaling pathways, termed the Toll and Imd pathways, are conserved between invertebrates and vertebrates. Cellular immunity is defined by encapsulation or engulfment of infective agents by circulating hemocytes. It has been less well characterized at the genetic level, although some genes that mediate cellular recognition and trigger phagocytic engulfment of microbes have been identified. A distinct process, RNA silencing (RNAi), allows specific detection and eradication of RNA viruses (Wang et al., 2006a). It is expected that functional diversity within the immune response will translate into variation in the selective pressures on different components of the defense response. This chapter will examine the evolutionary genetics of immune defense, interpreting molecular evolutionary patterns in light of protein function to draw insight into how the immune response

adapts to pathogen pressures.

1.2 Evolutionary Patterns in the Antimicrobial Immune Response

Immune genes tend to evolve more quickly and adaptively than non-immune genes in both vertebrates and insects (Murphy, 1993; Schlenke and Begun, 2003; Nielsen et al., 2005; Waterhouse et al., 2007; Sackton et al., 2007). This adaptive evolution is evidenced both by elevated rates of amino acid substitution between species and by elevated rates of duplication within gene families. The availability of whole genome sequences allows for quantitative contrasts to be made between immune and non-immune genes, as well as for comparisons between functional classes of immune response genes. The recent complete genome sequencing of twelve species of fruit flies in the genus *Drosophila* has allowed particularly fine measurement of rates of substitution and genomic rearrangements between closely related species. More distant comparative genomic analyses can be achieved by comparing genome sequences of *Drosophila*, the mosquitoes *Anopheles gambiae* and *Aedes aegypti*, the honey bee *Apis mellifera* and the red flour beetle *Tribolium castaneum* (Figure 1.2).

Genome comparisons between species reveal the distinct selective pressures acting on each species through its unique life history. For example, the honey bee *Apis mellifera* has apparently reduced copy number in immune-related gene families, perhaps reflecting decreased emphasis on immunological defense due to hygienic behavior in the hive (Evans et al., 2006). Mosquitoes have expansions in gene families thought to play defensive roles against pathogens borne

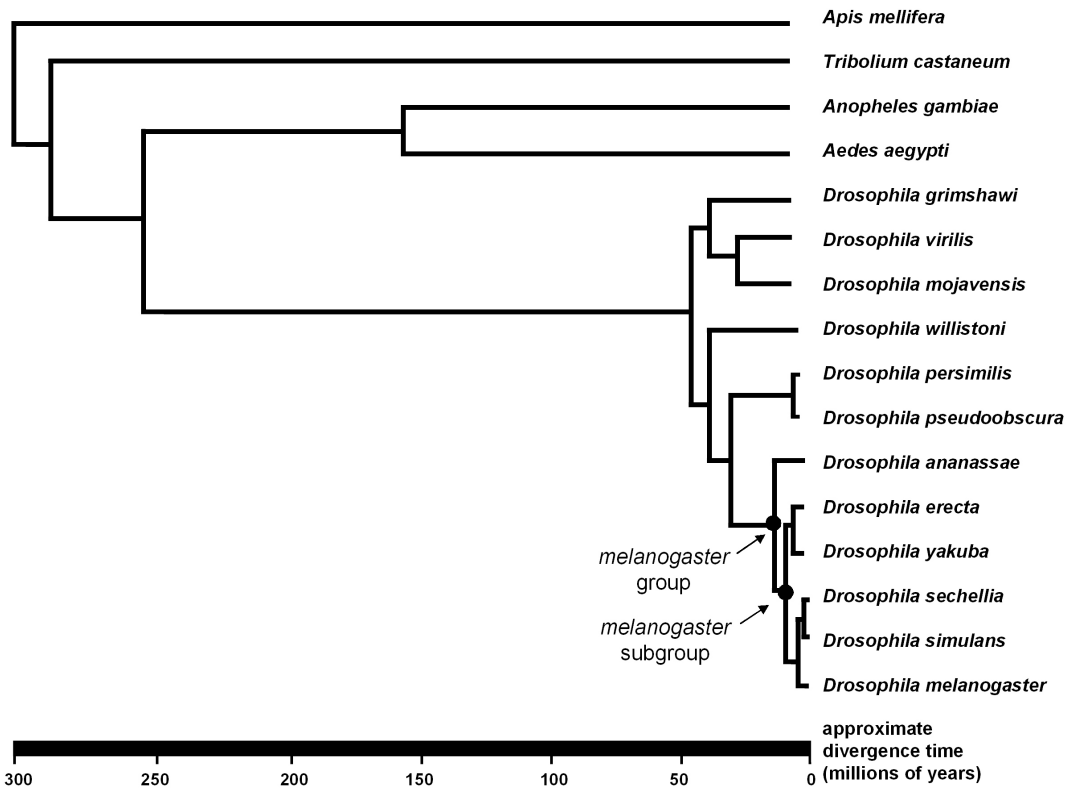


Figure 1.2: Phylogeny of select insect species with sequenced genomes. The *melanogaster* species group and *melanogaster* species subgroup are indicated. Gene family expansions and contractions were evaluated among *Drosophila* (fruit flies), *Anopheles gambiae* (African malaria mosquito), *Aedes aegypti* (yellow fever mosquito), *Apis mellifera* (honey bee) and *Tribolium castaneum* (red flour beetle) and within the genus *Drosophila*. Adaptive amino acid evolution measurement, which requires shorter phylogenetic distances, was primarily performed in the *melanogaster* species group within *Drosophila* (Tamura et al., 2004; Savard et al., 2006; Clark et al., 2007).

in vertebrate blood (Christophides et al., 2002; Waterhouse et al., 2007). Interpretation of these comparisons is often limited, however, because identification of most immune genes in insects stems from functional characterization in only a few species, and primarily in *D. melanogaster*. Novel defense mechanisms in functionally uncharacterized organisms will not be detected through homology searching of genome sequences if they are too divergent to be detected by similarity at the DNA sequence level. Additionally, extremely rapidly evolving genes may diverge too quickly to be identified in comparisons between distantly related species. Genomic comparisons will gain power with increasing functional characterization of non-model systems and the accumulation of whole-genome sequences for phylogenetically dispersed organisms.

Comparative genomic and molecular evolutionary analyses have revealed that not all genes in the immune system evolve along the same trajectories. Genes in broadly defined functional categories differ in evolutionary mode, suggesting contrasting selective pressures based on gene function. The supporting data and potential selective pressures that drive these evolutionary patterns will be considered in detail.

1.2.1 Toll and Imd Signaling Pathways

Nearly all core signaling proteins in the Imd and Toll pathways are maintained as strict orthologs among *Drosophila* species (Sackton et al., 2007) and between *Drosophila* and mosquitoes (Christophides et al., 2002; Waterhouse et al., 2007), honey bees (Evans et al., 2006), and *Tribolium* (Zou et al., 2007). Despite this maintenance of orthology, however, these signaling genes show unexpectedly

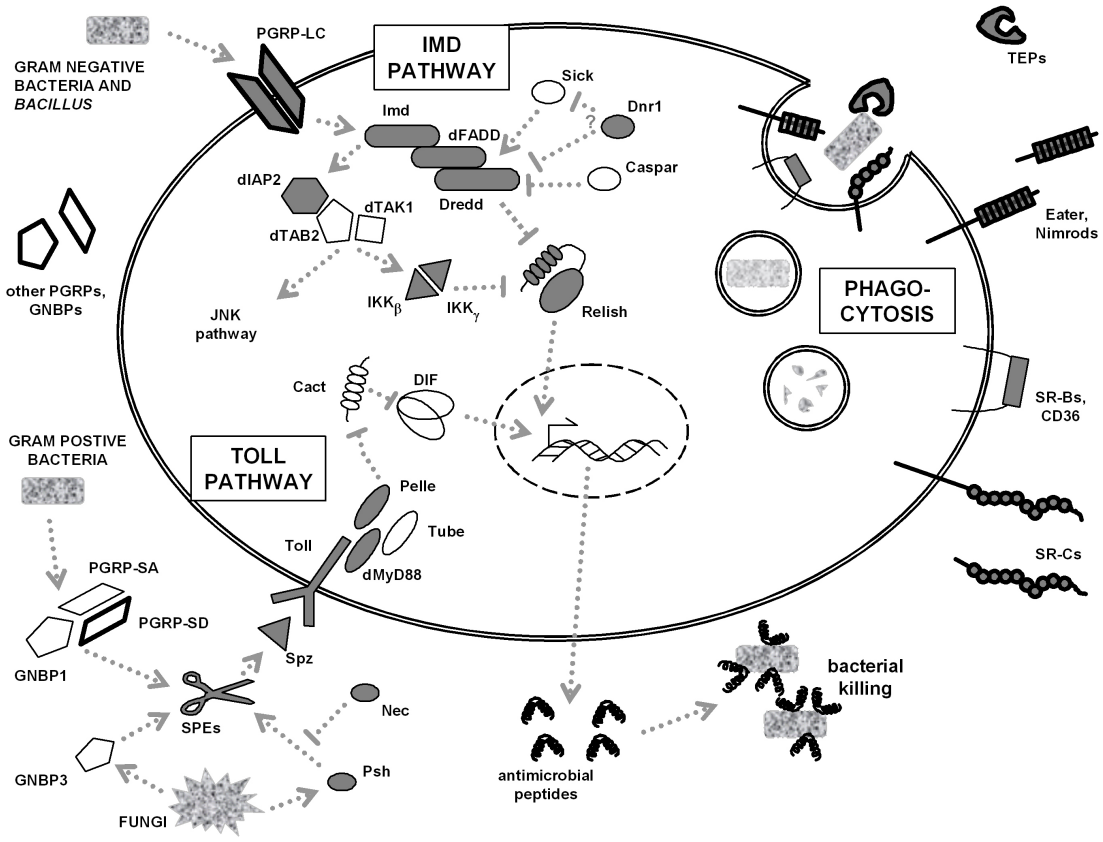


Figure 1.3: A schematic illustration of an idealized *D. melanogaster* immune responsive cell illustrating prominent proteins required for the activation of a humoral immune response and receptors involved in defensive phagocytosis. Proteins whose gene families have experienced considerable genomic turnover within the genus *Drosophila* and among *Drosophila*, *Anopheles*, *Aedes*, *Apis* and *Tribolium* are outlined in heavy black. Gray shaded proteins have been implicated as evolving adaptively at the amino acid sequence level in *D. melanogaster* and/or *D. simulans*. (Reproduced with permission from Lazzaro (2008).)

high levels of amino acid divergence between *D. melanogaster* and mosquitoes and considerable evidence of adaptive evolution within *Drosophila* (Schlenke and Begun, 2003; Jiggins and Kim, 2007; Sackton et al., 2007; Waterhouse et al., 2007; Kafatos et al., 2009) (Figure 1.3).

The adaptive evolution of innate immune signaling pathways is dramatically illustrated by proteins in the Relish cleavage complex of the Imd signaling pathway (Figure 1.4). Relish is a NF-κB family transcription factor that is cytoplasmically bound in the absence of infection. Activation of the Imd signaling pathway leads to phosphorylation of Relish, caspase-mediated cleavage of the Relish inhibitory domain, and translocation of the activated transcription factor to the nucleus. Several proteins in the cleavage complex (Dredd, dFADD, IKK_b, IKK_γ, and Relish itself) appear to be evolving adaptively in *D. melanogaster*, *Drosophila simulans*, and/or the *melanogaster* species group. Adaptive mutations are disproportionately located in protein domains important for the release of activated Relish: the Relish autoinhibitory domain and cleaved linker, the Dredd caspase domain, the dFADD death domain, and the IKK_b kinase domain (Begun and Whitley, 2000; Schlenke and Begun, 2003; Jiggins and Kim, 2007; Sackton et al., 2007) (Figure 1.4). Adaptive evolution of the Relish complex is not universal among *Drosophila*, but is restricted to certain species in the *melanogaster* group (Levine and Begun, 2007; Sackton et al., 2007). In an interesting parallel, the *Relish* gene of *Nasutitermes* termites also evolves adaptively, again with positively selected mutations localized in and around the caspase cleavage site and linker (Bulmer and Crozier, 2006), suggesting convergence of selective pressures in these distantly related insects. Nor is adaptive evolution in *Drosophila* restricted to the Relish complex. Many other signal transduction genes in the Imd and Toll pathways (*imd*, *spirit, persephone*, *Toll, dorsal, necrotic*) also show evidence of rapid evolution in *Drosophila* (Schlenke and Begun, 2003; Jiggins and Kim, 2007; Sackton et al., 2007).

One hypothesis to explain the preponderance of adaptive mutations in signaling genes is that at least some pathogens may actively interfere with host

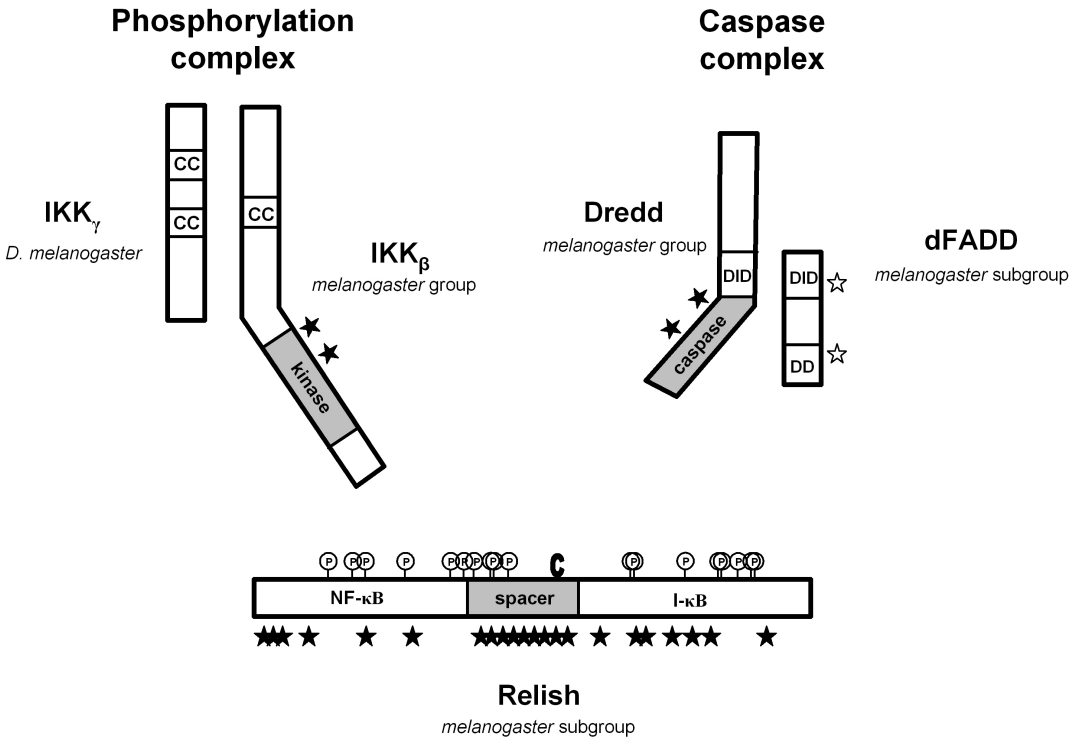


Figure 1.4: Adaptive evolution in the Relish complex. Caspase cleavage of the phosphorylated Relish spacer region allows the NF-kB domain to be translocated to nucleus, where it drives expression of immune response genes. IKK$_\gamma$ and IKK$_\beta$form a complex through interaction at coiled-coil domains, and IKK$_\beta$ phosphorylates Relish. The caspase Dredd is activated by dFADD via interaction at death inducing domains and forms a complex with Relish. Putative Relish activation domains are indicated in grey. Positively selected sites (posterior probability > 0.75) are indicated (=significant at P<0.01; =significant at P<0.02) and reflect selection along the *D. melanogaster* branch (*Relish*, *IKK$_\beta$*, or *dFADD*) or across the *melanogaster* (*Dredd*) species group (Sackton et al., 2007). Taxonomic lineages where these genes appear to have evolved adaptively are indicated beneath each gene name (Begun and Whitley, 2000; Schlenke and Begun, 2003; Jiggins and Kim, 2007; Sackton et al., 2007). (CC=coiled-coil domain; DID=death inducing domain; DD=death domain; P=phosphorylation site; C=caspase cleavage site)

immune signaling (Begun and Whitley, 2000). Such pathogens could include bacteria that inject immunomodulatory molecules into host cells, immunosuppressive fungi and parasitoid mutualistic polydnaviruses (reviewed in Schmid-Hempel (2008)). In the Relish example, pathogen interference with the assembled cleavage complex could drive co-evolutionary adaptation in several proteins. Alternatively, interference with a single important member of the complex could drive adaptation in that member while promoting compensatory adaptations in the interacting proteins to retain host function. Such compensatory mutations may occur throughout the signaling pathway, amplifying the evidence of natural selection in this gene set (DePristo et al., 2005). The convergence of adaptive evolution of genes within the Relish complex in different insect species suggests that some of these genes are common targets of pathogens.

1.2.2 Antimicrobial Peptides (AMPs)

The humoral immune response culminates in the production of effector molecules that kill invading microbes. One well-studied class of effector molecules is antimicrobial peptides. Most AMPs are short cationic peptides whose microbicidal activity is mediated by direct interaction with the negatively charged lipid membranes of bacteria and fungi (Zasloff, 2002; Lemaitre and Hoffmann, 2007; Yeaman and Yount, 2007). Antimicrobial peptides drew early attention as potential sites of host-pathogen co-evolution (Clark and Wang, 1997; Ramos-Onsins and Aguadé, 1998; Date et al., 1998) because of their direct role in the lysis and targeted killing of pathogens. However, systematic study of AMP genes, first in *D. melanogaster* and more recently across six *Drosophila* species, has failed to uncover evidence of adaptive evolution at the amino acid

level (Lazzaro and Clark, 2003; Jiggins and Kim, 2005; Sackton et al., 2007). *Drosophila* AMP genes do, however, show extremely high rates of gene family expansion and contraction (Sackton et al., 2007). This high rate of genomic turnover extends to other taxa and is characteristic of most AMPs (Figure 1.3). In fact, the majority of *Drosophila* AMPs have no identifiable homologs in the genomes of mosquitoes, honey bees or *Tribolium* (Christophides et al., 2002; Evans et al., 2006; Waterhouse et al., 2007; Zou et al., 2007). Instead, these insects each have their own unique peptide families (Bulmer and Crozier, 2004; Waterhouse et al., 2007; Evans et al., 2006; Zou et al., 2007). In some cases, AMP families in different species independently converge on similar tertiary structures and presumably functions (Broekaert et al., 1995). Thus, while antimicrobial peptides as a functional class of protein are ubiquitous among higher eukaryotes, there appears to be little homologous retention of peptides over evolutionary time.

The levels of sequence constraint seen in *Drosophila* do not characterize AMP evolution in all taxa. Genomic duplication of antimicrobial peptide genes is occasionally coupled with adaptive diversification at the amino acid level, presumably reflecting functional divergence (Tennessen, 2005; Yeaman and Yount, 2007). Genes encoding a termite-specific class of AMPs, *termicins*, have independently duplicated or triplicated in several termite species, with one duplicate typically sustaining mutations that decrease the polarity of the peptide (Bulmer and Crozier, 2004). These changes, which are driven by positive selection on amino acid sequence, result in peptides with divergent charges. Similarly, the mosquito *A. gambiae* has duplicated members within the *defensin* family (Dassanayake et al., 2007). Again, expansion is coupled with elevated rates of amino acid substitutions that change polarity, suggesting adaptive value to having two

defensins with slightly different polar affinities. Previous studies in vertebrate AMP families have also found evidence of duplication coupled with positive selection, although in these cases peptide charge is maintained (reviewed in Tennessen (2005) and Yeaman and Yount (2007)). There is compelling evidence from insects and vertebrates that gene family expansion can sometimes allow adaptive diversification of peptide function (Tennessen, 2005).

AMPs are remarkably efficient at combating infection. Resistance in microbes is seldom observed in nature, and when it is, it tends to arise in specialized pathogens that are likely to be under strong selective pressure to resist this form of defense (see Samakovlis et al. (1990) and Zasloff (2002)). There are several possible explanations for why it may be difficult for most bacteria to evolve resistance. One common AMP mechanism is to disrupt membrane integrity though biochemically simple mechanisms, such as forming open pores (Zasloff, 2002; Yeaman and Yount, 2007). The ability of microbes to evolve resistance to such activities may be limited. However, heritable variation for resistance can be created and selected upon in microbial populations in the laboratory (Perron et al., 2006). In natural contexts, hosts simultaneously produce an array of AMPs that differ in charge, hydrophobicity, structure and activity, probably ensuring that most pathogens are susceptible to at least a subset of them. This is conceptually similar to the application of multiple antibiotics in clinical settings and may serve to delay or eliminate the evolution of resistance (Yeaman and Yount, 2007). If pathogens are slow or fail to evolve resistance to peptides, there may be little selective pressure on insect hosts to adapt their AMPs at the amino acid level over modest evolutionary time. However, divergent bacteria and fungi display a range of susceptibilities to individual peptides (Zasloff, 2002), so diversification in AMP function may be selectively favored in instances when a

host shifts to a new ecological niche and is immediately presented with a novel and distinct set of pathogen pressures.

1.2.3 Recognition Molecules in the Humoral Response

The humoral immune response is activated when circulating recognition factors are stimulated by highly conserved microbial compounds. Gram-negative binding proteins (GNBPs) and peptidoglycan recognition proteins (PGRPs) activate the humoral response after recognizing microbial cell wall peptidoglycans and β-glucans. Some members of the PGRP family downregulate the immune response by degrading free peptidoglycan into non-immunogenic monomers (Lemaitre and Hoffmann, 2007).

PGRP and *GNBP* gene families generally evolve under purifying selection over short evolutionary time, but have undergone substantial genomic turnover on the lineages that separate *Drosophila* from mosquitoes, honey bees and *Tribolium* (Evans et al., 2006; Waterhouse et al., 2007; Zou et al., 2007; Kafatos et al., 2009) (Figure 1.3). Most *GNBP*s and *PGRP*s do not appear to have experienced recent adaptive evolution in *Drosophila* (Schlenke and Begun, 2003; Jiggins and Hurst, 2003; Jiggins and Kim, 2006; Sackton et al., 2007), mosquitoes (Little and Cobbe, 2005), or the crustacean *Daphnia* (Little et al., 2004). A notable exception, however, is a *Drosophila PGRP* which shows strong indications of adaptive evolution. *PGRP-LC*, an alternatively spliced gene that sits atop the Imd signaling cascade, has sustained a two amino acid insertion in the PGRP-LCa isoform in species of the *melanogaster* subgroup. This insertion is predicted to alter the binding specificity of that isoform, and appears to have been positively

selected in conjunction with several additional adaptive substitutions (Sackton et al., 2007). Interestingly, the alternatively spliced binding domains of *PGRP-LC* show evidence of either recent independent duplication or concerted evolution in *D. melanogaster* and *A. gambiae* (Christophides et al., 2002). These patterns potentially reflect lineage-specific selection for recognition of distinct microbes. In another exception, limited positive selection was also detected in *GNBP* genes of *Nasutitermes* termites (Bulmer and Crozier, 2006). In this case, it was hypothesized that adaptation of recognition capability was driven by a shift in ecology as previously herbivorous termite species adapted to feed on decaying matter, exposing them to a novel community of pathogens.

One potential explanation for the observation that *PGRP*s and *GNBP*s tend to exhibit little indication of adaptive amino acid evolution is that these proteins recognize highly conserved pathogen sugar moieties. The cell wall components recognized by these proteins are indispensable for most microbes, and, generally speaking, may not be easily modifiable. There thus may be little pressure on these genes to adapt over short time periods. Additionally, these recognition proteins are active against molecules that are conserved across a wide range of microbial taxa. There are, however, a limited number of examples of positive selection on *PGRP*s and *GNBP*s. Coupled with the observations of gene family duplication and divergence among species, instances of positive selection may reflect bursts of diversification as recognition function is fine-tuned to species-specific selective pressures.

1.2.4 Recognition Molecules in the Cellular Response

Recognition is also a necessary prerequisite for pathogen clearance via cellular immunity, and several gene families have been identified that encode membrane-bound phagocytic receptors. Phagocytosis is also promoted by ”tagging” of microbes with extracellularly secreted opsonins. Several genes encoding both phagocytic receptors and opsonins show evidence of adaptive amino acid evolution within the genus *Drosophila* (Sackton et al., 2007) and frequent genomic turnover within *Drosophila* and between *Drosophila* and other insects (Evans et al., 2006; Zou et al., 2007; Waterhouse et al., 2007; Sackton et al., 2007) (Figure 1.3). In *Drosophila*, recognition genes are significantly more likely to show evidence of positive selection than genes with signaling or microbicidal functions (Sackton et al., 2007). This difference is largely driven by recognition genes that trigger the cellular response, with nine of ten recognition genes that yield significant evidence of positive selection having been either experimentally confirmed to be involved in phagocytosis or homologous to known phagocytosis genes. Specifically, these are genes encoding thioester-containing proteins (Jiggins and Kim, 2006; Sackton et al., 2007), the *eater* and *nimrod* families (Sackton et al., 2007), the class C scavenger receptors (Lazzaro, 2005), and the CD36 homolog *epithelial membrane protein* (*emp*) (Sackton et al., 2007).

Thioester-containing proteins (TEPs) have been directly implicated as opsonins mediating the cellular clearance of microbes including bacteria and malaria-causing *Plasmodium* in *Drosophila* and *Anopheles* (Levashina et al., 2001; Blandin and Levashina, 2004; Stroschein-Stevenson et al., 2006). Proteolytic cleavage of a hypervariable spacer, or ”bait,” domain exposes the thioester motif, which then covalently binds microbes and labels them for phagocytosis.

TEPs appear to be hotspots of adaptation in several species. In *D. melanogaster*, there are six *Tep* genes, four of which have intact thioester domains and thus are likely to function as opsonizing agents (Blandin and Levashina, 2004). One of four of the intact *Tep*s show evidence of adaptive divergence between *D. melanogaster* and *D. simulans* and three show evidence for directional selection in the *melanogaster* species group (Jiggins and Kim, 2006; Sackton et al., 2007) (Table 1.1). Interestingly, one of the adaptively evolving *Tep* genes is constitutively expressed at higher levels in European than African populations of *D. melanogaster*, suggesting that expression of this *Tep* may be locally adapted (Hutter et al., 2008). *Tep* genes in mosquitoes and the more distantly related crustacean *Daphnia* also show evidence of adaptive amino acid evolution (Little et al., 2004; Little and Cobbe, 2005). In all cases, positively selected amino acid mutations are overrepresented in the bait domain that is cleaved to expose the thioester motif. It is unknown whether the proteases that cleave TEPs are produced by host or pathogen, so it is not yet possible to say whether adaptation in this domain is due to co-evolution with pathogen proteases or with pathogen molecules that interfere with host proteolysis.

Tep gene families are expanded in mosquitoes, with 13 *Tep* genes found in the *Anopheles gambiae* genome and 8 in the *Aedes aegypti* genome (Christophides et al., 2002; Waterhouse et al., 2007). The expansions in size of the *Tep* gene family appear to have been independent in each of these two taxa and potentially reflect elevated pressure on cellular immunity. The *A. gambiae Tep1* gene is segregating for two sharply divergent alleles, one of which, when homozygous, confers absolute resistance to experimental infection with the rodent malaria *Plasmodium berghei* (Blandin and Levashina, 2004; Baxter et al., 2007). Individuals homozygous for the susceptible allele sustain robust *P. berghei* infections.

Table 1.1: Evolutionary genetics of the *Tep* gene family of phagocytic recognition molecules in *Drosophila*

	Tep 1	*Tep 2*	*Tep 3*	*Tep 4*	*Tep 5*	*Tep 6 (Mcr)*	**Reference**
Functional data:							
Overview	Upregulated in response to infection	Upregulated in response to infection		Upregulated in response to infection	Not expressed; likely to be a pseudogene	Lacks a thioester domain	Blandin and Levashina (2004)
Phagocytic activity			Required for efficient phagocytosis of the bacterium *Escherichia coli*	Required for efficient phagocytosis of the bacterium *Staphylococcus aureus*		Required for efficient phagocytosis of the fungus *Candida albicans*	Stroschein-Stevenson et al. (2006)
Species divergence:							
d_N/d_S (Figure 1.1)	Exceptionally elevated d_N/d_S between *D. melanogaster* and *D. simulans* clustered around the bait domain; elevated d_N/d_S in the *melanogaster* species subgroup	Elevated d_N/d_S in the *melanogaster* species subgroup	Not significant	Not significant			Jiggins and Kim (2006)
	Elevated d_N/d_S in the *melanogaster* group with trend towards an excess of positively selected sites at the bait domain	Elevated d_N/d_S across the entire gene in the *melanogaster* species group	Not significant	Elevated d_N/d_S in the *melanogaster* species group with an excess of positively selected sites at the bait domain		Not significant	Sackton et al. (2007)
McDonald-Kreitman test (Figure 1.1)	Elevated amino acid replacements across entire gene in *D. melanogaster*	Not significant		Not significant			Jiggins and Kim (2006)
Population divergence:							
Differential expression		Expression levels significantly higher in European rather than in African *D. melanogaster* populations	Not significant	Not significant			Hutter et al. (2008)

"Not significant" indicates genes that were included in the referenced studies but did not depart from null expectations. Empty cells indicate that no information has been obtained.

These two alleles differ by multiple amino acid substitutions, including several that are clustered around the thioester domain. It is currently unclear which substitutions cause the phenotypic differences in susceptibility, or whether it is an epistatic phenotype involving substitution in multiple domains of the protein. Both alleles are found at high frequencies in natural populations (Obbard et al., 2008), suggesting selective forces maintain these two alleles in the wild. This system provides a tantalizing opportunity to understand the mechanisms that lead to the maintenance of immune response polymorphisms in a natural context.

Whole-genome comparisons within the genus *Drosophila* indicate that, in striking contrast to recognition molecules that trigger the humoral response, recognition molecules that initiate the cellular response show abundant evidence of adaptive evolution. Deeper investigation of the *Tep* gene family reveals that adaptive evolution extends beyond *Drosophila* to include mosquitoes and *Daphnia*, and demonstrates extant functional variation in a mosquito *Tep* gene. The signals of adaptive evolution suggest that these recognition molecules interact with evolutionarily labile pathogen motifs or that, like signaling molecules in humoral defense, they are potentially subject to interference by pathogen-produced proteins.

1.2.5 Summary

The diverse evolutionary trajectories of various genes in the insect immune response (Figure 1.3) can be interpreted in light of their molecular functions and interactions with pathogens. Pathogen recognition molecules that stimulate the

humoral response interact with highly conserved microbial cell wall components. Although obligate pathogens are sometimes able to reduce their cell walls to escape detection, most microbes are evolutionarily constrained because they must also be able to persist in non-infectious environments. Similarly, there may be few ways in which microbes can evolve resistance to antimicrobial peptides, especially when host insects simultaneously employ multiple peptides with distinct activities. If there is little adaptation in pathogens to escape host humoral recognition and antibiotic killing, then it may be expected that there would be little indication of adaptive amino acid evolution in the host genes over short evolutionary time. Both humoral recognition factors and antimicrobial peptides exhibit rapid rates of genomic duplication and deletion, and in some taxa, duplication is coupled with a burst of amino acid diversification that presumably increases breadth of function.

In contrast, signal transduction proteins in the humoral immune response are largely maintained in strict orthology across insect species, but frequently show indications of adaptive amino acid evolution within species. A hypothesized explanation is that the strong maintenance of orthology in these pathways makes them attractive targets for immune suppression by generalist pathogens. This may be a particularly successful strategy for microbes that are unable to evade or resist the recognition and microbicidal stages of humoral immunity. Gene duplication and diversification are not commonly observed here, perhaps because this is not a successful strategy for escaping pathogen interference. Genomic retention of a duplicated gene that can be manipulated by pathogens would be detrimental because host signaling function would be impaired. Instead, rapid fixation of amino acid "escape" variants in signaling genes seems to be the most effective host strategy, and coordinate compensatory mutation in

physically interacting proteins may amplify the signal of adaptive evolution in this functional category.

Recognition factors and opsonins in the cellular immune response evolve both by adaptive amino acid evolution and frequent genomic turnover. In general, little is known about the specific activities and recognition profiles of these genes, making it difficult to interpret the evolutionary patterns in a functional context. The evolutionary genetics, however, do lead to functional predictions, including that the cellular recognition factors bind evolutionarily labile pathogen epitopes or are subject to pathogen interference, both of which could drive rapid amino acid evolution. At the moment, virtually nothing is known about the molecular evolution or population genetics of host genes that drive phagocytosis after pathogen recognition. Microbes are capable of manipulating host cells both to promote and inhibit phagocytic uptake (Schmid-Hempel, 2008), leading to the prediction that genes encoding the machinery of phagocytosis will, like genes in humoral signaling pathways, show abundant evidence of adaptive evolution.

1.3 Evolutionary Patterns in the Antiviral Immune Response

Early characterization of the immune response primarily focused on antimicrobial defense. Antiviral defense is at least partially distinct from that against microbes, and currently is only poorly understood. Both the Toll and Imd pathways are activated during the course of some viral infections; however only the Toll pathway seems to confer protection (Lemaitre and Hoffmann, 2007). RNA interference (RNAi) provides an independent mechanism of defense that

is specific against RNA viruses (Wang et al., 2006a). Viruses are formidable opponents for the immune system. They are capable of rapid evolution owing to their fast generation times, large population sizes, high mutation rates and obligate pathogen lifestyles. These factors hint that the evolutionary patterns of antiviral defense genes will be different from those described previously for the antimicrobial defense.

Short term evolution of an antiviral defense gene has been studied at the *D. melanogaster* locus *ref(2)P*, which is proposed to function in the Toll pathway (Avila et al., 2002). This locus is polymorphic for alleles that explain a large component of the variation in susceptibility or resistance to the rhabdovirus Sigma (Contamine et al., 1989; Bangham et al., 2007, 2008). A single domain, termed PB1, of *ref(2)P* is required for viral replication (Carre-Mlouka et al., 2007). Sigma is infective if a permissive allelic variant of this domain is present, but not with a restrictive allele or genetic knock-out of the domain. This domain has an excess of amino acid polymorphisms (Wayne et al., 1996), consistent with natural selection acting to maintain allelic diversity. A random sample of ten phenotypically random alleles identified six amino acid polymorphisms in the PB1 domain (Wayne et al., 1996). A single complex mutation, with a single glycine residue substituted for glutamine and asparagine residues, was found on restrictive but not permissive alleles. The remaining polymorphisms are shared by both restrictive and permissive alleles. The frequency of the complex mutation varies between populations, ranging from absent in some African and European populations to 23% in some North American populations (Bangham et al., 2007). There is greatly reduced variation in the restrictive haplotype in a North American population, suggesting that it has recently risen to high frequency by directional selection (Figure 1.1). This indicates that selection is act-

ing on localized spatial scales, likely in concert with Sigma virus, which also varies in frequency and genotype between populations (Carpenter et al., 2007).

The fact that there is an excess of nonsynonymous polymorphism in *ref(2)P* PB1 domain but that only a single complex mutation separates restrictive and permissive alleles suggests that current Sigma virus populations have become adapted to some of the remaining polymorphisms. Indeed, analysis of all combinations of polymorphisms on the restrictive allele in artificially generated constructs indicates that no fewer than two of the three mutations are required to create a restrictive allele (Carre-Mlouka et al., 2007). These data suggest a model wherein novel mutations have been driven to high frequency by directional selection, but that the sweeps are incomplete because the virus quickly adapts to the increasingly common allele before it fixes in the population. Host resistance then requires the repeated reintroduction of novel restrictive mutations. The most escalated rates of evolution are expected when host and pathogen are co-evolving, such that host adaptations to "escape" infection are met by a gene-for-gene pathogen adaptation to maintain virulence (Dawkins and Krebs, 1979). Over the evolutionary long term, there is evidence for elevated amino acid substitution at this domain, with more adaptive mutations becoming fixed in *D. melanogaster* when compared with *D. simulans*, a species in which Sigma infection is rare or absent (Wayne et al., 1996). Restrictive polymorphisms that are driven to high frequencies during partial selective sweeps will fix by genetic drift more often than mutations that are selectively neutral over their entire evolutionary history, which may lead to elevated amino acid divergence between species.

A distinct pathway using RNAi presents an important defense against RNA

viruses. In *D. melanogaster*, double-stranded viral RNA (dsRNA) is recognized and cleaved into small interfering RNA (siRNA) by Dicer-2 (Wang et al., 2006b). These siRNAs then guide cleavage of matching RNA via formation of a RNA-induced silencing complex (RISC). Some viruses produce proteins that suppress RNA silencing. For example, *Drosophila* picornavirus C produces a dsRNA-binding protein that interferes with Dicer-2 activity and promotes viral establishment and proliferation (van Rij et al., 2006). *Dicer-2*, along with RISC genes *R2D2* and *Argonaute-2*, are amongst the most rapidly evolving genes in the *D. melanogaster* genome. These anti-viral genes, but not their paralogs with housekeeping regulatory function, show indications of adaptive evolution by recurrent fixation of novel amino acid mutations (Obbard et al., 2006).

The unique patterns of evolution of antiviral defense yield a useful system for integrating measures of short term and long term evolution. In the case of *ref(2)P* in *D. melanogaster*, rapid evolution is driven by a gene-for-gene interaction between host and virus, and is evidenced by reduced genetic variation within the selectively favored allele in the short term and increased amino acid divergence in the long term. Rates of long term evolution in RNAi antiviral genes in *D. melanogaster* are dramatically higher than the genome average. Evidence suggests that the selective pressures are different from those that act on antimicrobial defense, leading to elevated rates of evolution. This may reflect either rapid viral evolution or high host specificity in viruses, either of which would facilitate co-evolution. Like humoral signaling pathways in the antimicrobial defense, RNAi pathways are also subject to pathogen interference to overcome host defenses, indicating that they too are a potential site of direct conflict. Thus, evidence from both types of defense suggests that sites of pathogen interference display elevated evolutionary rates. As antiviral defense

becomes better characterized at the molecular level, this system will yield further insights into genetic adaptation to pathogen pressures and serve as a comparison for evolutionary patterns observed in antimicrobial defense.

1.4 From Genotype to Phenotype

All the patterns discussed thus far have pertained to the long term evolution of the immune system. It is important to remember, however, that all adaptive evolution is based on phenotypic polymorphism that segregates in populations at some point in time. Indeed, extant natural populations harbor considerable genetic variation for immunocompetence. This segregating phenotypic variation is the substrate for short term evolution. Understanding its genetic basis and the forces governing its persistence is essential for predicting the evolutionary response to natural or artificial perturbations in infectious pressure in natural populations.

In organisms with well characterized genomes, it is possible to directly test the phenotypic effects of allelic variation in pre-chosen "candidate" genes though genotype-phenotype association mapping. These studies have been employed most effectively in *D. melanogaster*. For instance, natural allelic variation in the *ref(2)P* gene clearly determines resistance to the vertically-transmitted Sigma virus in *D. melanogaster* females in an almost purely Medelian fashion (Contamine et al., 1989; Bangham et al., 2008). Genetic variation in Sigma viral transmission through males, however, does not map to *ref(2)P* (Bangham et al., 2008). Variation in the ability of *D. melanogaster* to suppress bacterial infection has been mapped to polymorphisms in pathogen recognition factors and

signaling genes within the Toll and Imd pathways (Lazzaro et al., 2004, 2006). Expression levels, but not polymorphisms, of antimicrobial peptides are also associated with resistance to infection (Sackton et al., 2010). These observations, coupled with evaluation of transcriptional activity of the immune system, indicate that signaling flux through the Toll and Imd pathways is a tremendously important determinant of resistance to bacterial infection. In contrast to the antiviral resistance determined by *ref(2)P*, polymorphisms mapped in the antibacterial association studies each make relatively small contributions to variance in the resistance phenotype, suggesting that resistance to bacterial infection is a combinatorial function of multiple genes of individually small effect. Even in sum, the mapped antibacterial factors do not explain the entirety of the genetic variance, indicating that other unstudied genes also contribute to variation in resistance.

