SOME THINGS I HAVE WRITTEN ABOUT FLIES

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Susan M. Rottschaefer

January 2015

ABSTRACT

Host-pathogen interactions shape the evolution of host immune defense. Genetic variation in genes of the immune system influences resistance to infection, but host immune defenses can additionally be influenced by environmental and physiological factors, including the presence of symbiotic microbes. In this thesis I examine both non-genetic and genetic factors involved in host-pathogen interactions. In the first chapter I ask whether the antibacterial immune response of *D. melanogaster* is influenced by the endosymbiotic bacterium *Wolbachia pipientis*, and in the second chapter I investigate whether immune system genes of the malaria mosquito *Anopheles coluzzii* show signatures of pathogen-driven coevolution.

The endosymbiotic bacterium *Wolbachia pipientis* confers *D. melanogaster* and other insects with resistance to infection by RNA viruses. I investigated whether *Wolbachia* infection plays a role in protecting *D. melanogaster* against secondary bacterial infection, and in particular against pathogenic intracellular bacteria. I found no evidence that *Wolbachia* alters resistance to any of the bacterial pathogens tested, irrespective of how they colonize the host.

The ability to resist and/or survive infection is a critical determinant of host fitness, so natural selection is predicted to drive the evolution of resistance mechanisms in response to novel or coevolving pathogens. I used molecular population genetic analyses to investigate how natural selection operates on the immune system of *Anopheles coluzzii*. I found evidence of rapid adaptive evolution in *STAT-B*, and long-term balancing selection in *CTLMA2*. In contrast to these *Anopheles*-specific immune genes, we found that genes encoding the Imd immune pathway, which are orthologously conserved across insects, exhibit patterns of genetic variation consistent with relaxed purifying selection. These results indicate that adaptive coevolution between *A. coluzzii* and its pathogens is more likely to involve novel or lineage-specific molecular mechanisms than the canonical humoral immune pathways.

BIOGRAPHICAL SKETCH

Susan's first week of work in the Lazzaro lab involved spending several hours a day in a small dark walk-in chamber at a microscope with red light filters, poking flies with a pin. It was weird. But the people seemed nice, so she decided to stick around.

ACKNOWLEDGMENTS

I would like to thank Brian Lazzaro for allowing me to do this master's degree through the employee degree program while working in the lab, and for his patience throughout the process. Additionally, I would like to thank my committee member Rick Harrison for thoughtful comments and suggestions on the thesis.

I am grateful to members of the Lazzaro Lab, past and present, who helped me with the work presented in this thesis. I would particularly like to thank Jacob Crawford for his guidance in population genetic analysis, and for taking the time to provide meaningful feedback and discussion of the manuscript. I am also extremely thankful to Mark Jandricic, who performed thousands of colony PCRs and countless plasmid preps for me in the course of collecting the sequence data used in chapter 2. I would also like to thank Rob Unkless for helping me with Linux, and Sarah Short for her help with statistical analysis of bacterial load data.

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH iv

ACKNOWLEDGEMENTS v

CHAPTER 1: NO EFFECT OF *WOLBACHIA* ON RESISTANCE TO INTRACELLULAR INFECTION BY PATHOGENIC BACTERIA IN *DROSOPHILA MELANOGASTER*

Abstract 1
Introduction 2
Methods 4
Results 9
Discussion 10
Tables 14
Figures 17
References 18

CHAPTER 2: POPULATION GENETICS OF *ANOPHELES COLUZZII* IMMUNE PATHWAYS AND GENES

Abstract 20
Introduction 22
Methods 25
Results 29
Discussion 38
References 43
Tables 48
Figures 53
Supplemental Information 57

CHAPTER 1

NO EFFECT OF *WOLBACHIA* ON RESISTANCE TO INTRACELLULAR INFECTION BY PATHOGENIC BACTERIA IN *DROSOPHILA MELANOGASTER*[1]

Abstract

Multiple studies have shown that infection with the endosymbiotic bacterium *Wolbachia pipientis* confers *Drosophila melanogaster* and other insects with resistance to infection by RNA viruses. Studies investigating whether *Wolbachia* infection induces the immune system or confers protection against secondary bacterial infection have not shown any effect. These studies, however, have emphasized resistance against extracellular pathogens. Since *Wolbachia* lives inside the host cell, we hypothesized that *Wolbachia* might confer resistance to pathogens that establish infection by invading host cells. We therefore tested whether *Wolbachia*-infected *D. melanogaster* are protected against infection by the intracellular pathogenic bacteria *Listeria monocytogenes* and *Salmonella typhimurium*, as well as the extracellular pathogenic bacterium *Providencia rettgeri*. We evaluated the ability of flies infected with *Wolbachia* to suppress secondary infection by pathogenic bacteria relative to genetically matched controls that had been cured of *Wolbachia* by treatment with tetracycline. We found no evidence that *Wolbachia* alters host ability to suppress proliferation of any of the three pathogenic bacteria. Our results indicate that *Wolbachia*-induced antiviral protection does not result from a generalized response to intracellular pathogen.

[1] This paper has been published with the citation as follows:
Rottschaefer SM, Lazzaro BP (2012) No Effect of *Wolbachia* on Resistance to Intracellular Infection by Pathogenic Bacteria in *Drosophila melanogaster*. PLoS ONE 7(7): e40500.

Introduction

Wolbachia is a genus of maternally inherited, obligate intracellular bacteria that infect a wide range of arthropods and filarial nematodes. It has been estimated that as many as 70% of all insect species may be infected [1]. Extensive horizontal transfer is credited with introducing *Wolbachia* to such a large number of host species. Once introduced, the successful spread of *Wolbachia* throughout host populations can be explained in large part by the ability to act as reproductive parasites, manipulating or disrupting the host reproductive biology in such a way to promote their own transmission. In many species, *Wolbachia* induces cytoplasmic incompatibility (CI), which causes high egg mortality in crosses between infected males and uninfected females, resulting in a relative fitness advantage for infected females and driving *Wolbachia* spread once *Wolbachia* infection has reached a critical threshold in the population [2]. Natural selection could help *Wolbachia* reach that threshold and facilitate further spread if the bacterium provides an additional selective advantage to infected hosts. In one example of such an advantage, *Drosophila melanogaster* infected with *Wolbachia pipientis* show dramatic resistance to infection by RNA viruses [3,4]. This antiviral protection appears robust in *D. melanogaster*, having been observed across multiple host genotypes and *Wolbachia* strains [3,4]. Similar antiviral protection is observed when *D. simulans* is infected with certain *Wolbachia* strains, although other *Wolbachia* strains infecting *D. simulans* do not alter resistance [5]. These observations indicate that *Wolbachia* infection can influence host immunity, but the mechanism of pathogen resistance remains unknown. Previous work in *Drosophila* suggesting that *Wolbachia* infection does not confer protection against secondary bacterial infection has focused on extracellular bacterial pathogens [6]. Like viruses, however, some pathogenic bacteria

establish infection by invading host cells where *Wolbachia* is resident. To date, there have been no published tests of whether *Wolbachia* can confer resistance to intracellular bacterial infection.