If pathogen infection can be so detrimental to the condition of the host, and host alleles that confer high resistance to infection exist in natural populations, why then does resistance not spread to all individuals? Genetic tradeoffs, whereby immunocompetence comes at a cost to another phenotype within an organism, can constrain natural selection from fixing resistant genotypes (Roff and Fairbairn, 2007). Potential costs of resistance include direct damage to host tissues due to immune activity and correlated reduction in investment in other physiological traits, including alternative immune functions, metabolism, and reproduction. Which investment strategy is most favorable will depend on the strength of pathogen pressures and on selection acting on other fitness traits of the organism.

An experimental approach that has been used to study genetic tradeoffs is

artificial selection for increased resistance to infection and subsequent measurement of correlated changes in other fitness traits. This method identifies costs of resistance, defined as changes in traits that reduce fitness in selected lines compared with unselected lines. Artificially selecting the Indian meal moth, *Plodia interpunctella*, for increased resistance to granulosis virus infection led to correlated increases in larval development time and pupal weight and a decrease in egg viability in selected lines (Boots and Begon, 1993). Selection in *D. melanogaster* for resistance to parasitoid or fungal infection led to a correlated decreases in larval competitive ability and adult fecundity, respectively, in the absence of infection (Kraaijeveld and Godfray, 1997, 2008). Costs that are measured in artificial selection lines should be interpreted with caution, however, as selection experiments can sometimes result in the fixation of rare alleles with large phenotypic effects that are not representative of functional genetic variation in natural contexts. For example, *A. gambiae* mosquitoes selected for refractoriness to *Plasmodium* infection achieve this through an increased melanization response (Collins et al., 1986) and high levels of cellular oxidative free radicals that are extremely damaging to host cells (Kumar et al., 2003). Natural resistance in wild populations of *A. gambiae*, however, is generally accomplished with a melanization-independent mechanism (Riehle et al., 2006), and is likely to be less costly or damaging than mechanisms seen in laboratory-selected lines.

A more relevant, but much subtler, measurement of genetic tradeoffs is obtained by measuring genetic correlations between traits in naturally occurring, unselected genotypes. This is commonly done by measuring phenotypes in genetic clones or in individuals' with known genetic relatedness and estimating the genetic contribution to the phenotype. In *D. melanogaster*, genotypes with high resistance to bacterial infection had low fecundity in the absence of

infection in a food-limited environment (McKean and Nunney, 2008). In the pea aphid *Acyrthosiphon pisum*, clonal lines with high resistance to attack by the parasitoid wasp *Aphidius ervi* had reduced fecundity (Gwynn et al., 2005). However, in this case, resistance to parasitoids can be conferred by bacterial endosymbionts, so the genetic basis for this tradeoff may be mediated by factors outside the host genome. In both examples, the cost of resistance is a decrease in reproductive fitness.

The ultimate goal is to identify the genetic architecture underlying tradeoffs. Quantitative trait locus (QTL) mapping has been used to locate these genetic regions. This approach relies on contrived crosses between chosen parents to established phenotypically variable recombinant progeny. Genetic markers are then genotyped at periodic intervals across the genome, allowing the localization of genomic regions encoding the phenotypic variation without relying on *a priori* "candidate" genes. QTL mapping, however, lacks the resolution to identify specific genes or alleles. In the red flour beetle *T. casteneum* and in the bumble bee *Bombus terrestis*, simultaneous mapping of immune and fitness traits found that loci associated with immune phenotypes occasionally co-localized with QTL involved in fecundity, viability and body size (Zhong et al., 2005; Wilfert et al., 2007a). There are two potential genetic mechanisms that could cause genetic correlations between immune and fitness traits. Genetic correlations can be caused by pleiotropy, where a single gene influences multiple traits. Tradeoffs are due to antagonistic pleiotropy, where a single allelic variant of a gene has a positive effect on one trait but a negative effect on the other. Alternatively, allelic variants of distinct genes affecting the two traits may be in linkage disequilibrium due to physical proximity on a chromosome, and thus these variants are coordinately passed to the offspring. Selection acts simultaneously on traits

that are correlated by either pleiotropy or linkage disequilibrium. However, only antagonistic pleiotropy places a long term constraint on selection because recombination can degrade correlations based on linkage disequilibrium. QTL mapping relies on experimentally generated linkage disequilibrium that spans much greater physical distances than are observed in natural populations, so it is relevant to follow QTL-based studies of genetic correlations with field-based studies to determine whether the traits co-segregate in nature.

Tradeoffs have also been identified within the immune response. For example, in *B. terrestis*, lines selected for increased resistance to trypanosome infection also had a higher investment in a phenoloxidase response coupled with a lower investment in antimicrobial peptide response (Wilfert et al., 2007b). The Egyptian cotton leafworm, *Spodoptera littoralis*, demonstrated positive genetic correlations amongst hemocyte density, cuticular melanization, and phenoloxidase activity, but a negative genetic correlation between hemocyte density and lysozyme-like antibacterial activity (Cotter et al., 2004). A different result is obtained from females of the mealworm beetle *Tenebrio molitor*, where cuticular melanization shows a negative genetic correlation with hemocytes and phenoloxidase, suggesting that the genetic architecture of these correlations can vary between species (Rolff et al., 2005). These results demonstrate that increased investment in one component of the immune response can come at a cost to other immune functions, and indicate the potential for tradeoffs within the immune response to place constraints on the evolution of global resistance.

Thus far, all resistance measures have been considered only in a single environment; however, the optimal immune strategy can be expected to vary based on environmental conditions (Lazzaro, 2008). Selective pressures are heteroge-

neously distributed in the environment. Abiotic factors such as daylength, temperature, and moisture vary between populations, affecting development time, metabolic flux, and other traits, and also altering the composition of pathogen communities and nutrient availability. Allelic variants in some genes respond differently to changes in the environment, termed genotype-by-environment interactions. If a genotype is particularly favored in certain conditions, local adaptation to the proximate environment can occur. Temperate and tropical populations of *D. melanogaster* varied significantly in their resistance to the generalist fungal pathogen *Beauveria bassiana* (Tinsley et al., 2006) and bacterial pathogen *Providencia rettgeri* (Lazzaro et al., 2008). Considerable genotype-by-environment interaction was observed in resistance of *D. melanogaster* to *P. rettgeri* infection across multiple temperatures. Despite that observation, temperate populations were on average more resistant to *P. rettgeri* than the tropical one at lower temperatures, which potentially reflects adaptation to the local temperature. Spatial heterogeneity in the environment can lead to the maintenance of multiple resistance alleles if local adaptation is sufficiently strong to withstand erosion by migration and gene flow.

The magnitude, or even the existence, of genetic tradeoffs can also vary between environments. In natural and laboratory settings, infestation by the mite *Macrocheles subbadius* negatively affects the fertility and body size of its host, *Drosophila nigrospiracula* (Luong and Polak, 2007). There is genetic variation for resistance to mites, which in this case is mediated by an avoidance behavior. It has been demonstrated that, similar to *D. melanogaster* selected for parasitoid resistance, lines selected for mite resistance also suffer a cost in terms of decreased larval competitive ability. Manipulating the environment with high temperatures and increased larval density to create stressful conditions tends

to increase costs of resistance. For instance, in previously considered examples from *D. melanogaster*, resistance to bacterial infection was correlated with low fecundity only in a food-limited environment (McKean and Nunney, 2008), and larval success of parasitoid-resistant larvae was compromised only under crowded conditions (Kraaijeveld and Godfray, 1997). In all of these cases, selection can act independently on the traits in a non-stressful environment but the traits are constrained to each other under resource-limited conditions. Genetic variation for different allocations of resources between resistance and fitness traits can be maintained by environmental heterogeneity since the optimal investment strategy will be context-dependent (Roff and Fairbairn, 2007). Selection on these variants will be inefficient because tradeoffs will only be apparent in certain conditions.

The host immune response faces a special obstacle in evolving immunity: the immune system must respond to living organisms that are themselves free to evolve. Its pathogen 'environment' is capable of rapid evolution, often much more quickly than the host. Analogous to genotype-by-environment interactions, a genotype-by-genotype interaction occurs when the efficacy of a host resistance genotype is dependent on the genotype of the pathogen. Antagonistic pleiotropy can occur in this context if resistance to one pathogen genotype comes with susceptibility to another. The specificity of these interactions can allow for temporal fluctuations in host and parasite genotypes in a frequency dependent manner. Such fluctuations are generally difficult to measure experimentally, but have been observed in natural populations of the snail host *Potamopyrgus antipodarum* and trematode parasite *Microphallus* sp. as well as in the crustacean host *Daphnia magna* and bacterial parasite *Pasteuria ramosa* (Dybdahl and Lively, 1998; Decaestecker et al., 2007). In both cases, resistant host geno-

types are at an advantage when they are rare because their infective parasite genotypes are also rare, allowing resistant host genotypes to then to rise in frequency. This leads to a time-lagged increase in the infective parasite genotype, causing the host advantage to decline, subsequently reducing the frequency first of the host genotype and then the parasite genotype. This type of co-evolution is probably rare, occurring only when a parasite infects a narrow species range of hosts, allowing for specific, reciprocal adaptation, and when the parasite greatly reduces the fitness of the host such that selective pressure on resistance is high. In reality, many parasites are likely adapting to multiple hosts and impose only small reductions of fitness, placing more diffuse selective pressures on their hosts.

Environmental heterogeneity in pathogens and pathogen genotypes can lead to spatial adaptation to local pathogen pressures (Woolhouse et al., 2002). Genotype-by-genotype interactions between hosts and pathogens allow for adaptation to proximate pathogen pressures. Experimental evolution has been used to demonstrate the potential for local adaptation. In an experiment where *P. ramosa* was serially passaged for several generations on *D. magna*, it evolved high levels of infectivity on the host used for passage and in some cases lost virulence on non-passaged hosts (Little et al., 2006). This indicates that parasites can adapt to current hosts, perhaps at a cost of infecting alternate hosts, in only a few generations. Spatial variation in resistance can be detected by comparing the success of infection between host-parasite combinations that are either sympatric (local) or allopatric (foreign). Although most theoretical models predict that the parasite should be most successful in sympatric infections, in practice both parasite local adaptation and maladaption are observed (Woolhouse et al., 2002). In *A. gambiae*, a locus that was found to control encapsulation response

to the malaria parasite *Plasmodium falciparum* was strongest against allopatric infections (Niare et al., 2002). Another locus restricting infection intensity was strongest against sympatric infections. Despite the opposite directions of these responses, both findings demonstrate population variation in resistance. In some cases, host resistance and parasite virulence have been observed to covary. The parasitoid *Asobara tabida* has been reported to have the highest virulence in the Mediterranean and lower virulence in northern Europe (Kraaijeveld and Godfray, 1999). *D. melanogaster*, the host, was observed to have the highest resistance in the Mediterranean and southern Europe, and low resistance in northern Europe, evidence of adaptation to local parasitoid pressures.

Tremendous variation in immunocompetence exists in extant natural populations. Tradeoffs within the immune response and between immunocompetence and other fitness components constrain the ability of natural selection to drive resistant genotypes to fixation. Variation in tradeoffs is maintained in part by environmental variation, whereby the costs associated with a particular genotype are context-dependent. Genotype-by-environment interactions and local adaptation can potentially lead to the maintenance of multiple polymorphisms in heterogeneous environments. Furthermore, the pathogen "environment" is itself evolving. These forces in combination oftentimes limit the evolution of a single globally resistant genotype.

1.5 Conclusion

Genes involved in the immune response show signals of rapid evolution, with the precise evolutionary mode varying among components of the immune sys-

tem. Extant populations harbor tremendous genetic and phenotypic variation in resistance, providing the substrate upon which selection acts. Examination of both evolutionarily ancient and current patterns has only rarely been performed. The most complete example is from the *ref(2)P* locus in *D. melanogaster*, which is polymorphic for the ability to permit or restrict Sigma virus infection. In natural populations, this locus shows evidence for elevated polymorphism, partial selective sweeps, and spatial heterogeneity in allele frequencies, all of which reflect an on-going battle between host and pathogen. These polymorphisms also often become fixed, driving long term adaptive amino acid evolution. Other parts of the immune system could be equivalently studied, such as a polymorphic locus in the mosquito *A. gambiae* that confers resistance to malaria. In general, characterization of forces that facilitate or inhibit the spread of host resistance through populations, combined with genome-scale comparisons between species, will allow the linkage of short term and long term patterns to fully define the lability and constraint on adaptive evolution across the immune system.

Understanding the factors that influence the evolution of the immune response has important ramifications for diverse fields of study. Evaluation of the feasibility of applications such as the proposed engineering of transgenic disease vector insects to control transmission and the use of pathogens to implement biological control of pest populations benefits from the most complete understanding possible of how resistance arises and propagates through natural populations. These are inherently evolutionary biological questions. The evolutionary dynamics of insect-pathogen interactions also has clinical importance insofar as insects can serve as model hosts for humans. Evolutionary inferences about how pathogens interact and interfere with different components

of the immune system inform studies in molecular immunology. Advances in immunology, in turn, will test these predictions and identify new sets of genes and pathways in a wider range of organisms, further broadening the field of evolutionary genetics.

1.6 Acknowledgements

We thank Madeline Galac for helpful advice during the preparation of this manuscript and Jacob Crawford, Sarah Short, and Gerardo Marquez for feedback on the manuscript. We thank Tim Sackton for providing portions of the data illustrated in Figure 1.4. Work in the Lazzaro lab is funded by grants from the National Science Foundation and National Institutes of Health.

CHAPTER 2

DIVERSITY OF BACTERIA ASSOCIATED WITH THE HEMOLYMPH OF WILD-CAUGHT *DROSOPHILA MELANOGASTER*

2.1 Abstract

Although *Drosophila melanogaster* is one of the genetically best-characterized organisms in science, knowledge of its ecology lags far behind. This is a major handicap for understanding evolution and providing context for function of the immune system system. Identifying the bacteria that are naturally associated with these flies can potentially tell us a great deal about whether the immune response evolved to fight pathogens, to maintain symbionts, or both. To this end, I surveyed bacteria from the hemolymph of wild-caught *D. melanogaster* from an Ithaca, New York, population over three years. I developed culture- and PCR-based detection methods that could be applied to the hemolymph of individual flies and demonstrate that culturing methods can successfully be used to detect hemolymph-borne bacteria. I found that between 0.3% and 2.0% of flies had infections that could be detected by culturing methods. Infections were caused by a taxonomically diverse array of bacteria that varied in virulence in subsequent reinfection experiments. No bacterium other than the intracellular symbiont *Wolbachia pipientis* was detected regularly by PCR, suggesting the absence of any other prevalent bacteria in the hemolymph, although limitations in the application of the PCR-based method precluded drawing firm conclusions about overall infection frequency. My data suggest that the *D. melanogaster* population sampled lacks bacterial specialist pathogens and symbionts other than *W. pipientis*, and reveal that bacterial infection of the hemolymph of wild *D.*

melanogaster is rare and caused by an assortment what appear to be opportunistic pathogens.

2.2 Introduction

Recent years have seen a great expansion in our understanding of the genetics of insect immune responses, particularly focusing on antibacterial immunity in *Drosophila melanogaster*, but our understanding of the natural pathogens that shape the evolution of this response has grown at a much slower rate. From studies in other insects, we know that a wide variety of bacteria-host interactions exist, ranging from commensal to mutualist to pathogenic. Some of these bacteria are maintained and some are cleared, in part due to the activity of the host immune response (Reynolds and Rolff, 2008). Characterization of the bacteria that are naturally associated with *D. melanogaster* will lead to a better understanding of the function and evolution of its immune system. This study was aimed at identifying bacteria from the hemolymph of individual wild flies using a combination of culture-dependent and culture-independent methods.

Studies in other insects reveal a number of different types of bacteria-host interactions. In mutualistic interactions, the presence of bacteria benefits the host nutritionally, defensively, or otherwise, and the bacteria benefits by using the host as a habitat and as a source of nutrients. In some cases, mutualists are highly adapted to their host to the extent that they are obligate, having specialized or reduced genomes that specifically code genes that aid the host (Reynolds and Rolff, 2008). Mutualistic bacteria can sometimes be found concentrated within specialized host cells called bacteriocytes, although they can often also

be found in the hemolymph (Feldhaar and Gross, 2009). Pathogenic bacteria, which cause harm to the host, are also known and may have specific adaptations to increase virulence. Commensal bacteria, on the other hand, may benefit from their hosts while causing neither harm nor benefit. Little is known about identities of bacteria that are associated with *D. melanogaster* in nature and about the types of interactions in which they participate.

Some effort has been placed in recent years on increasing our knowledge of wild and lab bacteria associated with *D. melanogaster*. Corby-Harris et al. (2007) looked at the diversity of microbes internally and externally associated with whole flies from different geographic locations. A total of 74 OTU's belonging to *Proteobacteria*, *Firmicutes*, and *Bacteroidetes* were identified in flies from 11 North American populations (7-30 OTU's per population) of wild *D. melanogaster*. A study by Cox and Gilmore (2007) found 25 OTU's from the above three phyla plus *Actinobacteria* associated with wild *D. melanogaster* from a single North American population. They also found a high prevalence of *Enterococcus* species in laboratory-reared flies. These *Enterococcus* species are commonly thought to be commensal but Cox and Gilmore (2007) found that pathogenicity could be induced by overexpressing a single gene. In both studies cited above, the predominant bacteria in wild-caught flies were *Proteobacteria*. Both of these studies were performed on pools of 5 or 10 flies so the properties of individual flies are unknown.

Similar studies from other insect species reveal differences in the diversity and structure of microbial communities associated with various insects. A study of *Anopheles* mosquitoes found several bacteria in the midgut that were intracellular and related to known insect symbionts (Lindh et al., 2005). Although it is

not known whether these bacteria provide a nutritional benefit to mosquitoes, studies of a variety of insects show that symbionts are most likely to be found in insects that feed on specialized or nutritionally-poor diets such as blood or plant phloem (Feldhaar and Gross, 2009). This suggests that there would be differences between the microbial communities associated with mosquitoes and those found with *D. melanogaster*, which is not surprising given their dramatically different life histories. A survey of the midguts of gypsy moth found that bacterial species diversity was highly influenced by food substrate (Broderick et al., 2004), again pointing to the importance of environment in structuring microbial communities. Bacterial community complexity is substantially higher in humans than in insects, with hundreds of phylotypes predominantly from the phyla *Bacteriodetes* and *Firmicutes* (Eckburg et al., 2005) being found in the human gut compared with only several dozen phylotypes seen in some insects (Broderick et al., 2004; Lindh et al., 2005; Cox and Gilmore, 2007; Corby-Harris et al., 2007).

Rigorous comparison among microbial survey studies is difficult because of the multitude of different approaches used. Many studies use a combination of culture-based methods and/or non-culture-based methods (Broderick et al., 2004; Lindh et al., 2005; Eckburg et al., 2005; Cox and Gilmore, 2007; Corby-Harris et al., 2007). Culture-based methods are useful because they allow isolation of bacteria for subsequent experiments. However, it is estimated that less than 1% of bacteria can be cultured from some environments, especially those that greatly differ from the culture medium (Schloss and Handelsman, 2006). Non-culture-based methods also have problems. Sequencing of cloned PCR products introduces uncertainty due to nucleotide incorporation errors during amplification, ligation biases, and primer-dependent amplification success.

High-throughput sequencing is prone to a higher rate of sequencing error than Sanger-based methods (Johnson et al., 2006). DNA extraction methods can influence the relative recoveries of Gram-positive versus Gram-negative bacteria because of the relative difficulty in lysing the cell walls of Gram-positive bacteria (Broderick et al., 2004; Corby-Harris et al., 2007). These caveats must be kept in mind and data should be interpreted within the context of how it was collected.

In this study, I conducted a survey of the bacteria present in the hemolymph of wild-caught *D. melanogaster* from a single population across three different years. I isolated bacteria using a culture-based method in which the cultured bacteria were identified by direct sequencing and retained for subsequent characterization of virulence. Bacteria were also identified by a culture-independent nested-PCR method that partly but not completely distinguished between bacteria in the hemolymph versus on the cuticle of the fly. Wild *D. melanogaster* were occasionally found to carry the bacterial symbiont *Wolbachia pipientis* and were found to have a low level of infection in the hemolymph from a wide range of other bacteria.

2.3 Materials and Methods

Wild *D. melanogaster* were collected from Little Tree Orchards in Newfield, Tompkins County, New York, or from residential areas in Ithaca, New York. All males were identified to species level using morphological characters. Females could not definitively be excluded from being *D. simulans* using morphology. However, male *D. simulans* were only observed in October and were always at

Table 2.1: Number of infected flies and flies sampled by month and year.

Year	Month	Number of infected flies	Number of flies sampled
2005	June	1	37
	July	0	85
	August	2	264
	September	6	229
	October	0	0
	Year Total	*9*	*615*
2006	June	4	138
	July	11	521
	August	7	297
	September	1	233
	October	2	94
	Year Total	*25*	*1283*
2008	June	0	0
	July	0	0
	August	0	173
	September	1	191
	October	0	0
	Year Total	*1*	*364*

a very low frequency, and this month was only minimally represented in one year of sample collection (Table 2.1). Collections were always performed in the mornings over piles of discarded apples or apple pomace using aerial sweep nets. Flies were placed in empty *Drosophila* vials (25x95 mm; Laboratory Products Sales #L284051) with slightly damp cotton to avoid dessication and stored on ice to slow bacterial growth and to prevent over-heating. The time between collection and the extraction of hemolymph was from one to four hours. Samples were collected from June to September 2005, June to October 2006, and August to September 2008 (Table 2.1).

In the lab, flies were anesthetized using carbon dioxide and then immobilized on index cards using mucilage glue (Ross Kraft). Groups of 15-30 flies were surface sterilized for 20 minutes using UV irradiation in a Spectrolinker XL-1000 UV Crosslinker (Spectronics Corporation) at default settings. Hemolymph

(~0.05 μl per fly) was extracted from individual flies using pulled glass microcapillary tubes and placed into 50 μl of ddH_2O in 200 μl PCR tubes (USA Scientific #1402-2900). All water used in this experiment was filter sterilized (pore size 0.2 μm), and all equipment was cleaned with 10% bleach between uses. Between 16 and 79 flies were sampled on any given day. A total of 2,262 flies were surveyed during the course of the study. Between 1 and 11 negative controls were included on each sampling day where empty microcapillary tubes were evacuated into water. Hemolymph samples were plated on tryptic soy agar (BD Difco) 100x15mm plates. A 25 μl aliquot of hemolymph was pipetted onto each plate and spread using a glass spreader. The remaining 25 μl of hemolymph was stored at -80°C until PCR could be performed.

The sampling methodology varied slightly among years. In 2005, a micromanipulator was used to withdraw hemolymph by capillary action and nitrogen gas was used to evacuate hemolymph into PCR tubes. In 2006 and 2008, hemolymph was extracted manually instead of using a micromanipulator and was evacuated from the microcapillary tubes using a microcapillary bulb (VWR #53507-268) instead of using nitrogen gas, which greatly increased the speed of extractions. Between experimental days, UV irradiation (20 minutes) was used in addition to bleach to remove amplifiable DNA from water and all equipment used for the extractions. In 2008, flies were not incapacitated using glue. Instead, flies were anesthetized during UV irradiation by using ether or by using carbon dioxide emanating from dry ice. Instead of using a glass spreader, each hemolymph sample was spread by holding the sample vertically so that the sample would flow downwards on the plate in a straight line. This method required that each plate be opened and exposed to air only once as opposed to the two openings when using a glass spreader and thus further minimized the risk

of contamination.

2.3.1 Culture-Dependent Survey

Plates with hemolymph samples were stored upside down and allowed to grow aerobically at room temperature (~24°C) for 10 days with light. After 10 days, the number of colony forming units (CFUs) was recorded along with any notes about colony location or appearance. Colonies with unique morphologies (including size, shape, pigmentation, and texture) on each plate were restreaked onto fresh tryptic soy plates and grown at 24°C. Each isolate was used for PCR reactions and Taq-amplified using primer set 1A in years 2005 and 2006 and primer set 1B in 2008 (Table 2.2). PCR products were prepared for sequencing by one hour of incubation at 37°C with 0.5 μl Exonuclease I and 0.5 μl shrimp alkaline phosphatase and then directly sequenced. DNA sequences were assembled and aligned in CodonCode Aligner (CodonCode Corp.). Identities were determined by submitting sequences to Sequence Match in Michigan State's Ribosomal Database Project (Cole et al., 2009). 16S rDNA sequences offer little resolution at the species level, thus genus level assignments were made by matching sequences to the genus of the nearest type species. In several cases, sequence identities below 97% were observed, indicating potential novel species (Cox and Gilmore, 2007; Corby-Harris et al., 2007). These sequences were assigned to the nearest genera for identification purposes. When multiple highly similar sequences were obtained, isolates were assigned to the same Operation Taxonomic Units (OTUs) if they were at least 97% similar.

Colonies with unique morphologies from each hemolymph sample plate

were used to inoculate tryptic soy broth (BD Difco) and grown for 24-72 hours with shaking at 24°C. Liquid cultures were frozen at -80°C with a final concentration of 15% glycerol. In two cases, bacteria were unable to grow in liquid culture and thus were not retained. One representative from each OTU was chosen for characterization of bacterial virulence in the wild type Oregon R strain of D. melanogaster. Bacteria were grown in liquid culture for two days at 24°C and diluted to 1.0 O.D. (600 nm) using a spectrophotometer. Gram-negative bacteria grew to higher densities (~10^9 CFUs/ml) than Gram-positive bacteria (~10^8 CFUs/ml) at an optical absorbance of 1 O.D., so Gram-positive bacteria were concentrated to 10 O.D. prior to infection. Groups of ten flies aged three to five days were infected by being pricked by a minuten pin that had been dipped into bacterial culture. Groups of flies were maintained at room temperature in fly vials, and mortality was scored five days post-infection. Two to three replicates were performed with each bacterium.

2.3.2 Culture-Independent Survey

A protocol was developed to directly amplify bacteria from the hemolymph of individual flies. The polymerase *Tth* was previously found to perform well in the presence of inhibitors (Panaccio and Lew, 1991), and it was able to amplify bacterial DNA directly from the hemolymph. A nested PCR protocol was employed to allow for efficient amplification of small amounts of starting template DNA (referred to as PCR protocol version 1 from this point on and used in 2005 and 2006). For Round 1 of PCR, 21 μl ddH_2O with hemolymph was combined with 2.5 μl 10X Tth PCR buffer (Roche), 0.5 μl 10 μM forward and reverse primers from primer pair1A (Table 2.2), 0.5 μl 10 mM dNTPs, and 0.125 μl Tth

Table 2.2: List of primer pairs used for PCR

Primer Set	Forward Primer	Reverse Primer
1A	8F[1] AGAGTTTGATCCTGGCTCAG	1492R[1] GGTTACCTTGTTACGACTT
2A	8F[1] AGAGTTTGATCCTGGCTCAG	806R[2] GGACTACCAGGGTATCTAAT
3A	515Fm[3] GTGCCAGCMGCCGGGGTA	1492R[1] GGTTACCTTGTTACGACTT
1B	8Fm[4] AGAGTTTGATCMTGGCTCAG	1492Ry[5] GGYTACCTTGTTACGACTT
2B	8Fm[4] AGAGTTTGATCMTGGCTCAG	806R[2] GGACTACCAGGGTATCTAAT
3B	515Fm[3] GTGCCAGCMGCCGGGGTA	1492Ry[5] GGYTACCTTGTTACGACTT

Primer pairs ending in "A" were used for the PCR protocol version 1 and those ending in "B" were used for PCR protocol version 2. All primers are given in the 5′ to 3′ orientation.
[1] Weisburg et al., 1991
[2] Nikkari et al., 2002
[3] Frey et al., 2006
[4] Schütte et al., 2008
[5] Schloss and Handelsman, 2006

polymerase (Roche). PCR cycling conditions were 1 cycle of 2 minutes at 94°C, 10 cycles of 94°C for 30 seconds, 53°C for 30 seconds, and 72°C for 2 minutes, 15 cycles of 94°C for 30 seconds, 53°C for 30 seconds, and 72°C for 2.5 minutes, and 1 cycle of 72°C for 7 minutes. First round PCR products were treated with 1 μl of a 1:1 solution of Exonuclease I and shrimp alkaline phosphatase for 60 minutes at 37°C and 15 minutes at 80°C to remove primers from the first round. For round 2 of PCR, 2.5 μl of treated product from the first round was added to 20.75 μl ddH_2O, 2.5 uL 10X Thermopol PCR buffer, 0.5 μl 10 μl forward and reverse primers from primer pair 2A or 3A (Table 2.2), 0.5 μl 10 mM dNTPs, 0.25 μl Taq polymerase. PCR cycling conditions were 1 cycle of 2 minutes at 95°C, 25 cycles of 95°C for 30 seconds, 53°C for 30 seconds, and 72°C for 2 minutes.

In 2008, a few modifications were made to increase representation of Gram-positive bacteria (referred to as PCR protocol version 2 from this point on). A lysozyme incubation step was included before the first round of PCR, with hemolymph samples incubated in water with 1 μl of 0.01 mg/ml lysozyme at 37°C for 30 minutes immediately before beginning PCR. Primer pairs 1B replaced primer pair 1A, 2B replaced 2A, and 3B replaced 3A. The replacement primers were degenerate at specific positions that allowed annealing to a wider range of bacterial sequences. The amount of primer added to each reaction was also doubled.

In 2006, samples were analyzed by terminal restriction fragment length polymorphism (t-RFLP), which allows the simultaneous detection of multiple 16S rDNA sequences present in a single sample (Frey et al., 2006). Combinations of primer pairs and restriction enzymes were identified that produced the maximum number of unique terminal fragment sizes predicted from the culture-based results. Samples were amplified using PCR protocol version 1 except the nested PCR reaction used fluorescent-labeled primers NED-8F or VIC-1492R (Applied Biosystems). Excess primers were removed using Sephadex columns. Primer set 2A was digested with DpnII (5 units; New England Biosystems) and primer set 3A was digested with MspI (5 units; New England Biosystems) in the recommended digestion buffers. Samples were digested overnight at 37°C and then ethanol precipitated. Samples were loaded with LIZ1200 size standard (Applied Biosystems) and formamide and sent for fragment analysis to the Cornell Life Sciences Core Laboratories Center. In 2006, many samples analyzed using this approach were found to contain a single 16S rDNA sequence and therefore in 2008, sample PCRs were directly sequenced.

2.3.3 Controls

Several controls were performed to test the accuracy and sensitivity of these methods. I isolated *Providencia alcalifaciens*, *Serratia marcescens*, *Staphylococcus pasteuri*, *Exiguobacterium acetylicum*, or *Pseudomonas fluorescens* from wild-caught flies using the previously described procedures and used them for control tests because they represent a diversity of phyla. To test the ability of UV irradiation to remove viable bacteria and bacterial DNA from the surfaces of flies, lab-reared flies were allowed to walk around on bacterial lawns for 24-48 hours in groups of 20. Flies were then subjected to 0 or 20 minutes of UV irradiation before having their hemolymph extracted. Hemolymph was plated as described earlier and/or retained for PCR.

To test the sensitivity of hemolymph extractions, Oregon R flies were artificially infected with test bacteria and then hemolymph was extracted 12 to 96 hours post-infection. Individual fly carcasses were homogenized in 500 µl tryptic soy broth, and an aliquot was plated with a WASP spiral plater (Don Whitley Scientific Ltd.) to measure whole-fly bacterial load of flies that also had hemolymph extracted. The number of colonies per fly was estimated using a counter associated with the plater. In some cases, plates with fly carcass homogenates were incubated at 37°C to reduce growth of bacteria endogenous to fly stocks in the laboratory. *S. pasteuri* and *E. acetylicum* rarely grew to high densities in Oregon R flies. *seml* mutant flies (obtained from L. Pham and D. Schneider), which are immunocompromised because they lack a functional *PGRP-SA* bacterial recognition gene, were used to test growth of these two bacteria in susceptible flies. Between 6 to 18 flies were sampled with each bacterium (Table 2.3).

2.4 Results

2.4.1 Controls

Hemolymph extractions performed on artificially infected flies were found to be sensitive for detecting cultivable bacteria at above 10^5 CFU per fly (Table 2.3). In three out of five cases, *S. pasteuri* at a density of 10^3-10^4/fly could be detected. In all cases, the bacteria that were recovered from the hemolymph were the same bacteria that were used for infection. In one out of 29 cases, a fly carcass that had greater than 10^5 CFU had no bacteria in the hemolymph extract. In one out of 6 cases, a fly carcass that had no test bacteria had a single colony recovered from the hemolymph.

UV irradiation was found to be highly effective at removing cultivable bacteria from the surface of flies prior to hemolymph extraction (Table 2.4). Cultivable bacteria were frequently collected in hemolymph samples of lab flies that were exposed to bacteria but not surface sterilized. This presumably reflects surface contamination being collected with the hemolymph sample. Cultivable bacteria were never collected in hemolymph samples of healthy lab flies that were externally exposed to bacteria and subsequently treated with UV irradiation.

In control tests, colonies were occasionally found around the edges of culture plates. These are likely to be contaminants since hemolymph was plated in the center of the plates and also because these spurious colonies were occasionally found on negative control plates. However, out of over 150 control plates, negative controls were never observed to have more than 5 colonies and had a

Table 2.3: Testing the sensitivity of hemolymph extractions for recovering cultivable bacteria from hemolymph of artificially infected flies

Bacterium	**Fly Genotype**	**CFU/fly**	**CFU/25 uL hemolymph sample**
P. fluorescens	Oregon R (n=6)	TNTC(6)	30,100,TNTC(4)
P. alcalifaciens	Oregon R (n=9)	TNTC(9)	15,44,˜100(3),TNTC(4)
S. marcescens	Oregon R (n=7)	TNTC(7)	TNTC(7)
E. acetylicum	Oregon R (n=3)	0,0,1630	0(3)
E. acetylicum	seml (n=7)	0,0,TNTC(5)	0,0,0,149,TNTC(3)
S. pasteuri	Oregon R (n=6)	10(2),30,90,7200,0	0(5),1
S. pasteuri	seml (n=12)	0,10(2),30,190,360,5800,4815,2040,41600,TNTC(2)	0(7),4,7,77,93,TNTC

CFU/fly is given in the same order as the CFU/hemolymph sample to allow for comparison between bacterial load in the whole fly versus the hemolymph. The number of isolates of a particular density is indicated in parentheses if the number is greater than 1. TNTC (too numerous to count) per fly indicates that bacterial load is greater than 10^5 CFU and TNTC per hemolymph sample indicates that bacterial load is greater than ˜200 CFU.

Table 2.4: Testing the ability of UV irradiation (20 minutes) to remove cultivable bacteria from the surface of flies

Bacterium	**# colonies in hemolymph sample (+UV)**	**# colonies in hemolymph sample (-UV)**
P. fluorescens	0(8)	0(3),1,2
P. alcalifaciens	0(4)	0,1,TNTC
S. marcescens	0(3)	0,1,TNTC
E. acetylicum	0(4)	0,1,5
S. pasteuri	0(15)	0(6),3,4,30,60,TNTC(3)

The number of isolates with a particular # of colonies in the hemolymph sample is indicated in parentheses if the number is greater than 1. TNTC indicates that bacterial load is greater than ˜200 CFU.

single colony over 80% of the time. Thus, a minimum of 5 colonies was required to be considered a true infection. This is a conservative criterion set to ensure that contamination was never called a true infection.

PCR protocol version 2 was able to detect all test bacteria in hemolymph samples whereas PCR protocol version 1 was biased against detection of certain Gram-positive bacteria. In particular, the incorporation of a lysozyme digestion step and the increase in primer concentration in PCR protocol version 2 were found to be necessary for detecting the Gram-positive bacteria *E. acetylicum*. PCR protocol version 2 was found to be relatively sensitive for detecting bacterial DNA in the hemolymph (Table 2.5). Ten of twelve infections with over 100 CFUs were PCR positive, whereas 5 of 8 infections below 100 CFUs were positive. This suggests that the ability of PCR to detect bacteria in the hemolymph may have a slight dependence on the level of infection.

UV irradiation was not efficient at removing bacterial 16S rDNA from the surface of flies (Table 2.6). In two instances, 16S rDNA sequences of test bacteria were recovered from flies after UV irradiation. In one case, the sequence con-

Table 2.5: Testing the sensitivity of hemolymph extractions for recovering bacterial 16S rDNA from hemolymph of individual artificially infected flies with different densities of infections

	PCR +		PCR -	
Bacterium	**5-100 colonies**	**>100 colonies**	**5-100 colonies**	**>100 colonies**
P. alcalifaciens (9 flies)	2	3	3	1
P. fluorescens (6 flies)	2	4	0	0
E. acetylicum (2 flies)	0	2	0	0
S. pasteuri (3 flies)	1	1	0	1

tained mutations consistent with what would be expected after UV radiation, which crosslinks thymine bases. Also, there was a high level of background bacterial 16S rDNA from an unknown source (see Table 2.6 PCR/Sequencing Results), which could be bacteria on the surface or inside the fly or contamination from PCR reagents. This result makes it difficult to say conclusively that 16S amplicons are derived from the hemolymph of the fly.

2.4.2 Culture-Dependent Survey

A total of 2,262 flies were sampled in 2005, 2006, and 2008, yielding a total of 35 isolates of bacteria that were found in high density in the culture step (Figure 2.1). *Staphylococcus* sp. 1 (which was used as a test bacterium for control experiments and is referred to as its nearest match (>99.5%) at the species level, *S. pasteuri*) was recovered the most frequently (9 times), but the majority of bacterial OTUs were recovered once or twice. The bacteria that were recovered represented 3 phyla: *Firmicutes*, *Proteobacteria*, and *Actinobacteria*. The frequency of infection varied between 0.27% in 2008 and 1.95% in 2006. Years 2006 and

Table 2.6: Testing the ability of UV irradiation (20 minutes) to remove bacterial 16S rDNA from the surface of flies

Bacterium on fly's surface	**UV**	**Culture Results**	**PCR/Sequencing Results**	**# of times observed**	**interpretation**
P. alcalifaciens	Y	-	PCR negative	6	UV successful
		-	*P. alcalifaciens*	1	UV failure
			Staphylococcus sp.	1	?
			Pseudomonas sp.	1	?
			sequence failure	2	?
	N	1-4 colonies	*P. alcalifaciens*	2	control successful
		>5 colonies	*P. alcalifaciens*	1	control successful
			sequence failure	1	control successful
		-	*P. alcalifaciens*	3	control successful
			sequence failure	1	?
P. fluorescens	Y	-	*P. fluorescens**	1	UV failure
			Ralstonia sp.	1	?
			sequence failure	2	?
	N	-	*P. fluorescens*	1	control successful
			Wolbachia pipientis	1	control successful
E. acetylicum	Y	-	PCR negative	1	UV successful
			Sphingobacterium sp.	2	?
	N	1-4 colonies	*Providencia* sp.	1	?
		>5 colonies	*Providencia* sp.	1	?
		-	PCR negative	1	control failure

The results shown are from using PCR protocol version 2.
Y indicates UV was applied.
N indicates UV was not applied.
"sequence failure" indicates a positive PCR reaction and a failure of the sequencing reaction
"UV successful" indicates that no sequence matching the bacterium on the fly's surface was obtained after UV irradiation.
"UV failure" indicates that a sequence matching the bacterium on the fly's surface was obtained after UV irradiation.
"control successful" indicates that the technique successfully detected bacteria on the fly's surface or *Wolbachia pipientis* within the fly in the absence of UV irradiation.
"control failure" indicates that the technique did not detect bacteria on the fly's surface in the absence of UV irradiation.
? indicates a sequence of unknown origin.
*UV mutation

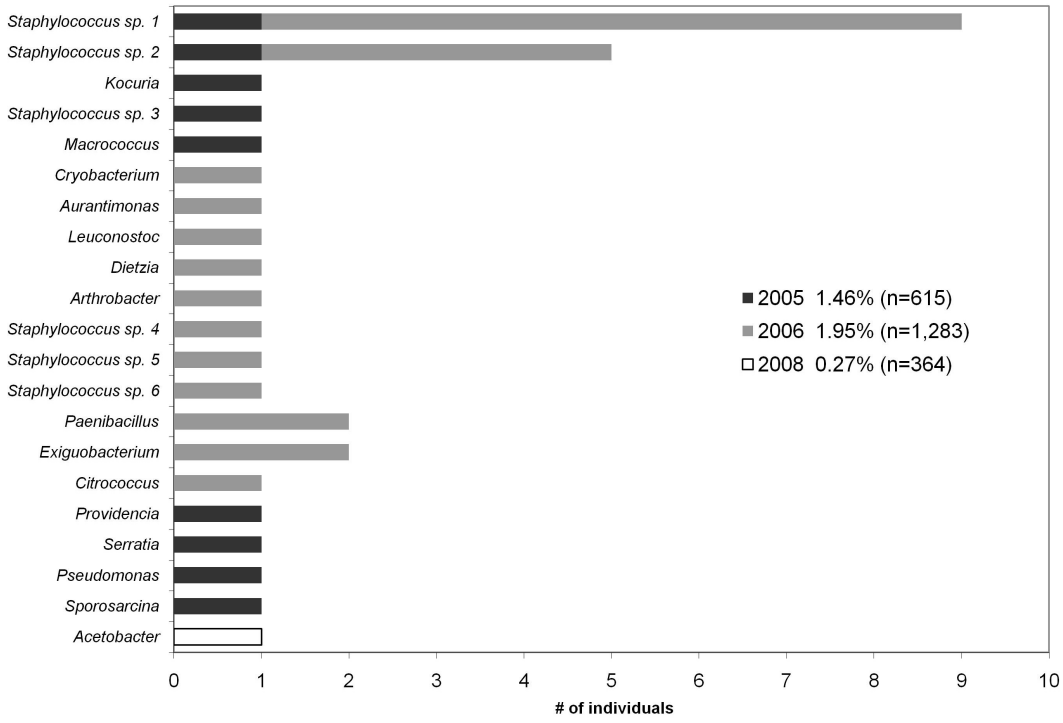


Figure 2.1: Genera of cultivable bacteria obtained from the hemolymph of wild-caught flies in 2005, 2006, and 2008

2008 had significantly different rates of infection (Fisher's exact test, p=0.0174), although this difference was not significant after Bonferroni correction. In 2005, the highest number of isolates was obtained in September while in 2006 the highest number of isolates was obtained in July (Figure 2.2). There was no significant difference in the number of isolates obtained from flies with and without *W. pipientis*, for cases where the *W. pipientis* infection status was known.

The virulence of bacteria recovered from the hemolymph of wild-caught flies varied greatly upon experimental reinfection, with some bacteria causing less than 10% mortality and others causing greater than 40% mortality in experimentally infected flies (Figure 2.3). Bacteria that were recovered with the greatest frequency were the least virulent.

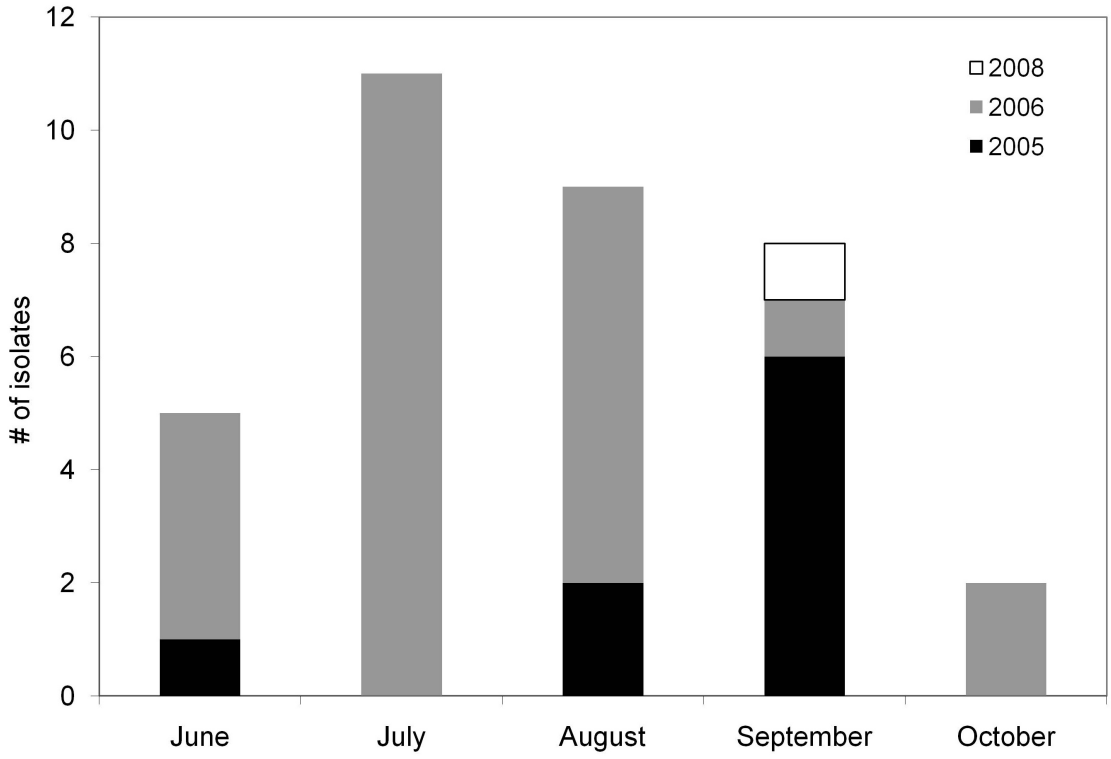


Figure 2.2: Number of isolates of cultivable bacteria obtained from the hemolymph of wild-caught flies in 2005, 2006, and 2008

A phylogenetic tree of cultivable bacteria in the phylum *Firmicutes* shows the relationship of samples obtained from the hemolymph of wild-caught flies to their nearest type species and to cultivable bacteria obtained from negative controls (Figure 2.4). In two cases, sequences obtained from negative controls closely match those obtained from the hemolymph of wild-caught flies.

2.4.3 Culture-Independent Survey

Samples collected in 2005 were not analyzed by PCR because the use of non-UV irradiated water for sample collection was found to introduce DNA contami-

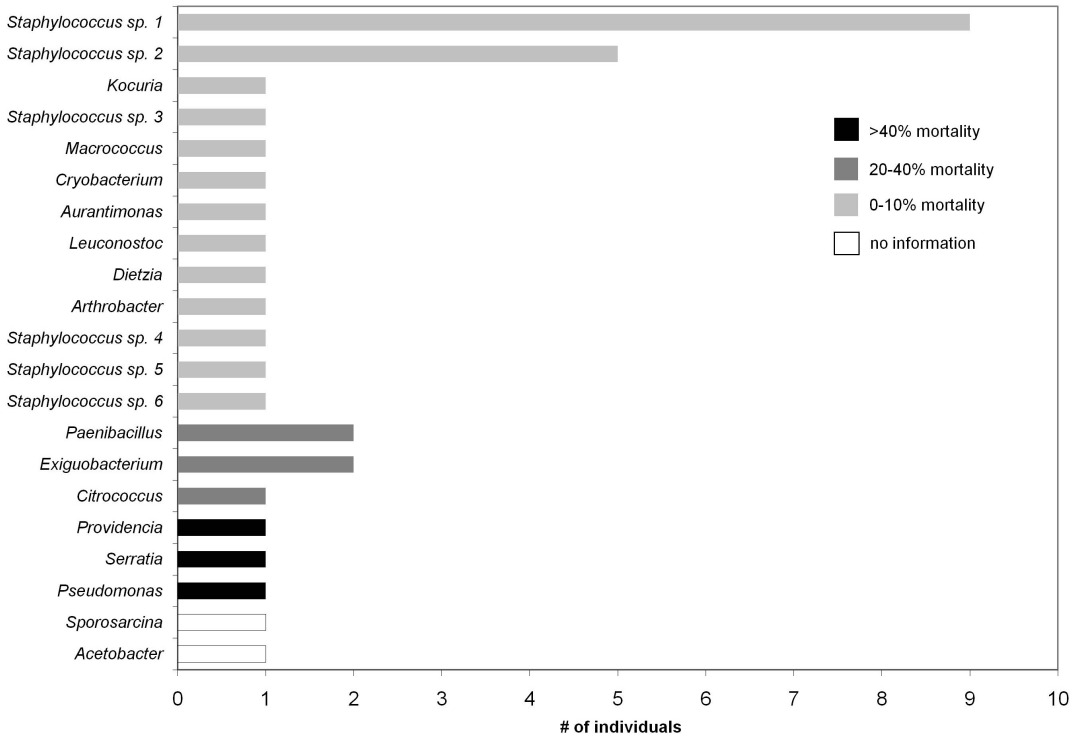


Figure 2.3: Virulence of cultivable bacteria obtained from the hemolymph of wild-caught flies. Oregon R lab flies were artificially infected for each isolate, and bacteria were classified as causing low (0-10%), intermediate (20-40%) or high (>40%) levels of mortality.

nation. Samples collected in 2006 were analyzed using PCR protocol version 1 and t-RFLP. The only bacterium that was identified using this approach was *W. pipientis*, which was present in 64% of males and 57% of females that were sampled (n=99 males, n=54 females). This frequency of infection by *W. pipientis* has previously been seen in other populations of *D. melanogaster* (Hoffmann et al., 1998). The primers used in PCR protocol version 1 were a perfect match to *Wolbachia* and this result probably reflects a bias towards the recovery of this bacterium.

The poor concordance of PCR results with culture-based results led to the

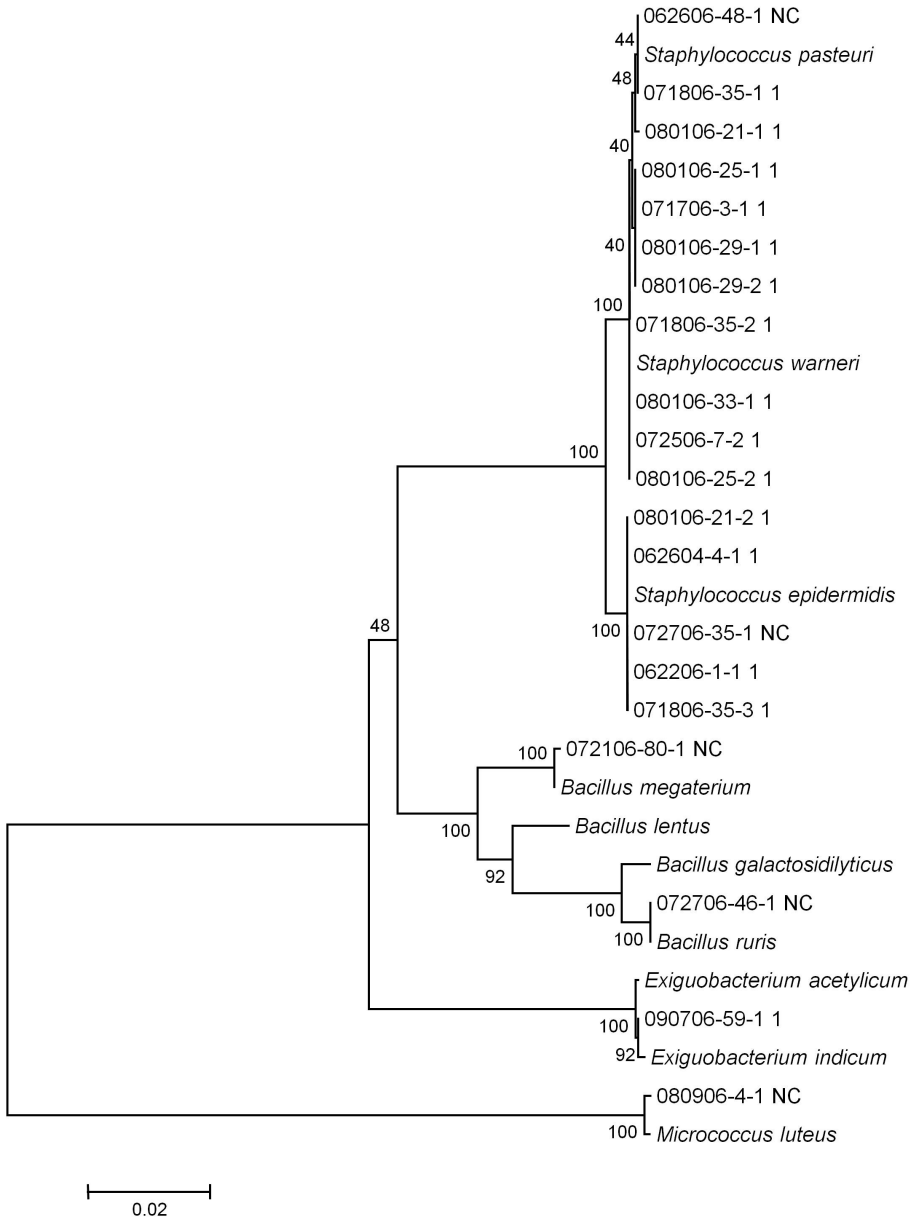


Figure 2.4: Neighbor-joining tree of the bacterial phylum *Firmicutes* showing the relationship of cultivable isolates obtained from the hemolymph with nearest type species and negative control isolates. All isolates are labeled with a code that specifies the date of collection, the fly number on the day of collection, and the isolate number within the fly. Negative controls (NC), which occasionally yielded fewer than fives colonies, are indicated to show the relationship of contaminants to hemolymph isolates. The bar represents 2% sequence divergence.

Table 2.7: Bacteria associated with wild-caught flies and identified using PCR. A total of 77 flies were sampled on August 16, 2008. Fourteen hemolymph samples were PCR positive and could be directly sequenced (shown here). Ten hemolymph samples were PCR positive but were mixed or degraded and could not be directly sequenced.

Phylum	**Family**	**Genus**	**# of samples**
Proteobacteria	*Rhodocyclaceae*	?	1
	Pseudomonadaceae	*Pseudomonas*	2
	Anaplasmataceae	*Wolbachia*	4
	Incertae sedis	*Enhydrobacter*	1
	Enterobacteriaceae	*Serratia*	1
	Enterobacteriaceae	?	1
	Burkholderiaceae	*Ralstonia*	1
	Incertae sedis 5	?	1
Actinobacteria	*Dermabacteraceae*	*Brachybacterium*	1
Firmicutes	*Lactobacillaceae*	*Lactobacillus*	1

? indicates that bacterium could not be identified to the genus level.

development of PCR protocol version 2 that was much more efficient at recovering test bacteria in control experiments. The primers used in this protocol were degenerate and thus less likely to be biased towards *W. pipientis* detection. Samples collected in 2008 were analyzed using PCR protocol version 2 and direct sequencing. There was a high rate of contamination in the negative controls of samples analyzed from this year which likely reflects the increased sensitivity of this technique. The frequency and identities of the bacteria that were identified using this method are shown in Table 2.7 from a single collection day (August 16, 2008) where all 11 negative controls were blank. None of the samples collected on this day yielded culture-positive results so no conclusion can be drawn about the performance of PCR at recovering cultivable bacteria. On this day, 56 samples were PCR negative, 14 samples were PCR positive and could be directly sequenced, and 9 samples were PCR positive but were mixed or degraded and could not be identified by direct sequencing. Because

the control experiments showed that DNA sequences recovered by PCR cannot be conclusively said to be coming from the hemolymph of flies, some of these DNA sequences may be coming from the surface of the fly. The recovery of *W. pipientis*, which cannot be a surface contaminant, suggests that at least some of the other sequences came from the hemolymph. *Lactobacillus*, *Pseudomonas*, and *Serratia* species have previously been reported in association with the gut or whole body of *D. melanogaster* (Cox and Gilmore, 2007; Corby-Harris et al., 2007), and I recovered *Pseudomonas* and *Serratia* species using the culture-based methods.

PCR protocol 2 recovered a wider array of bacteria than PCR protocol 1, suggesting that the modifications that were made improved the procedure. The frequency of *W. pipientis* recovered from field samples in 2006 using PCR protocol 1 was between 57% and 64%. Using PCR protocol 2, *W. pipientis* was confirmed to be present in only 5% of samples collected in 2008, but may have been present in as much as 18% of the samples if mixed or degraded PCR positive samples that were not identifiable by direct sequencing also contained *W. pipientis*. Fluctuations in *W. pipientis* prevalence are not unusual in natural populations of *D. melanogaster* (Hoffmann et al., 1998), and thus it is not clear if the difference in *W. pipientis* frequencies between 2006 and 2008 reflects variations in years or protocols. However, PCR protocol 2 appears to be substantially less biased towards *W. pipientis* recovery than PCR protocol 1.

2.5 Discussion

This study identified cultivable bacteria representing 21 OTUs from the hemolymph of wild-caught flies. The performance of the culturing method was found to be sensitive for detecting a wide range of bacteria provided that they were present in high densities within the fly. The culturing method was able to distinguish between bacteria inside and outside the fly. I developed methods for detecting bacteria from hemolymph samples using a nested PCR approach. The final protocol was sensitive for detecting bacteria in the hemolymph. Unfortunately, the sterilization technique was not 100% efficient at removing bacterial DNA from the surface of flies so bacteria identified by PCR could not conclusively be demonstrated to be coming from the hemolymph of flies.

The culturing method is likely to be good at detecting many of the bacteria that are present in the hemolymph. Environments such as soil and water with notoriously low rates of concordance between culture and non-culture-based experiments are generally nutrient poor (Handelsman, 2004). In contrast, insect hemolymph is nutrient rich and should be well represented by the artificial media that were used for collecting cultivable bacteria. In a survey of bacteria from the midguts of gypsy moths using culture and PCR methods, (Broderick et al., 2004) found that roughly two thirds of the bacteria identified were culturable. Surveys of bacteria causing blood infections in humans find moderate to high concordance between culture and PCR methods (Bloos et al., 2010; Tsalik et al., 2010). Thus, it is likely that my culturing technique recovered many of the bacteria that are present in the hemolymph of the flies sampled here.

The PCR method developed here is very sensitive for recovering bacterial

DNA from hemolymph samples of individual flies. Lysozyme digestion of cell walls and use of degenerate primers for PCR amplication greatly increased the breadth of recovery of diverse bacterial phyla. Unfortunately, UV irradiation alone was not adequate for removing bacterial DNA from the cuticle of flies, and therefore we were unable to distinguish between the hemolymph and outside of flies. If a more robust sterilization technique were developed, perhaps using hypochlorite in combination with UV irradiation, then this PCR technique could be used to identify bacterial DNA localized in the hemolymph. Also, since this technique allows amplification of bacterial DNA from complex samples containing PCR inhibitors, it would be useful for other applications where the sample volume is too small to allow DNA extraction.

Over 98% of all *D. melanogaster* sampled in the course of this study did not have cultivable bacteria in their hemolymph. Between 0.27% and 1.95% of flies were infected in any given year, and there was a significant difference in the infection rate between two of the years which may represent variation in environmental conditions that influenced bacterial growth. The low infection rates observed are consistent with the idea that the hemolymph is free of bacteria in healthy insects (Bakula, 1969). Although this idea has long been held as fact, recent studies have provided evidence that this conventional wisdom is not always true. For example, cultivable *Bacillus* and *Staphylococcus* species are routinely recovered from the hemolymph of larval and adult *Solenopsis* fire ants (Tufts and Bextine, 2009). Another study identified cultivable *Novosphingobium*, *Escherichia*, *Pseudomonas*, and other species from the hemolymph of lab colonies of *Anopheles gambiae* (Garver et al., 2008). *D. melanogaster domino* mutants, which lack hemocytes, have high levels of bacteria, including *Staphylococcus* and *Providencia* species, in their hemolymph (Braun et al., 1998). This

suggests that perhaps the presence of bacteria in the hemolymph is associated with reduced immune competence. Our study implies that the hemolymph only rarely harbors non-*W. pipientis* bacteria in *D. melanogaster* and suggests that the immune system does an adequate job at keeping the hemolymph free of infection.

The virulence estimates of the cultivable bacteria suggest that the sampling strategy recovered more weakly virulent than highly virulent bacteria. This most likely represents a bias in our study for the recovery of less pathogenic bacteria since our sampling strategy required that flies be in flight and therefore relatively healthy. Presumably, flies infected with very virulent bacteria would become less active and thus not be sampled as often. It is also possible that the virulence levels obtained from experimentally infecting lab-reared flies do not reflect actual pathology experienced by flies living in the wild, which may be subject to a multitudes of additional stresses.

We identified bacteria belonging to the phyla *Actinobacteria* (n=5), *Firmicutes* (n=11), and *Proteobacteria* (n=5). Two previous studies have thus far also examined the bacteria associated with wild-caught flies using mainly PCR-based methods (Cox and Gilmore, 2007; Corby-Harris et al., 2007) (Figure 2.5). Both studies found a prevalence of *Proteobacteria* and also identified *Bacteriodetes* associated with the guts and whole bodies of flies. (Corby-Harris et al., 2007) did not find *Actinobacteria* associated with the wild-caught flies. The low proportion of *Proteobacteria* identified in our study relative to previous studies may reflect differences between sampling techniques. For example, DNA extraction methods and PCR amplification can bias against recovery of Gram-positive bacteria such as *Actinobacteria* and *Firmicutes* (Corby-Harris et al., 2007). Alternatively,

several of the *Proteobacteria* identified in our study included *Serratia*, *Providencia*, and *Pseudomonas*, which were all found to be highly virulent and thus are less likely to be sampled using our sweep net sampling strategy. Of the 16 genera identified in the culture-based survey, *Staphylococcus*, *Leuconostoc, Providencia*, *Serratia*, *Pseudomonas*, and *Acetobacter* have previously been reported in association with either the gut or whole body of *D. melanogaster*. The concordance of the bacteria between the hemolymph, gut and whole fly suggests that these bacteria are being acquired from the environment, either through the gut or the cuticle. This also suggests that the community of bacteria to which flies are exposed to depends on the environment it lives in. The fact that infection is rare suggests that these bacteria opportunistically cause infections when the host is immunocompromised.

The array of bacteria that was identified in this study suggests that wild flies face risk of infection from a diverse rather than narrow spectrum of bacteria. We have also not found any evidence for a prevalent co-evolving bacterial pathogen such as *Pasteuria*, which is found in *Daphnia*, or of bacteria closely related to previously described mutualists of other insects, other than the previously described *W. pipientis*. Both of these findings have implications for our expectations about the evolution of immune response in *D. melanogaster*. The diversity of potential pathogens suggests selection for a broad rather than specific antibacterial immune response. Recognition proteins within the host humoral immune system respond to a vast array rather than a specific subset of potential pathogens as is expected given the breadth of potential pathogens that they face. Host-pathogen co-evolutionary arms races are expected to lead to high rates of adaptive evolution. The lack of a tightly co-evolving bacterial pathogen is perplexing given the observation that the immune response is rapidly evolv-

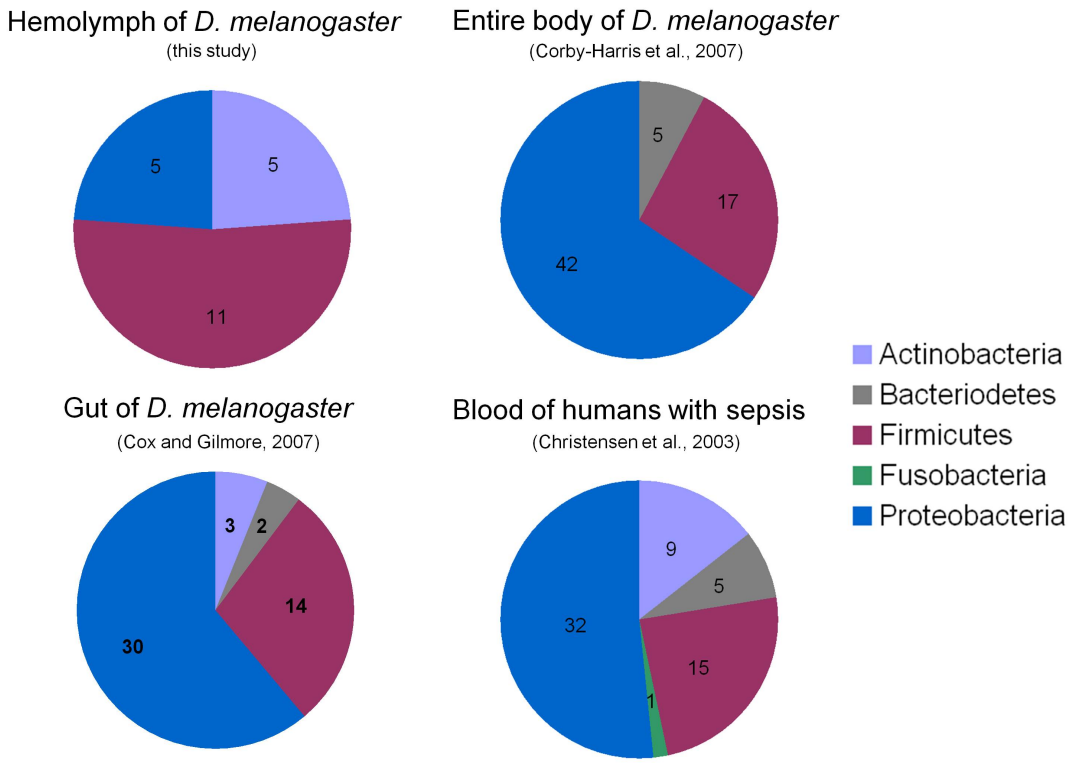


Figure 2.5: Distribution of bacterial phyla recovered in association with *D. melanogaster* and from the blood of humans with sepsis. The numbers in the pie charts indicate the number of isolates belonging to each phylum.

ing (Sackton et al., 2007), and suggests that viral and eukaryotic pathogens and parasites with more narrow host ranges may be important drivers of rapid evolution in the immune system. Our understanding of the evolution of the antibacterial immune response is limited by our lack of knowledge of how natural pathogens interact with hosts, and further study of these natural pathogens should elucidate mechanisms of adaptation in the host immune response.

2.6 Acknowledgements

Thanks to Julie Frey for advice on t-RFLP and for providing primers and primer sequences. Thanks to Dennis Hartley for allowing me to collect flies at Littletree Orchard and for providing entertaining conversation and apple cider donuts. Thanks to Ann Hajek for helpful comments and suggestions on this chapter.

CHAPTER 3

PROVIDENCIA SNEEBIA SP. NOV. AND *P. BURHODOGRANARIEA* SP. NOV., NOVEL SPECIES ISOLATED FROM WILD *DROSOPHILA MELANOGASTER**

3.1 Abstract

Multiple isolates of the genus *Providencia* were obtained from the hemolymph of wild-caught *Drosophila melanogaster* fruit flies. Sixteen isolates were distinguished from the six previously described species based on 16S rDNA sequence. These isolates belong to two distinct groups, which we propose each comprise previously undescribed species. Two isolates, designated A^T and B^T, were characterized by DNA sequence at the *fusA*, *lepA*, *leuS*, *gyrB*, and *ileS* housekeeping genes, whole-genome DNA-DNA hybridizations with their nearest relatives, and utilization of substrates for metabolism. The closest phylogenetic relatives of A^T are B^T (86.9% identity at the housekeeping genes) and *P. stuartii* DSMZ 4539^T (86.0% identity). The closest phylogenetic relatives of B^T are A^T (86.9% identity) and *P. stuartii* DSMZ 4539^T (86.6% identity). Described species in this genus share between 84.1% and 90.1% identity. DNA-DNA relatedness between A^T-B^T, A^T-*P. stuartii*, and B^T-*P. stuartii* all resulted in less than 25% hybridization. In addition, patterns of utilization of amygdalin, arbutin, esculin, salicin, D-sorbitol, D-trehalose, D-inositol, D-adonitol and D-galactose distinguish A^T and B^T from other members of this genus. A^T and B^T therefore represent novel species, for which the names *Providencia sneebia* sp. nov. (A^T =DSM 19967

* Presented with minor modifications from the originally published article "Juneja, P., Lazzaro, B. P., 2009. *Providencia sneebia* sp. nov. and *P. burhodogranariea* sp. nov., novel species isolated from wild *Drosophila melanogaster*. *International Journal of Systematic and Evolutionary Microbiology* **59**(5):1108-11."

=ATCC BAA-1589) and *P. burhodogranariea* sp. nov. (B^T=DSM 19968 =ATCC BAA-1590) are proposed.

3.2 Introduction

The genus *Providencia*, in the family Enterobacteriaceae, currently has six described species. Members of this genus have repeatedly been found in association with humans, insects, and many other vertebrate and invertebrate animals in both pathogenic and nonpathogenic contexts (Penner and Hennessy, 1979; Muller et al., 1986; Yoh et al., 2005; Somvanshi et al., 2006). We describe here two novel *Providencia* species isolated as from the hemolymph of field-captured *Drosophila melanogaster* fruit flies.

3.3 Materials and Methods

D. melanogaster were collected in State College, Pennsylvania, USA, in 1998 and 2001. Individual flies were surface sterilized by UV irradiation prior to hemolymph extraction with pulled glass microcapillary needles. The hemolymph was used to inoculate 1 mL liquid cultures of brain heart infusion (BHI). Liquid cultures were grown aerobically for 24 hours at 37°C. Enriched cultures were then streaked on BHI agar plates, and individual colonies were selected for identification. Seventeen out of 337 total *D. melanogaster* yielded bacterial isolates assignable to the genus *Providencia* based on sequence at the 16S rDNA, amplified using primers fd1 and rp2 as described by Weisburg et al. (1991). These isolates clustered into four groups based on 16S rDNA sequence

(978 nucleotides), termed "A", "B", "C" and "D" (Figure 3.1). The sole isolate belonging to group "C" was subsequently identified as *P. rettgeri* based on metabolic profile and DNA sequence at housekeeping genes (described below). The remaining isolates did not closely match described species. Isolates A^T, A75, A91, A101, A102, B^T, B18, B97, D, D37 were chosen for further characterization. These isolates were compared to *Providencia* type strains *P. stuartii* DSM 4539^T, *P. alcalifaciens* DSM 30120^T, *P. heimbachae* DSM 3591^T, *P. rustigianii* DSM 4541^T, *P. rettgeri* DSM 4542^T, and *P. vermicola* DSM 17385^T, all obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ; Braunschweig, Germany).

Metabolic characteristics were determined from single colonies grown for 24 hours at 37°C on Luria broth agar plates. Metabolic profiles were determined with API 20E and API 50CH test strips (BioMérieux, Inc., Marcy-l'Etoile, France). Inocula for API20E strips were prepared in distilled water, and inocula for API 50CH strips were prepared using the API 50 CHB/E medium. Assays were interpreted after 30 hour incubations at 25°C. Test strips were run at least twice for A^T, B^T, and all described type species except *P. rustigianii*. Test strips were run once for *P. rustigianii* and non-type strains of the novel species.

3.4 Results and Discussion

The metabolic profiles of novel isolates were distinct from each other and from all described *Providencia* species (Table 3.1). Results for previously described species varied slightly from published reports by (Hickman-Brenner et al., 1983; Farmer 3rd et al., 1985; Somvanshi et al., 2006). These deviations may be due to

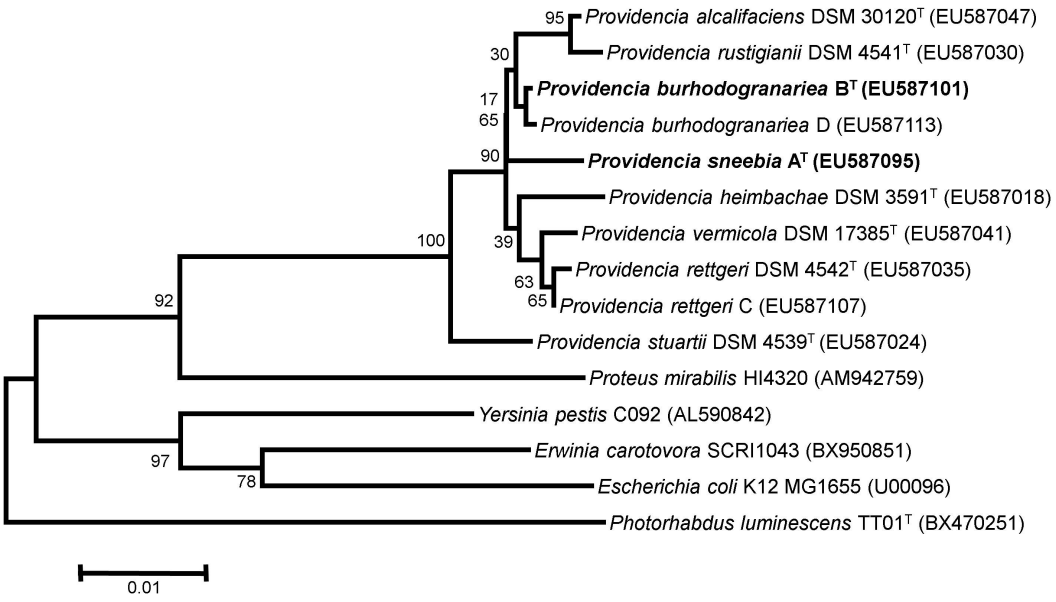


Figure 3.1: Phylogenetic tree based on sequence from the 16s rDNA (978 nucleotides), showing the positions of novel type species within the genus *Providencia*. Sequences for *Proteus mirabilis*, *Photorhabdus luminescens*, *Yersinia pestis*, *Erwinia carotovora*, and *Escherichia coli* were obtained from GenBank. Bootstrap values on each node are based on 1000 replicates. The positions of *P. heimbachae* and *P. stuartii* in this phylogeny are reciprocally exchanged relative to those in a previously published tree based on this locus (Somvanshi et al., 2006), though the topology of the trees is the same. Sequences directly obtained in this study match those deposited into GenBank by Somvanshi et al. (2006). The discrepancy therefore appears to result from a clerical error in the construction of the tree by Somvanshi et al. (2006). Bar indicates 1% sequence divergence.

differences in the test methods, temperatures, and reaction durations employed by different authors. Isolates A^T, A75, A91, A101, and A102 all had identical metabolic profiles. Isolates B^T, B18, B97, D, and D37 had identical metabolic profiles, except for amygdalin, L-rhamnose, and D-sorbitol, for which there was variation among isolates in their substrate utilization (Table 3.1). The *Providencia* species represented by A^T is uniquely able to utilize D-sorbitol and D-xylose. In contrast to all previously described species of *Providencia*, isolates A^T and B^T are able to utilize D-trehalose.

Partial sequences of five housekeeping genes (*fusA*, 616 nucleotides; *lepA*, 735 nucleotides; *leuS*, 412 nucleotides; *gyrB*, 817 nucleotides; *ileS*, 920 nucleotides) were obtained from the ten novel isolates and from the six described *Providencia* type species. Housekeeping genes were amplified and sequenced using a combination of degenerate primers described by (Santos and Ochman, 2004) and *Providencia*-specific custom primers (Table 3.2). PCR products were prepared for sequencing by one hour incubation with the enzymes exonuclease I (USB Corporation) and shrimp alkaline phosphatase (USB Corporation) and sequenced using ABI BigDye Terminator chemistry on an Applied Biosystems Automated 3730 DNA Analyzer. In some cases, amplicons were agarose gel purified prior to sequencing. Sequences were aligned using CodonCode Aligner (CodonCode Corporation). Phylogenetic analysis was performed using MEGA version 3.1 (Kumar et al., 2004) both on alignments of individual genes and on an alignment of a concatenation of all six genes. Distances were calculated based on Jukes-Cantor corrected percent divergence, and clustering was performed by neighbor-joining. Bootstrap values from 1,000 replications were used to assess confidence at each node.

Table 3.1: Differentiation of *Providencia* strains based on metabolic substrate reactions

All bacteria were tested under aerobic conditions. Multiple isolates of *P. sneebia* (A^T, A75, A91, A101 and A102) and *P. burhodogranariea* (B^T, B18, B97, D, and D37) were tested and yielded virtually identical results (see main text). Each strain was tested at least twice, except *P. rustigianii* and non-type strains of the novel species, which were each tested once. All strains were positive for tryptophane deaminase‡ and oxidation/fermentation of D-glucose‡§, glycerol§, D-ribose§, D-fructose§, D-mannose§, N-acetylglucosamine§, and potassium gluconate§. All strains were negative for β-galactosidase‡, arginine dihydrolase‡, lysine decarboxylase‡, ornithine decarboxylase‡, H_2S production‡, acetoin production‡ (given as positive for all strains in (Somvanshi et al., 2006), gelatinase‡, and oxidation/fermentation of D-melibiose‡§, D-arabinose§, L-arabinose§, L-xylose§, methyl β-D-xylopyranoside§, L-sorbose§, dulcitol§, methyl α-D-mannopyranoside§, methyl α-D-glucopyranoside§, amygdalin§, D-cellobiose§, D-lactose§, D-saccharose§, inulin§, D-melezitose§, D-raffinose§, amidon§, glycogen§, gentiobiose§, D-turanose§, D-tagatose§, D-fucose§, L-fucose§, and potassium 5-ketogluconate§. Test strips did not give repeatable results within isolates for citrate utilization‡ and indole production‡ so results from these assays are not presented. (Key: +, positive; –, negative; v, variable between strains; ?, variable within isolate; w, weak)

Characteristic	1	2	3	4	5	6	7	8	9
urease	+	-	-	?*	+	-	-	-	-
utilization of:									
L-arabinose‡	-	+	+	+	w*	-	w*	-	-
D-adonitol§	-	+	+	+	+	-	-	+	+
amygdalin‡	+	-	v	-	+	-	-	-	-
D-arabitol§	+	+	+	+	+	-	-	+	-
L-arabitol§	-	-	-	+	+	-	-	+*	-
arbutin§	+	-	-	-	+	-	-	-	-
erythritol§	-	-	-	+	?*	-	-	+*	-
esculin§	+	-	-	-	+*	-	-	-	-
D-galactose§	-	-	-	+	+	+	+	+	-
inositol‡	-	+	+	+	+	w	-	+	-
D-inositol§	-	+	+	+	+*	+	-	+	-
2-ketogluconate§	-	w	+	+	+*	-	-	+*	-
D-lyxose§	-	-	-	-*	-*	+*	-	-*	-
D-maltose§	-	-	-	-	-	-	-	+*	-
D-mannitol‡	+	+	+	+	+	-	-	-	-
D-mannitol§	+	+	+	+	+	-	-	-*	-
L-rhamnose‡	-	-	v	-	-	+	-	-	+
L-rhamnose§	-	-	-	-	+	-	-	+	-
salicin§	+	-	-	-	+	-	-	-	-
D-sorbitol‡	+	v	v	-	-	-	-	-	-
D-sorbitol§	+	-	-	-	-	-	-	-	-
D-sucrose‡	-	?	?	-	-	-	-*¶	-	-*
D-trehalose§	+	+	+	-	-	-#¶	-	-	-
xylitol§	-	-	-	-	-	+	-	-	-
D-xylose§	w	-	-	-	-	-	-	-	-

‡API 20 E test strip

§API 50 CH test strip

*results differ from those previously published by Somvanshi *et al.*, 2006

#results differ from those previously published by Farmer *et al.*, 1985

¶results differ from those previously published by Hickman-Brenner *et al.*, 1983

Table 3.2: PCR primer sequences for specific amplification of *Providencia* sp., *P. burhodogranariea*, or *P. sneebia* housekeeping genes

Gene	Target Organism	Primer Name	Primer Sequence (5′-3′)
leuS	*Providencia* sp.	leuS-prov-F	TGCTGGCGYTGTGAYAC
		leuS-prov-R	AAACACCCCAGTCACG
	P. burhodogranariea	leuS-provB-F	ATCACCTTTGATGTCGCTGA
		leuS-provB-R	GAAACACCCCAATCACGTAAA
	P. sneebia	leuS-provA-F	GTTTACACAACGCGTCCAGA
		leuS-provA-R	AACACCCCAATCACGTAAGC
fusA	*Providencia* sp.	fusA-prov-F	GGACTGGATGGAGCAGGA
		fusA-prov-R	TGCAGAACCACAGGTAACCA
lepA	*P. burhodogranariea*	lepA-provB-F	AATGGCGTCTCAGGTTCTTG
		lepA-provB-R	GCATCACGAAATGATTCATAA
	P. sneebia	lepA-provA-F	CCGTATTATTCAGATTTGTGGTG
		lepA-provA-R	TGACTGGGAATAAACCTGCAT
ileS	*Providencia* sp.	ileS-prov-F	CCGATTGAACACAAAGTTGAA
		ileS-prov-R	AGATCCCACCATGCTTGA
gyrB	*Providencia* sp.	gyrB-prov-F	TATCGGTGATACCGACGATGG
		gyrB-prov-R	CGCARTTTATCTGGGTT

The percent divergence among described *Providencia* species across the concatenated housekeeping genes ranged from 9.9% between *P. alcalifaciens* and *P. rustigianii* to 15.9% between *P. stuartii* and *P. heimbachae* (Figure 3.2). Within the ”A” group of novel isolates, the maximum observed divergence was 0.3%. Based on this level of sequence similarity, A^T, A75, A91, A101, and A102 are considered to belong to the same species. Within the ”B” and ”D” groups of isolates, the maximum observed divergence was 0.2%. The ”B” and ”D” groups of isolates differed by only 6.3%, lower than the minimum divergence observed between described *Providencia* type species. Based on their sequence similarity and nearly identical metabolic profiles, B^T, B18, B97, D, and D37 are considered to belong to the same species. B^T differed from A^T by 13.1% and *P. stuartii* by 13.4%. A^T differed from *P. stuartii* by 14.0%. These percent divergences fall well above the minimum percent divergence observed between described *Providencia* species, supporting the hypothesis that A^T and B^T represent distinct and

novel *Providencia* species.