It has been hypothesized that *Wolbachia* alters the systemic immune response of the host, increasing the ability to quickly detect and mount a response to the infection. In *Aedes aegypti*, for example, *Wolbachia*-induced resistance to a range of pathogens including filarial nematodes, Gram-negative bacteria, and Dengue virus is associated with increased basal expression of immune genes [7,8,9]. Microarray analysis of Drosophila S2 cells showed slight upregulation of some genes involved in the Toll and IMD pathways in the presence of *Wolbachia* infection [10], although other studies of selected immune genes in whole flies have found that *Wolbachia* does not alter expression in *D. melanogaster* [6] or *D. simulans* [11]. If *Wolbachia* is able to alter the systemic immune response of *D. melanogaster*, we would expect to see increased resistance against bacterial pathogens in addition to viruses. *Wolbachia* infection does not confer *D. melanogaster* or *D. simulans* with resistance against the pathogenic bacteria *Pseudomonas aeruginosa*, *Serratia marcescens*, and *Erwinia carotovora* [6] which are all extracellular pathogens. Another hypothesis is that *Wolbachia* infection increases resistance specifically to intracellular pathogens. Intracellular pathogen surveillance could be heightened as a consequence of *Wolbachia* infection, allowing for rapid detection and elimination of pathogens invading the cytoplasm. Additionally, since *Wolbachia* resides within host cells, it could limit the success of an intracellular pathogen through competition for resources within the host cytoplasm. In either of these cases, increased resistance would only be observed when *Wolbachia*-infected individuals are challenged with an intracellular pathogen.

We investigated whether *Wolbachia* infection alters *D. melanogaster* defense against secondary bacterial infection, and in particular against pathogenic intracellular bacteria. We

specifically focus in this paper on resistance, defined as the ability to minimize pathogen burden [12]. We compared the ability to suppress secondary pathogen infection of flies from five isofemale lines of *D. melanogaster* that are naturally infected with *Wolbachia* to the ability of those same lines to suppress pathogenic infection after removal of the *Wolbachia* with tetracycline. In order to control for the effect of the tetracycline, we also evaluated tetracycline treatment in five naturally *Wolbachia*-uninfected isofemale lines. In order to determine whether *Wolbachia* infection influences generalized resistance to multiple pathogens or a more specific response to intracellular pathogens, we tested infection with *Salmonella typhimurium* and *Listeria monocytogenes*, which are intracellular bacterial pathogens, and *Providencia rettgeri*, an extracellular pathogen. We found no evidence that *Wolbachia* alters resistance to any of the three bacterial pathogens tested.

Methods

Flies and Antibiotic Treatment: The *D. melanogaster* isofemale lines used in this experiment were established from field-inseminated females collected in Newfield, New York, USA, in 2005. Each individual female was placed in media-containing vials immediately after collection, and her resulting progeny were allowed to sib-mate. These isofemale lines have been maintained since then by recurrent mass sib-mating. Genetic variation observed among the isofemale lines therefore reflects variation in the natural population from which they were sampled. A diagnostic PCR which amplified *wsp* was used to determine *Wolbachia* infection status of the lines [13]. Antiviral protection has been observed in *D. melanogaster* infected with the *Wolbachia* strains *w*Mel, *w*MelCS and *w*MelPop [3,4]. There is evidence that *w*Mel is the predominant variant infecting field populations [14], so it is likely to be the strain present in our recently founded

lines, although we did not explicitly test this. We randomly chose 5 infected [WOLB(+)] and 5 uninfected [WOLB(-)] lines to use for the experiment. *D. melanogaster* can be experimentally cured of *Wolbachia* by treatment with the antibiotic tetracycline [15]. Flies from both naturally infected and naturally uninfected lines were treated with tetracycline as described below, resulting in four treatment groups to be contrasted for resistance to pathogenic bacterial infection: WOLB(+)TET(-), WOLB(+)TET(+), WOLB(-)TET(-), and WOLB(-)TET(+).

Flies were reared on the standard Cornell Drosophila medium (8.3% w/v glucose, 8.3% w/v brewer's yeast, 1% w/v agar) throughout the experiment. For the antibiotic treatment, the flies were reared for three generations on the standard Cornell medium with 50ug/ml tetracycline added [4]. After each generation on tetracycline supplemented medium, eight flies from each line were screened for the presence of *Wolbachia* using the PCR assay described above [13]. Approximately 50% of the flies screened after one generation of tetracycline treatment were cured of *Wolbachia* and approximately 90% were cured after two generations of treatment. After three generations of tetracycline treatment, *Wolbachia* was not detected in any of the flies tested and the isofemale lines were then returned to the standard medium without antibiotic for all subsequent generations. Flies in all treatments were maintained at 25ºC with 12h light, 12h dark. Flies were infected 1-5 hours after "dawn". All males used for infections were aged 3-5 days.

Bacterial strains: *Providencia rettgeri* strain Dmel is a Gram-negative extracellular pathogen isolated from wild caught *D. melanogaster* that causes moderate mortality in the fly [16]. *Salmonella enterica* serotype Typhimurium S5520 (obtained from Dr. Martin Wiedmann, Cornell University) is a Gram-negative bacterium which is able to establish an intracellular infection causing mortality in *D. melanogaster*, although the bacteria do not replicate to high

numbers [17]. *Listeria monocytogenes* 10403S (obtained from Dr. Martin Wiedmann, Cornell University) is a Gram-positive intracellular bacterium which is able to invade and replicate to high numbers within the cells of *D. melanogaster*, causing moderate mortality [18].

Infections: Since residual effects of tetracycline may persist multiple generations after treatment [19], we measured systemic bacterial load 2, 4, and 6 generations after ending tetracycline treatment. Three sets of infections were done in a day (one for each pathogen) and were repeated on three replicate days for each generation tested. For the infections, 15 males from each line and treatment were anesthetized on CO_2 and pricked in the thorax with a 0.1 mm pin dipped into a bacterial culture. *P. rettgeri* cultures were grown in LB at 37°C with shaking overnight and diluted to A_{600}=1 immediately before infections. *L. monocytogenes* cultures were grown in BHI liquid overnight at 37°C with shaking. To prepare the inocula, 2ml of liquid culture with A_{600}=1 was spun down and the supernatant removed, and the pellet was resuspended in 200µl of BHI. *S. typhimurium* cultures were grown in BHI liquid overnight at 37°C without shaking. To prepare the inocula, 2ml of liquid culture with A_{600}=1 was spun down and the supernatant removed, and the pellet was resuspended in 200µl of BHI.

To measure systemic bacterial load, 3 pools of 5 flies from each line were homogenized and plated approximately 24 hours after infection. Flies infected with *P. rettgeri* were homogenized in 500µl LB, and the homogenate was diluted 1:100 prior to plating on LB plates. Flies infected with *L. monocytogenes* and *S. typhimurium* were homogenized in 250µl BHI. The *L. monocytogenes* homogenate was diluted 1:10 in BHI prior to plating on BHI plates, and the *S. typhimurium* homogenate was not diluted prior to plating on BHI plates. A spiral plater (Don Whitley Scientific) was used to plate 50µl of each sample over a continuous exponential dilution.

Plates were grown at 37°C overnight. The bacteria used for experimental infections grow into visible colonies during this period, while gut commensal bacteria do not appear as visible colonies on the plates until approximately 24 hours later. Thus, we can be certain that the colonies we count reflect systemic pathogen load. Every plate was visually inspected to verify that the color and morphology of all colonies were consistent with that of the experimental bacteria, and any plates with contaminating colonies were discarded. The resulting colonies were counted using the ProtoCOL plate counter associated with the spiral plater to determine the systemic pathogen load of the flies.

Statistical analysis: To assess the effect of *Wolbachia* infection, tetracycline treatment, and time since tetracycline treatment on resistance to each pathogen, we performed a mixed-model analyses of variance (ANOVA) on the natural log transformed bacterial load data using the following model:

$$Y_{ijklm} = \mu + \text{line(WOLB)}_i + \text{WOLB}_j + \text{TET}_k + \text{GEN}_l + \text{REP(GEN)}_m + \text{WOLB}_j*\text{TET}_k + \text{TET}_k*\text{GEN}_l + \text{WOLB}_j*\text{TET}_k*\text{GEN}_l + \text{GEN}_l*\text{line(WOLB)}_i + \text{TET}_k*\text{line(WOLB)}_i + \text{TET}_k*\text{GEN}_l*\text{line(WOLB)}_i + \varepsilon_{ijklm}$$

where Y is the natural log of the bacterial load, line(WOLB) (i=1,5) represents the effect of *Drosophila* genetic line within each level of the model factor WOLB, WOLB (j=1,2) represents the *Wolbachia* infection status of each line prior to antibiotic treatment, TET (k=1,2) represents whether or not flies were treated with tetracycline, GEN (l=1,3) represents whether the experiment was performed 2,4, or 6 generations after tetracycline treatment, and REP(GEN) (m=1,3) is the random effect of the replicate day on which the data were collected within each generation. The factor $\text{WOLB}_j*\text{TET}_k$ tests for differential effects of tetracycline treatment on

Wolbachia-infected and *Wolbachia*-uninfected lines, which allows us to distinguish the effect of removing *Wolbachia* from the overall effect of tetracycline. The factor $WOLB_j$*TET_k*GEN_l tests whether effects of tetracycline on *Wolbachia*-infected and uninfected lines are consistent across the successive generations. The factor GEN_l*line$(WOLB)_i$ tests whether the lines within each WOLB level behave consistently across the generations. The factor TET_k*line$(WOLB)_i$ tests whether the effect of tetracycline treatment varies among lines within each WOLB level. The factor TET_k*GEN_l*line$(WOLB)_i$ tests whether tetracycline treatment has genotype-dependent effects that vary across generations.

To further elucidate the nature of the observed effect of the line$(WOLB)_i$*TET_k *GEN_l interaction on resistance to *P. rettgeri* (see Results), we performed an additional mixed ANOVA for each generation separately. This model takes the form:

$$Y_{ijkl} = \mu + \text{line}(WOLB)_i + WOLB_j + TET_k + REP_l + WOLB_j*TET_k + TET_k*\text{line}(WOLB)_i + \varepsilon_{ijkl}$$

where Y is the natural log of the bacterial load, line(WOLB) (i=1,5) represents the effect of genotype nested within each level of the factor WOLB, WOLB (j*=1,2)* represents the *Wolbachia* infection status of each line prior to antibiotic treatment, TET (j=1,2) represents whether or not flies were treated with tetracycline, and REP (k=1,3) is the random effect of the replicate day on which the experiment was performed. The factor $WOLB_j$*TET_k tests for differential effects of tetracycline treatment on *Wolbachia*-infected and *Wolbachia*-uninfected lines. The factor TET_j*line$(WOLB)_i$ tests whether tetracycline treatment has genotype-dependent effects.

All of the model factors described in the text above are also listed in Table 1. Removal of various non-significant factors from the model does not change the qualitative outcome of any of

our analyses, so we present here the full models in order to provide the most complete information. All analyses were performed using SAS 9.3 (SAS Institute).

Results

When flies were infected with *P. rettgeri*, we observed significant differences in bacterial load across the isofemale lines ($p < 0.0001$, Table 2), but no difference in bacterial load owing to the initial *Wolbachia* status of those lines ($p = 0.0873$). Systemic pathogen load of tetracycline-treated flies, considered across genotypes, did not differ from that of untreated flies ($p = 0.2062$). The effect of tetracycline treatment on *Wolbachia*-infected lines was not different from the effect of tetracycline treatment on *Wolbachia*-uninfected lines (WOLB*TET, $p = 0.086$, Table 2 and Figure 1A), indicating that removal of *Wolbachia* does not influence ability to suppress *P. rettgeri* infection. Interestingly, we find a nearly significant TET*line(WOLB) interaction ($p = 0.061$, Table 2), which suggests that the effects of tetracycline may be stronger in some genetic backgrounds than others. Additionally, the three-way TET*GEN* line(WOLB) interaction is significant ($p = 0.0117$, Table 2). This three way interaction indicates that the genotype-specific effect of tetracycline treatment varies across generations, but it does not provide any direct information about the nature of this complex interaction. We decided to investigate this three-way interaction further by running a separate analysis for each of the three generations tested. Interestingly, we find a significant TET*line(WOLB) interaction in response to *P. rettgeri* two generations after treatment ($p = 0.0017$, Table 3) whereas this interaction is not significant in the subsequent generations. Taken together, these results suggest that an effect of tetracycline may persist in some, but not other genetic backgrounds two generations after treatment, but that the effect does not persist for four or more generations in any of the genetic backgrounds.

When flies were infected with *L. monocytogenes*, we observed significant differences in bacterial load across the isofemale lines ($p < 0.0001$, Table 2), but no difference in *L. monocytogenes* load owing to the initial *Wolbachia* status of those lines ($p = 0.288$) or to tetracycline treatment ($p = 0.3117$). The effect of tetracycline treatment on *Wolbachia*-infected lines was not different from the effect of tetracycline treatment on *Wolbachia*-uninfected lines (WOLB*TET, $p = 0.7254$, Table 2 and Figure 1B), indicating that removal of *Wolbachia* does not influence ability to suppress *L. monocytogenes* infection. In contrast to infection with *P. rettgeri*, there was no genotype-by-treatment interaction in response to *L. monocytogenes* infection ($p = 0.1857$), nor was there any indication of a three way genotype-by-treatment-by-generation interaction ($p = 0.191$).

When flies were infected with *S. typhimurium*, we observed significant differences in bacterial load across the isofemale lines ($p = 0.0222$, Table 2), but no effect of initial *Wolbachia* status ($p = 0.302$) or tetracycline treatment ($p = 0.374$). The effect of tetracycline treatment on *Wolbachia*-infected lines was not different from the effect of tetracycline treatment on *Wolbachia*-uninfected lines (WOLB*TET, $p = 0.2548$, Table 2 and Figure 1C), indicating that removal of *Wolbachia* does not influence ability to suppress infection by *S. typhimurium.* As with infection by *L. monocytogenes*, there was no genotype-by-treatment interaction in response to *S. typhimurium* infection ($p = 0.5767$) and no three way genotype-by-treatment-by-generation interaction ($p = 0.091$).

Discussion

In this experiment we used two intracellular bacterial pathogens and one extracellular bacterial pathogen to investigate whether *Wolbachia* infection influences *D. melanogaster*

resistance to pathogenic bacteria. Unfortunately there are no known natural intracellular bacterial pathogens of *D. melanogaster*, so for this experiment we used the human pathogens *Listeria monocytogenes* and *Salmonella typhimurium.* Although these are not natural pathogens of *D. melanogaster*, both are able to invade and replicate within the cells of *D. melanogaster* and have been used to study intracellular infection in *D. melanogaster* [17,18]. We did not find evidence that *Wolbachia* confers protection against either of the intracellular bacteria. When *Wolbachia*-infected *D. melanogaster* are infected with DCV or Nora virus, both of which are natural pathogens, survival is increased and viral proliferation is inhibited [4]. Increased survival is also observed in *Wolbachia*-infected flies infected with the non-natural pathogen FHV, but in this case viral proliferation does not appear to be inhibited [4]. This observed disconnection between virus proliferation and host mortality suggests that the mechanisms by which *Wolbachia* confers protection involve both host immunity and host tolerance, the effects of which may be specific to particular pathogens or natural host-pathogen pairs. It therefore may be interesting to discover with future studies whether similar infection phenotypes are observed with natural intracellular bacterial pathogens.

Likewise, further studies might investigate the effects of *Wolbachia* on host fitness over the course of an infection. In this experiment we measured resistance, defined as the ability to minimize pathogen burden [12], because we were specifically interested in whether the presence of *Wolbachia* influences the host ability to suppress secondary bacterial infection. *Wolbachia* infection could conceivably also increase host tolerance of infection, such that *Wolbachia*-infected flies might survive longer or have higher reproductive success than uninfected flies despite similar pathogen infection loads. However, *Wolbachia* infection has previously been

reported to have no effect on mortality in *D. melanogaster* after infection with extracellular bacterial pathogens [6].