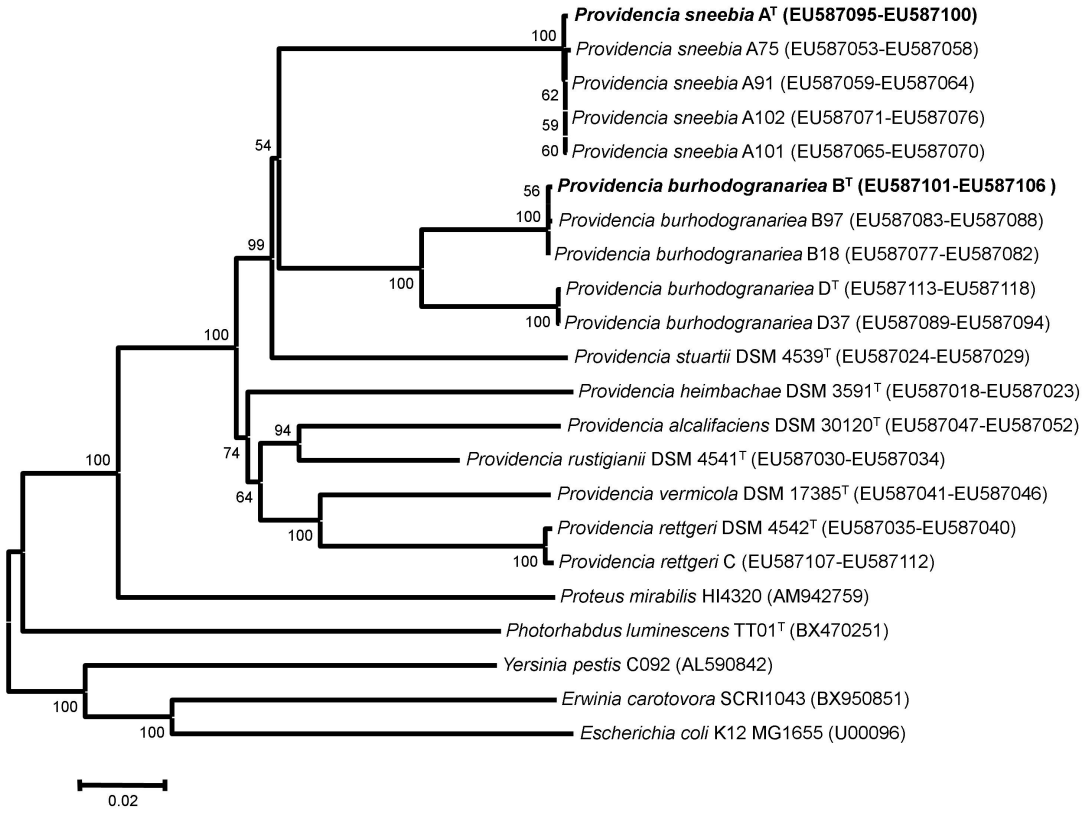


Figure 3.2: Phylogenetic tree based on concatenated sequence from the 16S rDNA, *fusA*, *lepA*, *leuS*, *gyrB*, and *ileS* loci (4478 nucleotides) showing the positions of novel type species within *Providencia*. Sequences for *Proteus mirabilis*, *Photorhabdus luminescens*, *Yersinia pestis*, *Erwinia carotovora*, and *Escherichia coli* were obtained from GenBank, and the accession numbers are given in parentheses. Bootstrap values on each node are based on 1000 replicates. Bar indicates 2% sequence divergence.

DNA-DNA hybridizations were performed between all pairs of A^T, B^T, and *P. stuartii* DSM 4539^T, which was inferred to be the nearest described relative based on DNA sequence at the housekeeping genes. DNA isolations and DNA-DNA hybridizations were performed by the DSMZ following the methods described by (De Ley et al., 1970; Cashion et al., 1977; Huss et al., 1983). The re-

association value for the A^T-*P. stuartii* DSM 4539^T pairing was 13.0% (average of 12.7% and 13.3% obtained from two separate measurements). The reassociation value for the B^T-*P. stuartii* DSM 4539^T pairing was 18.8% (average of 13.8% and 23.8% obtained from two separate measurements). The reassociation values for the A^T-B^T pairing was 13.1% (average of 7.6% and 18.6% obtained from two separate measurements). Previous studies have reported reassociation values that range from 22 to 49% between species in this genus (Hickman-Brenner et al., 1983; Muller et al., 1986; Somvanshi et al., 2006). The results from the DNA-DNA hybridizations fall well below the 70% reassociation threshold recommendation of Wayne et al. (1987) for designation of a new species. These results indicate that A^T and B^T are significantly distinct from each other and from their nearest described relative in *Providencia*.

The results from the sequence, hybridization, and metabolic analysis meet the requirements outlined by Wayne et al. (1987) for designating a bacterial species as novel. Based on these results, A^T, A75, A91, A101, and A102 belong to a single novel species for which the name *Providencia sneebia* sp. nov. is proposed. Likewise, B^T, B18, B97, D, and D37 belong to a single novel species for which the name *Providencia burhodogranariea* sp. nov. is proposed.

3.5 Description of *Providencia sneebia* sp. nov.

Providencia sneebia (snee'bia. N.L. fem. adj. *sneebia* of "S.N.E.E.B.," the name of a series of informal academic gatherings at Cornell University where properties of these bacteria were extensively discussed).

This species, like others in the genus *Providencia*, is a Gram-negative rod-

shaped bacterium. Colonies grown on LB agar for 48 hours at 37°C are up to 4 mm in diameter, white, opaque, glossy, and convex. Growth occurs faster at 37°C than at 25°C. *P. sneebia* is unique in the genus *Providencia* for being able to produce acid from amygdalin, arbutin, esculin, salicin, D-xylose, D-sorbitol and D-trehalose but not from D-inositol, D-adonitol or D-galactose.

The type strain is of *Providencia sneebia* is A^T (=ATCC BAA-1589^T =DSM 19967^T). This and other non-type strains were isolated from *Drosophila melanogaster* captured in an apple orchard in State College, Pennsylvania.

3.6 Description of *Providencia burhodogranariea* sp. nov.

Providencia burhodogranariea (bu.rho.do.gran.ar'iea. L. n. *granaria* barn; pref. *bu-*, *rhodo-* meaning big, red; Big Red Barn being the name of the building where academic discussions of these bacteria were held).

This species, like others in the genus *Providencia*, is a Gram-negative rod-shaped bacterium. Colonies grown on LB agar for 48 hours at 37°C are up to 4 mm in diameter, white, opaque, glossy, and convex. Growth occurs faster at 37°C than at 25°C. After 24 to 48 hours of growth, *P. burhodogranariea* colonies express brown pigmentation in their centers. *P. burhodogranariea* is unique in the genus *Providencia* for being able to produce acid from D-adonitol, D-trehalose and D-inositol but not from D-galactose.

The type strain of *Providencia burhodogranariea* is B^T(=ATCC BAA-1590^T =DSM 19968^T). This and other non-type strains were isolated from *Drosophila melanogaster* captured in an apple orchard in State College, Pennsylvania.

3.7 Acknowledgements

We would like to thank Madeline Galac and Martin Wiedmann for helpful discussion and advice. We would like to thank the *Proteus mirabilis* Sequencing Group at the Sanger Institute for early access to the *P. mirabilis* genome sequence. This work was supported by NSF grant DEB-0415851.

CHAPTER 4

HAPLOTYPE STRUCTURE AND EXPRESSION DIVERGENCE AT THE *DROSOPHILA* CELLULAR IMMUNE GENE *EATER**

4.1 Abstract

The protein Eater plays an important role in microbial recognition and defensive phagocytosis in *Drosophila melanogaster*. We sequenced multiple alleles of the *eater* gene from an African and a North American population of *D. melanogaster* and found signatures of a partial selective sweep in North America that is localized around the second intron. This pattern is consistent with local adaptation to novel selective pressures during range expansion out of Africa. The North American sample is divided into two predominant haplotype groups, and the putatively selected haplotype is associated with a significantly higher gene expression level, suggesting that gene regulation is a possible target of selection. *eater* alleles contain from 22 to 40 repeat units that are characterized by the presence of a cysteine-rich NIM motif. NIM repeats in the structural stalk of the protein exhibit concerted evolution as a function of physical location in the repeat array. Several NIM repeats within *eater* have previously been implicated in binding to microbial ligands, a function which in principle might subject them to special evolutionary pressures. However, we find no evidence of elevated positive selection on these pathogen-interacting units. Our study presents an instance where gene expression rather than protein structure is thought to drive the adaptive evolution of a pathogen recognition molecule in the immune sys-

* Presented with minor modifications from the originally published article "Juneja, P., Lazzaro, B. P., 2010. Haplotype structure and expression divergence at the *Drosophila* cellular immune gene *eater*. *Molecular Biology and Evolution* **27**(10):2284-99." Permission to reproduce article obtained from the Society of Molecular Biology and Evolution.

tem.

4.2 Introduction

All living organisms have a vital need to protect themselves against pathogenesis, and hosts are thus constantly being forced to adapt their defenses to novel and reciprocally evolving pathogens and parasites (Ebert, 2000). Population genetic analyses can answer questions about the role of local adaptation in driving rapid evolution and the geographic distribution of selected alleles, as well as help determine the relative importance of selection on standing genetic variation versus on novel variants introduced by mutation. Additionally, population geneticists studying host-pathogen relationships can localize the specific targets of selection within proteins and determine whether these correspond to domains that interact directly with pathogens. In the present paper, we address these questions with respect to the evolution of the *eater* gene of *Drosophila melanogaster*.

The gene *eater* encodes a recognition receptor that is critical for defensive phagocytosis (Kocks et al., 2005), an important first line of protection against invading microbes. In *D. melanogaster*, *eater* is expressed solely in hemocytes and is thought to be a cell-surface bound molecule that binds to microbial compounds and stimulates phagocytosis (Kocks et al., 2005). Ablation of this single gene with RNAi knockdown can decrease phagocytosis by 55-70% (Kocks et al., 2005). *eater* is part of the recently described *nimrod* superfamily of cellular recognition molecules that also includes multiple *nimrod* homologues and *draper* (Kocks et al., 2005; Kurucz et al., 2007; Somogyi

et al., 2008). Proteins in the nimrod superfamily all have similar compositions, each containing a signal peptide, a CCxGY amino acid motif, and at least one cysteine rich NIM domain (Somogyi et al., 2008). NIM domains are defined by a consensus sequence motif (CxPxCxxxCxNGxCxxPxxCxCxxGY), which is closely related to the epidermal growth factor (EGF) consensus motif (xxxxCx$_{2-7}$Cx$_{1-4}$(G/A)xCx$_{1-13}$ttaxCx-CxxGax$_{1-6}$GxxCx) (Kurucz et al., 2007). Genes in the *nimrod* superfamily are found in syntenic clusters in *D. melanogaster* as well as in other *Drosophila* species, the honey bee (*Apis mellifera*), a mosquito (*Anopheles gambiae*), and the red flour beetle (*Tribolium castaneum*) (Kurucz et al., 2007; Somogyi et al., 2008).

NIM units occur as tandem repeats in some members of the *nimrod* superfamily, including *eater*. Repeated motifs of highly similar sequence often exhibit concerted evolution due to mispairing and unequal crossing over between homologous chromosomes and to gene conversion between non-homologous repeats. This type of evolution results in repeat arrays where paralogous repeat units are more similar to each other within species than they are to homologous units among species (Charlesworth et al., 1994). Of genes in the *nimrod* superfamily, *eater* is the only member whose NIM repeats show evidence of concerted evolution (Somogyi et al., 2008). NIM repeats in the interior of the gene appear to be evolving concertedly (Somogyi et al., 2008) and are thought to provide a structural "stalk" between the microbe binding units and the hemocyte cell membrane (Kocks et al., 2005). The first four NIM repeat units in *eater*, which have been shown to be necessary for microbial binding (Kocks et al., 2005), show no signs of concerted evolution (Somogyi et al., 2008).

In a molecular evolutionary comparison among *Drosophila* species, *eater* and

three *nimrod* family genes were found to be evolving under positive selection (Sackton et al., 2007). In one *nimrod* gene, *nimC1*, the positively selected sites are clustered within putative microbial binding domains, which suggests that pathogen interactions drive this rapid evolution. In contrast, adaptive mutations in *eater* are scattered throughout the gene, including outside domains known to interact directly with pathogens (Sackton et al., 2007). Selective pressures on the immune system are geographically variable, corresponding to heterogeneity in pathogen identity or abundance and other environmental factors. Immune system genes therefore may show evidence of local adaptation that can be detected with population genetic statistics. For instance, immune system genes display elevated differences in allele frequencies among populations relative to the genome average (Ryan et al., 2006; McEvoy et al., 2009). Recent selection can also be detected by examining patterns of genetic variation within populations. Strong positive selection leads to a rapid rise in the frequency of an adaptive mutation, incidentally dragging neutral variants linked to the target of selection upward in frequency. This leads to excess linkage disequilibrium (Kelly, 1997; Sabeti et al., 2002), decreased nucleotide diversity (Smith and Haigh, 1974), and too many high and low frequency polymorphisms (Tajima, 1989; Fu, 1997; Fay and Wu, 2000) relative to expectations under selective neutrality. Analyses of these properties can easily be applied to coding and noncoding regions, allowing us to detect selection on regulatory gene regions. We can potentially also identify the specific trait on which selection acts by linking genetic diversity patterns and phenotypes.

In the current work, we have sequenced the complete upstream and nonrepetitive coding region of *eater* in a North American and an African population of *D. melanogaster*. We find that both populations harbor substantial polymor-

phism in the number of NIM repeats and therefore for the overall size of the protein. We confirm the patterns of concerted evolution in NIM repeats that have been previously reported but also find evidence for varying degrees of concerted evolution between units. There is extensive linkage disequilibrium in the second intron of *eater* that extends through the upstream and 5′, non-repetitive gene region in the North American population, with the major haplotypes at the second intron significantly associated with gene expression level. Additional analysis suggests that one of these haplotypes has recently risen to high frequency in North America, which we interpret to reflect adaptation of the immune response to the novel pathogen environment that was encountered after emigration from Africa.

4.3 Materials and Methods

4.3.1 Fly Strains

D. melanogaster strains used for DNA sequence analysis in this study came from Zimbabwe or the United States. Strains ZW09, ZW139, ZW140, ZW142, ZW144, ZW149, ZW155, ZW184, ZW185, and ZW190 were originally collected in 2002 by J.W.O. Ballard from Victoria Falls, Zimbabwe. Strains I01, I03, I04, I06, I07, I13, I16, I17, I22, I23, I24, I26, I29, I31, I33, I34, I35 and I38 were originally collected in 2004 by E. M. Hill-Burns and B. P. Lazzaro from Ithaca, New York, United States. Additional collections from China (Beijing, courtesy of X. Huang and R. Roush via A. G. Clark; (Begun and Aquadro, 1995)), the Netherlands (Houten, courtesy of Z. Bochdanovits via A. G. Clark; (Bochdanovits and

de Jong, 2003)), Australia (Tasmania, courtesy of A. A. Hoffman, via A. G. Clark), and the United States (Athens and Blairsville, Georgia, courtesy of V. Corby-Harris and D. Promislow; (Lazzaro et al., 2008)) were used for PCR to measure the size of *eater*. Each line was initiated by intercrossing the progeny of a single, field-inseminated female and has been maintained by mass sib-mating in the lab since collection. The African lines in particular still segregate for residual heterozygosity.

eater is located on the right arm of chromosome 3 at cytological band 97E2. To isolate single *eater* alleles for sequencing, an individual male from each stock was crossed to virgin females from the deficiency line Df(3R)Tl-P, e^1 ca^1/TM3, Ser^1 (Bloomington Drosophila Stock Center stock number 1910). Single male progeny from this cross with the genotype Df(3R)Tl-P, e^1 ca^1/+ were crossed to virgin females from the original deficiency line. Males and virgin females from the second cross that had the genotype Df(3R)Tl-P, e^1 ca^1/+ were crossed to each other to isolate a single wild type allele from the original isofemale line along with the deficiency chromosome. Only flies that were either homozygous for a single wild type allele or hemizygous over the deficiency were sequenced.

4.3.2 PCR and DNA Sequencing

PCR amplifications of genomic DNA were performed using iProof high-fidelity polymerase (BioRad) or Taq polymerase (New England Biolabs). iProof-derived products were prepared for sequencing using PCR purification columns (Invitrogen). Taq-derived products were prepared for sequencing using Exonuclease I (USB Corp.) and shrimp alkaline phosphatase (USB Corp.). PCR products

were then directly sequenced. DNA sequences for the United States and Zimbabwe populations were collected for all non-repetitive *eater* coding regions, all introns, 5′ and 3′ untranslated regions, and an approximately 2kb region upstream of the transcriptional start site. Complete sequence could not be obtained for some alleles with large numbers of repetitive internal repeats. In these cases, the length of the repetitive regions was determined by amplifying the repeat region using primers that anneal to the flanking non-repetitive regions and sizing the products on 1% (1-2.4 kb) or 0.6% (>2.4kb) agarose gels. This genotyping of repeat region length was done for all populations. All primers are available by request. Nucleotide sequences have been deposited in GenBank (HM165155-HM165182). Outgroup sequence were obtained from the reference genomes of *D. simulans* (Release 1.0) and *D. yakuba* (Release 2.0) (Begun et al., 2007).

4.3.3 DNA Sequence Analysis

DNA sequences were assembled in CodonCode Aligner (CodonCode Corp.). NIM repeat units were identified by the 26 amino acid consensus sequence CxPxCxxxCxNGxCxxPxxCxCxxGY (Somogyi et al., 2008). An alignment was built of NIM repeats using this conserved motif as in (Kurucz et al., 2007) and (Somogyi et al., 2008) since the nucleotides within this sequence could be aligned for all NIM repeat units from all sampled alleles and both outgroups. Alignments based on the NIM consensus sequence were used to build neighbor-joining unrooted trees. Trees were constructed in MEGA 4.0 (Tamura et al., 2007) using an amino acid model with a Poisson correction and uniform substitution rates among all sites. Five hundred bootstrap replicates were performed to indicate support of each node. In agreement with (Somogyi et al., 2008), we will refer

to a repeat as "independently evolving" if its sequence is found just once per individual allele and it is more closely related to homologous units in *D. yakuba* and *D. simulans* than to repeat units at non-homologous positions within the *D. melanogaster eater* gene. Independently evolving units were numbered 1 to 11 (Figure 4.1).

We compared the evolutionary patterns of the four NIM repeats that have previously been shown to be important for microbial binding (Kocks et al., 2005) to those where no such functional assignment has been made. We calculated K_A, the rate of amino acid substitution, and K_S, the rate of silent substitution for each NIM unit 1 to 11 independently (Nei and Gojobori, 1986). Fixations were polarized using *D. yakuba* and *D. simulans* as outgroups and only fixations that occurred along the D. melanogaster lineage were considered. Wilcoxon rank sum tests were used to test for differences in substitution rates between microbial binding versus all other NIM repeats.

We calculated population genetic statistics on all gene regions that were not evolving concertedly. Nucleotide diversity, Tajima's D, and linkage disequilibrium were calculated using DnaSP v. 5.0 (Librado and Rozas, 2009) and scripts written in the programming language R R Development Core Team (2006). Nucleotide diversity (π) was measured both as the average pairwise differences between sequences per locus and per site with a Jukes-Cantor correction applied. Tajima's D (Tajima, 1989) was calculated using all mutations. Tajima's D measures the difference between two different estimates of the population genetics parameter ($4N_e\mu$), one of which measures nucleotide diversity (θ_π) and the other which relies on the number of segregating sites (θ_w). The f statistic, which is the nucleotide diversity in a putatively selected allele divided by the

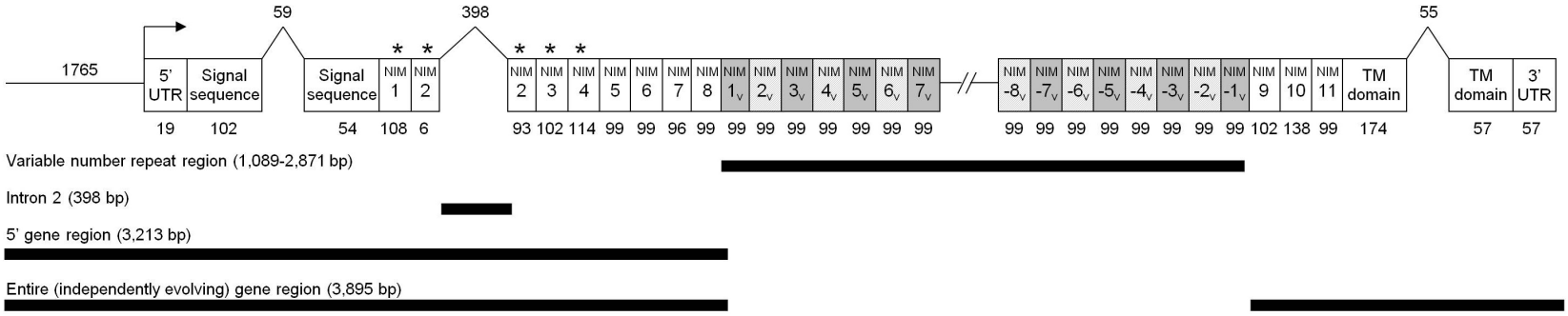


Figure 4.1:
Gene structure and survey region. Variable number NIM repeat units were excluded when calculating population genetic statistics. The signal sequence, NIM 2, and the transmembrane (TM) domain are interrupted by introns, which are indicated by up-carats with the size of the intron (in base pairs) given above the carat. The numbers below sequence domains indicate their size in base pairs. The forward pointing arrow indicates the transcriptional start site. The 1765 base pair region immediately upstream of the transcriptional start site is indicated and includes the minimal enhancer region (Tokusumi et al. 2009). The boxes labeled 5′ and 3′ UTR are untranslated regions. * indicates NIM repeats that have previously been implicated in microbial binding (Kocks et al., 2005). Dark lines below the gene schematic indicate various survey regions considered in different components of this paper. Variable number repeat units are indicated with a 'v' subscript. "NIM 8-like" repeats are shown with gray and white diagonal lines and "alternate" repeats are shown in gray. (NIM 8-like consensus motif: CKPICSxxCENGxCxAPEKCSCNGY, "alternate" consensus motif: CxxVCxxGCKNGFCxAPxKCSCxxxx.) Between 0 and 15 repeats were not sequenced in the interior of the gene (shown with hatched lines).

total nucleotide diversity (Macpherson et al., 2008), was calculated in R. Linkage disequilibrium was measured using the ZnS statistic (Kelly, 1997), which is a standardized average of all calculations of the D statistic between all pairs of segregating sites. The standardized measure of linkage disequilibrium between pairs of sites, r^2, was plotted using the LDheatmap package in R (Shin et al., 2006). Extended haplotype homozygosity (EHH), another measure of linkage disequilibrium, is the probability at a given position that two sampled alleles with the same pre-defined core genotype are identical by descent (Sabeti et al., 2002). EHH was calculated using a script written in R with the core genotype defined at the center of the second intron since linkage disequilibrium was most extreme in this region (see Results). We calculated Z_{nS}, EHH, nucleotide diversity (π), Tajima's D, and f on the entire eater gene region (3,895 bp) with the variable number repeat units excluded to look for selection over the entire locus. We also calculated these statistics on the second intron (398 bp) since the most extreme values of the population genetic statistics should be near the site of selection. Lastly, we calculated the above set of statistics on the 5′ gene region (3,213 bp), which included the entire upstream region, 5′ UTR, NIM 1-8, and two introns, since this is the region used for simulations (see Coalescent Simulations).

4.3.4 Coalescent Simulations

We used coalescent simulations run in the *ms* program (Hudson, 2002) to build null distributions of our test statistics under various neutral demographic scenarios. Our empirically determined test statistics were then compared to the null distributions to test for deviations from neutrality that could be attributed

to selection and to assess statistical significance of empirically observed patterns. Polymorphism data were simulated for a recombining, neutrally evolving locus of length 3,106 base pairs intended to represent the majority of the non-repetitive coding and non-coding 5′ end of the gene with insertion or deletion events considered as single base pair mutations. The 682 base pairs at the 3′ end of the gene were not included in the simulations since recombination distance across the intervening repetitive region could not be accurately incorporated into the simulations. For these simulations, a fixed number of 86 segregating sites (our empirical observation at *eater*) was assumed. We also simulated patterns of polymorphism at the second intron, which was modeled as a recombining locus that was 398 base pairs long with 15 segregating sites. The local recombination rate was estimated using the *D. melanogaster* Recombination Rate Calculator (Singh et al., 2005) and was estimated to be 1.77 cM/Mbp for this locus. The effective population size was assumed to be 10^6, and the mutation rate was assumed to be 1.5×10^{-9} bp^{-1} gen^{-1} (Li, 1997).

Our simplest demographic scenario assumes a panmictic population of constant size. Our two other scenarios account for the bottleneck that the North American population underwent when it was founded from an ancestral African population (David and Capy, 1988). The details of this bottleneck have recently been inferred in detail by two separate analyses of datasets from the Netherlands and East Africa (Thornton and Andolfatto, 2006; Li and Stephan, 2006). Previous work has shown that all non-African populations were derived from a single colonization event (Baudry et al., 2004; Schlötterer et al., 2006) so it is appropriate to apply the parameters inferred from these data to North American populations (Macpherson et al., 2008). The exact parameters of the bottleneck that were inferred differ between studies (Macpherson et al., 2008).

Thornton and Andolfatto (2006) estimated three parameters of the bottleneck (referred to as the TA scenario from this point forward): the timing of the population size reduction (T_b), the timing of recovery (T_r), and the ratio of the population size during the bottleneck to the size before and after (R_b). The best estimate from their approximate Bayesian methods suggests that T_b was 16,000 years ago, T_r was 3,000 years ago, and that R_b was 0.029. Assuming 10 generations a year (Thornton and Andolfatto, 2006), this corresponds to a T_b of 0.022 * $4N_e\mu$ generations ago and a T_r of 0.0042 * $4N_e\mu$ generations ago. Li and Stephan (2006) used a maximum likelihood procedure and estimated that T_b was 15,800 years (0.0367 * $4N_e\mu$ generations ago) and lasted only a few hundred years (T_r equal to 0.0360 * $4N_e\mu$ generations ago), and that R_b was 0.002. In addition, they estimate that previous to the bottleneck out of Africa, the African population underwent an expansion in population size. They estimated the expansion (T_e) occurred 60,000 years ago (0.1395 * $4N_e\mu$ generations ago) and that the ratio of the expansion size to the current population size (R_e) was 8.0. We will refer to these parameters as the LS scenario. Simulations under the LS scenario that incorporate this expansion of the ancestral population are a significantly better fit to overall genomic patterns of polymorphism than simulations under the TA scenario (Li and Stephan, 2006), and an ancient expansion explains the excess of rare derived mutations that are observed within African populations.

Our *eater* sequences were divided into two distinct clades that we hypothesize may have adaptive significance, so we only retained simulations that matched the empirically observed topology (see Results). Specifically, we required that the final coalescence event occur at the span representing the second intron and divide the data into two clades of sizes 8 and 10, where the clade of size 8 represents an allele experiencing a partial selective sweep. In this way, we

simulated coalescent trees with the same topology as seen in our dataset and generated a distribution of each test statistic that would be expected for this given topology in the absence of selection. Population genetic statistics were calculated for each simulated dataset with the entire population included, as well as separately with only individuals containing the putatively adaptive allele included. The distribution of each simulated test statistic was determined, and statistical significance was defined as the number of simulated datasets that had a value of the test statistic equal to or more extreme than that observed for *eater*. We conducted 2-tailed tests on empirical estimates of *eater* from the entire population to test for deviations from neutrality. Empirical estimates at *eater* were considered significant if they fell into the 2.5% tails of the simulated distributions. We conducted 1-tailed tests on the putatively adaptive haplotype group to test for a selective sweep in this class. Haplotype number, nucleotide diversity, Tajima's D and f were considered extreme if the observed value was in the lower tail of the simulated values. Z_{nS} was considered extreme if the observed value was in the upper tail of the simulated values.

4.3.5 Gene Expression

We sequenced a set of 3rd chromosome substitution lines from a Pennsylvania, United States, collection of *D. melanogaster* (Fiumera et al., 2007) at the second intron of the *eater* locus and measured *eater* gene expression in the 19 lines that were a perfect match to either the "A" or "B" haplotype between base pair 390 and 488 (see Results). These substitution lines had been previously backcrossed for eight generations to remove variation on the 2nd, 4th, and sex chromosomes, and thus only vary at the 3rd chromosome (Fiumera et al., 2007). This should re-

duce the amount of *trans*-regulatory variation in *eater* expression. We designed primers specific to *eater* and to a house-keeping gene, *rp49*, which was used to control for variation in the efficiency of RNA extraction and cDNA synthesis. Transcript abundance was measured using Power SYBR Green (Applied BioSystems). Replicate samples of 10 males aged 3 to 5 days post-eclosion were taken from each of two individual fly vials per line, RNA was extracted using a modified Trizol protocol (Invitrogen), and all quantitative PCR reactions were run in duplicate. This procedure was done twice, on separate days and for different fly generations. Significance of the "A" or "B" haplotype to predict *eater* transcript abundance was assessed using Proc Mixed in SAS (SAS Institute, Cary, NC) after accounting for the random effect of experiment day, line nested within genotype, vial nested within line and genotype, and random variance among replicate samples drawn from the same vial, as well as the fixed effect of the estimated abundance of *rp49* transcripts.

4.3.6 Analysis of Variable Number Repeat Units

Repeat units between NIM 8 and NIM 9 have high sequence similarity at the nucleotide and amino acid level and have previously been shown to be evolving concertedly (Somogyi et al., 2008). We found individual *D. melanogaster* to be polymorphic for the number of repeats of this type (see Results). These variable number repeat units, which are 99 base pairs in length, cluster together into two types ("NIM 8-like" core consensus motif (78 base pairs / 26 amino acids): CKPICSxxCENGxCxAPEKCSCNGY; "alternate" core consensus motif: CxxVCxxGCKNGFCxAPxKCSCxxxx) which are always found in tandem (Figure 4.1; (Somogyi et al., 2008)). We labeled these units starting with the ones

closest to NIM 8 or NIM 9 and counting inwards towards the center of the array. Units immediately 3′ of NIM 8 were numbered starting with 1v and units immediate 5′ of NIM 9 were numbered starting with -1_v (Figure 4.1).

We obtained an average of 1393 bp (~14 units) of sequence across the variable number repeat units per individual and measured the physical size of PCR products in this region for all individuals. The software package R_{ST} *calc* (Goodman, 1997) was used to calculate genetic differentiation between populations. Because R_{ST} *calc* requires diploid samples and our fly lines were artificially made haploid, we randomly assigned alleles into diploid combinations to create artificial "genotypes." Statistical significance of pairwise comparisons between populations was determined by permuting the sequenced alleles among subpopulations and recalculating R_{ST} 10,000 times to determine an empirical null distribution. To determine the statistical significance of the worldwide value of R_{ST}, we ran 10,000 bootstrap simulations to determine a confidence interval of our observed R_{ST} value. We measured genetic distance between all pairs of variable number repeat units using Kimura's 2-parameter model in the *ape* package in R (Paradis, 2004).

4.4 Results

4.4.1 Summary Population Genetic Statistics

We sought to determine whether the *eater* gene, which is required for immunological phagocytosis, shows signs of recent adaptation at the molecular population genetic level. We sequenced multiple *eater* alleles from a Zimbabwe and a

United States population of *D. melanogaster* (Figure 4.2) and estimated sequence diversity and linkage disequilibrium for each population (Table 4.1; Figure 4.3). Linkage disequilibrium (Figure 4.3; second intron is outlined with a black triangle) and diversity (Figure 4.4) are highest in and around the second intron in the United States population but not the Zimbabwe population. We simulated 1000 coalescent genealogies of a neutrally evolving equilibrium population under the estimated recombination rate, none of which showed linkage disequilibrium at the second intron that was as high as we observed in the United States population ($p < 10^{-3}$; Table 4.2). These patterns reflect the presence of two high frequency haplotype groups that are substantially diverged from each other. The presence of two high frequency haplotypes could in principle be explained by a partial selective sweep, balancing selection, or some non-equilibrium demographic scenarios. However, the lack of variation within each haplotype class is contrary to the expectation for an ancient balanced polymorphism (e.g., (Hudson and Kaplan, 1988)), rendering a partial sweep or non-equilibrium demography as more plausible explanations.

To see if non-equilibrium patterns extended beyond the second intron, we compared patterns of nucleotide diversity and the site frequency spectrum at the second intron with those across the rest of the gene region. When interpreting our results, we considered only the 5′ end of the gene (3,213 bp) since simulations could not be performed across the entire gene region (3,895 bp) due to the variable number repeat units (see Materials and Methods). Tajima's D (Tajima, 1989), nucleotide diversity, and linkage disequilibrium (Z_{nS}; (Kelly, 1997)) are all elevated in the second intron relative to the rest of the 5′ gene region in the North American population (second intron: Tajima's D=+2.6697, π=0.01867, Z_{nS}=0.7474; 5′ gene region: Tajima's D=+0.2906, π=0.00861, Z_{nS}=0.1560; Ta-

Table 4.1: Population genetic summary statistics for independently evolving regions of *eater*.

	n	**h**	**S**	**bp**	π	π_{ns}	π_{s}	**Tajima's *D***	$\mathbf{Z}_{nS}$
Zimbabwe									
Entire gene	10	9	142	3895	0.01188	0.00367	0.02075	-0.5199	0.1321
5′ gene region	10	9	123	3213	0.01271	0.00356	0.01934	-0.4621	0.1341
Intron 2	10	8	30	398	0.02903	n/a	n/a	0.2622	0.1796
New York, US									
Entire gene	18	16	94	3895	0.00768	0.00343	0.00654	0.2665	0.1462
5′ gene region	18	16	86	3213	0.00861	0.00345	0.00862	0.2906	0.1560***
Intron 2	18	5***	15	398	0.01867***	n/a	n/a	2.6697***	0.7474*
New York, US ("A" group)									
Entire gene	8	6	64	3895	0.00564	0.00386	0.00589	-0.7286	0.2987
5′ gene region	8	6**	59	3213	0.00635	0.00434	0.00913	-0.7129***	0.312***
Intron 2	8	2*	3	398	0.00188**	n/a	n/a	-1.4475*	1*
New York, US ("B" group)									
Entire gene	10	10	56	3895	0.00477	0.00284	0.00493	-0.3730	0.187
5′ gene region	10	10	49	3213	0.00499	0.00285	0.00509	-0.4352	0.1869
Intron 2	10	3	4	398	0.00201	n/a	n/a	-1.6671	0.5062

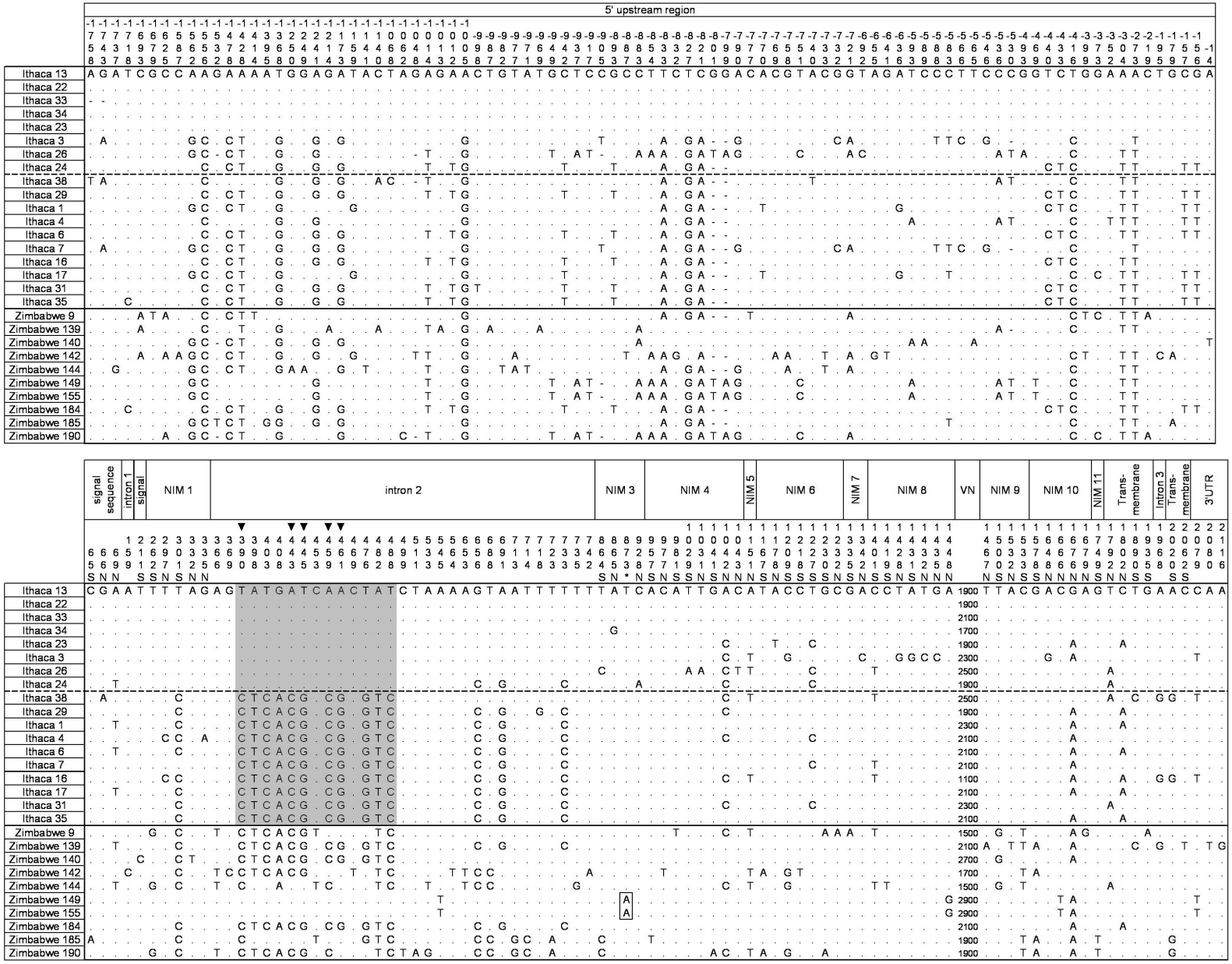
The United States population was both analyzed as a whole and split into haplotype groups. n, number of alleles sampled; h, number of haplotypes; S, segregating sites; bp, sequence length in base pairs' π is nucleotide diversity per base pair (ns=non-synonymous, s=synonymous); n/a, not applicable. Tajima's *D* (Tajima, 1989) includes all mutations. Z_{nS} (Kelly, 1997) is linkage disequilibrium based on segregating sites. Statistically significant values (*:0.01=p<0.05; **: 0.001=p<0.01; ***: p<0.001) under a null demographic model are indicated (*) for the 5′ gene region and intron 2 in the New York, US, population and the "A" haplotype group (see Table 4.2).