In addition to investigating the effect of *Wolbachia* on resistance to bacterial pathogens, we examined the residual effect of tetracycline on flies multiple generations after treatment. Reduced mitochondrial metabolism and increased mtDNA density have been reported in *D. simulans* two generations after treatment with tetracycline [19], and antibiotic treatment additionally eliminates commensal gut microbes. Gut microbes have important regulatory effects on the immune system in the gut, and the presence or absence of individual microbes can disrupt gut homeostasis [20]. For example, aseptically reared *Anopheles gambiae* are more susceptible to *Plasmodium falciparum* infection than are non-sterile mosquitoes [21]. Two generations after cessation of tetracycline treatment, we found a significant line-by-tetracycline interaction on the ability of flies to suppress infection by the Gram-negative extracellular pathogen *P. rettgeri.* This suggests that residual effects of tetracycline may persist in some, but not other, genetic backgrounds for multiple generations. We speculate that there may be genetic variation for the number of generations required to recover from the effects of tetracycline treatment, perhaps resulting from differences in the ability to reacquire commensal gut microbes and return gut homeostasis or to differences in the rate of mitochondrial recovery. Further experimentation is required to further elucidate the nature of this interaction.

In summary, it is well established that *Wolbachia* provides protection against RNA viruses in *Drosophila* [3,4] so we sought to determine whether the *Wolbachia*-induced resistance to viruses could be generalized to other intracellular pathogens. We measured the abilities of *Wolbachia*-infected and uninfected *D. melanogaster* to suppress infection by the intracellular pathogenic bacteria *L. monocytogenes* and *S. typhimurium* and the extracellular pathogenic bacterium *P.*

rettgeri, but we observed no effect of *Wolbachia* on resistance to infection by any of the three, irrespective of how they colonize the host.

Acknowledgements

We are thankful to Mark Jandricic and Chloe Ota for invaluable help collecting the data, and to Sarah Short and Madeline Galac for helpful discussion and comments on the manuscript.

Table 1. Description of Factors Tested in Analyses of Variance

Factor	Type	Effect Measured
line(WOLB)	fixed	effect of each genetic line nested within the factor WOLB
WOLB	fixed	*Wolbachia* status of each line prior to tetracycline treatment
TET	fixed	whether or not flies were treated with tetracycline
GEN	fixed	number of generations since tetracycline treatment (2, 4, 6)
REP(GEN)	random	replicate day on which the experiment was performed
WOLB*TET	fixed	differential effect of tetracycline on flies with and without *Wolbachia*
TET*GEN	fixed	differential effect of tetracycline across the generations tested
GEN*line(WOLB)	fixed	differential effect of line across the generations tested
TET*line(WOLB)	fixed	differential effect of tetracycline on flies of each line
WOLB*TET*GEN	fixed	differential effect of tetracycline on flies with and without *Wolbachia* across generations
TET*GEN*line(WOLB)	fixed	differential effects of tetracycline on flies of each line and across generations

Table 2. Analyses of variance for fixed effects relating genotype, *Wolbachia* status, tetracycline treatment, and generation to bacterial load.

		P. rettgeri		*L. monocytogenes*		*S. typhimurium*	
Factor	d.f.	F-ratio	P-value	F-ratio	P-value	F-ratio	P-value
line(WOLB)	8	21.10	<0.0001	14.01	<0.0001	2.26	0.0222
WOLB	1	2.94	0.0873	1.13	0.2880	1.07	0.3020
TET	1	1.60	0.2062	1.03	0.3117	0.79	0.3740
GEN	2	3.93	0.0811	0.19	0.8324	0.02	0.9800
WOLB*TET	1	2.96	0.0860	0.12	0.7254	1.30	0.2548
WOLB*TET*GEN	2	0.01	0.9892	0.48	0.6206	0.09	0.9099
TET*GEN	2	2.11	0.1219	0.15	0.8623	0.56	0.5705
GEN* line(WOLB)	16	1.37	0.1543	0.71	0.7841	0.49	0.9514
TET* line(WOLB)	8	1.88	0.0609	1.42	0.1857	0.83	0.5767
TET*GEN*line(WOLB)	16	2.01	0.0117	1.30	0.1910	1.51	0.0910

Table 3. Analyses of variance relating fixed effects of genotype, *Wolbachia* status, and tetracycline treatment to bacterial load when infected with *P. rettgeri* 2, 4, and 6 generations after tetracycline treatment.

		generation 2		generation 4		generation 6	
Factor	d.f.	F-ratio	P-value	F-ratio	P-value	F-ratio	P-value
line(WOLB)	8	8.33	<0.0001	6.69	<0.0001	9.24	<0.0001
WOLB	1	1.02	0.3146	1.64	0.2029	0.39	0.5311
TET	1	0.68	0.4117	3.91	0.0497	1.12	0.2918
TET*WOLB	1	1.08	0.2999	0.83	0.3646	1.01	0.3176
TET* line(WOLB)	8	3.30	0.0017	0.61	0.7707	1.91	0.0619

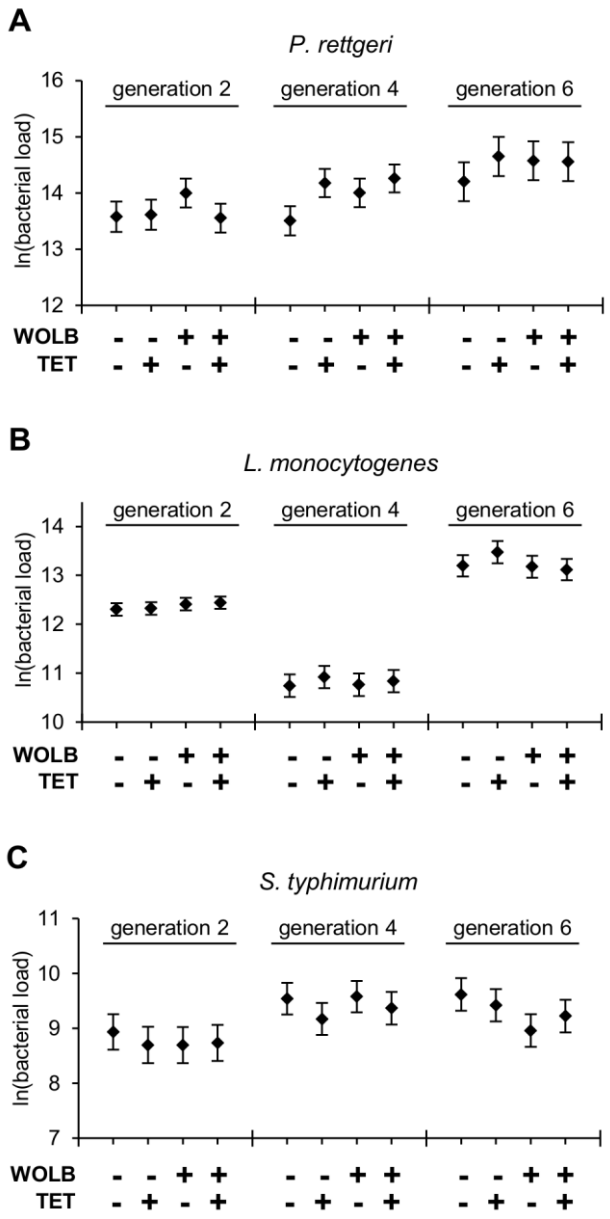


Figure 1. Systemic bacterial load is not influenced by *Wolbachia* infection. Least squares mean bacterial load (± 1SE) of five *Wolbachia*-infected lines [WOLB(+)TET(-)] and their genetically matched lines that have been cured of *Wolbachia* [WOLB(+)TET(+)], as well as five *Wolbachia*-uninfected lines [WOLB(-)TET(-)] and genetically paired tetracycline treated lines[WOLB(-)TET(+)]. Note that the "WOLB" category on the x-axis refers to initial *Wolbachia*-infection status prior to antibiotic treatment, rather than infection status at the time of experimental infections. Bacterial load was measured 24 hours after infection with the pathogenic bacteria (A) *P. rettgeri* (B) *L. monocytogenes* and (C) *S. typhimurium*. Assays were performed 2, 4, and 6 generations after ending tetracycline treatment, with three replicates in each generation. For each replicate, bacterial load was measured in 3 pools of 5 flies from every line.

REFERENCES

1. Jeyaprakash A, Hoy MA (2000) Long PCR improves *Wolbachia* DNA amplification: wsp sequences found in 76% of sixty-three arthropod species. Insect Mol Biol 9: 393–405.

2. Siozios S, Sapountzis P, Ioannidis P, Bourtzis K (2008) *Wolbachia* symbiosis and insect immune response. Insect Sci 15: 89–100.

3. Hedges LM, Brownlie JC, O'Neill SL, Johnson KN (2008) *Wolbachia* and virus protection in insects. Science 322: 702.

4. Teixeira L, Ferreira A, Ashburner M (2008) The bacterial symbiont *Wolbachia* induces resistance to RNA viral infections in *Drosophila melanogaster*. PLoS Biol 6:e1000002.

5. Osborne SE, Leong YS, O'Neill SL, Johnson KN (2009) Variation in antiviral protection mediated by different *Wolbachia* strains in *Drosophila simulans*. PLoS Pathog 5:e1000656.

6. Wong ZS, Hedges LM, Brownlie JC, Johnson KN (2011) *Wolbachia*-mediated antibacterial protection and immune gene regulation in Drosophila. PLoS ONE 6:e25430.

7. Bian G, Xu Y, Lu P, Xie Y, Xi Z (2010) The endosymbiotic bacterium *Wolbachia* induces resistance to Dengue virus in *Aedes aegypti*. PLoS Pathog 6:e1000833.

8. Kambris Z, Cook PE, Phuc HK, Sinkins SP (2009) Immune activation by life-shortening *Wolbachia* and reduced filarial competence in mosquitoes. Science 326: 134–136.

9. Moreira LA, Iturbe-Ormaetxe I, Jeffery JA, Lu G, Pyke AT, et al. (2009) A *Wolbachia* symbiont in *Aedes aegypti* limits infection with dengue, Chikungunya, and Plasmodium. Cell 139: 1268-1278.

10. Xi ZY, Gavotte L, Xie Y, Dobson SL (2008) Genome-wide analysis of the interaction between the endosymbiotic bacterium *Wolbachia* and its *Drosophila* host. BMC Genomics 9:1.

11. Bourtzis K, Pettigrew MM, O'Neill SL (2000) *Wolbachia* neither induces nor suppresses transcripts encoding antimicrobial peptides. Insect Mol Biol 9: 635–639.

12. Raberg L, Sim D, Read AF (2007) Disentangling genetic variation for resistance and tolerance to infectious diseases in animals. Science 318: 812–814.

13. Zhou W, Rousset F, O'Neil S (1998) Phylogeny and PCR-based classification of *Wolbachia* strains using *wsp* gene sequences. Proc Biol Sci 265: 509–515.

14. Riegler M, Sidhu M, Miller WJ, O'Neill SL (2005) Evidence for a global *Wolbachia* replacement in *Drosophila melanogaster*. Curr Biol 15: 1428–1433.

15. Hoffmann AA, Turelli M, Simmons GM (1986) Unidirectional incompatibility between populations of *Drosophila simulans*. Evolution 40: 692–701.

16. Galac M, Lazzaro BP (2011) Comparative pathology of bacteria in the genus *Providencia* to a natural host, *Drosophila melanogaster*. Microbes Infect 13:673-683.

17. Brandt SM, Dionne MS, Khush RS, Pham LN, Vigdal TJ, et al. (2004) Secreted Bacterial Effectors and Host-Produced Eiger/TNF Drive Death in a Salmonella-Infected Fruit Fly. PLoS Biol 2:e418.

18. Mansfield BE, Dionne MS, Schneider DS, Freitag NE (2003) Exploration of host-pathogen interactions using *Listeria monocytogenes* and *Drosophila melanogaster*. Cell Microbiol 5: 901–911.

19. Ballard JW, Melvin RG (2007) Tetracycline treatment influences mitochondrial metabolism and mtDNA density two generations after treatment in *Drosophila*. Insect Mol Biol 16: 799-802.

20. Ryu JH, Kim SH, Lee HY, Bai JY, Nam YD, et al. (2008) Innate immune homeostasis by the homeobox gene caudal and commensal-gut mutualism in *Drosophila*. Science 319: 777-782.

21. Dong Y, Manfredini F, Dimopoulos G (2009) Implication of the mosquito midgut microbiota in the defense against malaria parasites. PLoS Pathog 5:e1000423.

CHAPTER 2

POPULATION GENETICS OF *ANOPHELES COLUZZII* IMMUNE PATHWAYS AND GENES[2]

Abstract

Natural selection is expected to drive adaptive evolution in genes involved in host-pathogen interactions. In this study, we use molecular population genetic analyses to understand how natural selection operates on the immune system of *Anopheles coluzzii* (formerly *A. gambiae* "M form"). We analyzed patterns of intraspecific and interspecific genetic variation in 20 immune-related genes and 17 non-immune genes from a wild population of *A. coluzzii*, and asked if patterns of genetic variation in the immune genes are consistent with pathogen-driven selection shaping the evolution of defense. We found evidence of a balanced polymorphism in *CTLMA2*, which encodes a C-type lectin involved in regulation of the melanization response. The two *CTLMA2* haplotypes, which are distinguished by fixed amino acid differences near the predicted peptide cleavage site, are also segregating in the sister species *A. gambiae* ("S form") and *A. arabiensis*. Comparison of the two haplotypes between species indicates that they were not shared among the species through introgression, but rather that they arose before the species divergence and have been adaptively maintained as a balanced polymorphism in all three species. We additionally found that *STAT-B*, a retroduplicate of *STAT-A*, shows strong evidence of adaptive evolution that is consistent with neofunctionalization after duplication. In contrast to the striking patterns of adaptive evolution observed in these *Anopheles*-specific immune genes, we found no evidence of adaptive evolution in the Toll and Imd innate immune pathways that are orthologously conserved throughout insects. Genes encoding the Imd pathway exhibit high rates

[2] This paper has been accepted at Genes, Genomes, Genetics with the authors SM Rottschaefer, JE Crawford, MM Riehle, WM Guelbeogo, A Gneme, N Sagnon, KD Vernick, and BP Lazzaro.

of amino acid divergence between *Anopheles* species, but also display elevated amino acid diversity that is consistent with relaxed purifying selection. These results indicate that adaptive coevolution between *A. coluzzii* and its pathogens is more likely to involve novel or lineage-specific molecular mechanisms than the canonical humoral immune pathways.

INTRODUCTION

Anopheles mosquitoes live in pathogen-rich environments where survival requires an effective immune system. Population genetic studies in *Anopheles gambiae* have generally focused on putative anti-malaria genes (Obbard *et al*. 2007, 2008, 2009a, Slotman *et al*. 2007, Cohuet *et al*. 2008, Parmakelis *et al.* 2008, White *et al*. 2011, Rottschaefer *et al*. 2011). However, mosquitos are also exposed to diverse microbes, microsporidia, and other pathogens during their larval stages when they live in septic standing water (Muspratt 1946, Bargielowski and Koella 2009, Wang *et al*. 2011), and these larval pathogens probably impose stronger selection on the immune system than the malaria parasite. Regardless of the proximal selective agent, any pathogen-driven evolution of the immune system is likely to shape the efficacy of resistance to parasites of public health relevance, including human malaria, through pleiotropic effects on cross-resistance (Mitri *et al*. 2009, White *et al*. 2011, Rottschaefer *et al*. 2011). In this study, we look broadly at the *Anopheles* immune system to determine targets of natural selection and identify potential instances of pathogen-driven evolution that may shape general resistance to infection.