Table 4.2: Distribution of summary statistics at *eater* based on simulations under three different demographic models.

	h	h_A	π	π_A	f	D	D_A	Z_{nS}	Z_{nSA}
Second intron									
observed:									
eater (US)	5	2	7.431	0.750	0.1009	2.6697	-1.4475	0.7474	1.0
simulations:									
null model	11.31 (9, 14)	4.91 (3, 8)	4.56 (3.24, 5.85)	2.94 (0.93, 5.5)	0.64 (0.21, 1.15)	0.18 (-0.68, 1.29)	0.08 (-1.47, 1.43)	0.14 (0.09, 0.26)	0.33 (0.08 ,0.81)
p-value	0 *	0.017 *	0 *	0.001 *	0 *	0 *	0.031 *	0 *	0.016 *
TA	5.74 (2, 12)	2.56 (1, 6)	6.34 (2.92, 7.84)	1.06 (0, 4.64)	0.19 (0, 0.75)	1.72 (-1.26, 3.03)	-0.53 (-1.74, 1.91)	0.55 (0.06, 1)	0.54 (0.02, 1)
p-value	0.547	0.534	0.872	0.549	0.5	0.1	0.272	0.235	0.261
LS	9.01 (5, 13)	3.88 (1, 7)	5.92 (4.01, 7.18)	2.3 (0, 5.46)	0.39 (0, 0.88)	1.36 (-0.31, 2.45)	-0.31 (-1.64, 1.44)	0.34 (0.17, 0.65)	0.43 (0.04, 1)
p-value	0.076	0.182	0 *	0.201	0.156	0 *	0.091	0.018 *	0.077
5′ gene region									
observed:									
eater (US)	16	6	26.74	19.75	0.74	0.2906	-0.7129	0.1560	0.312
simulations:									
null model	17.39 (15, 18)	7.8 (7, 8)	25.47 (22.46, 28.75)	23.27 (17.64, 28.32)	0.91 (0.77, 1.06)	0.08 (-0.43, 0.63)	0.1 (-0.51, 0.91)	0.08 (0.07, 0.1)	0.18 (0.15, 0.24)
p-value	0.107	0.005 *	0.782	0.089	0.018 *	0.218	0 *	0 *	0 *
TA	14.94 (12, 17)	6.59 (4, 8)	34.01 (22.6, 40.13)	19.45 (3.75, 32.46)	0.58 (0.11, 1.09)	1.51 (-0.4, 2.53)	-0.01 (-1.87, 1.53)	0.31 (0.19, 0.52)	0.44 (0.22, 0.85)
p-value	0.851	0.396	0.059	0.495	0.713	0.941	0.238	1	0.851
LS	17.28 (14, 18)	7.58 (6, 8)	31.46 (26.35, 36.38)	26.1 (17.18, 32.68)	0.83 (0.6, 1.07)	1.08 (0.22, 1.9)	0.36 (-0.56, 1.22)	0.12 (0.09, 0.15)	0.23 (0.17, 0.29)
p-value	0.139	0.04 *	0.06	0.069	0.188	0.97	0.02 *	0.06	0 *

Empirical estimates of population genetic statistics are given for the United States population and the "A" haplotype group. For the simulated distributions, the means and 95% confidence intervals (in parentheses) under 3 different demographic models are indicated. TA refers to the model based on Thornton and Andolfatto (2006), and LS refers to the model based on Li and Stephan (2006). π is nucleotide diversity, or average pairwise differences, per locus. f is nucleotide diversity in the putatively selected allele ("A") divided by the total nucleotide diversity. D is Tajima's D including all mutations. P-values indicate the proportion of simulations where the simulated values were more extreme than the empirical estimate at *eater*. Tests of h, π, D, Z_{nS} were 2-tailed, and the means were standardized around 0 to calculated p-values. f, h_A, π_A, D_A, Z_{nSA}, were used to test for a partial selective sweep, and simulated values were considered extreme if they were less than empirical estimates of f, h_A, π_A, and D_A or greater than empirical estimates of Z_{nSA}.
*p values less than 0.05.

Figure 4.2: Polymorphic sites for the *eater* locus. The United States population (Ithaca, New York) is divided into ″A″ and ″B″ type haplotypes based on sequence between base pairs 390 and 488 (highlighted in gray). ″A″ haplotypes are above the dotted line and ″B″ haplotypes are below. Non-synonymous (N) and synonymous polymorphisms (S) in coding regions are indicated. Stop codons were found segregating in two individuals from Zimbabwe (boxed). CF2-II motif polymorphisms are shown (▼; see Figure 4.6). Variable number repeat units between NIM 8 and NIM 9 could not be aligned with confidence and are not shown, but the approximate length of that region is shown in base pairs (VN). Sites with alignment gaps were considered if there was a polymorphism. Base pair position within the gene corresponds with Figure 4.1 with the first position being the transcriptional start site and the 5′ upstream region indicated as -1765 to -1.


5' upstream region
signal sequence
intron 1
signal
NIM 1
intron 2
NIM 3
NIM 4
NIM 5
NIM 6
NIM 7
NIM 8
VN
NIM 9
NIM 10
NIM 11
Trans-membrane
Intron 3
Trans-membrane
3'UTR
Ithaca 13
Ithaca 22
Ithaca 33
Ithaca 34
Ithaca 23
Ithaca 3
Ithaca 26
Ithaca 24
Ithaca 38
Ithaca 29
Ithaca 1
Ithaca 4
Ithaca 6
Ithaca 7
Ithaca 16
Ithaca 17
Ithaca 31
Ithaca 35
Zimbabwe 9
Zimbabwe 139
Zimbabwe 140
Zimbabwe 142
Zimbabwe 144
Zimbabwe 149
Zimbabwe 155
Zimbabwe 184
Zimbabwe 185
Zimbabwe 190

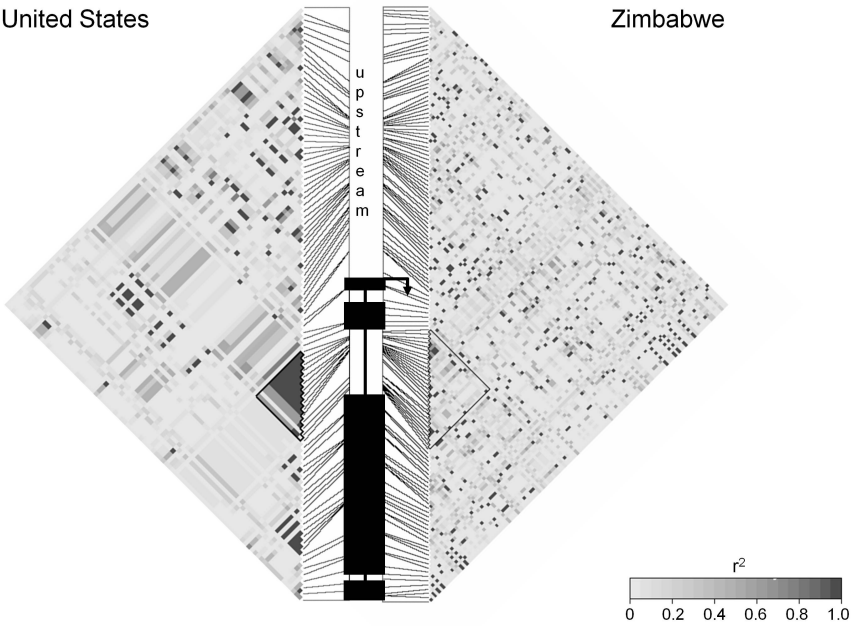


Figure 4.3: Linkage disequilibrium (r^2) plotted across the concatenated gene region. Each pixel represents r^2 plotted between a pair of segregating sites. Exons are shown with black boxes, with the transcriptional start site indicated with an arrow. Introns are shown as lines between the exons. The black triangle indicates the block of high linkage disequilibrium in the second intron of the United States (Ithaca, New York) population, and is indicated in the Zimbabwe population for comparison.

ble 4.1). These patterns of diversity are extremely unlikely under the standard neutral null model (Table 4.1). Tajima's D is significantly positive at the second intron ($p<0.01$), and linkage disequilibrium is significantly high at both the second intron ($p<0.001$) and in the 5' gene region ($p<0.001$).

To test whether this reflects the pooling of two intermediate frequency, disparate allelic classes, we calculated the statistics separately for each haplotype group in the North American population (Table 4.1). Only two sequence haplotypes were observed between base pairs 390 and 488 (99 base pairs) within

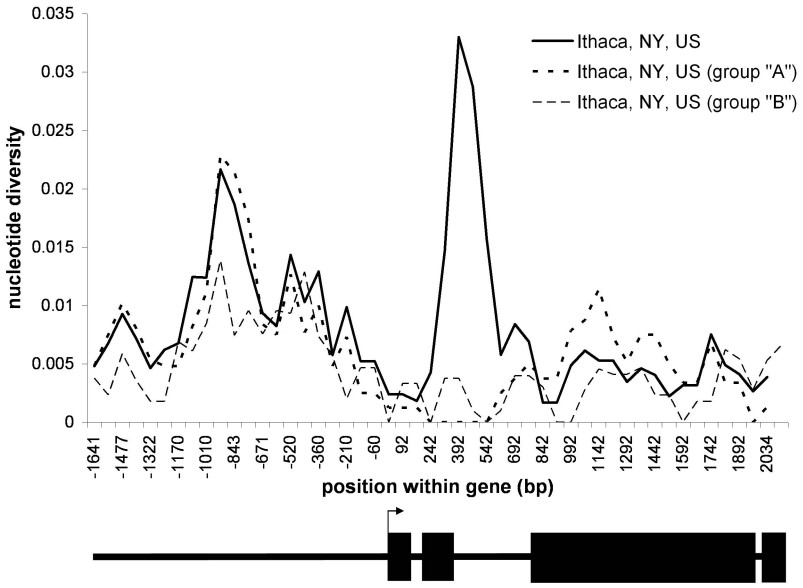


Figure 4.4: Plot of nucleotide diversity in the North American population by haplotype group. Nucleotide diversity is plotted for sliding windows with a window length of 200 sites and a step size of 75 sites. A schematic of the gene is shown below the graph with exons indicated as black boxes and the transcription start site indicated with an arrow. A spike in apparent diversity is seen over intron 2, where two divergent haplotypes (groups ″A″ and ″B″) are segregating. There is no excess diversity within either haplotype.

the second intron (398 base pairs), so the alleles were divided into group ″A″ or group ″B″ based on this sequence (Figure 4.2; region in gray). The ″A″ group is named because it is a perfect match to the reference genome of *D. melanogaster* (Adams et al., 2000). Across the remainder of the second intron, these haplotypes each have very little variation within haplotype group (π_A=0.00188; π_B=0.00201) but 11 of 15 segregating sites in this 398 bp window are fixed differences between the two groups ($\pi_{combined}$=0.01867; Table 4.1; Figure 4.2). The reduction in nucleotide diversity extends approximately 800 base

pairs in the "A" haplotype (Figure 4.4). Across the 5′ gene region, the levels of nucleotide diversity are similar in each haplotype group (π_A: 0.00635, π_B: 0.00499). The "A" haplotype has a higher level of linkage disequilibrium and a more negative value of Tajima's D than the "B" haplotype does (Z_{nS} A: 0.312, B: 0.1869; Tajima's D A: -0.7129, B: -0.4352). Compared to the standard null neutral model, the "A" haplotype group has a significantly low value of Tajima's D ($p<0.001$), too few haplotypes ($p=0.005$), excess linkage disequilibrium ($p<0.001$), and a low value of f ($p=0.018$) indicating a deficit in nucleotide diversity in haplotype "A" relative to the entire population (Tables 4.1, 4.2). The observed reduction in nucleotide diversity and excess linkage disequilibrium in the "A" haplotype is consistent with a recent rise to high frequency due to a partial selective sweep.

The Zimbabwe population, in contrast, did not show evidence for haplotype structuring or a recent selective sweep around the second intron. The patterns of diversity are compatible with our expectations for a neutrally evolving African population. The Zimbabwe population harbors substantially more diversity than the United States population ($\pi_{Zimbabwe}$: 0.01188, π_{US}: 0.00768; Table 4.1), reflecting the larger effective size of this population, which is presumed to be ancestral to the United States population (David and Capy, 1988). Two individuals in the Zimbabwe population have a stop codon in the third NIM repeat that presumably results in a truncated version of Eater (Figure 4.2). Such potentially deleterious mutations are expected to occur at low frequencies in populations that are in mutation-selection balance. For both populations, the diversity at synonymous sites exceeded that at non-synonymous sites in *eater* (Table 4.1), as would be expected if purifying selection acts to remove deleterious amino acid variation.

4.4.2 Standard Neutral and Bottleneck Simulations

The observed population genetic statistics are highly suggestive of the haplotype "A" having recently risen to high frequency in the United States population due to positive selection. These patterns are not compatible with a standard neutral null model of evolution. However, selectively neutral demographic processes such as population expansions or bottlenecks can often lead to patterns that mimic those expected under natural selection. It is believed that all non-African populations have only recently been founded from Africa (David and Capy, 1988; Baudry et al., 2004). Models that incorporate bottlenecks similar to what these populations underwent as they expanded their population range are a better fit to genomewide patterns of diversity (Li and Stephan, 2006) than the standard neutral null model. Two recent analyses have described bottleneck models that can be applied to the North American population (Thornton and Andolfatto, 2006; Li and Stephan, 2006), one with a prolonged bottleneck that ended recently and the other with a short, ancient bottleneck that was preceded by a population expansion. Simulations under these models can give us a mean and range of values for the number of haplotypes, nucleotide diversity, linkage disequilibrium, and Tajima's *D* that we can expect to observe at the *eater* locus in the absence of selection. If our empirically observed statistics fall into the tails of the null distributions (see Methods) then we infer that selection may have occurred.

The TA model (Thornton and Andolfatto, 2006) describes a hypothesized prolonged bottleneck that ended recently. The population genetic statistics that we observed at the *eater* locus are all consistent with the TA demographic scenario (Table 4.2; Figure 4.5a,b). The LS model (Li and Stephan, 2006) presumes a

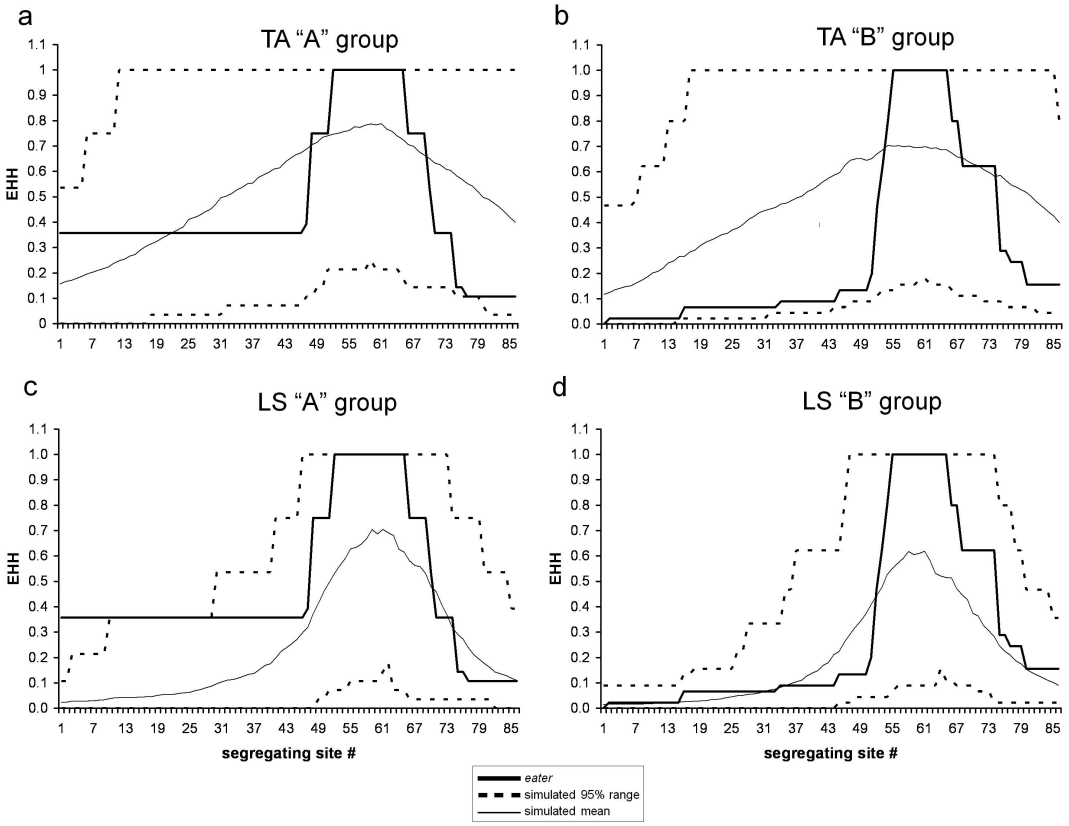


Figure 4.5: Extended haplotype homozygosity. EHH at the *eater* locus is compared with values obtained from simulations under various demographic scenarios. The mean and the 2.5% and 97.5% tails of the simulated distributions are shown. a, b) EHH at *eater* and simulated EHH values obtained from the TA demographic scenario for the ″A″ and ″B″ haplotypes respectively based upon Thornton and Andolfatto (2006) c, d) EHH at *eater* and simulated EHH values obtained from the LS demographic scenario for the ″A″ and ″B″ haplotypes respectively based upon Li and Stephan (2006). EHH at the *eater* locus is not consistent with distributions of EHH obtained from simulations of the locus under the LS demographic scenario

bottleneck that was ancient, brief, and was preceded by a population expansion in the ancestral African population. Using this model, there are expected to be significantly more haplotypes (p=0.04), a higher value of Tajima's *D* (p=0.02), less linkage disequilibrium (p<0.001), and a shorter extent of EHH in the "A" group than is actually observed at the *eater* locus (Table 4.2; Figure 4.5c,d). This indicates that the values observed at the *eater* locus cannot be explained by an ancient and brief bottleneck such as the one proposed by (Li and Stephan, 2006). The ranges of the distributions of most test statistics are much wider when modeling the TA scenario than under the LS scenario (Table 4.2) and would attribute demographic explanations to all but the most extreme instances of positive selection. The TA scenario is likely to be too conservative for the detection of less radical selective pressures. Of the three neutral models presented, the LS scenario best explains genomic patterns of polymorphism in derived populations of *D. melanogaster* (Li and Stephan, 2006), and thus we favor interpretation of our results in light of this scenario.

4.4.3 Gene Expression Differences between Haplotypes and Potential Targets of Selection

If natural selection has indeed shaped patterns of variation at *eater*, then we might expect to see a phenotypic difference associated with the high frequency haplotypes which could be the target of selection. We therefore examined the sequences of the "A" and "B" groups to find candidate sequence differences that could give us insight into the nature of a potential phenotypic difference. The excess linkage disequilibrium in the "A" haplotype extends

through NIM 1 and NIM 2, two repeat units that are implicated in microbial binding (Kocks et al., 2005). However, no fixed non-synonymous differences were found between the haplotype groups in these NIM repeats (Figure 4.2). Since the population genetic statistics were most extreme at the second intron of *eater*, we evaluated group "A" and "B" sequences at this intron with a sequence-motif finder against insect motifs within the library TRANSFAC at GenomeNet (http://motif.genome.jp/). Four putative chorion transcription factor 2 (CF2-II) binding regions were found in the "A" haplotype (Figure 4.6) using the search motif sequence GTATATATA. All four regions had polymorphisms in the "B" haplotype that made them poorer matches to the consensus motif. This sequence motif can be either an enhancer or a suppressor during *D. melanogaster* oogenesis and embryonic muscle development (Hsu et al., 1996; García-Zaragoza et al., 2008) and is a suppressor of expression of the antimicrobial peptide *gloverin* in the silkworm *Bombyx mori* (Mrinal and Nagaraju, 2008). We therefore hypothesized that the second intron might contain one or more regulatory sequences and that transcriptional differences between the alleles is the target of selection.

To test the hypothesis that sequence variation between the "A" and "B" haplotype groups results in differing expression levels of the *eater* gene, we measured constitutive expression of *eater* in adult males in 19 *D. melanogaster* isogenic lines that were homozygous for either the "A" or "B" haplotype. We found that lines bearing the "A" haplotype express significantly more *eater* than "B" haplotype lines ($p=0.0417$), exhibiting an average of 69% higher expression (Table 4.3). Although this observation does not directly test the function of the putative CF2-II binding sequences that are present in the "A" haplotype but absent in the "B" haplotype, it is consistent with the hypothesis that these or other

Haplotype	Score	Position								
		385	386	387	388	389	390	391	392	393
A	100	G	T	A	T	A	T	A	T	A
B	88	•	•	•	•	•	C	•	•	•
		394	393	392	391	390	389	388	387	386
A	91	T	T	A	T	A	T	A	T	A
B	79	•	•	•	•	G	•	•	•	•
		455	456	457	458	459	460	461	462	463
A	88	G	A	A	T	A	T	A	T	A
B	70	•	•	•	•	C	•	G	•	•
		438	439	440	441	442	443	444	445	446
A	87	G	T	A	T	A	A	A	T	A
B	77	•	•	•	•	•	C	•	G	•

Figure 4.6: Polymorphisms in putative CF2-II transcription factor recognition motifs in positions within the second intron of North American haplotypes ″A″ and ″B.″ Motif GTATATATA is considered a perfect match. The score indicates how well the input sequence matches the motif. The position within the gene region is indicated above each nucleotide. Haplotype group ″A″ has four high score matches to the CF2-II motif.

unidentified regulatory sequences cause functional differentiation between the two haplotypes and may be targets of selection.

Table 4.3: Factors and p-values of a linear model used to show that US haplotype group "A" expresses *eater* at a higher level than group "B."

Factor Name	Effect Type	df	Z-value or F-value	p-value
Line(Haplotype)[a]	Random		2.34	0.0097
Vial(Line*Haplotype)[b]	Random		0.35	0.3634
Sample(Line*Haplotype*Vial)[c]	Random		2.12	0.0170
Day[d]	Random		0.69	0.2448
Extraction(Day)[e]	Random		0.95	0.1711
Residual	Random		9.96	<.0001
Rp49[f]		1	1508.9	<.0001
Haplotype[g]	Fixed	1	4.83	0.0417

df, degrees of freedom
[a]Background variation due to genetic line
[b]Variation due to rearing vial
[c]Random variation among replicate samples of flies within the vial
[d]Variation due to day of experiment
[e]Variation due to RNA extraction
[f]Variation due to amount of RNA, measured as expression of housekeeping gene
[g] Variation due to haplotype

4.4.4 Evolutionary Patterns of NIM Repeats

The first four NIM repeat units (NIM 1-4) have previously been implicated in microbial binding (Kocks et al., 2005). We considered that these repeats specifically might participate in host-pathogen co-evolutionary interactions that the other NIM repeats would not. To test this hypothesis, we compared the rate of non-synonymous substitution (K_A) and of synonymous substitution (K_S) between NIM 1-4 and the remaining independently evolving NIM repeats (NIM 5-11) (Table 4.4). We found no evidence for any difference in the evolutionary patterns between the two sets of repeats, with K_A and K_S not significantly differing between the two groups (Wilcoxon signed rank test, K_A p-value=0.6202, K_S p-value=0.2183; Table 4.4). We also examined the phylogenetic relationship

of each NIM repeat among *D. melanogaster*, *D. simulans*, and *D. yakuba*. The accepted relationship among these species places *D. melanogaster* and *D. simulans* as sister species and *D. yakuba* as the outgroup (Begun et al., 2007). Six NIM repeats had phylogenetic relationships that deviated from this pattern, which could indicate elevated selective pressures along particular branches. However, these repeats were evenly distributed between NIM 1-4 (microbial binding) and NIM 5-11 (unknown function) (Table 4.4). Nucleotide diversity levels were not different between the two sets of functionally distinct repeats in either Zimbabwe or United States populations ($\pi_{Zimbabwe}$ p-value=0.7879, π_{US} p-value=0.7748). NIM 2, a unit with a putative role in microbial binding, had no polymorphism in either the Zimbabwe and United States populations (Table 4.4), nor in additional populations sampled from Australia, the Netherlands, or China (not shown). The second intron lies within this NIM repeat, so the deficit in diversity of NIM 2 may be linked with the unusual evolutionary patterns of the intron.

4.4.5 Properties of the Variable Number Repeat Units

The number of repeats in the region between NIM 8 and NIM 9 is polymorphic and ranges between 11 and 29 (Figure 4.7a). The Zimbabwe and China populations have the highest variation in the number of repeat units. Worldwide R_{ST}, a measure of genetic differentiation that ranges from 0 for completely undifferentiated to 1 for complete isolation, was 0.00388 (95% confidence interval: -0.0488, 0.4082) which indicates a lack of differentiation between populations. Pairwise comparisons between individual populations ranged from -0.0971 to 0.3939 (Figure 4.7b), and all were non-significant (p>0.05) after a Bonferroni cor-

Table 4.4: Evolutionary patterns of independently evolving *eater* NIM repeat units.

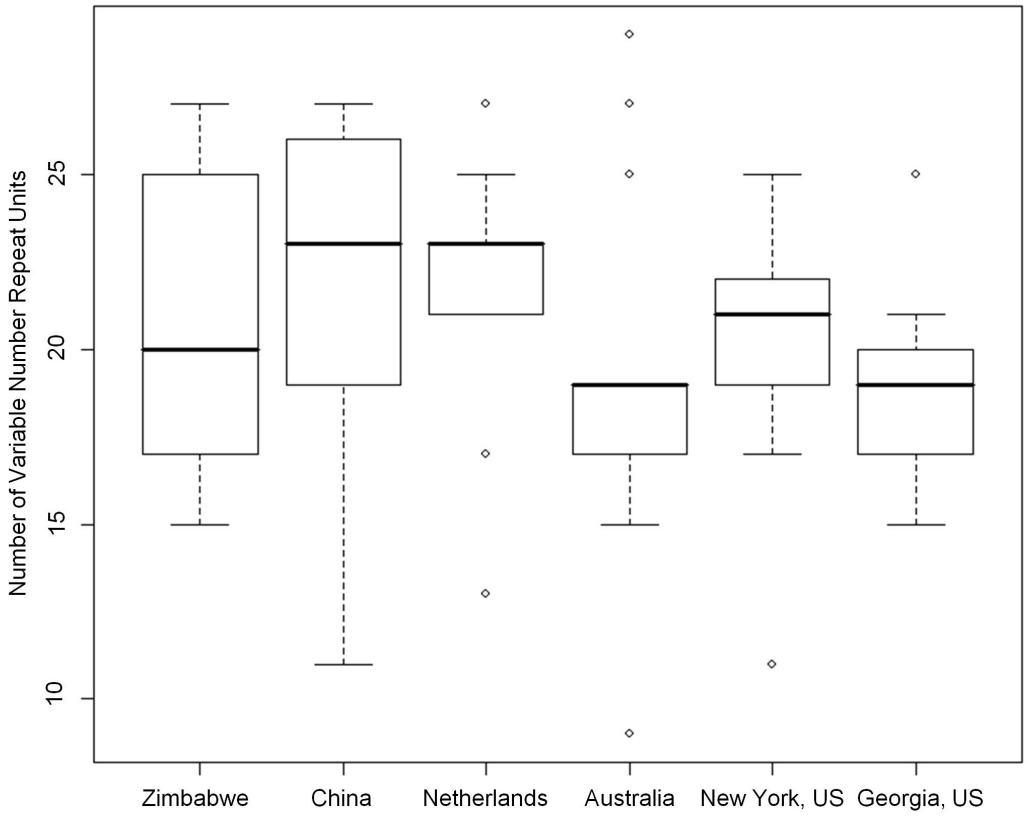
NIM #	Microbial binding?[a]	K_A[b]	K_S[b]	K_A/K_S	$\pi_{zimbabwe}$[c]	π_{US}[c]	Tree structure[d]
1	yes	0.0602	0.2326	0.259	0.00947	0.00769	$yak_{83}(mel_{64}sim)$
2	yes	0	0.1073	0	0	0	$yak_{91}(mel_{92}sim)$
3	yes	0	0.0632	0	0.00697	0.00310	$sim_{40}(mel_{27}yak)$[e]
4	yes	0.0475	0.1145	0.415	0.01111	0.00739	$mel_{93}(yak_{53}sim)$
5	unknown	0	0	NA	0.00539	0.00354	mel/sim/yak
6	unknown	0.0347	0	NA	0.01594	0.00673	$sim_{87}(mel_{48}yak)$
7	unknown	0.0347	0.0556	0.306	0.00208	0.00110	$yak_{77}(mel_{63}sim)$
8	unknown	0	0.0589	0	0.00920	0.00780	yak_{42}(mel/sim)
9	unknown	0.0167	0.1253	0.133	0.01047	0.00449	$yak_{50}(mel_{81}sim)$
10	unknown	0.0360	0.0690	0.522	0.01047	0.00449	$yak_{98}(mel_{63}sim)$[e]
11	unknown	0	0.1337	0	0.00359	0	yak_{99}(mel/sim)

[a] Kocks et al. 2005
[b] K_A and K_S are the rates of amino acid or silent substitution respectively polarized along the *D. melanogaster* branch using *D. yakuba* and *D. simulans* as outgroups.
[c] π indicates nucleotide diversity calculated as the average pairwise differences between sequences per base pair.
[d] Neighbor joining trees were constructed for all NIM repeat units. Subscript numbers indicate bootstrap support for each node based on 500 replicates.
[e] One *D. melanogaster* allele was an outlier from the pattern indicated here.

rection for multiple tests. We therefore find no evidence that the overall length of the variable number repeat region is geographically differentiated or locally adapted.

At the nucleotide sequence level, the variable number repeat units do not tightly cluster phylogenetically based on physical location in the array, in contrast with the conserved-number NIM 1-11, whose nearest phylogenetic neighbors are always physically homologous repeats in alleles isolated from different individuals and from the outgroup species (Somogyi et al., 2008). This suggests

A



B

		Zimbabwe	China	Netherlands	Australia	New York, US
Zimbabwe	(n=10)					
China	(n=14)	-0.0466				
Netherlands	(n=16)	-0.0971	-0.0278			
Australia	(n=18)	-0.0307	0.1138	0.0466		
New York, US	(n=18)	-0.0681	0.0932	0.0215	-0.0456	
Georgia, US	(n=12)	0.1289	0.3585	0.3939*	-0.0114	0.1427

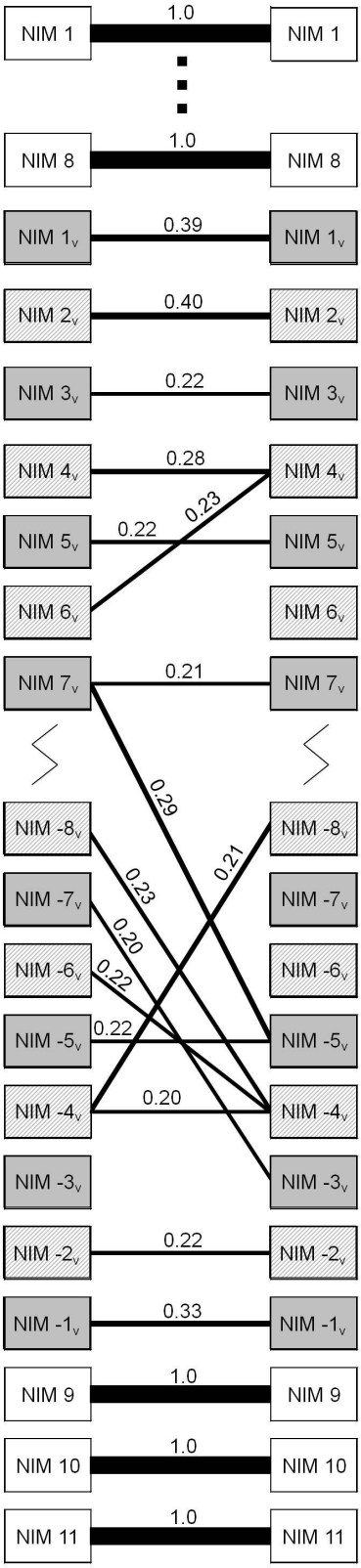
Figure 4.7: Absence of genetic differentiation (R_{ST}) between populations in variable number repeat sizes. a) Boxplots of the distribution of sizes of variable number repeat region by population. b) Pairwise R_{ST} values between populations. *: p=0.0287 (not significant after a Bonferroni correction); p>0.05 for all other pairwise comparisons

that the variable number repeat units are evolving concertedly by birth-and-death of repeat units and gene conversion across paralogous units within the variable number repeat array. This is in contrast with the conserved-number repeat units, which are evolving independently. We find evidence that there is also variation in the evolutionary patterns within the variable number repeat region. For units on the periphery of the variable number repeat region, the lowest genetic distances between units was generally found when comparing between units at the same homologous position (Figure 4.8). No such pattern existed for units in the interior of the array. This indicates a higher degree of independent evolution in units in the periphery than in interior units and suggests that the birth-and-death process that gives rise to new alleles is most likely to occur in the interior of the gene.

4.5 Discussion

The patterns of genetic diversity and the divergence in gene expression between two high frequency haplotypes give strong support for a partial selective sweep at the *eater* locus in a North American population. One haplotype group, labeled the "A" group, has a high level of linkage disequilibrium, extended haplotype homozygosity (EHH) that reaches over a long genomic distance, and a negative value of Tajima's *D*. These extreme values reach statistical significance under two of three previously-described demographic null models (neutral and LS models) of selective neutrality. The model to which the *eater* data can be fit (the TA model) is so general that it provides little resolution between selective and neutral scenarios. Overall, our data are consistent with the "A" haplotype having recently risen to high frequency in North America due to an incomplete

Figure 4.8: Nearest genetic neighbors between NIM repeat units. Genetic distances were calculated between all pairwise combinations of NIM repeats from different individuals. The thickness of the connecting lines and the number on the line indicate the proportion of times that the nearest neighbor of a particular repeat unit was the indicated NIM repeat. Genetic distances were calculated with the Kimura 2-parameter model using an alignment of the 78 base pair NIM consensus motif that is conserved between all repeat units. NIM 1 through NIM 8 all showed the same pattern, so the intervening repeats are not shown (region indicated with dots). Variable number repeat units are shaded ("NIM 8-like" = gray, "alternate" = gray and white stripes). Some variable number repeats units were not sequenced (region indicated with a jagged line).


NIM 1
1.0
NIM 1
NIM 8
1.0
NIM 8
NIM 1v
0.39
NIM 1v
NIM 2v
0.40
NIM 2v
NIM 3v
0.22
NIM 3v
NIM 4v
0.28
NIM 4v
0.23
NIM 5v
0.22
NIM 5v
NIM 6v
NIM 6v
NIM 7v
0.21
NIM 7v
0.29
NIM -8v
NIM -8v
0.23
0.21
NIM -7v
NIM -7v
0.20
NIM -6v
0.22
NIM -6v
NIM -5v
0.22
NIM -5v
NIM -4v
0.20
NIM -4v
NIM -3v
NIM -3v
NIM -2v
0.22
NIM -2v
NIM -1v
0.33
NIM -1v
NIM 9
1.0
NIM 9
NIM 10
1.0
NIM 10
NIM 11
1.0
NIM 11

selective sweep. Notably, the expression of *eater* in isogenic lines with the "A" allele is on average 69% higher than in lines with the "B" allele. This expression difference offers a phenotypic basis upon which selection could act.

The "A" group does not display the strong deficit in nucleotide diversity that could be expected if it had recently and rapidly reached high frequency. This may, however, be a consequence of our assignment of individual alleles to haplotype groups. Of the eight alleles in the "A" group, seven are identical across the entire second intron and five have an average of 2.4 pairwise differences among them for the entire length of the gene, compared with the average of 21.4 pairwise differences among alleles in the "A" group as a whole. The eighth "A" allele brings in the majority of sites segregating in that haplotype group and appears to be a recombinant between the "A" and "B" haplotype groups. A less conservative assignment that excluded this eighth allele from haplotype group "A" would have led to a much more extreme deficit of nucleotide diversity in the "A" group. The 99 base pairs that we used to define the "A" haplotype are perfectly conserved in two lines from Zimbabwe, suggesting that this allele was present in the ancestral population prior to founding of the North American population (c.f., (Pool et al., 2006)), although we cannot exclude the possibility that the haplotype was reintroduced back into the African population by back-migration. Selective events that act on standing genetic variation leave much less dramatic signatures than those seen when selection strongly favors novel mutations (Przeworski et al., 2005). The fact that we are able to see any distortions to the site frequency spectrum at all suggests strong positive selection at this locus.

It is striking that the two "A" haplotypes found in the Zimbabwe popu-

lation are identical across the entire 4kb region in those two individuals (Figure 4.2). This sample size is too small to do simulations similar to those we did with the United States population, but this observation begs the question of whether or not selective sweeps involving these two haplotypes are happening in other populations or if this selective sweep is unique to the United States population. Geographically restricted selective sweeps could potentially stem from adaptation of populations to their local environments (Aminetzach et al., 2005; Macpherson et al., 2008). We surveyed 8 alleles from each of two additional populations from the Netherlands and China at the second intron (data not shown), but found no evidence of the "A" haplotype being present in either of these populations. This suggests that, of the derived populations, the selective sweep involving the "A" haplotype is a local phenomenon restricted to the North American population. In contrast, we find no evidence of genetic differentiation (R_{ST}=0.00388) in the total number of NIM repeat units among populations around the world. The lack of differentiation indicated by this R_{ST} value suggests that the number of repeats is not free to drift to different frequencies in individual populations, and certainly is not adaptively diverging among subpopulations, but instead that the number of repeats is subject to purifying selection.

We have hypothesized that enhancer motifs present in the second intron of "A" group haplotypes but absent in "B" group alleles result in higher expression of *eater* "A" haplotypes, and have noted polymorphism in putative CF2-II binding sites as candidates for responsibility. The CF2-II zinc finger transcription factor is an alternatively spliced variant of the CF2 transcription factor that was first identified in *D. melanogaster* and has been shown to be important during oogenesis and in embryonic muscle tissue development (Hsu et al., 1996;

García-Zaragoza et al., 2008), where it can act as either an enhancer (García-Zaragoza et al., 2008) or repressor (Hsu et al., 1996). In the silkworm *B. mori*, CF2 was found to act as a repressor of expression of the antimicrobial peptide *gloverin* (Mrinal and Nagaraju, 2008). The ancestral member of the *gloverin* family has a CF2 motif in an intron in the 3′ UTR, and a deletion of this intron in other members of the gene family has been associated with the gain of expression of *gloverin* in embryos. Although the haplotype structuring and expression association we identified was centered around CF2-II motifs in the second intron of *eater*, this does not prove that the CF2-II sites are responsible for the expression difference, and does not preclude the role of a different sequence motif either inside or outside this intron. Sequence important for *eater* expression has been identified in the 5′ upstream region of the gene (Tokusumi et al., 2009), and it is possible that a still unidentified region of the gene is responsible for the expression differences between haplotypes.

Increased expression of *eater* and other genes involved in cellular and humoral immunity has previously been reported in *D. melanogaster* selected for increased resistance to the bacterial pathogen *Pseudomonas aeruginosa* (Ye et al., 2009). This supports the hypothesis that higher expression of *eater* is beneficial in the face of pathogen pressure. Artificially selected lines rapidly lost resistance when the selective pressure was removed, suggesting that resistance is costly to maintain. We report evidence of a partial selective sweep at the *eater* locus in a North American population but not in an African population. *D. melanogaster* was likely to have encountered novel pathogens as the population range expanded out of Africa. Geographically restricted selective sweeps can occur if selective pressures such as bacterial species and frequencies vary across different areas. The selective sweep may be ongoing which is why the allele as-

sociated with higher *eater* expression is not fixed in the population, or the costs related to increased expression may inhibit the fixation of this allele.

eater is a cellular recognition gene, a class which shows evidence of rapid evolution between species (Sackton et al., 2007). Like *eater*, other genes in this class have previously shown evidence of selection at the population level. Thioester-containing proteins (TEPs), which are thought to function as opsonins and label microbes for phagocytosis, show evidence of adaptive evolution in *Drosophila*, *Anopheles* mosquitoes, and the crustacean *Daphnia* (Little et al., 2004; Little and Cobbe, 2005; Jiggins and Kim, 2006). In *Tep* genes, positively selected sites are often clustered around putative sites of interaction between host and pathogen, suggesting that co-evolutionary arms races drive their rapid evolution. Single *Tep* genes show evidence of recent selection within an African population of *D. melanogaster* (Jiggins and Kim, 2006) and divergence in gene expression levels between populations (Hutter et al., 2008). Class C scavenger receptor (SR-Cs) proteins are implicated in the internalization of microbial compounds (Ramet et al., 2001), and some members of this family display evidence of adaptive amino acid replacement between species of *Drosophila* (Lazzaro, 2005). SR-Cs show unusual patterns of nucleotide diversity and haplotype structuring within one North American population of *D. simulans* which suggests a recent and rapid rise to high frequency of putatively selected haplotypes (Lazzaro, 2005; Schlenke and Begun, 2005). These previous studies suggest that, although cellular recognition molecules evolve rapidly as a class, unique evolutionary patterns and pressures drive the evolution of individual genes.