Since host fitness depends on the ability to combat infection, pathogens are expected to impose significant selection pressure and immune system genes are often observed to evolve rapidly (Schlenke and Begun 2003, Nielsen *et al.* 2005, Obbard *et al.* 2006, Sackton *et al.* 2007, McTaggart *et al.* 2012). Comparative genomic studies show conserved gene orthology in insect innate immune systems, particularly for genes in intracellular signaling pathways, yet amino acid divergence between species is often elevated in these orthologs relative to non-immune genes (Sackton *et al.* 2007; Waterhouse *et al.* 2007). Immunological divergence can additionally occur through lineage- and species-specific gene family expansions and contractions, which may

reflect differences in selection pressures imposed by distinct pathogenic environments. This process is particularly pronounced in recognition and effector genes (Sackton *et al.* 2007; Waterhouse *et al.* 2007). These comparative genomic studies have provided insight into how the immune system evolves between species over longer evolutionary time, but they do not provide information about how natural selection operates to shape the immune system within a species over shorter timescales. Additionally, comparative genomic analyses are only able to reveal adaptive divergence, but cannot detect the adaptive maintenance of polymorphism within species. Signatures of pathogen-driven evolution over shorter time scales can be detected more effectively from patterns of intraspecific polymorphism relative to divergence in immune system genes.

Insect immune systems can be broadly divided into categories of humoral and cellular responses. The humoral immune response involves recognition of pathogens by pattern recognition molecules, which initiate intracellular signaling cascades that stimulate transcriptional activation of anti-microbial effector molecules. The Toll and Imd immune signaling pathways have been most extensively studied for their role in anti-bacterial and anti-fungal immunity in *Drosophila,* but these pathways have also been shown to play a role in immunity against a variety of pathogens in other species. In *A. gambiae,* the Imd pathway has been implicated in defense against both Gram-positive and Gram-negative bacteria (Meister *et al.* 2005, 2009, Garver *et al.* 2009) and has been shown to play an important role in mediating protection against the human malaria parasite *Plasmodium falciparum* (Mitri *et al.* 2009, Garver *et al.* 2009, 2012, Meister *et al.* 2009). The *A. gambiae* Toll pathway is involved in defense against bacteria and rodent malaria (Barillas-Muray *et al.*1996, Frolet *et al.* 2006, Riehle *et al.* 2008, Garver *et al.* 2009). The JAK-STAT pathway also plays a role in the response to bacterial

infection in *A. gambiae* (Barillas-Muray *et al.* 1999, Gupta *et al.* 2009). While most dipterans have only a single STAT gene, there are two STAT genes in *A. gambiae*, which appear to be the result of a retroduplication event along the Anopheline lineage. *STAT-A,* the ancestral gene, is most closely related to STAT genes in other insects, and *STAT-B* is a derived retrocopy (Gupta *et al.* 2009). The JAK-STAT pathway in *A. gambiae* is less well studied than the Toll and Imd pathways, but studies indicate that it may play a role in killing *P. falciparum* parasites at the oocyst stage in *A. gambiae* (Gupta *et al.* 2009), and has also been shown to limit *Plasmodium vivax* infection in *Anopheles aquasalis* (Bahia *et al.* 2011).

In the cellular immune response, pathogen recognition leads to encapsulation, phagocytosis, or melanization of the pathogen by hemocytes. The melanization response in particular has been studied in *A. gambiae* for its role in killing plasmodium parasites (Collins *et al.* 1986). Two C-type lectins, CTL4 and CTLMA2, have been shown to play a role in regulating the melanization response in *A. gambiae*. CTL4 and CTLMA2 primarily exist in the form of a heterodimer, which is secreted into the hemolymph (Schnitger *et al.* 2009). RNAi experiments indicate an important role for CTL4 and CTLMA2 in anti-bacterial immunity, as silencing of either gene reduces the number of Gram-negative bacteria melanized (Schnitger *et al*. 2009). In contrast, silencing of either CTL4 or CTLMA2 results in an increase in the number of melanized *Plasmodium berghei* ookinetes, suggesting that the CTL4/CTLMA2 heterodimer acts as an agonist for *P. berghei* parasites, preventing their melanization (Osta *et al*. 2004).

In this study, we examine patterns of genetic variation at 20 immune genes in a population of *Anopheles coluzzii* (formerly *A. gambiae* "M form"; Coetzee *et al*. 2013) from Burkina Faso. The set of immune genes consists of genes belonging to the Toll, Imd, and JAK-STAT pathways, as well as other mosquito-specific immune factors. To control for demography

and the effects of genomic location, we additionally sequenced 17 non-immune control genes located physically nearby in the genome to our genes of interest. We found evidence that two distinct haplotypes in *CTLMA2* arose before the divergence of *A. coluzzii, A. gambiae* ("S form") and *A. arabiensis,* and have been adaptively maintained as a balanced polymorphism in all three species. We also found evidence of rapid adaptive evolution in *STAT-B*, suggesting an important role for the JAK-STAT pathway in *A. coluzzii*. We observe higher overall rates of amino acid divergence in the immune genes relative to the control genes. Genes involved in the Imd pathway show particularly high amino acid divergence, but also display elevated amino acid diversity that is consistent with relaxed purifying selection.

METHODS

Mosquito collection and DNA isolation

The *Anopheles coluzzii* individuals used in this study were collected in the village of Goundry, Burkina Faso (coordinates 12°30'N, 1°20'W) in September 2008. Freshly fed females were captured indoors by manual aspirator catch and DNA was extracted from individual carcasses using DNAzol (Invitrogen). Diagnostic PCRs were performed to confirm the species (Scott *et al.* 1993, Favia *et al.* 1997), and whole genome amplification of each sample was performed using the GenomiPhi V2 DNA Amplification Kit (GE Healthcare). *Anopheles merus* DNA from the OPHANSI colony was obtained from Malaria Research and Reference Reagent Resource Center (MR4).

Loci analyzed

We sequenced a set of 20 immune genes including the Toll pathway genes *GNBPB1*, *TOLL1A, TUBE, PELLE, TRAF6, CACT* and *REL1*, the Imd pathway genes *IMD*, *FADD,*

CASPL1, *TAK1*, *IAP2*, *IKK1*, *IKK2*, and *REL2*, the JAK-STAT transcription factors *STAT-A* and *STAT-B*, and the *Anopheles* specific immune factors *CTL4*, *CTLMA2*, and *LRIM1*. In order to control for demography and background effects of genomic location, we additionally sequenced 17 non-immune control genes located nearby our genes of interest. These controls are located within 40-100KB of their 'matched' immune genes, and are similar in size and structure to the immune gene they are matched to. The names and relative chromosomal locations of all loci are shown in Figure S1. All PCR primers were designed based on the published *A. gambiae* genome sequence (Vectorbase, *A. gambiae* genome, version P3). Loci were sequenced from a set of 20 *A. coluzzii* females. Some loci could not be amplified in all 20 individuals, but all loci were sequenced in a minimum of 18 *A. coluzzii* individuals as well as in *A. merus*.

PCR and Sequencing

Each gene was amplified in a single amplicon from whole genome amplified DNA using iProof high fidelity DNA Polymerase (BioRad). PCR products were run out on a 1% agarose gel and the product fragments were excised and purified using EZNA gel extraction kits (Omega BioTek). Adenosine tails were added to the purified products by incubating for 20 minutes at 72° with PCR buffer, dATP and Taq polymerase. Products were then cloned using the either TOPO or TOPO XL cloning kits (both from Invitrogen). Colonies to be sequenced were grown overnight at 37° in liquid Luria-Bertani broth supplemented with 20 mg/ml kanamycin, and the plasmids were isolated using the Qiaprep spin miniprep kit (Qiagen). The products were sequenced directly from the plasmids using the BigDye Terminator Cycle Sequencing Kit v3.1(ABI). The sequences were assembled using Sequencher (Gene Codes Corp.). All sequences have been deposited in GenBank under accession numbers KP274100-KP274844.

Only one of the two alleles at each gene was sequenced from any given mosquito in the study. To correct sequencing errors, all singleton polymorphisms were verified by re-amplification and direct sequencing of heterozygous PCR products. For singleton validation, the entire gene was amplified directly from the whole genome amplified DNA using iProof high fidelity DNA Polymerase (BioRad) and this full-length amplicon was then used as template for a secondary PCR that used internally nested primers to robustly amplify the gene region containing the singleton to be validated. Unincorporated primers and dNTPs were inactivated from these secondary amplification products by incubation with ExoI and SAP (both manufactured by USB) and amplification products were sequenced using the BigDye Terminator Cycle Sequencing Kit v3.1 (ABI).

Population genetic analysis

Average pairwise genetic diversity (π) was calculated for all sites, and also separately for synonymous (π_s) and nonsynonymous (π_a) sites using DnaSP v.5 (Librado and Rozas 2009). The Tajima's *D* statistic (Tajima 1989) was also calculated in DnaSP using silent sites only. The average pairwise genetic divergence at synonymous (K_S) and nonsynonymous (K_A) sites, and their ratio (K_A/K_S) were calculated in DnaSP using *A. merus* as an outgroup. Maximum likelihoood multi-locus HKA tests were implemented using the mlhka program (Wright and Charlesworth 2004) using synonymous sites only. Multi-locus McDonald Kreitman tests were performed using the software MKtest v2.0 (Welch 2006, Obbard *et al*. 2009). The multi-locus tests were performed on the full dataset, as well as on the Toll and Imd pathway genes separately. Three models, which varied only in the parameter α, were implemented for each dataset. In the first model (M0), α was fixed at zero for all loci. In the second model (M1) a single α value estimated from the data was shared by all loci, and in the third model (M2) α was

estimated separately for the immune and control loci. For all three models, the expected neutral divergence ($\lambda=\mu t$) and neutral diversity ($\theta=4N_e\mu$) each took a single value at all loci, and selective constraint was allowed to vary between loci. Maximum likelihood ratio tests and Akaike weighting were used to assess model fit.

To identify distinct haplotype clades in CTLMA2, neighbor-joining gene trees were constructed in MEGA5 (Tamura 2011) using the maximum composite likelihood method and uniform substitution rates, with 1,000 bootstrap replicates. To determine if the CTLMA2 haplotype clades were present in other species we compared our data to published CTLMA2 sequences of *A. gambiae* (GenBank accession numbers EF519453- EF519450 and EF519463 – EF519478) and *A. arabiensis* (GenBank accession numbers EF519419 – EF519428), both collected from Kenyan populations, as well as the outgroup species *A. quadriannulatus* (GenBank accession number EF519435) (Obbard *et al*. 2007). To test for introgression of the CTLMA2 haplotypes among *A. coluzzii, A. gambiae* and *A. arabiensis*, we calculated the average number of pairwise differences (D_{xy}) within and between each clade in DnaSP. To determine typical values of D_{xy} in the genomic region of CTLMA2, we calculated *coluzzii-arabiensis* D_{xy} and *gambiae-arabiensis* D_{xy} from published data for the nearby loci AGAP005540 (GenBank accession numbers EF519480 – EF519501) and APL2 (GenBank accession numbers EF519504 – EF519528) (Obbard *et al.* 2007).

No position matched control gene was sequenced for one immune locus (*TRAF6*), and in two instances a pair of immune loci were located very near to each other (*TAK1* and *PELLE* on chromosome 2R, *CTL4* and *CTLMA2* on chromosome 2L) so a single position control was sequenced for each pair (see Figure S1). Differences in divergence between the immune and control groups were assessed using Mann-Whitney U-tests. When analyzing the full dataset, each

control gene was included only once, but as *TAK1* and *PELLE* are involved in different signaling pathways (Imd and Toll, respectively), the shared control gene was included in the control group for analysis of each individual pathway. Differences in nucleotide polymorphism between the immune and control groups were assessed using paired Wilcoxon tests. *TRAF6*, which lacks a position control, was excluded from these comparisons, and the shared control genes were included twice as they were paired to each immune locus. Mann Whitney U-tests and paired Wilcox tests for differences in divergence and diversity were implemented in R (R Development Core Team 2011).

RESULTS

Reduced purifying selection in the IMD pathway

To test the hypothesis that *A. coluzzii* immune genes might evolve under positive selection, we analyzed patterns of intraspecific and interspecific genetic variation in 20 immune genes and 17 control genes from a single population in Burkina Faso. Population genetic statistics for each locus are listed in Table S1. Across all 37 genes, the average per-site nucleotide diversity was 1.3% at all sites (π), 2.6% at synonymous sites (π_s), and 0.3% at nonsynonymous sites (π_a). Nucleotide diversity was lower for genes on the X chromosome compared to the autosomes (X mean π =0.4%, π_s =0.7%, π_a = 0.1%; autosome mean π = 1.5%, π_s = 2.9%, π_a = 0.4%), in keeping with the lower effective population size on the X (Cohuet *et al.* 2008). Nucleotide divergence was measured using *A. merus* as an outgroup. The average per-site nucleotide divergence was 3.3% at all sites (K), 6% at synonymous sites (K_S), and 0.9% at nonsynonymous sites (K_A). These estimates for diversity and divergence are comparable to previously published estimates in *A. coluzzii* (e.g., Cohuet *et al.* 2008).

Genes which are the target of recurrent positive selection are expected to have elevated rates of amino acid evolution, which can be measured using the ratio of nonsynonymous to synonymous divergence (K_A/K_S). The average K_A/K_S ratio of the immunity genes is significantly higher than that of the position matched control genes (immune K_A/K_S = 0.175, control K_A/K_S = 0.077, Mann-Whitney *U*-test *p*=0.001, Table 1). This difference is driven by divergence at nonsynonymous sites (K_A), which is 3 times higher in the immune genes than controls, while divergence at synonymous sites (K_S) is not significantly different between the immune and control groups (Table 1). When we looked at genes involved in the Toll and Imd pathways separately, we find that genes involved in the Imd pathway have higher average K_A/K_S values than their controls (Imd immune K_A/K_S =0.197, Imd control K_A/K_S =0.068, Mann-Whitney *U*-test *p*=0.003; Table 1). This difference is not driven by one or a few outliers, since the difference remains significant even after removing the highest three K_A/K_S ratios. The average K_A/K_S value in Toll pathway genes is also higher than their controls, although the difference is not statistically significant (Toll immune K_A/K_S =0.109, Toll control K_A/K_S =0.056, Mann-Whitney *U*-test *p*=0.063; Table 1). In both the Toll and Imd pathways, divergence at synonymous sites is not different between the immune and control groups, so the higher K_A/K_S ratio in immune genes is attributable to elevated replacement divergence in the immune genes (Table 1). When all genes in the Imd pathway are excluded from the analysis, immune genes still have marginally higher average K_A/K_S relative to the control genes (0.16 vs. 0.084, Mann-Whitney *U*-test *p*=0.048; Table 1), indicating that the Imd pathway is not the only driver of high amino acid divergence found in the immune genes.