Partial selective sweeps have been invoked to explain the presence of high

frequency haplotypes with low genetic diversity, and previous studies have identified *D. melanogaster* loci with similar patterns of genetic variation as we see at *eater* (Hudson et al., 1997; Aminetzach et al., 2005). The *Doc1420* long interspersed element (LINE)-like transposon is a polymorphic insertion in *D. melanogaster* which results in a truncated version of a protein and confers organophosphate pesticide resistance (Aminetzach et al., 2005). There are fewer haplotypes, reduced variation, and excess linkage disequilibrium in the group of alleles containing the element, and the transposon insertion is found in high frequency in derived populations but only low frequency in ancestral African populations (Aminetzach et al., 2005). At the *Sod* locus, two haplotype groups, one within a fast electromorph group and one containing all slow electromorphs, each have very little or no nucleotide diversity (Lee et al., 1981). A complex pattern of selection where the fast haplotype group underwent a partial selective sweep and then a subsequent mutation led to the slow haplotype, which is different by only one amino acid, is the most likely explanation for patterns of variation at this locus (Hudson et al., 1997). It should be noted that in both these examples, the excess linkage disequilibrium and reduced genetic diversity extended as far away as 10kb and therefore these loci may have been subject to stronger or more recent selection.

The coding regions of *eater* are largely composed of NIM repeat units. These repeats in *eater* have been previously identified as evolving either independently or concertedly (Somogyi et al., 2008). Four of the eleven independently evolving repeats have been implicated in microbial binding, and it has been hypothesized that repeats evolving concertedly compose a structural ”stalk” between the ligand binding NIM repeats and the phagocyte membrane (Kocks et al., 2005). Co-evolutionary arms races between pathogens and the host im-

mune response can drive unusual patterns such as accelerated rates of amino acid substitution, selective sweeps, or balancing selection. To look for evidence of pathogen-imposed selection on NIM repeats with functional evidence of microbe binding, we compared evolutionary patterns of these four repeats with the seven other independently evolving repeats. We found no evidence of a difference in the rate of amino acid substitution, patterns of genetic diversity, or phylogenetic relationships with outgroup species. This is consistent with previous evidence that, although *eater* potentially shows evidence of positive selection between *Drosophila* species, selection is not concentrated around pathogen interaction domains (Sackton et al., 2007).

Sequence similarity between NIM repeats is especially high in the interior of the gene and has led to concerted evolution. Concerted evolution can arise because of unequal crossing over due to non-homologous pairing during recombination or because of gene conversion (Charlesworth et al., 1994). We present evidence that the repeat units in the periphery of the variable number repeat region show signs of independent evolution and that the internal repeats are truly evolving concertedly. This is indicated by the observation that units on the periphery are more likely to be most closely related to units in the same physical location in different individuals, whereas units in the interior show no such concordance between physical location and genetic distance. This is also strong evidence that the duplication and deletion of repeat units is more likely to occur in the internal repeats than in the external repeats, in part because non-homologous pairing becomes less likely as the genetic distance between sequences increases (Stephan, 1989). Polymorphism in repeat number like we observe at *eater* can only be caused by unequal crossing over (Smith, 1976). Gene conversion is likely also driving concerted evolution in this region.

In one instance, we observed patterns of concerted and independent evolution within a single NIM repeat unit. The 78 base pairs that define the core consensus "NIM" motif are evolving concertedly in NIM -1_v (Figure 4.8). In contrast, the last 15 base pairs of this repeat are evolving independently (data not shown). This pattern can only be driven by gene conversion and indicates that multiple factors contribute to concerted evolution within *eater*.

The data we have collected at the *eater* locus support a model wherein this recognition molecule, which is critical in the cellular immune response of *D. melanogaster*, is subject to distinctive evolutionary pressures. However, unlike observations for other genes and contrary to our expectations, this selection is not centered around pathogen interaction domains. Instead, selection appears to be acting on gene expression level in a geographically restricted subpopulation. Further experimentation will be required to determine the organismal fitness consequences of variation in *eater* expression. Novel mutations that are selectively advantageous in local environments have a chance to rapidly rise to high frequency and may eventually serve as the basis for between species divergences. Unlike comparisons between species that have found evidence of amino acid adaptation in cellular immune response genes, our data implicates non-coding regulatory changes as playing an important role in the evolution of *eater*.

4.6 Acknowledgements

We thank Ann Hajek, Andy Clark, Sarah Short, Jacob Crawford, Jennifer Comstock, Madeline Galac, and two anonymous reviewers for comments on the

manuscript. We thank Sean Hackett and Andy Clark for providing some of the lines used in this study. Part of this work was carried out by using the resources of the Computational Biology Service Unit from Cornell University which is partially funded by Microsoft Corporation. This work was supported by grants from the National Science Foundation (DEB-0415851) and the National Institutes of Health (AI064950).

CHAPTER 5

SHORT AND LONG TERM PATTERNS OF EVOLUTION WITHIN IMMUNE RESPONSE GENES OF *DROSOPHILA MELANOGASTER*

5.1 Abstract

Immune response genes frequently show evidence of elevated rates of adaptive evolution between species. The precise pattern of evolution depends on the function of the gene within the immune system. Studies of natural populations reveal that tremendous genetic and phenotypic variation in immune function exists within species. We sequenced a large pool of alleles from a *Drosophila melanogaster* population with the goal of understanding how genetic variation within a population is related to long term evolutionary patterns. We find that genes with the highest rates of adaptive evolution between species have low levels of variation within a population. This pattern suggests that the rapid long term evolution of this group of genes is driven by strong directional selection, which results in a short term reduction in nucleotide diversity. Functional classes of immune genes also differed in their levels of within population variation in a manner consistent with their between species evolutionary patterns.

5.2 Introduction

Tremendous genetic and phenotypic variation exists within natural populations, some of which is likely to contribute to differences between species (Lewontin and Hubby, 1966). Some mutations are deleterious and ex-

pected to be found at low frequencies. Some variants with small or neutral effects occur as a result of genetic drift and mutation. An unknown proportion of variation is maintained by positive selection. These processes are expected to leave different signatures across the genome, which can be distinguished using population genetics.

Variation in immune response has important fitness consequences, and genes in the immune system show both evidence of positive selection at the species level (Nielsen et al., 2005; Sackton et al., 2007) and abundant genetic and phenotypic polymorphism within species (Lazzaro et al., 2004). From studies of various insects, it is known that different functional classes within the immune response vary in their long term evolutionary dynamics (Christophides et al., 2002; Sackton et al., 2007). Within populations, there is evidence for phenotypic tradeoffs within the immune system and amongst immunity, reproduction and other fitness traits (Boots and Begon, 1993; Gwynn et al., 2005; Wilfert et al., 2007b; McKean and Nunney, 2008), which leads to maintenance of variation in immune function rather than the fixation of immune resistance. Less is known of how genetic variation in the immune system within a species relates to long evolutionary patterns.

Much of our understanding of the distribution of genetic variation within populations comes from studies of small numbers of genes. Hard selective sweeps reduce genetic diversity within species as the selected variant rises in frequency and fixes in the population, dragging with it linked variants that also rise in frequency due to genetic hitch-hiking (Smith and Haigh, 1974). It has recently been argued that hard selective sweeps are rare and that partial selective sweeps, where the selected mutation does not completely replace variation in

the population, or soft selective sweeps, which occur on standing variation in the population, are responsible for most of adaption within populations (Burke et al., 2010; Pritchard et al., 2010). Partial and soft selective sweeps lead to less dramatic decreases in nucleotide variation than hard selective sweeps. Examples of hard selective sweeps (Obbard et al., 2006, 2010) and partial or soft sweeps (Bangham et al., 2007; Juneja and Lazzaro, 2010) are found in genes involved in the immune response. Genetic diversity can increase in a population by balancing selection, where multiple beneficial alleles are maintained. Balancing selection on immune response genes has not been described in insects but is known in plants (Tian et al., 2002), frogs (Tennessen and Blouin, 2008), and humans (Piertney and Oliver, 2006). Purifying selection, which acts to remove deleterious new mutations from the population, somewhat reduces genetic diversity within a species and also constrains divergence between species. Many immune response genes evolve by purifying selection (Sackton et al., 2007; Mukherjee et al., 2009; Lehmann et al., 2009; Mendes et al., 2010).

Here we undertake a systematic approach to understanding how genes and gene classes within the immune response evolve in *D. melanogaster* with the goal of contrasting short and long term evolutionary patterns. Much progress has been made recently at studying the distribution of genetic variation in populations at a genomewide scale using high-throughput sequencing technologies (Kolaczkowski et al., 2010; Durbin et al., 2010). We took advantage of a target enrichment and high-throughput sequencing approach using a custom-designed Nimblegen SeqCap array and Illumina sequencing to collect population genomic data from immune response genes. With this technique, we were able to sequence a large pool of individuals from a North American population and to measure population genetic parameters. We found that some patterns

of adaptive evolution between species were linked with levels of genetic variation within species, with the most rapidly evolving genes having low levels of nucleotide variation within species suggesting they are subject to hard selective sweeps.

5.3 Methods

5.3.1 Sequence Collection

We employed a target gene enrichment strategy using a Nimblegen sequence capture array (Hodges et al., 2007) that allowed high-throughput sequencing of selected immune response and metabolism genes. Each array contains over 385,000 probes with lengths over 60 base pairs that match target regions in the reference *D. melanogaster* genome (Adams et al., 2000). DNA regions that match our desired targets bind to the array, non-targets are rinsed and then target DNA eluted, yielding a concentrated sample that is enriched for areas of interest. I selected 257 immune response genes whose products recognize, transduce signal, and clear microbial and viral compounds in the cellular and humoral immune systems of *D. melanogaster*. Another 250 genes with metabolic or other non-immune functions were selected which here serve as a set of control genes. A total of 12 additional genes fell into both categories. These genes were included in analyses of immune response genes and as a separate category when comparing between immune response and metabolism genes. We designed a custom sequence capture array which tiles 2,616,934 bp of exonic, intronic, and non-coding sequence in *D. melanogaster*. For each gene region, we sequenced all

exons, 500 bp of upstream sequence, and some or all of the introns.

To simultaneously collect population genetic data for a large number of alleles, DNA extracted from a pool of flies was hybridized to sequence capture arrays. *D. melanogaster* male flies were collected in September and October of 2006 over piles of apples at Little Tree Orchards, Newfield, Tompkins County, New York, USA. Females were excluded because they cannot definitively be identified to the species level using morphological characters. Whole flies were stored at -80°C until DNA was extracted. DNA was extracted *en masse* from a pool of 299 flies using a PureGene DNA Purification Kit (Gentra Systems) according to manufacturer's instructions. DNA was nebulized to ˜250 bp pieces and Illumina amplification and sequencing primers were ligated onto the sheared DNA. The ligation products were hybridized to Nimblegen arrays following standard protocols by the Microarray Core Facility at Cornell University. Eluted product was sequenced on the Illumina Solexa GAII platform with 60 bp paired end reads at the Life Sciences Core Laboratories Center at Cornell University.

5.3.2 Sequence Alignment and Analysis Pipeline

A total of 2 x 16,533,784 reads were obtained for a total of 1.98 billion base pairs of sequence. No trimming was applied prior to read mapping. Fastaq sequences were converted from fastaq-Illumina format to fastaq-Sanger format using the SeqIO module in Biopython. Sequence reads were aligned to reference *D. melanogaster* sequence using paired end mapping in BWA version 0.8.5 (Li et al., 2009) with default settings except that during the 'aln' command a maximum of 6 mismatches was specified using the '-n 6' command. The reference

sequences were target capture regions from the *D. melanogaster* genome (Adams et al., 2000) with 500 bp of flanking sequence added to each end.

Alignments generated by BWA were subject to quality filtering using scripts written in R (R Development Core Team, 2006). Reads with a mapping quality score below 40 or an insert size greater than 500 base pairs were discarded to reduce potential mapping errors. Reads with insert sizes less than 120 bp resulted in overlapping sequencing sequence being obtained from mate pairs. In these cases, reads were trimmed to eliminate overlap and CIAGR mapping codes recalculated. In some cases, the same amplicon was sequenced more than once resulting in an overrepresentation of that sequence in allele frequency estimates. These events are identifiable because of shared mapping start positions at both ends of the paired-end sequence. One of these redundant sequences was randomly selected and retained for downstream analysis, and all others were discarded.

Pileup files, which are compilations of base calls and sequence and mapping qualities for all reads that mapped to our target regions, were generated using SAMtools version 0.1.7a (Li et al., 2009). The number of sequences that covered each base pair varied because Illumina sequences DNA at random in each sample. The population genetic analyses performed here require equal coverage, so pileup files were randomly thinned to a fixed coverage of 60X with a minimum quality score of 20. Coverage of 60X was chosen because we wanted as high coverage as possible to allow for robust estimation of allele frequencies while still covering a large percentage of the target region. Given this depth of coverage and random sampling, we expect that on average 54 of 60 reads will represent unique alleles within the pool of 598. Sites with coverage below the

cutoff were not analyzed. Twenty-four base pair windows around indels identified by SAMtools were also excluded from analysis. These represent areas that introduce sequence hybridization bias and mapping error, which was corroborated by drops in sequence coverage around putative indels. The bacteriophage phiX174 was sequenced at the same time as our sample as a control, and this analysis pipeline was repeated on those sequences. Because the genome sequence is known for phiX174, we were able to use this control to estimate the overall rate of error after quality filtering was applied.

A pipeline was also created for annotating each base pair on the capture array. Each position was labeled as appropriate as coding, upstream, exonic, intronic, 5'UTR, 3'UTR, and plus/minus strand using the *D. melanogaster* genome release 5.29. Polymorphisms were labeled as synonymous or non-synonymous. This information was used to label polymorphisms as synonymous or non-synonymous. Coding sequence alignments of *D. simulans*, *D. sechellia*, *D. yakuba*, and *D. erecta* used by Clark et al. (2007) were obtained from FlyBase. These alignments were used to identify synonymous and non-synonymous fixations between species. They were also used to identify the derived allele in *D. melanogaster* for generating the site frequency spectra using either *D. simulans* or *D. sechellia* as the outgroup. Because the alignments with *D. sechellia* covered a larger portion of the genome, this species was used as the outgroup for the remainder of the analysis. Sites segregating with greater than two states were regarded by the frequency of the reference allele. Local meiotic recombination rates were estimated using the *Drosophila melanogaster* Recombination Rate Calculator: Version 2.1 (Fiston-Lavier et al., 2010).

5.3.3 Population Genetics

Nucleotide variation metrics θ_S and θ_π were calculated per nucleotide (Hartl and Clark, 2007). θ_s is equal to S/La_n, where S is the total number of segregating sites, L is the length of the sequence in base pairs, n is the number of sequences, and $a_n = \sum_{i=1}^{n} 1/i$. θ_π is equal to π/L, where π is the average of the number of pairwise differences between samples. Illumina sequencing has a relatively high rate of error which resulted in an abundance of singletons, or variants that were present at a frequency of 1/60 or 59/60. For this reason, revised estimates of θ_S and θ_π were also calculated. $\theta_{S_{-\eta_1}}$, where $S_{-\eta_1}$ is the number of segregating sites without singletons, was calculated

$$\hat{\theta}_{S_{-\eta_1}} = \frac{S_{-\eta_1}}{a_n - n/(n-1)}$$

as derived by Achaz (2008). $\theta_{\pi_{JS}}$ was calculated by revising π based on the observed error rate and was calculated as

$$\theta_{\pi_{JS}} = \frac{\frac{2}{n(n-1)} \sum_{i<j}^{n} \pi_{ij_o} - \mathrm{E}[p_2]}{1 - \mathrm{E}[p_1] - \mathrm{E}[p_2]}$$

where $\mathrm{E}[p_1] = (2/3)\varepsilon(1 - \varepsilon) + (2/9)\varepsilon^2$ and $\mathrm{E}[p_2] = 2\varepsilon(1 - \varepsilon) + (2/3)\varepsilon^2$ and ε is the average error probability as derived by Johnson and Slatkin (2008). X chromosome values were multiplied by 4/3 to correct for smaller effective population size relative to autosomes (assuming equal numbers of males and females). Simulations were performed using the program *ms* (Hudson, 2002) to assess the performance of these statistics assuming a neutral bottleneck model similar to the one undergone by North American populations of *D. melanogaster* (Li and Stephan, 2006).

5.4 Results

5.4.1 Recovery of Target Genes

A total of 1,137,496 base pairs (43.5%) of the target region had at least 60X coverage after quality filtering was applied, meaning that 68 million base pairs of sequence were analyzed for this study. Of the 512 genes on the sequence capture array, 492 were represented post-quality filtering with over 100 base pairs of sequence and 256 were represented with over 1,357 base pairs of sequence (Table 5.1). This high depth of coverage of hundreds of genes spanning over a megabase of the genome supports the idea that population genomic data can efficiently be collected using Nimblegen sequence capture and Illumina sequencing.

5.4.2 Quality Checking

The average error rate for the phiX control lane was 0.67% before post-alignment quality filters described in the methods were applied. After quality filtering, the error rate drops to 0.10%, or ~10 errors per 1,000 base pairs. One concern is that the sequence hybridization procedure may bias towards recovery of alleles that most closely match the reference genome used to design the array. To establish the accuracy of our methods, sequence capture data were compared with data from previous studies for a subset of the genes on the array. In the previous studies, 12 isogenic lines derived from a Pennsylvania population were Sanger sequenced at the *SR-CIII/I*, *SR-CII*, *SR-CIV*, *Defensin*, *Attacin C*, and *Metchnikowin* loci (Lazzaro and Clark, 2001; Lazzaro, 2005). There is a

Table 5.1: Coverage of the sequence capture target region by gene category. Many of the genes analyzed here have previously been analyzed among *Drosophila* species (Sackton et al., 2007; Larracuente et al., 2008) and within *D. melanogaster* (Obbard et al., 2009).

	# of genes		
	On Sequence Capture Array	On Array & Sequenced to 60X at >100 bp [post-quality filtering]	On Array & Sequenced to 60X at >1,357 bp [post-quality filtering]
Immunity genes	**257**	**235**	**110**
Sackton et al., 2007	233	224	104
Larracuente et al., 2008	139	139	71
Obbard et al., 2009	116	109	39
Metabolism genes	**250**	**245**	**138**
Sackton et al., 2007	0	0	0
Larracuente et al., 2008	152	150	82
Obbard et al., 2009	6	6	4
Immunity/metabolism genes	**12**	**12**	**8**
Sackton et al., 2007	12	12	8
Larracuente et al., 2008	10	10	7
Obbard et al., 2009	4	4	4

significant correlation ($p<0.01$) between the allele frequency estimates in previous studies and those obtained here (Figure 5.1), and no substantial bias towards recovery of the reference allele is seen. The estimates of allele frequencies are not identical between the two studies, which may be attributed to using different populations, to misestimating allele frequencies using Sanger sequencing with a small number of lines, or to variation introduced by sequence capture or Illumina sequencing.

A total of 553 sites out of 5,381 aligned nucleotides were polymorphic in either dataset. The ability to detect low frequency polymorphisms depends on the depth of coverage, which was 12 for the Sanger data and 60 for the Illumina

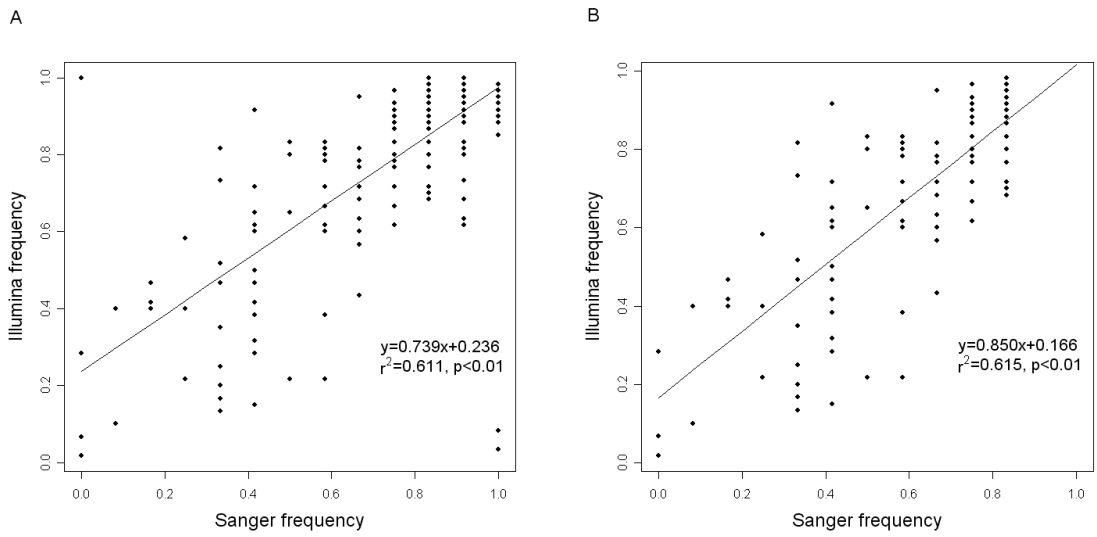


Figure 5.1: Frequency of the reference allele as estimated by Sanger sequencing of a Pennslyvania population (Lazzaro and Clark, 2001; Lazzaro, 2005) or by sequence capture and Illumina sequencing of a New York population. A) All polymorphic sites, B) Excluding singleton or fixed sites that were at a frequency of 11/12 or 12/12 in the Sanger data or 59/60 or 60/60 in the Illumina data.

data. For the Sanger data, in which all singletons were verified for accuracy, 162 polymorphic sites were detected. Of these, 151 were also detected by Illumina. Notably, the SNPs undetected by Illumina were all at a low frequency of either 1/12 or 2/12 in the Sanger data. For the Illumina data, a total of 540 polymorphic sites were detected. The nucleotide diversity (θ_S) estimated from the Sanger data was 0.0100 and from the Illumina data was 0.0215. Calculating nucleotide diversity without the low-confidence singletons ($\theta_{S_{-\eta_1}}$) gives an estimate of 0.0100 for the Illumina data. The similarity between this estimate and the one obtained from the Sanger data supports the idea that $\theta_{S_{-\eta_1}}$ is more accurate than θ_S for Illumina data. These data demonstrate that Illumina data can be used to discover SNPs and estimate their frequency, especially when SNPs have

intermediate population frequencies.

Nucleotide variation ($\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$) was calculated for each individual gene region, which includes the coding and non-coding sequence between the 5'UTR or first exon and 3'UTR or last exon. The variance in $\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$ was higher for shorter gene regions (Figure A.1), which is expected due to the sampling variance being higher when fewer sites are involved (Figure A.2). Based on these data and simulations, only genes whose lengths were above the median length of 1,357 base pairs for gene regions, 862 base pairs for coding regions, and 264 base pairs for introns were analyzed to minimize this noise. Although short and long genes are known to differ somewhat in their evolutionary properties (Begun et al., 2007), we are not biasing our results since we never make comparisons between short and long genes.

We assessed the performance of our measures of nucleotide variation in several ways. Simulations suggest a high correlation between traditional statistics θ_S and θ_π measured in simulated samples and corrected statistics $\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$ measured after sequencing error was introduced (Figure A.3). The correlation is slightly higher between θ_π and $\theta_{\pi_{JS}}$. However, this approach has the downside of requiring an external estimate of the rate of sequencing error, which we obtained by estimating error in the phiX control and is not necessarily an accurate reflection of error in our data. This correction changes the absolute value of diversity but not the rank order and therefore is especially useful for comparisons between genes and gene classes. Simulations demonstrate that both metrics give robust estimates of nucleotide variation and thus results from both are presented for comparison.

Nucleotide variation measured by $\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$ are highly correlated

(r^2=0.895, p<0.001) (Figure A.4). $\theta_{S_{-\eta_1}}$ tends to be slightly higher than $\theta_{\pi_{JS}}$, as expected since $\theta_{S_{-\eta_1}}$ is more influenced by the increase in low frequency sites that occurs in an expanding recently founded population such as this one. Nucleotide variation tends to be higher in introns than in coding regions (intron mean $\theta_{\pi_{JS}}$=0.0050, mean $\theta_{S_{-\eta_1}}$=0.0060; coding region mean $\theta_{\pi_{JS}}$=0.0042, mean $\theta_{S_{-\eta_1}}$=0.0050) (Figure A.5), which indicates greater constraint on coding regions and is supported by the literature (Andolfatto, 2005; Durbin et al., 2010). Andolfatto (2005) found that the mean θ_π was 0.0125 in introns and 0.0108 in coding regions on the X chromosome in a more genetically diverse African population of *D. melanogaster*, which is approximately the same ratio of variation that we observe between coding and introns. Glinka et al. (2003) found that the mean θ_π was 0.0046 and the mean θ_S was 0.0044 in introns on the X chromosome in a cosmopolitan population similar to ours, which is comparable to the levels of variation that we observe in introns in our population. The results from simulations and the agreement of these data with previous observations suggest that $\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$ provide accurate estimates of nucleotide variation.

5.4.3 Short versus Long Term Patterns of Evolution

To gain insight into the relationship between short and long term evolution, we compared our within population estimates of variation with estimates of the rate of adaptive evolution (ω) along the *D. melanogaster* species lineage (Larracuente et al., 2008) (Table 5.1, Figure 5.2). ω is the ratio of the rate of non-synonymous substitutions (d_N) to synonymous substitutions (d_S), with high ω values, especially those over one, taken to indicate adaptive evolution and low ω values, especially those less than 0.1, taken to indicate purifying selection.

It is a conservative measure of adaptive evolution because a high ω requires many nonsynonymous substitutions and would not be influenced by a small number of substitutions with large effects. We expected that genes with high ω values would have low levels of variation ($\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$) if adaptive evolution is driven by recurrent hard selective sweeps, intermediate levels if driven by partial or soft selective sweeps, and high levels if driven by balancing selection. We divided the data into high and low amounts of variation based on the median ($\theta_{S_{-\eta_1}}$=0.00473 or $\theta_{\pi_{JS}}$=0.00392) and high and low rates of adaptive evolution based on an ω of 0.15 and tested to see if data points were homogenously distributed amongst all quadrants (Figure 5.2). We observed significant heterogeneity in the distribution (Fisher's exact test, p<0.005 for $\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$) that is consistent with an excess of genes with high rates of evolution and low amounts of variation. Similar heterogeneities were found using ω=0.2, 0.3 or 0.4, but not with ω=0.1, suggesting that only rapidly evolving genes have deficits in nucleotide variation. This pattern indicates that the adaptively evolving genes are subject to hard selective sweeps within species.

Local recombination rate has previously been shown to be correlated with levels of polymorphism (Begun and Aquadro, 1992). We see a similar pattern in our data, with significantly higher levels of nucleotide variation ($\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$) seen in genes that are found in regions of high meiotic recombination ($\theta_{S_{-\eta_1}}$ Spearman ρ=0.417, p<0.001; $\theta_{\pi_{JS}}$ Spearman ρ= 0.427, p<0.001) (Figure 5.3). No evidence was found for a correlation between recombination rate and the rate of adaptive evolution (ω) (Figure 5.4), and no heterogeneity was found in the rates of adaptive evolution (ω) of genes with high and low rates of recombination (Fisher's exact test, NS). This suggests that while recombination rate is correlated with the level of nucleotide variation, it is not responsible for driving the

relationship between nucleotide variation and the rate of adaptive evolution.

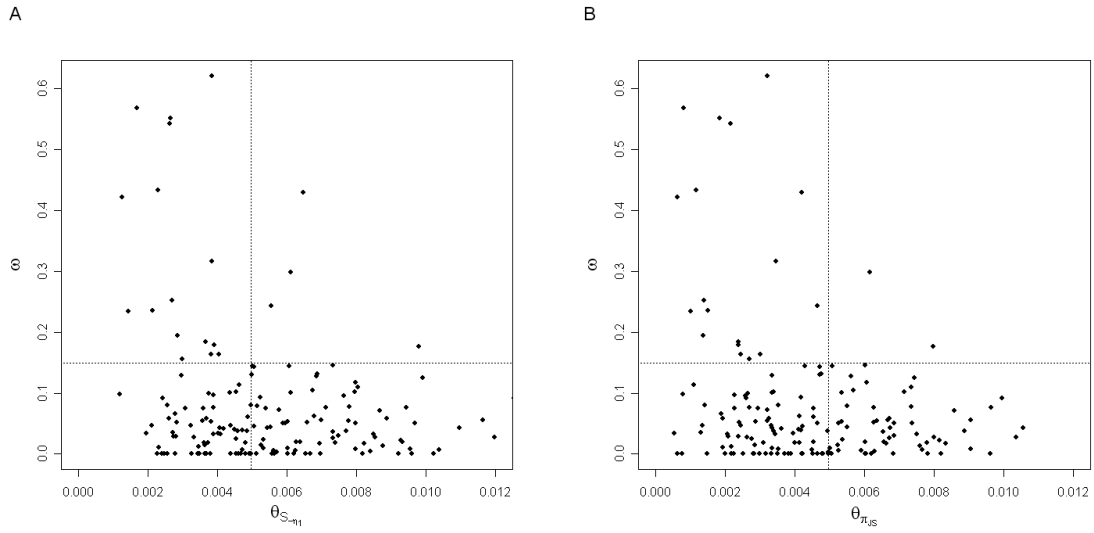


Figure 5.2: Genes with high rates of adaptive evolution (ω) along the *D. melanogaster* species lineage have low levels nucleotide variation within species suggestive of hard selective sweeps. The rate of adaptive evolution (ω) versus A) $\theta_{S_{-\eta_1}}$ and B) $\theta_{\pi_{JS}}$. Dotted lines separate nucleotide variation at the median ($\theta_{S_{-\eta_1}}$=0.00473 and $\theta_{\pi_{JS}}$=0.00392) and ω at 0.15. Fisher's exact tests indicate that genes are not homogenously distributed amongst quadrants ($\theta_{S_{-\eta_1}}$ p<0.005 and $\theta_{\pi_{JS}}$ p<0.005) with an apparent deficit of genes that have high ω values and high levels of nucleotide variation.

5.4.4 Evolutionary Patterns of Functional Gene Categories

Functional gene categories were compared to see if evolutionary patterns vary among genes with different roles. Levels of nucleotide variation ($\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$) did not vary between metabolism and immune response genes (Figure A.6). Within the immune system, genes with roles in the recognition or signal transduction of infection have significantly lower nucleotide variation ($\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$)

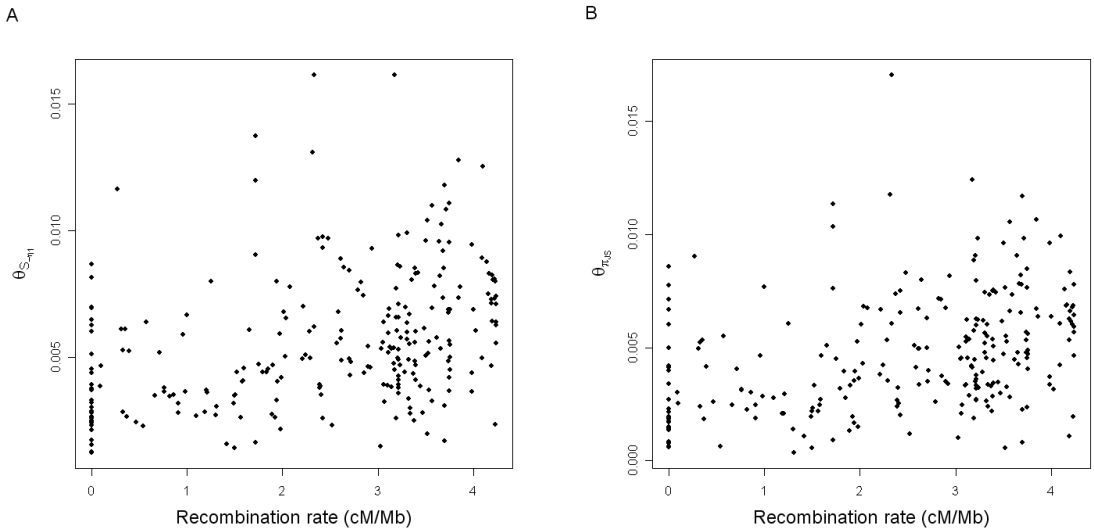


Figure 5.3: Recombination rate is positively correlated with estimates of nucleotide variation A) $\theta_{S_{-\eta_1}}$ (Spearman ρ=0.417, p<0.001) and B) $\theta_{\pi_{JS}}$ (Spearman ρ= 0.427, p<0.001).

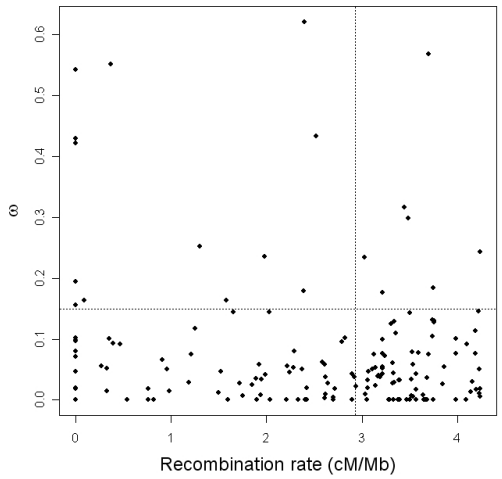


Figure 5.4: Recombination rate is not related to the rate of adaptive evolution (ω) (Spearman ρ= -0.075, p= 0.349). Dotted lines separate recombination rate at the median of 2.93 and ω at 0.15. A Fisher's exact test indicates that genes are homogenously distributed amongst quadrants.

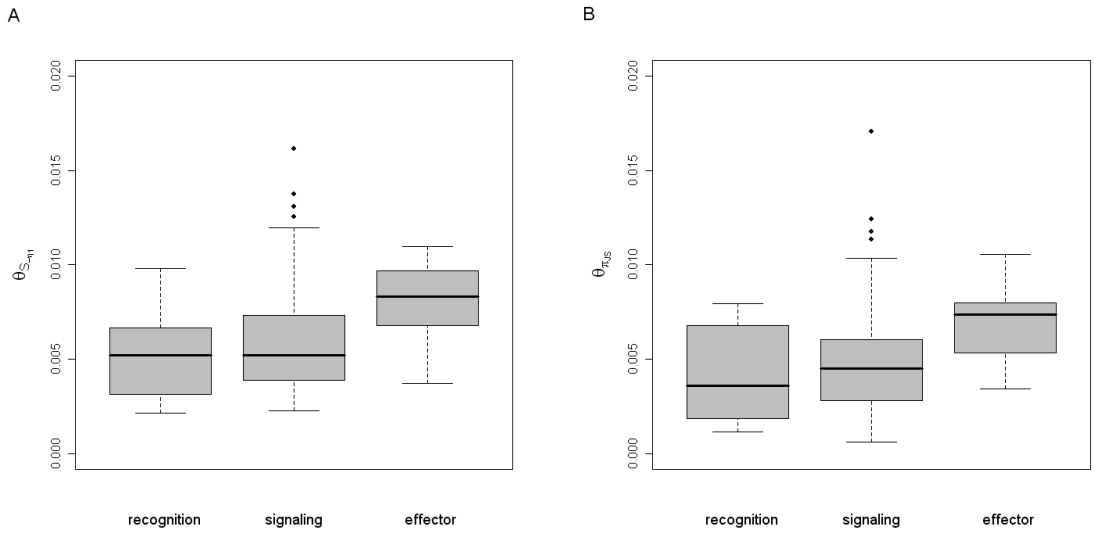


Figure 5.5: Nucleotide variation ($\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$) varies among functional categories of immunity genes. A total of 19 recognition genes, 82 signaling genes, and 18 effector genes were included in this analysis. Signaling and recognition genes both have significantly lower A) $\theta_{S_{-\eta_1}}$ (Wilcoxon rank sum test signaling-effector $p<0.005$, recognition-effector $p<0.005$) and B) $\theta_{\pi_{JS}}$ (Wilcoxon rank sum test signaling-effector $p<0.01$, recognition-effector $p<0.005$) than effector genes but are not significantly different from each other.

than effectors, or genes involved the clearance of infection (Figure 5.5). It has previously been demonstrated that a higher proportion of recognition and signaling genes are positively selected compared to the genomewide expectation (Waterhouse et al., 2007; Sackton et al., 2007). The low levels of nucleotide variation in these adaptively evolving gene categories suggest that these genes are subject to hard selective sweeps. We find no evidence that genes involved in the cellular versus humoral branches of the immune response differ in their levels of nucleotide variation (Figure A.7), although it has previously been observed that cellular recognition genes evolve faster than humoral recognition genes over the long-term (Sackton et al., 2007). We can identify potential can-

didate regions of selective sweeps or balancing selection by examining the extreme tails of the distributions of nucleotide variation (Tables 5.2, 5.3, 5.4, 5.5). Of the 25 genes with the lowest levels of nucleotide variation, all are either recognition or signaling genes and 10 have previously shown evidence for adaptive evolution in *Drosophila* species (Tables 5.2, 5.3). Two genes, *Dcr-2* and *AGO2*, have amongst the highest rates of adaptive evolution in the *D. melanogaster* genome and have previously been shown to have low levels of variation (Obbard et al., 2006, 2010). Effector genes are found amongst the 25 genes with the highest levels of nucleotide variation, of which only 5 have previously shown evidence of adaptive evolution (Tables 5.4, 5.5).

One concern is that methological biases could influence the patterns of variation we observed between gene classes. Effector genes and to some degree the recognition genes are often members of gene families and signaling genes are more likely to be conserved as one-to-one orthologues between species (Waterhouse et al., 2007; Sackton et al., 2007). Regions with high sequence similarity were excluded from the SeqCap array design. However, relatedness amongst members of gene families could lead to sequence reads being discarded during the assembly process if they match multiple regions of the genome and would lead to a drop in coverage from these regions. We found no significant differences in median sequence coverage pre-filtering for immunity versus metabolism genes, comparisons of immune functional classes, or single copy genes versus members of gene families. Alternatively, if sequence reads are being misassembled to paralogs instead of being discarded, this would artificially inflate nucleotide diversity estimates. We found no significant difference in nucleotide variation ($\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$) between genes that are only found once within *D. melanogaster* and genes that are members of gene families (Figure A.8). These

Table 5.2: List of genes with the 25 lowest values of $\theta_{S_{-\eta_1}}$

Rank	Symbol	Name	Synonyms	Chr	Immune Classification	$\theta_{S_{-\eta_1}}$	A	B	C
1	AGO2	Argonaute 2	.	3L	other	0.00163	?	-	-
2	Dcr-2	Dicer-2	.	2R	other	0.00168	+	+	+
3	cactin	.	.	X	signaling	0.00171	?	?	?
4	os	outstretched	upd	X	signaling	0.00197	-	-	?
5	crq	croquemort	.	2L	recognition	0.00215	+	-	?
6	Ras85D	Ras oncogene at 85D	Ras1	3R	signaling	0.00227	-	-	?
7	CG31217	.	.	3R	recognition	0.00230	-	+	?
8	emb	embargoed	Crm1	2L	signaling	0.00233	-	+	?
9	Atf-2	Activating transcription factor-2	.	2R	signaling	0.00240	?	-	?
10	emp	epithelial membrane protein	.	2R	recognition	0.00242	-	+	?
11	bhr	akirin	bhringi	3L	signaling	0.00248	-	?	?
12	brm	brahma	.	3L	signaling	0.00260	?	-	?
13	egr	eiger	.	2R	signaling	0.00261	-	?	?
14	ref(2)P	refractory to sigma P	.	2L	signaling	0.00263	+	-	-
15	Rel	Relish	.	3R	signaling	0.00265	+	-	-
16	CG11023	.	.	2L	signaling	0.00280	?	?	?
17	cact	cactus	.	2L	signaling	0.00282	-	-	-
18	MP2	melanization protease 2	.	3R	signaling	0.00286	-	-	-
19	SAE2	.	Uba2	3L	signaling	0.00296	-	-	?
20	Toll-9	Toll-9	.	3L	signaling	0.00300	-	-	-
21	POSH	Plenty of SH3s	.	2R	signaling	0.00308	-	-	?
22	TepI	Thiolester containing protein I	.	2L	recognition	0.00315	?	+	+
23	TepIV	Thiolester containing protein IV	.	2L	recognition	0.00318	?	+	-
24	Ulp1	.	.	X	signaling	0.00332	?	+	?
25	hep	hemipterous	MKK7	X	signaling	0.00340	?	-	-

Column A, data for adaptive evolution along the *D. melanogaster* branch (Larracuente et al. 2008) with an FDR<=0.05.
Column B, data for adaptive evolution in *Drosophila* (Sackton et al. 2007) with an FDR<=0.05.
Column C, data for adaptive evolution within *D. melanogaster* (Obbard et al. 2009) with an FDR<=0.05.
?, gene has not been tested for adaptive evolution
-, gene was tested but no evidence for adaptive evolution was found
+, gene was tested and evidence for adaptive evolution was found

Table 5.3: List of genes with the 25 lowest values of $\theta_{\pi_{JS}}$

Rank	Symbol	Name	Synonyms	Chr	Immune Classification	$\theta_{\pi_{JS}}$	A	B	C
1	cactin	.	.	X	signaling	0.00019	?	?	?
2	Ras85D	Ras oncogene at 85D	Ras1	3R	signaling	0.00064	-	-	?
3	Dcr-2	Dicer-2	.	2R	other	0.00081	+	+	+
4	AGO2	Argonaute 2	.	3L	other	0.00092	?	-	-
5	CG31217	.	.	3R	recognition	0.00118	-	+	?
6	CG11023	.		2L	signaling	0.00136	?	?	?
7	MP2	melanization protease 2	.	3R	signaling	0.00139	-	-	-
8	os	outstretched	upd	X	signaling	0.00142	-	-	?
9	emp	epithelial membrane protein	.	2R	recognition	0.00148	-	+	?
10	crq	croquemort	.	2L	recognition	0.00151	+	-	?
11	Atf-2	Activating transcription factor-2	.	2R	signaling	0.00170	?	-	?
12	hep	hemipterous	MKK7	X	signaling	0.00178	?	-	-
13	TepIV	Thiolester containing protein IV	.	2L	recognition	0.00183	?	+	-
14	Rel	Relish	.	3R	signaling	0.00185	+	-	-
15	emb	embargoed	Crm1	2L	signaling	0.00194	-	+	?
16	TepI	Thiolester containing protein I	.	2L	recognition	0.00194	?	+	+
17	egr	eiger	.	2R	signaling	0.00194	-	?	?
18	Stat92E	Signal-transducer and activator of transcription protein at 92E	.	3R	signaling	0.00208	-	-	-
19	cact	cactus	.	2L	signaling	0.00209	-	-	-
20	ref(2)P	refractory to sigma P	.	2L	signaling	0.00215	+	-	-
21	bhr	akirin	bhringi	3L	signaling	0.00219	-	?	?
22	brm	brahma	.	3L	signaling	0.00220	?	-	?
23	POSH	Plenty of SH3s	.	2R	signaling	0.00226	-	-	?
24	puc	puckered	.	3R	signaling	0.00228	-	+	?
25	Dredd	Death related ced-3/Nedd2-like protein	.	X	signaling	0.00237	+	-	-

Column A, data for adaptive evolution along the *D. melanogaster* branch (Larracuente et al. 2008) with an FDR<=0.05.
Column B, data for adaptive evolution in *Drosophila* (Sackton et al. 2007) with an FDR<=0.05.
Column C, data for adaptive evolution within *D. melanogaster* (Obbard et al. 2009) with an FDR<=0.05.
?, gene has not been tested for adaptive evolution
-, gene was tested but no evidence for adaptive evolution was found
+, gene was tested and evidence for adaptive evolution was found

Table 5.4: List of genes with the 25 highest values of $\theta_{S_{-\eta_1}}$

Rank	Symbol	Name	Synonyms	Chr	Immune Classification	$\theta_{S_{-\eta_1}}$	A	B	C
94	Pvf2	PDGF- and VEGF-related factor 2	Vegf	2L	signaling	0.00732	-	-	?
95	Stam	Signal transducing adaptor molecule	.	2L	signaling	0.00732	-	-	?
96	psh	persephone	.	X	signaling	0.00736	?	-	-
97	Mcr	.	TepVI	2L	recognition	0.00739	-	-	?
98	Tig	Tiggrin	.	2L	effector	0.00748	-	-	?
99	CG14225	.	.	X	signaling	0.00772	?	-	-
100	Nos	Nitric oxide synthase	.	2L	signaling	0.00777	-	+	-
101	kay	kayak	Fos	3R	signaling	0.00795	-	-	?
102	Pvr	PDGF- and VEGF-receptor related	Vegfr	2L	signaling	0.00797	-	-	+
103	CG8492	.	.	3L	effector	0.00804	-	-	?
104	Pvf3	PDGF- and VEGF-related factor 3	Vegf	2L	signaling	0.00808	?	-	?
105	Toll-4	Toll-4	.	2L	signaling	0.00830	?	?	?
106	PGRP-LD	Peptidoglycan recognition protein LD	.	3L	recognition	0.00831	?	-	+
107	Dox-A3	Diphenol oxidase A3	.	2R	effector	0.00831	?	-	?
108	Tsf2	Transferrin 2	.	3L	effector	0.00854	-	-	?
109	Duox	Dual oxidase	.	2L	effector	0.00864	?	-	?
110	Traf3	Traf3	.	X	signaling	0.00908	?	-	?
111	CG6361	CG6361	.	X	signaling	0.00909	?	-	?
112	Egfr	Epidermal growth factor receptor	.	2R	signaling	0.00956	+	-	?
113	Dnr1	Defense repressor 1	.	2R	signaling	0.00960	-	+	-
114	PGRP-LC	Peptidoglycan recognition protein LC	.	3L	recognition	0.00980	-	+	-
115	Pu	Punch	.	2R	effector	0.01023	-	-	?
116	Bc	Black cells	Dox-A1	2R	effector	0.01083	?	-	-
117	Tsf3	Transferrin 3	.	2R	effector	0.01097	-	-	?
118	gcm2	.	.	2L	signaling	0.01253	-	-	?

Column A, data for adaptive evolution along the *D. melanogaster* branch (Larracuente et al. 2008) with an FDR<=0.05.
Column B, data for adaptive evolution in *Drosophila* (Sackton et al. 2007) with an FDR<=0.05.
Column C, data for adaptive evolution within *D. melanogaster* (Obbard et al. 2009) with an FDR<=0.05.
?, gene has not been tested for adaptive evolution
-, gene was tested but no evidence for adaptive evolution was found
+, gene was tested and evidence for adaptive evolution was found

Table 5.5: List of genes with the 25 highest values of $\theta_{\pi_{JS}}$

Rank	Symbol	Name	Synonyms	Chr	Immune Classification	$\theta_{\pi_{JS}}$	A	B	C
94	CanA1	calcineurin A1	.	3R	signaling	0.00667	?	?	?
95	Stam	Signal transducing adaptor molecule	.	2L	signaling	0.00672	-	-	?
96	Tehao	.	Toll-5	2L	signaling	0.00674	-	-	-
97	Pvf3	PDGF- and VEGF-related factor 3	Vegf	2L	signaling	0.00677	?	-	?
98	Su(H)	Suppressor of Hairless	.	2L	signaling	0.00683	-	-	?
99	Pvr	PDGF- and VEGF-receptor related	Vegfr	2L	signaling	0.00685	-	-	+
100	Tig	Tiggrin	.	2L	effector	0.00685	-	-	?
101	Toll-4	Toll-4	.	2L	signaling	0.00689	?	?	?
102	Traf3	Traf3	.	X	signaling	0.00699	?	-	?
103	CG13079	Thiolester containing protein V	TepV	2L	recognition	0.00711	?	-	?
104	kay	kayak	Fos	3R	signaling	0.00715	-	-	?
105	CG8492	.	.	3L	effector	0.00735	-	-	?
106	Dox-A3	Diphenol oxidase A3	.	2R	effector	0.00743	?	-	?
107	PGRP-LD	Peptidoglycan recognition protein LD	.	3L	recognition	0.00755	?	-	+
108	Mcr	.	TepVI	2L	recognition	0.00780	-	-	?
109	Pu	Punch	.	2R	effector	0.00783	-	-	?
110	PGRP-LC	Peptidoglycan recognition protein LC	.	3L	recognition	0.00797	-	+	-
111	Tsf2	Transferrin 2	.	3L	effector	0.00799	-	-	?
112	Egfr	Epidermal growth factor receptor	.	2R	signaling	0.00906	+	-	?
113	Duox	Dual oxidase	.	2L	effector	0.00907	?	-	?
114	CG6361	CG6361	.	X	signaling	0.00959	?	-	?
115	Dnr1	Defense repressor 1	.	2R	signaling	0.00962	-	+	-
116	Bc	Black cells	Dox-A1	2R	effector	0.00984	?	-	-
117	gcm2	.	.	2L	signaling	0.00994	-	-	?
118	Tsf3	Transferrin 3	.	2R	effector	0.01056	-	-	?

Column A, data for adaptive evolution along the *D. melanogaster* branch (Larracuente et al. 2008) with an FDR<=0.05.
Column B, data for adaptive evolution in *Drosophila* (Sackton et al. 2007) with an FDR<=0.05.
Column C, data for adaptive evolution within *D. melanogaster* (Obbard et al. 2009) with an FDR<=0.05.
?, gene has not been tested for adaptive evolution
-, gene was tested but no evidence for adaptive evolution was found
+, gene was tested and evidence for adaptive evolution was found

data suggest that artifacts of experimental design do not contribute to the differences we observe between functional classes of genes.

5.5 Discussion

Some of the long term evolutionary patterns that are observed in genes of the *D. melanogaster* immune system and genomewide can be explained by patterns of variation within species. Genes with the highest rates of adaptive evolution between species have low rates of nucleotide diversity and polymorphism suggesting a role for recurrent selective sweeps in driving rapid evolution of these genes. In contrast, genes with high levels of variation have average rates of adaptive evolution between species suggesting that factors that maintain variation within species do not lead to rapid evolution in the long term.

Here we show that, like long term evolutionary patterns, within species variation depends on which functional class of the immune system is examined. Within species, immune genes involved in recognition and signaling have significantly lower levels of nucleotide variation than effector genes whose products clear infections. Recognition and signaling genes have previously shown evidence for being positively selected along the *D. melanogaster* species lineage (Sackton et al., 2007). Effector genes have higher levels of variation than signaling and recognition genes, and average rates of adaptive evolution between species. These results suggest that the short and long term evolutionary patterns of these gene classes are linked, with the rapidly evolving gene classes having lower levels of within-species variation.

Rapid evolution between species must arise as variation within species. In

principle, many different types of selection within populations, including hard and soft selective sweeps, partial selective sweeps, and balancing selection, can all lead to adaptive evolution between species. Hard selective sweeps are expected to cause the most dramatic decreases in nucleotide variation, and we find that adaptively evolving genes in our study have significantly lower levels of variation than other genes. This suggests that adaptive evolution is driven by novel variants arising and rapidly fixing in the population. Thus it appears that the most rapidly evolving genes have the least variation within a population and that the most genetically and potentially phenotypically diverse genes have little evidence for adaptive evolution at the species level.

Several genes with the lowest levels of variation (Table 5.2, 5.3) have previously been hypothesized to be evolving adaptively due to pathogen pressures. Pathogens such as *Salmonella typhimurium* and *Pseudomonas aeriginosa* actively suppress the imd pathway in an attempt to dampen the transcription of effector molecules (Lindmark et al., 2001; Apidianakis et al., 2005). The Relish complex is a group of interacting molecules immediately upstream of transcription in the imd pathway, and several genes in this complex, including *Relish* and *Dredd*, show evidence of rapid evolution in insects (Begun and Whitley, 2000; Bulmer and Crozier, 2006; Sackton et al., 2007; Obbard et al., 2009). *Dcr-2* is a known target of *Drosophila* picornavirus C when attempting to subvert the antiviral response (van Rij et al., 2006) and is amongst the most rapidly evolving genes in the *D. melanogaster* genome (Obbard et al., 2006). *ref(2)p* also shows evidence of adaptive evolution (Wayne et al., 1996) and is the only gene in *D. melanogaster* with alleles known to confer complete resistance or susceptibility to a virus infection (Contamine et al., 1989; Bangham et al., 2007, 2008). This suggests that adaptive evolution in some immune genes occurs when these genes are targeted

by pathogens attempting to escape the immune response.

Some genes which have not previously shown evidence of adaptive evolution in *D. melanogaster* also have low levels of nucleotide variation. This does not necessarily mean that these genes are not subject to directional selection within species. All the measures of adaptive evolution used here rely on frequent strong selection on coding regions. A single mutation with a strong phenotypic effect or changes to the regulation of a gene would not be detected using those methods. In principle, low nucleotide variation can also be caused by purifying selection or by neutral processes such as low recombination or mutation rates, and further research is required to distinguish amongst these processes. Regardless, the list of low diversity genes presents a candidate list of genes under selection within *D. melanogaster*. Several genes are directly upstream of transcription of major immune pathways, including *Stat92E* in the JAK-STAT pathway, *cactin* and *cactus* in the Toll pathway, *akirin* in the imd pathway, and *Atf-2* in the DUOX pathway. If genes that are immediately upstream of transcription are commonly targeted for immune suppression by pathogens as seen in the Relish complex, then these genes could similarly be targeted. Five recognition genes, *CG31217*, *emp*, *crq*, *TepI*, and *TepIV*, are involved in phagocytosis and are all evolving under positive selection in *Drosophila*. The clustering of certain types of genes within the low variation category strongly suggests a role for selection.

Nimblegen sequence capture combined with Illumina high-throughput sequencing provides a novel, rapid and efficient method of generating population genetic data. Our pooling approach allows a large population of alleles to be sequenced with a single capture array and one paired-end lane of sequenc-

ing. We are able to robustly generate the site frequency spectrum distribution of the number of mutations and calculate nucleotide polymorphism and diversity. Our allele frequency estimates were especially accurate when alleles were at intermediate frequency, suggesting that studies that rely on frequency estimates would benefit from this technique. The performance of the metrics of variation we used was highest when we were able to sequence over a kilobase in each gene region, demonstrating that optimal design of this strategy is to ensure a high depth of coverage. Our study suggests that genes under recurrent directional selection can be detected using species level comparisons when this leads to elevated adaptive evolution between species. Population level analysis can potentially also detect selection that does not elevate the rates of amino acid substitution. Thus, a combined approach that examines both the population and interspecific levels of selection is required to fully understand how the immune system evolves in response to pathogen pressures.

CHAPTER 6

RESEARCH SUMMARY

In my thesis, I have explored the short term evolution of the immune response of *Drosophila melanogaster* using several different approaches. Chapters 2 and 3 describe the natural bacteria that infect *D. melanogaster* in the wild with the goal of better understanding the selective pressures on the host immune system. Chapter 4 presents an in-depth study of the population genetics of the gene *eater*, which is part of an adaptively evolving class of immune response genes. In Chapter 5, over a hundred genes are surveyed at the population level and a connection is drawn between short and long term patterns of evolution. By studying the pathogens that infect *D. melanogaster*, selection on a single immune response gene, and adaptation across the entire immune system, I have been able to gain novel insight into how a host population adapts at the genetic level.

Genes involved in the immune response are rapidly evolving in many organisms (Murphy, 1993; Nielsen et al., 2005; Waterhouse et al., 2007), including *D. melanogaster* (Schlenke and Begun, 2003; Sackton et al., 2007). One hypothesis for this is that co-evolutionary arms races between hosts and pathogens cause reciprocal adaptation in host immune response genes and pathogen virulence genes, leading to an acceleration in the rates of adaptive evolution. I surveyed the hemolymph of wild caught *D. melanogaster* in an attempt to identify co-evolving bacterial pathogens but found no evidence for any. Instead, natural bacterial infections appear to largely be caused by a broad diversity of opportunistic pathogens, which are only pathogenic when they breach the barriers into the hemolymph and are not overcome by the host defenses. This per-

haps suggests that rapid evolution in the immune response in *D. melanogaster* is driven by non-bacterial pathogens such as viruses and parasitoids and that bacterial infections should select for a broad immune response.

Pathogen recognition molecules within the immune response are subject to particularly elevated rates of adaptive evolution in *Drosophila*. These molecules represent a specific point of interaction between the host immune response and the pathogen, and some genes are especially rapidly evolving at pathogen interaction domains. I examined the molecular population genetics of one particular recognition gene, *eater*, which has previously been shown to promote phagocytosis of both fungi and bacteria. I found evidence for a partial selective sweep at this locus in a North American population, with the putatively selected haplotype being associated with a significantly higher level of gene expression. Thus, at this locus, adaptation appears to be happening at the gene regulatory level rather than via adaptive substitutions at regions that interact with pathogens. Importantly, this type of regulatory change generally isn't detected using interspecies comparisons and requires population level analysis for detection.

A goal in evolutionary biology is understanding the genetic basis of adaptive evolution. For example, it is still not well understood what number and types of mutations are subject to selection and how these mutations spread within and between populations of a species. A partial selective sweep at a pigmentation locus in *D. melanogaster* has recently been associated with multiple cis regulatory changes (Pool and Aquadro, 2007; Rebeiz et al., 2009), suggesting that partial sweeps on regulatory variation may be more common than generally thought. One difficulty in this type of analysis is that regulatory variants are not as straightforward to identify as coding variants. At the *eater* locus, I

hypothesized that a particular region of the second intron was associated with a phenotypic difference in expression levels because of the unusual patterns of variation in this region, and I found support for this hypothesis. In this case, population genetics rather than reverse genetics was used to identify a putative enhancer region. An upstream enhancer region for *eater* has already been identified (Paul Kroeger, personal communication), and it is my hope that the second intron will also be demonstrated to have an enhancer region. Examination of variation in these enhancer regions in natural populations will lead insight into the nature of selection on regulatory regions. Ye et al. (2009) has recently found that expression of *eater* was elevated in Australia-derived *D. melanogaster* lines artificially selected for increased resistance to bacterial infection, and this elevation appears to be unrelated to the haplotype that I observed in a North American population (Yixin Henry Ye, personal communication). This suggests that multiple mutations of independent origin lead can lead to convergent phenotypes and presents a previously underappreciated mechanism of adaptation in the immune response.

Long term divergence between species arises as variation within populations that ultimately become fixed differences between species. How these variants arise, rise in frequency, and spread within and between populations is largely unknown. At the *eater* locus, the variant that rose in frequency in North America was also found in the founding African population and thus was not a new mutation. When selection acts on an existing mutation that is present at intermediate frequency in a population, this leads to a soft selection sweep because the selected variant has had time to recombine onto multiple genetic backgrounds and the signatures of selection are reduced. A hard selective sweep occurs when selection acts on a novel mutation and leaves more pro-

nounced signatures in the genome, including greatly reduced genetic variation and elevated linkage disequilibrium. A complete sweep leads to the fixation of the novel variant in the population, whereas in a partial sweep, like the one seen at *eater*, the variant only rises in frequency but never replaces all other variants. Recently, there has been debate in the literature about the relative roles of hard versus soft and complete versus partial selective sweeps in evolution within populations (Pritchard et al., 2010; Burke et al., 2010), with hard selective sweeps increasingly being thought of as uncommon. By studying individual examples in conjuction with genomic scale datasets, we can gain insight into the relative contributions of each type of process.

I examined the population genetics of over 100 genes involved in recognition, signal transduction and pathogen clearance within the immune system in order to relate short and long term patterns of evolution on a broader scale. I find that genes with evidence for adaptive evolution at the species level have significantly low levels of variation within-species. My data are consistent with rapidly evolving genes being subject to hard selective sweeps, with new alleles rapidly rise to high frequency and reducing variation temporarily in the population. Recognition and signal transduction genes are known to evolve rapidly, whereas pathogen clearance effector genes have rates of adaptive evolution on par with the genomewide average. The patterns of variation within populations in these gene classes reflect the rate of adaptive evolution, with rapidly evolving classes having low levels of variation. This further supports existing evidence that selection acts differently depending on the function of the gene.

High-throughput sequencing is a useful way to quickly and rapidly identify regions potentially under selection that can be subsequently be subject to more

detailed analysis. My large-scale analysis of immune response genes identified a number of genes that had low levels of variation in my analysis but that had not previously shown evidence for adaptive evolution. Adaptive evolution is typically measured only in coding regions and requires repeated amino acid substitutions to be significant, and therefore that analysis wouldn't detect selection in regulatory regions or single amino acid substitutions with strong effects. Thus, it is possible that the regions detected in my study represent genes under selection that cannot be detected by species level comparisons. Further characterization of these regions by Sanger sequencing would allow construction of haplotypes and calculation of additional population genetic statistics such as linkage disequilibrium to provide further evidence for directional selection. Detailed analysis of these genes could also address questions such as how often selection acts on coding regions versus regulatory regions and offer insight into what phenotypes are being selected.

Although much can be gained by studying a single population in detail, a more thorough understanding of evolution within species requires consideration of several populations. At the *eater* locus, I found evidence of a sweep that was localized to a single North American population, suggesting that local adaptation to a particular environment played an important role in shaping variation at this gene. In North America, where *D. melanogaster* has only been present for a few centuries, flies are still adapting to a new environment and to novel pathogen pressures. There has been some skepticism that selection could even be detected young, founded populations because noise from demographic processes can mask signals of selection (Thornton and Andolfatto, 2006). However, I was able to detect a relationship between rapid evolution and reduced genetic variation, which I believe is because this newly founded population is

subject to strong selection at a large number of loci. The target enrichment and high-throughput sequencing method developed here can be used to efficiently sequence our target genes in a number of different populations. Doing so will allow us to address questions that cannot be answered in a single population. For example, studies in multiple populations can distinguish between selection acting on standing variation that can be found in the ancestral African population and selection acting on new variants. It can also tell us how often selection occurs on the same genes and alleles in different populations. Answering these questions will further bridge the gap between studies of short and long term evolution.

My thesis addressed the question of what role selection within *D. melanogaster* plays in explaining the long term evolutionary patterns that have previously been observed. It represents the largest systematic study of the population genetics of immune response genes in *D. melanogaster* and clarifies the view that has been gained from studying individual genes in the past. Future studies of the genes and alleles identified in my work, perhaps using the methods established here, will further advance our understanding of adaptation of the immune response within populations.

APPENDIX A

SUPPLEMENTAL MATERIAL FOR CHAPTER 5: "SHORT AND LONG TERM PATTERNS OF EVOLUTION WITHIN IMMUNE RESPONSE GENES OF *DROSOPHILA MELANOGASTER*"

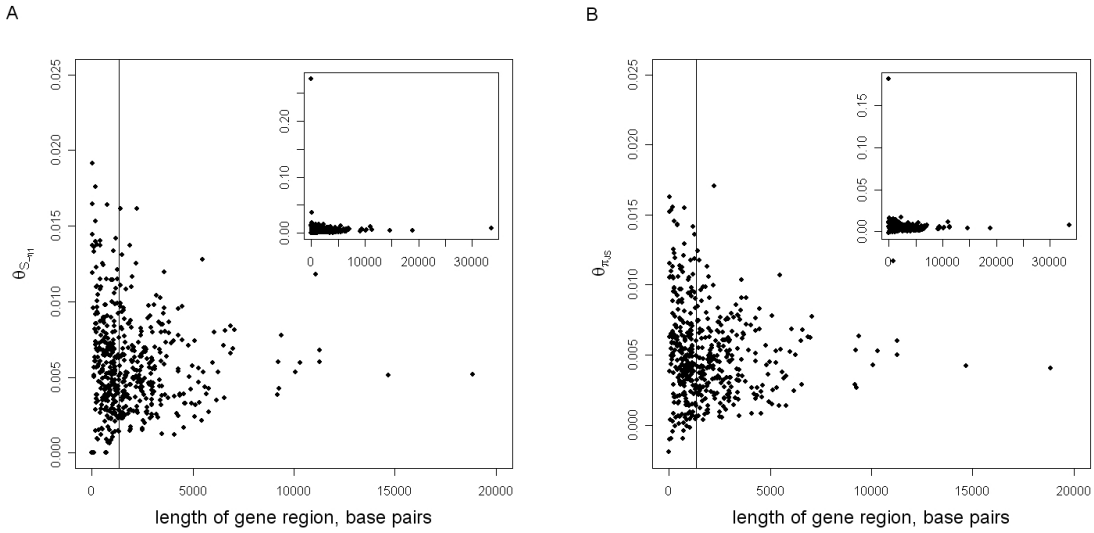


Figure A.1: Variance in nucleotide variation decreases as gene region length increases. The insets display the entire dataset and the main panels are a magnification. The lines indicate the median gene region length of 1,357 base pairs. A) $\theta_{S_{-\eta_1}}$ and B) $\theta_{\pi_{JS}}$.

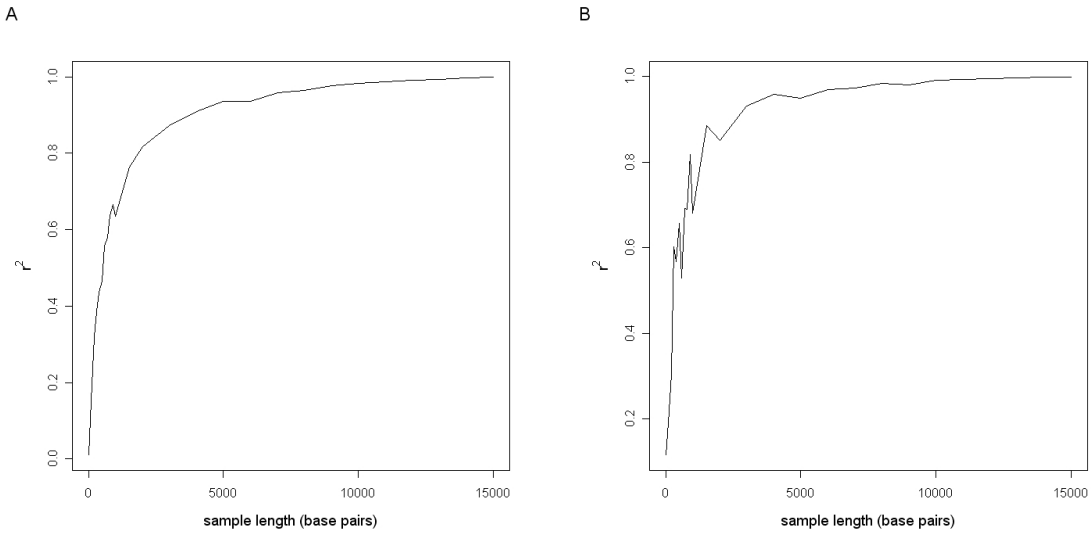


Figure A.2: The correlation (r^2) between the true sample variation and the measured variation (θ_S and θ_π) increases as the sample length increases. A) θ_S and B) θ_π

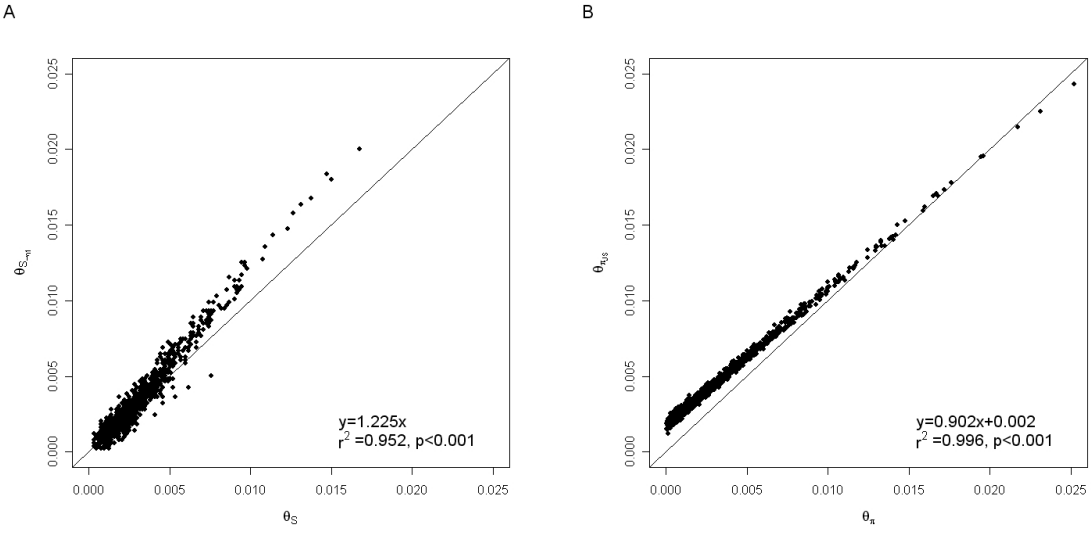


Figure A.3: Traditional measures of nucleotide variation (θ_S and θ_π) are highly correlated with corrected statistics ($\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$) measured after sequencing error is incorporated. A) $\theta_{S_{-\eta_1}}$ vs. θ_S and B) $\theta_{\pi_{JS}}$ vs. θ_π.

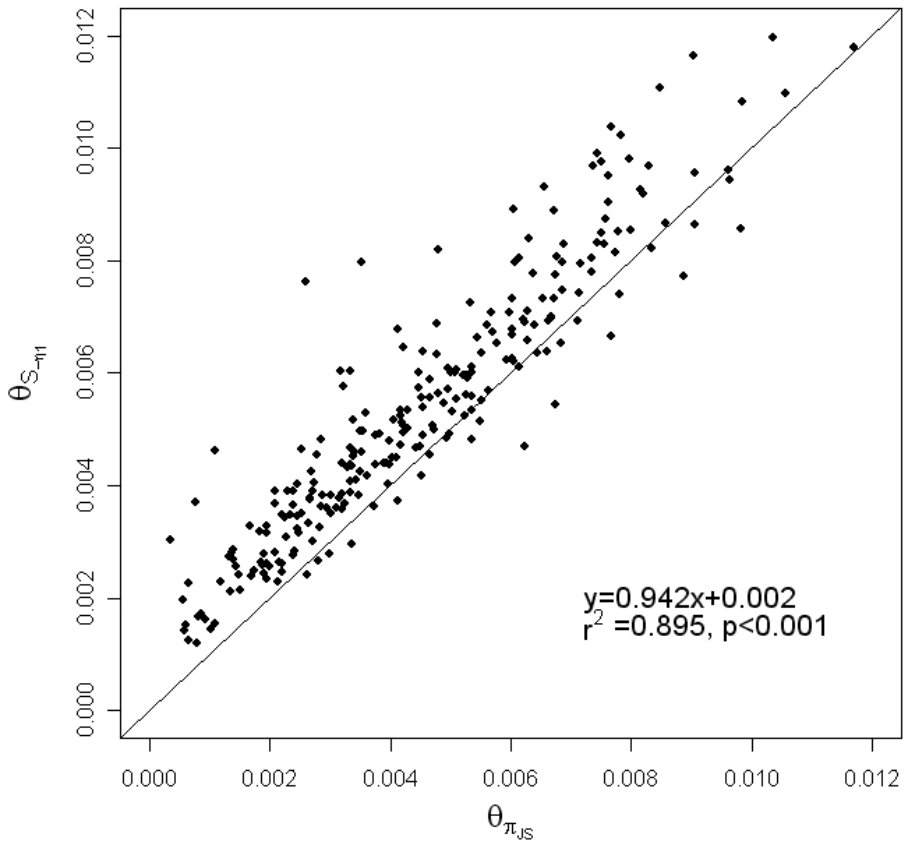


Figure A.4: Nucleotide variation measures $\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$ are highly positively correlated and $\theta_{S_{-\eta_1}}$ tends to be higher on average. Line indicates where $\theta_{S_{-\eta_1}} = \theta_{\pi_{JS}}$.

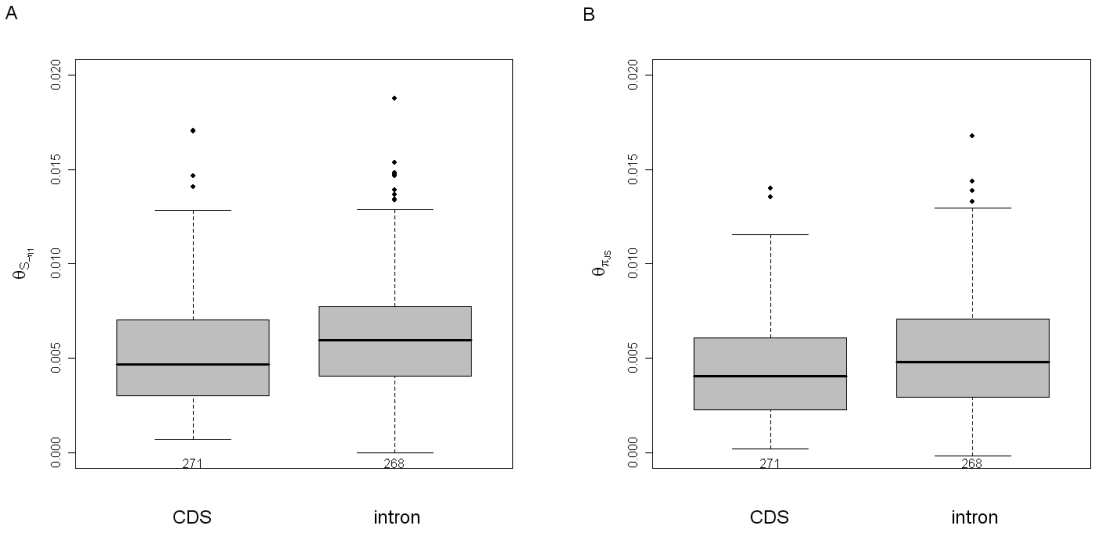


Figure A.5: Nucleotide variation ($\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$) is higher in introns versus in coding regions (CDS). A) $\theta_{S_{-\eta_1}}$ is significant at p<0.001 and B) $\theta_{\pi_{JS}}$ is significant at p<0.001 by Wilcoxon rank sum tests.

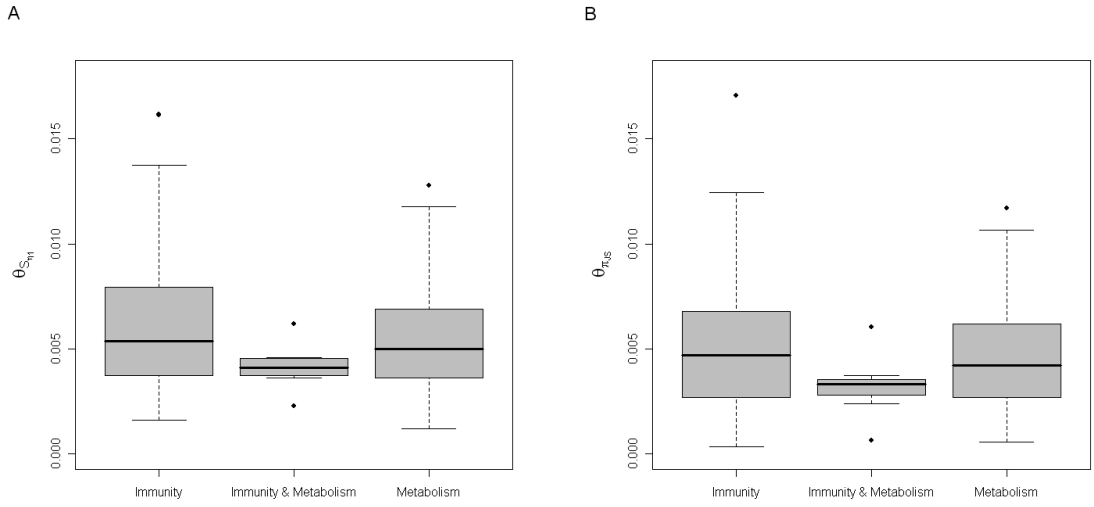


Figure A.6: Nucleotide variation ($\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$) does not vary between immunity and metabolism genes. A) $\theta_{S_{-\eta_1}}$ and B) $\theta_{\pi_{JS}}$.

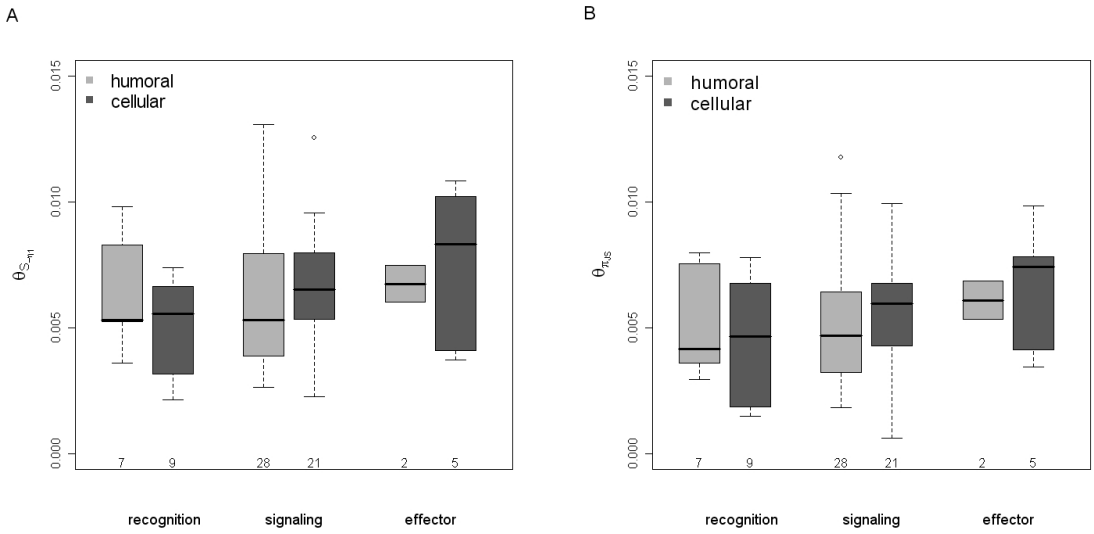


Figure A.7: Nucleotide variation ($\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$) does not vary between genes involved in the humoral and cellular branches of the immune response. The number of genes included in each boxplot is shown on the x-axis. Pairwise Wilcoxon rank sum tests were not significant for comparisons of humoral and cellular genes within the recognition, signaling and effector categories. A) $\theta_{S_{-\eta_1}}$ and B) $\theta_{\pi_{JS}}$.

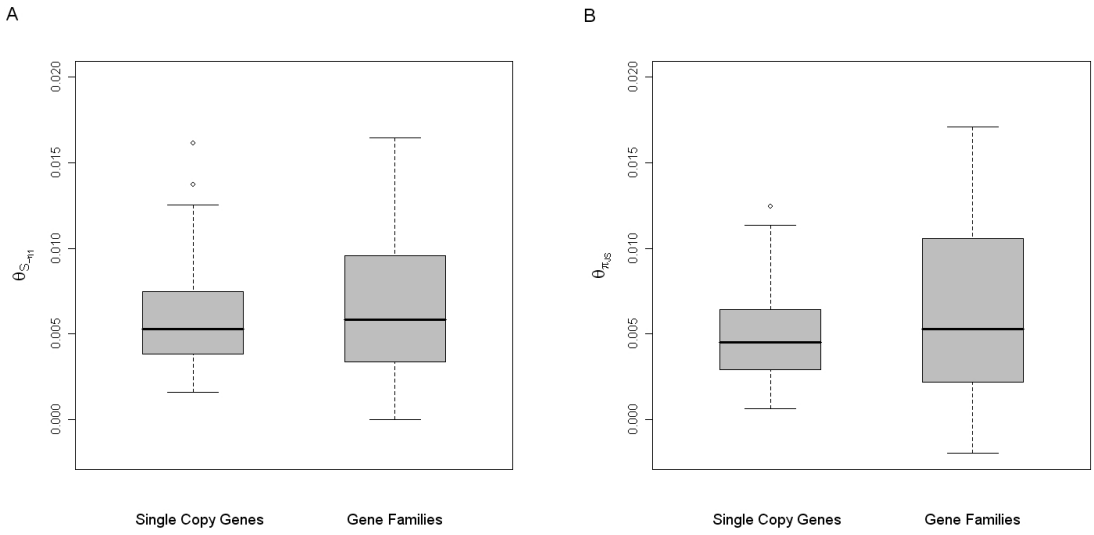


Figure A.8: Gene count does not affect nucleotide variation ($\theta_{S_{-\eta_1}}$ and $\theta_{\pi_{JS}}$). A total of 107 single copy genes and 36 genes belonging to gene families were included in this analysis. A Wilcoxon rank sum test did not indicate a significant difference in A) $\theta_{S_{-\eta_1}}$ and B) $\theta_{\pi_{JS}}$ for single copy genes versus those that are part of gene families.

BIBLIOGRAPHY

Achaz, G., 2008. Testing for neutrality in samples with sequencing errors. *Genetics*, **1424**(July):1409–1424.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.*, 2000. The genome sequence of *Drosophila melanogaster*. *Science*, **287**(5461):2185–2195.

Aminetzach, Y. T., Macpherson, J. M., and Petrov, D. A., 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. *Science*, **309**(5735):764–767.

Andolfatto, P., 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*, **437**(7062):1149–52.

Anisimova, M. and Liberles, D. A., 2007. The quest for natural selection in the age of comparative genomics. *Heredity*, **99**(6):567–79.

Apidianakis, Y., Mindrinos, M. N., Xiao, W., Lau, G. W., Baldini, R. L., Davis, R. W., and Rahme, L. G., 2005. Profiling early infection responses: Pseudomonas aeruginosa eludes host defenses by suppressing antimicrobial peptide gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(7):2573–2578.

Avila, A., Silverman, N., Diaz-meco, M. T., and Moscat, J., 2002. The Drosophila atypical protein kinase C-ref(2)P complex constitutes a conserved module for signaling in the toll pathway. *Molecular and Cellular Biology*, **22**(24):8787–8795.

Bakula, M., 1969. The persistance of a microbial flora during postembryogenesis of *Drosophila melanogaster*. *Journal of Invertebrate Pathology*, **14**:365–374.

Bangham, J., Kim, K.-W., Webster, C. L., and Jiggins, F. M., 2008. Genetic variation affecting host-parasite interactions: different genes affect different aspects of Sigma virus replication and transmission in *Drosophila melanogaster*. *Genetics*, **178**(4):2191–2199.

Bangham, J., Obbard, D. J., Kim, K.-W., Haddrill, P. R., and Jiggins, F. M., 2007. The age and evolution of an antiviral resistance mutation in *Drosophila melanogaster*. *Proceedings of the Royal Society B: Biological Sciences*, **274**(1621):2027–2034.

Baudry, E., Viginier, B., and Veuille, M., 2004. Non-African populations of *Drosophila melanogaster* have a unique origin. *Molecular Biology and Evolutioni*, **21**(8):1482–91.

Baxter, R. H. G., Chang, C.-I., Chelliah, Y., Blandin, S., Levashina, E. A., and Deisenhofer, J., 2007. Structural basis for conserved complement factor-like function in the antimalarial protein TEP1. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(28):11615–11620.

Begun, D. J. and Aquadro, C. F., 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, **356**:519–520.

Begun, D. J. and Aquadro, C. F., 1995. Molecular variation at the vermilion locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*. *Genetics*, **140**(3):1019–32.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., *et al.*, 2007. Popula-

tion genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol*, **5**(11):e310.

Begun, D. J. and Whitley, P., 2000. Adaptive evolution of Relish, a *Drosophila* NF-kappaB/IkappaB protein. *Genetics*, **154**(3):1231–1238.

Blandin, S. and Levashina, E. A., 2004. Thioester-containing proteins and insect immunity. *Molecular Immunology*, **40**(12):903–908.

Bloos, F., Hinder, F., Becker, K., Sachse, S., Mekontso Dessap, A., Straube, E., Cattoir, V., Brun-Buisson, C., Reinhart, K., Peters, G., *et al.*, 2010. A multicenter trial to compare blood culture with polymerase chain reaction in severe human sepsis. *Intensive Care Medicine*, **36**(2):241–247.

Bochdanovits, Z. and de Jong, G., 2003. Experimental evolution in *Drosophila melanogaster*: interaction of temperature and food quality selection regimes. *Evolution*, **57**(8):1829–36.

Boots, M. and Begon, M., 1993. Trade-offs with resistance to a granulosis virus in the Indian meal moth, examined by a laboratory evolution experiment. *Functional Ecology*, **7**(5):528–534.

Braun, A., Hoffmann, J. A., and Meister, M., 1998. Analysis of the *Drosophila* host defense in domino mutant larvae, which are devoid of hemocytes. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(November):14337–14342.

Broderick, N. A., Raffa, K. F., Goodman, R. M., and Handelsman, J., 2004. Census of the bacterial community of the gypsy moth larval midgut by using culturing and culture-independent methods. *Applied and Environmental Microbiology*, **70**(1):293–300.

Broekaert, W. F., Terras, F. R., Cammue, B. P., and Osborn, R. W., 1995. Plant defensins: novel antimicrobial peptides as components of the host defense system. *Plant Physiology*, **108**(4):1353–8.

Bulmer, M. S. and Crozier, R. H., 2004. Duplication and diversifying selection among termite antifungal peptides. *Molecular Biology and Evolution*, **21**(12):2256–2264.

Bulmer, M. S. and Crozier, R. H., 2006. Variation in positive selection in termite GNBPs and Relish. *Molecular Biology and Evolution*, **23**(2):317–326.

Burke, M. K., Dunham, J. P., Shahrestani, P., Thornton, K. R., Rose, M. R., and Long, A. D., 2010. Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature*, **467**(7315):587–590.

Carpenter, J. A., Obbard, D. J., Maside, X., and Jiggins, F. M., 2007. The recent spread of a vertically transmitted virus through populations of *Drosophila melanogaster*. *Molecular Ecology*, **16**(18):3947–54.

Carre-Mlouka, A., Gaumer, S., Gay, P., Petitjean, A. M., Coulondre, C., Dru, P., Bras, F., Dezelee, S., and Contamine, D., 2007. Control of sigma virus multiplication by the *ref(2)P* gene of *Drosophila melanogaster*: an *in vivo* study of the PB1 domain of ref(2)P. *Genetics*, **176**(1):409–419.

Cashion, P., Holder-Franklin, M., McCully, J., and Franklin, M., 1977. A rapid method for the base ratio determination of bacterial DNA. *Analytical Biochemistry*, **81**(2):461–466.

Charlesworth, B., Sniegowski, P., and Stephan, W., 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, **371**(6494):215–220.

Christophides, G. K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P. T., Collins, F. H., Danielli, A., Dimopoulos, G., *et al.*, 2002. Immunity related genes and gene families in *Anopheles gambiae*. *Science*, **298**(5591):159–165.

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., *et al.*, 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, **450**(7167):203–18.

Clark, A. G. and Wang, L., 1997. Molecular population genetics of *Drosophila* immune system genes. *Genetics*, **147**(2):713–724.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., *et al.*, 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, **37**:D141–5.

Collins, F. H., Sakai, R. K., Vernick, K. D., Paskewitz, S., Seeley, D. C., Miller, L. H., Collins, W. E., Campbell, C. C., and Gwadz, R. W., 1986. Genetic selection of a *Plasmodium*-refractory strain of the malaria vector *Anopheles gambiae*. *Science*, **234**(4776):607–610.

Contamine, D., Petitjean, A. M., and Ashburner, M., 1989. Genetic resistance to viral infection: the molecular cloning of a *Drosophila* gene that restricts infection by the rhabdovirus sigma. *Genetics*, **123**(3):525–33.

Corby-Harris, V., Pontaroli, A. C., Shimkets, L. J., Bennetzen, J. L., Habel, K. E., and Promislow, D. E. L., 2007. Geographical distribution and diversity of bacteria associated with natural populations of *Drosophila melanogaster*. *Applied and Environmental Microbiology*, **73**(11):3470–3479.

Cotter, S. C., Kruuk, L. E. B., and Wilson, K., 2004. Costs of resistance: genetic correlations and potential trade-offs in an insect immune system. *Journal of Evolutionary Biology*, **17**(2):421–429.

Cox, C. R. and Gilmore, M. S., 2007. Native microbial colonization of *Drosophila melanogaster* and its use as a model of *Enterococcus faecalis* pathogenesis. *Applied and Environmental Microbiology*, **75**(4):1565–1576.

Dassanayake, R. S., Silva Gunawardene, Y. I. N., and Tobe, S. S., 2007. Evolutionary selective trends of insect/mosquito antimicrobial defensin peptides containing cysteine-stabilized [alpha]/[beta] motifs. *Peptides*, **28**(1):62–75.

Date, A., Satta, Y., Takahata, N., and Chigusa, S. I., 1998. Evolutionary history and mechanism of the *Drosophila cecropin* gene family. *Immunogenetics*, **47**(6):417–29.

David, J. R. and Capy, P., 1988. Genetic variation of *Drosophila melanogaster* natural populations. *Trends in Genetics*, **4**(4):106–111.

Dawkins, R. and Krebs, J. R., 1979. Arms races between and within species. *Proceedings of the Royal Society B: Biological Sciences*, **205**(1161):489–511.

De Ley, J., Cattoir, H., and Reynaerts, A., 1970. The quantitative measurement of DNA hybridization from renaturation rates. *European Journal of Biochemistry*, **12**(1):133–42.

Decaestecker, E., Gaba, S., Raeymaekers, J. A. M., Stoks, R., Van Kerckhoven, L., Ebert, D., and De Meester, L., 2007. Host-parasite 'Red Queen/' dynamics archived in pond sediment. *Nature*, **450**:870–873.

DePristo, M. A., Weinreich, D. M., and Hartl, D. L., 2005. Missense meanderings

in sequence space: a biophysical view of protein evolution. *Nature Reviews Genetics*, **6**(9):678–87.

Durbin, R. M., Altshuler, D. L., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., *et al.*, 2010. A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319):1061–1073.

Dybdahl, M. F. and Lively, C. M., 1998. Host-parasite coevolution: evidence for rare advantage and time-lagged selection in a natural population. *Evolution*, **52**(4):1057–1066.

Ebert, D., 2000. Experimental evidence for rapid parasite adaptation and its consequences for the evolution of virulence. In Poulin, R., Morand, S., and Skorping, A., editors, *Evolutionary biology of host-parasite relationships: theory meets reality*, pages 163–184. Elsevier, Amsterdam.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E., and Relman, D. A., 2005. Diversity of the human intestinal microbial flora. *Science*, **308**(5728):1635–1638.

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., *et al.*, 2006. Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Molecular Biology*, **15**(5):645–656.

Farmer 3rd, J. J., Davis, B. R., Hickman-Brenner, F. W., McWhorter, A., Huntley-Carter, G. P., Asbury, M. A., Riddle, C., Elias, C., Fanning, G. R., Steigerwalt, A. G., *et al.*, 1985. Biochemical identification of new species and biogroups of

Enterobacteriaceae isolated from clinical specimens. *Journal of Clinical Microbiology*, **21**(1):46–76.

Fay, J. C. and Wu, C. I., 2000. Hitchhiking under positive Darwinian selection. *Genetics*, **155**(3):1405–13.

Feldhaar, H. and Gross, R., 2009. Insects as hosts for mutualistic bacteria. *International Journal of Medical Microbiology*, **299**(1):1–8.

Fiston-Lavier, A.-S., Singh, N. D., Lipatov, M., and Petrov, D. A., 2010. *Drosophila melanogaster* recombination rate calculator. *Gene*, **463**(1-2):18–20.

Fiumera, A. C., Dumont, B. L., and Clark, A. G., 2007. Associations between sperm competition and natural variation in male reproductive genes on the third chromosome of *Drosophila melanogaster*. *Genetics*, **176**(2):1245–60.

Frey, J. C., Angert, E. R., and Pell, A. N., 2006. Assessment of biases associated with profiling simple, model communities using terminal-restriction fragment length polymorphism-based analyses. *Journal of Microbiological Methods*, **67**(1):9–19.

Fu, Y.-X., 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147**:915–925.

García-Zaragoza, E., Mas, J. A., Vivar, J., Arredondo, J. J., and Cervera, M., 2008. CF2 activity and enhancer integration are required for proper muscle gene expression in *Drosophila*. *Mechanisms of Development*, **125**(7):617–630.

Garver, L. S., Xi, Z., and Dimopoulos, G., 2008. Immunoglobulin superfamily members play an important role in the mosquito immune system. *Developmental & Comparative Immunology*, **32**(5):519–531.

Glinka, S., Ometto, L., Mousset, S., Stephan, W., and De Lorenzo, D., 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics*, **165**(3):1269–78.

Goodman, S. J., 1997. Rst Calc : a collection of computer programs for calculating estimates of genetic differentiation from microsatellite. *Molecular Ecology*, **6**:881–885.

Gwynn, D. M., Callaghan, A., Gorham, J., Walters, K. F. A., and Fellowes, M. D. E., 2005. Resistance is costly: trade-offs between immunity, fecundity and survival in the pea aphid. *Proceedings of the Royal Society B: Biological Sciences*, **272**(1574):1803–8.

Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C., and Cristianini, N., 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, **15**(8):1153–1160.

Handelsman, J., 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, **68**(4):669–685.

Hartl, D. L. and Clark, A. G., 2007. *Principles of Population Genetics*. Sinauer Associates, Sunderland, MA.

Hickman-Brenner, F. W., Farmer 3rd, J. J., Steigerwalt, A. G., and Brenner, D. J., 1983. *Providencia rustigianii*: a new species in the family *Enterobacteriaceae* formerly known as *Providencia* alcalifaciens biogroup 3. *Journal of Clinical Microbiology*, **17**(6):1057–1060.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., Middle, C. M., Rodesch, M. J., Albert, T. J., Hannon, G. J., *et al.*, 2007. Genome-wide in situ exon capture for selective resequencing. *Nature Genetics*, **39**(12):1522–7.

Hoffmann, A. A., Hercus, M., and Dagher, H., 1998. Population dynamics of the Wolbachia infection causing cytoplasmic incompatibility in *Drosophila melanogaster*. *Genetics*, **148**(1):221–31.

Hsu, T., Bagni, C., Sutherland, J. D., and Kafatos, F. C., 1996. The transcriptional factor CF2 is a mediator of EGF-R-activated dorsoventral patterning in *Drosophila* oogenesis. *Genes & Development*, **10**(11):1411–1421.

Hudson, R. R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**(2):337–338.

Hudson, R. R. and Kaplan, N. L., 1988. The coalescent process in models with selection and recombination. *Genetics*, **120**:831–840.

Hudson, R. R., Saez, A. G., and Ayala, F. J., 1997. DNA variation at the Sod locus of *Drosophila melanogaster*: an unfolding story of natural selection. *Proceedings of the National Academy of Sciences of the United States of America*, **94**(15):7725–7729.

Huss, V. A. R., Festl, H., and Schleifer, K. H., 1983. Studies on the spectrophotometric determination of DNA hybridization from renaturation rates. *Systematic and Applied Microbiology*, **4**:184–192.

Hutter, S., Saminadin-Peter, S., Stephan, W., and Parsch, J., 2008. Gene expression variation in African and European populations of *Drosophila melanogaster*. *Genome Biology*, **9**(1):R12.

Jiggins, F. and Kim, K.-W., 2006. Contrasting evolutionary patterns in *Drosophila* immune receptors. *Journal of Molecular Evolution*, **63**(6):769–780.

Jiggins, F. M. and Hurst, G. D. D., 2003. The evolution of parasite recognition genes in the innate immune system: purifying selection on *Drosophila*

melanogaster peptidoglycan recognition proteins. *Journal of Molecular Evolution*, **V57**(5):598–605.

Jiggins, F. M. and Kim, K.-W., 2005. The evolution of antifungal peptides in *Drosophila*. *Genetics*, **171**(4):1847–1859.

Jiggins, F. M. and Kim, K. W., 2007. A screen for immunity genes evolving under positive selection in *Drosophila*. *Journal of Evolutionary Biology*, **20**(3):965–970.

Johnson, M. T. J., Lajeunesse, M. J., and Agrawal, A. A., 2006. Additive and interactive effects of plant genotypic diversity on arthropod communities and plant fitness.

Johnson, P. L. F. and Slatkin, M., 2008. Accounting for bias from sequencing error in population genetic estimates. *Molecular Biology and Evolution*, **25**(1):199–206.

Juneja, P. and Lazzaro, B. P., 2010. Haplotype structure and expression divergence at the *Drosophila* cellular immune gene *eater*. *Molecular Biology and Evolution*, **27**(10):2284–99.

Kafatos, F. C., Waterhouse, R. M., Zdobnov, E. M., and Christophides, G. K., 2009. Comparative genomics of insect immunity. In Rolff, J. and Reynolds, S., editors, *Insect infection and immunity: evolution, ecology, and mechanisms*, chapter 6, pages 86–105. Oxford University Press, New York.

Kelly, J. K., 1997. A test of neutrality based on interlocus associations. *Genetics*, **146**(3):1197–1206.

Kocks, C., Cho, J. H., Nehme, N., Ulvila, J., Pearson, A. M., Meister, M., Strom, C., Conto, S. L., Hetru, C., Stuart, L. M., *et al.*, 2005. Eater, a transmembrane

protein mediating phagocytosis of bacterial pathogens in *Drosophila*. *Cell*, **123**(2):335–346.

Kolaczkowski, B., Kern, A. D., Holloway, A. K., and Begun, D. J., 2010. Genomic differentiation between temperate and tropical Australian populations of *Drosophila melanogaster*. *Genetics*, **epub ahead**:1–59.

Kraaijeveld, A. R. and Godfray, H. C., 1997. Trade-off between parasitoid resistance and larval competitive ability in *Drosophila melanogaster*. *Nature*, **389**(6648):278–80.

Kraaijeveld, A. R. and Godfray, H. C. J., 1999. Geographic patterns in the evolution of resistance and virulence in *Drosophila* and its parasitoids. *American Naturalist*, **153**:S61–S74.

Kraaijeveld, A. R. and Godfray, H. C. J., 2008. Selection for resistance to a fungal pathogen in *Drosophila melanogaster*. *Heredity*, **100**:400–406.

Kumar, S., Christophides, G. K., Cantera, R., Charles, B., Han, Y. S., Meister, S., Dimopoulos, G., Kafatos, F. C., and Barillas-Mury, C., 2003. The role of reactive oxygen species on *Plasmodium* melanotic encapsulation in *Anopheles gambiae*. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(24):14139–44.

Kumar, S., Tamura, K., and Nei, M., 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Briefings in Bioinformatics*, **5**(2):150–63.

Kurucz, E., Markus, R., Zsamboki, J., Folkl-Medzihradszky, K., Darula, Z., Vilmos, P., Udvardy, A., Krausz, I., Lukacsovich, T., Gateff, E., *et al.*, 2007. Nim-

rod, a putative phagocytosis receptor with EGF repeats in *Drosophila* plasmatocytes. *Current Biology*, **17**(7):649–654.

Larracuente, A. M., Sackton, T. B., Greenberg, A. J., Wong, A., Singh, N. D., Sturgill, D., Zhang, Y., Oliver, B., and Clark, A. G., 2008. Evolution of protein-coding genes in *Drosophila*. *Trends in Genetics*, **24**(3):114–123.

Lazzaro, B. P., 2005. Elevated polymorphism and divergence in the class C scavenger receptors of *Drosophila melanogaster* and *D. simulans*. *Genetics*, **169**(4):2023–2034.

Lazzaro, B. P., 2008. Natural selection on the *Drosophila* antimicrobial immune system. *Current Biology*, **11**(3):284–289.

Lazzaro, B. P. and Clark, A. G., 2001. Evidence for recurrent paralogous gene conversion and exceptional allelic divergence in the *Attacin* genes of *Drosophila melanogaster*. *Genetics*, **159**(2):659–671.

Lazzaro, B. P. and Clark, A. G., 2003. Molecular population genetics of inducible antibacterial peptide genes in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **20**(6):914–923.

Lazzaro, B. P., Flores, H. A., Lorigan, J. G., and Yourth, C. P., 2008. Genotype-by-environment interactions and adaptation to local temperature affect immunity and fecundity in *Drosophila melanogaster*. *PLoS Pathogens*, **4**(3):e1000025.

Lazzaro, B. P., Sackton, T. B., and Clark, A. G., 2006. Genetic variation in *Drosophila melanogaster* resistance to infection: a comparison across bacteria. *Genetics*, **174**(3):1539–54.

Lazzaro, B. P., Sceurman, B. K., and Clark, A. G., 2004. Genetic basis of natural

variation in *D. melanogaster* antibacterial immunity. *Science*, **303**(5665):1873–1876.

Lee, Y. M., Misra, H. P., and Ayala, F. J., 1981. Superoxide dismutase in *Drosophila melanogaster*: biochemical and structural characterization of allozyme variants. *Proceedings of the National Academy of Sciences of the United States of America*, **78**(11):7052–5.

Lehmann, T., Hume, J. C. C., Licht, M., Burns, C. S., Wollenberg, K., Simard, F., and Ribeiro, J. M. C., 2009. Molecular evolution of immune genes in the malaria mosquito *Anopheles gambiae*. *PloS ONE*, **4**(2):e4549.

Lemaitre, B. and Hoffmann, J., 2007. The host defense of *Drosophila melanogaster*. *Annual Review of Immunology*, **25**:697–743.

Levashina, E. A., Moita, L. F., Blandin, S., Vriend, G., Lagueux, M., and Kafatos, F. C., 2001. Conserved role of a complement-like protein in phagocytosis revealed by dsRNA knockout in cultured cells of the mosquito, *Anopheles gambiae*. *Cell*, **104**(5):709–18.

Levine, M. T. and Begun, D. J., 2007. Comparative population genetics of the immunity gene, Relish: is adaptive evolution idiosyncratic? *PLoS ONE*, **2**(5):e442.

Lewontin, R. C. and Hubby, J. L., 1966. T o the. *Genetics*, **54**(August):595–609.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16):2078–9.

Li, H. and Stephan, W., 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genetics*, **2**(10):e166.

Li, W. H., 1997. *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Librado, P. and Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**(11):1451–2.

Lindh, J. M., Terenius, O., and Faye, I., 2005. 16S rRNA gene-based identification of midgut bacteria from field-caught *Anopheles gambiae* sensu lato and *A. funestus* mosquitoes reveals new species related to known insect symbionts. *Applied and Environmental Microbiology*, **71**(11):7217–7223.

Lindmark, H., Johansson, K. C., Stoven, S., Hultmark, D., Engstrom, Y., and Soderhall, K., 2001. Enteric bacteria counteract lipopolysaccharide induction of antimicrobial peptide genes. *The Journal of Immunology*, **167**(12):6920–6923.

Little, T. J. and Cobbe, N., 2005. The evolution of immune-related genes from disease carrying mosquitoes: diversity in a peptidoglycan- and a thioester-recognizing protein. *Insect Molecular Biology*, **14**(6):599–605.

Little, T. J., Colbourne, J. K., and Crease, T. J., 2004. Molecular evolution of *Daphnia* immunity genes: polymorphism in a *Gram-Negative Binding Protein* gene and an *α-2-Macroglobulin* gene. *Journal of Molecular Evolution*, **59**(4):498–506.

Little, T. J., Watt, K., and Ebert, D., 2006. Parasite-host specificity: experimental studies on the basis of parasite adaptation. *Evolution*, **60**(1):31–38.

Luong, L. T. and Polak, M., 2007. Costs of resistance in the *Drosophila-Macrocheles* system: a negative genetic correlation between ectoparasite resistance and reproduction. *Evolution*, **61**(6):1391–1402.

Macpherson, J. M., Gonzalez, J., Witten, D. M., Davis, J. C., Rosenberg, N. A., Hirsh, A. E., and Petrov, D. A., 2008. Nonadaptive explanations for signa-

tures of partial selective sweeps in *Drosophila*. *Molecular Biology and Evolution*, **25**(6):1025–1042.

McDonald, J. H. and Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, **351**(6328):652–4.

McEvoy, B. P., Montgomery, G. W., McRae, A. F., Ripatti, S., Perola, M., Spector, T. D., Cherkas, L., Ahmadi, K. R., Boomsma, D., Willemsen, G., *et al.*, 2009. Geographical structure and differential natural selection among North European populations. *Genome Research*, **19**(5):804–14.

McKean, K. A. and Nunney, L., 2008. Sexual selection and immune function in *Drosophila melanogaster*. *Evolution*, **62**(2):386–400.

Mendes, C., Felix, R., Sousa, A.-M., Lamego, J., Charlwood, D., do Rosário, V. E., Pinto, J., and Silveira, H., 2010. Molecular evolution of the three short PGRPs of the malaria vectors *Anopheles gambiae* and *Anopheles arabiensis* in East Africa. *BMC Evolutionary Biology*, **10**:9.

Mrinal, N. and Nagaraju, J., 2008. Intron loss is associated with gain of function in the evolution of the gloverin family of antibacterial genes in *Bombyx mori*. *Journal of Biological Chemistry*, **283**(34):23376–23387.

Mukherjee, S., Sarkar-Roy, N., Wagener, D. K., and Majumder, P. P., 2009. Signatures of natural selection are not uniform across genes of innate immune system, but purifying selection is the dominant signature. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(17):7073–8.

Muller, H. E., O'Hara, C. M., Fanning, G. R., Hickman-Brenner, F. W., Swenson, J. M., and Brenner, D. J., 1986. *Providencia heimbachae*, a new species of

Enterobacteriaceae isolated from animals. *International Journal of Systematic Bacteriology*, **36**(2):252–256.

Murphy, P. M., 1993. Molecular mimicry and the generation of host defense protein diversity. *Cell*, **72**(6):823–826.

Nei, M. and Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, **3**(5):418–426.

Niare, O., Markianos, K., Volz, J., Oduol, F., Toure, A., Bagayoko, M., Sangare, D., Traore, S. F., Wang, R., Blass, C., *et al.*, 2002. Genetic loci affecting resistance to human malaria parasites in a West African mosquito vector population. *Science*, **298**(5591):213–216.

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., *et al.*, 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, **3**(6):e170.

Nikkari, S., Lopez, F. A., Lepp, P. W., Cieslak, P. R., Ladd-Wilson, S., Passaro, D., Danila, R., and Relman, D. A., 2002. Broad-range bacterial detection and the analysis of unexplained death and critical illness. *Emerging Infectious Diseases*, **8**(2):188–94.

Obbard, D. J., Callister, D. M., Jiggins, F. M., Soares, D., Yan, G., and Little, T. J., 2008. The evolution of *TEP1*, an exceptionally polymorphic immunity gene in *Anopheles gambiae*. *BMC Evolutionary Biology*, **8**:274.

Obbard, D. J., Jiggins, F. M., Bradshaw, N. J., and Little, T. J., 2010. Recent and

recurrent selective sweeps of the antiviral RNAi gene *Argonaute-2* in three species of *Drosophila*. *Molecular Biology and Evolution*, **epub ahead**.

Obbard, D. J., Jiggins, F. M., Halligan, D. L., and Little, T. J., 2006. Natural selection drives extremely rapid evolution in antiviral RNAi genes. *Current Biology*, **16**(6):580–585.

Obbard, D. J., Welch, J. J., Kim, K.-W., and Jiggins, F. M., 2009. Quantifying adaptive evolution in the *Drosophila* immune system. *PLoS Genetics*, **5**(10):e1000698.

Panaccio, M. and Lew, A., 1991. PCR based diagnosis in the presence of 8% (v/v) blood. *Nucleic Acids Research*, **19**(5):1151.

Paradis, E., 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**(2):289–290.

Penner, J. L. and Hennessy, J. N., 1979. Application of O-serotyping in a study of *Providencia rettgeri* (*Proteus rettgeri*) isolated from human and nonhuman sources. *Journal of Clinical Microbiology*, **10**(6):834–40.

Perron, G. G., Zasloff, M., and Bell, G., 2006. Experimental evolution of resistance to an antimicrobial peptide. *Proceedings of the Royal Society B: Biological Sciences*, **273**(1583):251–256.

Piertney, S. B. and Oliver, M. K., 2006. The evolutionary ecology of the major histocompatibility complex. *Heredity*, **96**(1):7–21.

Pool, J. E. and Aquadro, C. F., 2007. The genetic basis of adaptive pigmentation variation in *Drosophila melanogaster*. *Molecular ecology*, **16**(14):2844–51.

Pool, J. E., Bauer DuMont, V., Mueller, J. L., and Aquadro, C. F., 2006. A scan of molecular variation leads to the narrow localization of a selective

sweep affecting both Afrotropical and cosmopolitan populations of *Drosophila melanogaster*. *Genetics*, **172**(2):1093–105.

Pritchard, J. K., Pickrell, J. K., and Coop, G., 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**(4):R208–15.

Przeworski, M., Coop, G., and Wall, J. D., 2005. The signature of positive selection on standing genetic variation. *Evolution*, **59**(11):2312–23.

R Development Core Team, 2006. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramet, M., Pearson, A., Manfruelli, P., Li, X., Koziel, H., Go, V., Chung, E., Krieger, M., and Ezekowitz, R. A. B., 2001. *Drosophila* scavenger receptor CI is a pattern recognition receptor for bacteria. *Immunity*, **15**:1027–1038.

Ramos-Onsins, S. and Aguadé, M., 1998. Molecular evolution of the *Cecropin* multigene family in *Drosophila* functional genes vs. pseudogenes. *Genetics*, **150**(1):157–71.

Rebeiz, M., Pool, J. E., Kassner, V. A., Aquadro, C. F., and Carroll, S. B., 2009. Stepwise modification of a modular enhancer underlies adaptation in a *Drosophila* population. *Science*, **326**(5960):1663–7.

Reynolds, S. and Rolff, J., 2008. Immune function keeps endosymbionts under control. *Journal of Biology*, **7**(8):28.

Riehle, M. M., Markianos, K., Niare, O., Xu, J., Li, J., Toure, A. M., Podiougou, B., Oduol, F., Diawara, S., Diallo, M., *et al.*, 2006. Natural malaria infection in *Anopheles gambiae* is regulated by a single genomic control region. *Science*, **312**(5773):577–579.

Roff, D. A. and Fairbairn, D. J., 2007. The evolution of trade-offs: where are we? *Journal of Evolutionary Biology*, **20**(2):433–47.

Rolff, J., Armitage, S. A. O., and Coltman, D. W., 2005. Genetic constraints and sexual dimorphism in immune defense. *Evolution*, **59**(8):1844–1850.

Ryan, A. W., Mapp, J., Moyna, S., Mattiangeli, V., Kelleher, D., Bradley, D. G., and McManus, R., 2006. Levels of interpopulation differentiation among different functional classes of immunologically important genes. *Genes and Immunity*, **7**(2):179–83.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., *et al.*, 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**(6909):832–837.

Sackton, T. B., Lazzaro, B. P., and Clark, A. G., 2010. Genotype and gene expression associations with immune function in *Drosophila*. *PLoS Genetics*, **6**(1):e1000797.

Sackton, T. B., Lazzaro, B. P., Schlenke, T. A., Evans, J. D., Hultmark, D., and Clark, A. G., 2007. Dynamic evolution of the innate immune system in *Drosophila*. *Nature Genetics*, **39**(12):1461–1468.

Samakovlis, C., Kimbrell, D. A., Kylsten, P., Engstrom, A., and Hultmark, D., 1990. The immune response in *Drosophila*: pattern of cecropin expression and biological activity. *EMBO Journal*, **9**(9):2969 – 2976.

Santos, S. R. and Ochman, H., 2004. Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environmental Microbiology*, **6**:754–759.

Savard, J., Tautz, D., Richards, S., Weinstock, G. M., Gibbs, R. A., Werren, J. H., Tettelin, H. A., and Lercher, M. J., 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Research*, **16**(11):1334–1338.

Schlenke, T. A. and Begun, D. J., 2003. Natural selection drives *Drosophila* immune system evolution. *Genetics*, **164**(4):1471–1480.

Schlenke, T. A. and Begun, D. J., 2005. Linkage disequilibrium and recent selection at three immunity receptor loci in *Drosophila simulans*. *Genetics*, **169**(4):2013–2022.

Schloss, P. D. and Handelsman, J., 2006. Toward a census of bacteria in soil. *PLoS Computational Biology*, **2**(7):e92.

Schlötterer, C., Neumeier, H., Sousa, C., and Nolte, V., 2006. Highly structured Asian *Drosophila melanogaster* populations: a new tool for hitchhiking mapping? *Genetics*, **172**(1):287–92.

Schmid-Hempel, P., 2008. Parasite immune evasion: a momentous molecular war. *Trends in Ecology & Evolution*, **23**(6):318–26.

Schütte, U. M. E., Abdo, Z., Bent, S. J., Shyu, C., Williams, C. J., Pierson, J. D., and Forney, L. J., 2008. Advances in the use of terminal restriction fragment length polymorphism (T-RFLP) analysis of 16S rRNA genes to characterize microbial communities. *Applied Microbiology and Biotechnology*, **80**(3):365–80.

Shin, J.-H., Blay, S., McNeney, B., and Graham, J., 2006. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *Journal of Statistical Software*, **16**:Code Snippet 3.

Singh, N. D., Arndt, P. F., and Petrov, D. A., 2005. Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. *Genetics*, **169**(2):709–22.

Smith, G. P., 1976. Evolution of repeated DNA sequences by unequal crossing over. *Science*, **191**(4227):528–535.

Smith, J. M. and Haigh, J., 1974. The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**:23–35.

Somogyi, K., Sipos, B., Penzes, Z., Kurucz, E., Zsamboki, J., Hultmark, D., and Ando, I., 2008. Evolution of genes and repeats in the *Nimrod* superfamily. *Molecular Biology and Evolution*, **25**(11):2337–2347.

Somvanshi, V. S., Lang, E., Straubler, B., Sproer, C., Schumann, P., Ganguly, S., Saxena, A. K., and Stackebrandt, E., 2006. *Providencia vermicola* sp. nov., isolated from infective juveniles of the entomopathogenic nematode *Steinernema thermophilum*. *International Journal of Systematic and Evolutionary Microbiology*, **56**(3):629–633.

Stephan, W., 1989. Tandem-repetitive noncoding DNA: forms and forces. *Molecular Biology and Evolution*, **6**(2):198–212.

Stroschein-Stevenson, S. L., Foley, E., O'Farrell, P. H., and Johnson, A. D., 2006. Identification of *Drosophila* gene products required for phagocytosis of *Candida albicans*. *PLoS Biology*, **4**(1):e4.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**(3):585–595.

Tamura, K., Dudley, J., Nei, M., and Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**(8):1596–9.

Tamura, K., Subramanian, S., and Kumar, S., 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Molecular Biology and Evolution*, **21**(1):36–44.

Tennessen, J. A., 2005. Molecular evolution of animal antimicrobial peptides: widespread moderate positive selection. *Journal of Evolutionary Biology*, **18**(6):1387–1394.

Tennessen, J. a. and Blouin, M. S., 2008. Balancing selection at a frog antimicrobial peptide locus: fluctuating immune effector alleles? *Molecular Biology and Evolution*, **25**(12):2669–80.

Thornton, K. and Andolfatto, P., 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics*, **172**(3):1607–1619.

Tian, D., Araki, H., Stahl, E., Bergelson, J., and Kreitman, M., 2002. Signature of balancing selection in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(17):11525–11530.

Tinsley, M. C., Blanford, S., and Jiggins, F. M., 2006. Genetic variation in *Drosophila melanogaster* pathogen susceptibility. *Parasitology*, **132**(06):767–773.

Tokusumi, T., Sorrentino, R. P., Russell, M., Ferrarese, R., Govind, S., and Schulz, R. A., 2009. Characterization of a lamellocyte transcriptional enhancer located within the misshapen gene of *Drosophila melanogaster*. *PLoS ONE*, **4**(7):e6429.

Tsalik, E. L., Jones, D., Nicholson, B., Waring, L., Liesenfeld, O., Park, L. P., Glickman, S. W., Caram, L. B., Langley, R. J., Velkinburgh, J. C. V., *et al.*, 2010. Multiplex PCR to diagnose bloodstream infections in patients admitted from the emergency department with sepsis. *Journal of Clinical Microbiology*, **48**(1):26–33.

Tufts, D. M. and Bextine, B., 2009. Identification of bacterial species in the hemolymph of queen *Solenopsis invicta* (Hymenoptera: Formicidae). *Environmental Entomology*, **38**(5):1360–1364.

van Rij, R. P., Saleh, M.-C., Berry, B., Foo, C., Houk, A., Antoniewski, C., and Andino, R., 2006. The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in *Drosophila melanogaster*. *Genes and Development*, **20**(21):2985–2995.

Wang, L., Weber, A. N. R., Atilano, M. L., Filipe, S., Gay, N. J., and Ligoxygakis, P., 2006a. Sensing of Gram-positive bacteia in *Drosophila*: GNBP1 is needed to process and present peptidoglycan to PGRP-SA. *The EMBO Journal*, **25**:5005–5014.

Wang, X.-H., Aliyari, R., Li, W.-X., Li, H.-W., Kim, K., Carthew, R., Atkinson, P., and Ding, S.-W., 2006b. RNA interference directs innate immunity against viruses in adult *Drosophila*. *Science*, **312**(5772):452–454.

Waterhouse, R. M., Kriventseva, E. V., Meister, S., Xi, Z., Alvarez, K. S., Bartholomay, L. C., Barillas-Mury, C., Bian, G., Blandin, S., Christensen, B. M., *et al.*, 2007. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science*, **316**(5832):1738–1743.

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O.,

Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E., Stackebrandt, E., *et al.*, 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic Bacteriology*, **37**(4):463–464.

Wayne, M. L., Contamine, D., and Kreitman, M., 1996. Molecular population genetics of *ref(2)P*, a locus which confers viral resistance in *Drosophila*. *Molecular Biology and Evolution*, **13**(1):191–199.

Weisburg, W. G., Barns, S. M., Pelletier, D. A., and Lane, D. J., 1991. 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology*, **173**(2):697–703.

Wilfert, L., Gadau, J., Baer, B., and Schmid-Hempel, P., 2007a. Natural variation in the genetic architecture of a host-parasite interaction in the bumblebee *Bombus terrestris*. *Molecular Ecology*, **16**(6):1327–1339.

Wilfert, L., Gadau, J., and Schmid-Hempel, P., 2007b. The genetic architecture of immune defense and reproduction in male *Bombus terrestris* bumblebees. *Evolution*, **61**(4):804–815.

Woolhouse, M. E. J., Webster, J. P., Domingo, E., Charlesworth, B., and Levin, B. R., 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genetics*, **32**(4):569–577.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M. K., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**(1):431–449.

Ye, Y. H., Chenoweth, S. F., and Mcgraw, E. A., 2009. Effective but costly,

evolved mechanisms of defense against a virulent opportunistic pathogen in *Drosophila melanogaster*. *PLoS Pathogens*, **5**(4):e1000385.

Yeaman, M. R. and Yount, N. Y., 2007. Unifying themes in host defence effector polypeptides. *Nature Review Microbiology*, **5**(9):727–740.

Yoh, M., Matsuyama, J., Ohnishi, M., Takagi, K., Miyagi, H., Mori, K., Park, K.-S., Ono, T., and Honda, T., 2005. Importance of *Providencia* species as a major cause of travellers' diarrhoea. *Journal of Medical Microbiology*, **54**(11):1077–1082.

Zasloff, M., 2002. Antimicrobial peptides of multicellular organisms. *Nature*, **415**(6870):389–395.

Zhong, D., Pai, A., and Yan, G., 2005. Costly resistance to parasitism: evidence from simultaneous quantitative trait loci mapping for resistance and fitness in *Tribolium castaneum*. *Genetics*, **169**(4):2127–2135.

Zou, Z., Evans, J., 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 Biology*, **8**(8):R177.