If the increased rate of amino acid evolution in the immune genes were due to recurrent fixation of beneficial alleles, we would also expect to see reduced diversity at physically linked

sites and a shift in the allele frequency spectrum (Nielsen 2005). We found that diversity at synonymous sites was similar in the immune and control genes overall (immune π_s = 2.83%, control π_s = 2.52%, paired Wilcox V=128, *p*=0.196, Table 3), but elevated in Imd pathway genes relative their controls (Imd immune π_s = 3.85%, Imd control π_s = 2.97%, paired Wilcox V=34, *p*=0.023, Table 3). We found no difference in Tajima's *D*, which measures shifts in the allele frequency spectrum, between the immune and control groups overall or when split by immune pathway (Table 3). We do, however, see significantly higher average polymorphism at nonsynonymous sites in the immune genes relative to controls (immune π_a = 0.48%, control π_a =0.16%, paired Wilcox V=172, *p*=0.001, Table 2), driven primarily by elevated π_a in the Imd pathway (Imd immune π_a = 0.66%, control π_a =0.18%, paired Wilcox V=36, *p*=0.008, Table 3). The Toll genes do not have significantly higher average π_a than their controls (Toll immune π_a = 0.31%, control π_a =0.14%, paired Wilcox V=19, *p*=0.219, Table 3).

The elevated Ka/Ks observed in immune genes could be consistent with positive selection driving adaptive evolution in these genes, but could also arise if immune genes are generally less constrained and therefore free to accumulate amino acid substitutions. Under a model of neutral evolution, the ratio of non-synonymous to synonymous divergence is expected to equal the ratio of non-synonymous to synonymous polymorphism. The McDonald-Kreitman test compares the number of polymorphisms and fixed differences at synonymous and nonsynonymous sites to detect deviation from the neutral expectation (McDonald and Kreitman 1991), and can be used to estimate the proportion of nonsynonymous fixations attributed to positive selection (α) (Smith and Eyre-Walker 2002). To determine whether the immune genes as a class show a higher proportion of adaptive substitutions than the control genes, we implemented a multi-locus MK test using the software MKtest v2.0 (Welch 2006, Obbard *et al*.

2009). We first implemented the test on the full dataset. Using a likelihood ratio test, the fit of a model allowing α to take the maximum likelihood value showed no significant improvement over the null model where α is fixed at zero (M1 vs M0 $2\Delta\log(L)=0.64$, p=0.4; Table 2). We then implemented the test on genes in the Toll and Imd pathways separately. In both cases, the model which allowed a single value of α estimated from the data showed a significant improvement over the null model where α is fixed at zero (Imd M1 vs M0 $2\Delta\log(L)=14.8$, p=0.0001; Toll M1 vs M0 $2\Delta\log(L)=5.34$, p=0.0208; Table 2). A model which allowed a separate α value to be estimated for the immune and control genes did not provide any additional improvement in model fit for either pathway (Imd M2 vs M1 $2\Delta\log(L)=2.6$, p=0.1069; Toll M2 vs M1 $2\Delta\log(L)=2.36$, p=0.1245; Table 2). We thus see a proportional increase in both nonsynonymous polymorphism and nonsynonymous divergence in immune genes, providing no support for the hypothesis that the immune genes experience more positive selection than the control genes.

The elevated nucleotide diversity at both synonymous and non-synonymous sites in the Imd pathway genes could be consistent with adaptive maintenance of polymorphism, or alternatively could indicate that purifying selection on these genes is weak relative to control genes, allowing deleterious nonsynonymous mutations to persist in the population as effectively neutral polymorphisms. The ratio of polymorphism to divergence is predicted be equivalent across neutrally evolving loci, whereas elevated polymorphism relative to divergence is predicted under adaptive maintenance of polymorphism. The HKA test, which compares the ratio of polymorphism relative to divergence across multiple loci, can help distinguish between these hypotheses. We used a multilocus HKA test in a maximum-likelihood framework to compare the polymorphism to divergence ratios of the 8 genes in Imd pathway genes to the other

29 genes in the dataset. Under this framework, a model allowing selection on Imd genes as a class did not show a significant improvement over a model that assumed all genes evolve neutrally ($X^2_{(8)}$=6.52, p= 0.6). These results indicate the elevated amino acid diversity observed in Imd pathway genes cannot be explained by a model of adaptive maintenance of polymorphism, but rather that these genes experience weakened purifying selection which allows deleterious nonsynonymous mutations to persist in the population as effectively neutral polymorphisms that may drift to fixation and contribute to divergence between species.

Adaptive evolution in *STAT-B*

STAT-B is an intronless duplicate of *STAT-A* that arose through retrotransposition of a *STAT-A* mRNA. The duplication event occurred more recently than the divergence of the Anopheline and Culicine lineages (145-200MYA, Kryzwinski 2009), yet the two STAT copies are remarkably divergent. STAT-A shows only 43% amino acid identity with STAT-B, while it retains 74% identity with *Aedes aegypti* STAT and 63% identity with *Culex quinquefasciatus* STAT. Despite the rapid divergence of *STAT-B*, there is no evidence of pseudogenization, and both STAT-A and STAT-B appear to play a role in immunity. STAT-A and STAT-B are differentially expressed in various developmental stages, with STAT-A being absent from the pupal stage when STAT-B is highly expressed and STAT-A being expressed at higher levels than STAT-B in adults (Gupta *et al* 2009). Both STAT-A and STAT-B appear to be involved in resistance to bacteria and *Plasmodium* parasites (Gupta *et al* 2009). Gene duplicates are often maintained in the genome because they acquire a function that is distinct from that of the ancestral gene. In this case, we might predict evidence of adaptive evolution in *STAT-B*, particularly when compared to *STAT-A*.

Synonymous site divergence between *A. coluzzii* and *A. merus* in *STAT-B* is typical of genes in our dataset (*STAT-B* K_S=6.3%, mean K_S=6%). Amino acid divergence, on the other hand, is remarkably high at this locus (*STAT-B* K_A= 3%, mean K_A= 0.9%), giving it the highest K_A/K_S ratio in the dataset (*STAT-B* K_A/K_S= 0.46, mean K_A/K_S= 0.13). Such an excess of replacement divergence, along with a deficit of intraspecific nucleotide diversity (*STAT-B* π=0.18%, mean π=1.3%), is consistent with a model of recent adaptive evolution at this locus. To test the hypothesis that positive selection is driving the rapid divergence of *STAT-B*, we used the maximum likelihood multi-locus HKA test to test for a departure from neutrality in *STAT-B* as compared to all other genes in the dataset. A model which hypothesized that *STAT-B* was evolving by directional selection fit the empirical data significantly better than the null model that assumed neutral evolution in all genes ($X^2_{(1)}$=9.1, p=0.0026). In contrast to the high amino acid divergence in *STAT-B*, *STAT-A* exhibits a complete lack of replacement divergence or polymorphism, consistent with a model of purifying selection. Estimates of total and synonymous polymorphism are low in *STAT-A*, but they are consistent with the estimates from other genes on the X chromosome. (X mean π=0.42%, π_S=0.7%; *STAT-A* π=0.42%, π_S =0.64%). This pattern of purifying selection in *STAT-A*, the ancestral gene, and rapid evolution in the derived duplicate *STAT-B*, is consistent with a model of neofunctionalization of *STAT-B*.

Balancing selection in *CTLMA2*

CTL4 and CTLMA2 play a role in regulating the melanization response in *A. gambiae* (Osta *et al*. 2004). They primarily exist in the form of a heterodimer, which is secreted into the hemolymph (Schnitger *et al.* 2009). *CTL4* and *CTLMA2* are located directly adjacent on the chromosome, and in our study were PCR amplified and cloned as a single fragment. While all

but three of the genes examined in this dataset have Tajima's *D* values that are negative or nearly zero, *CTLMA2* stands out with a Tajima's *D* value of positive 1.25 (Figure 1). A sliding window analysis of the entire *CTL4*/*CTLMA2* region shows that the high positive Tajima's *D* values are limited to the 5' end of *CTLMA2*, and that Tajima's *D* throughout the rest of the gene is similar to that of *CTL4* (Figure 2). Inspection of the sequences reveals the presence of two distinct haplotype clades, hereafter referred to as clades A and B. There are 11 fixed SNP differences between clades A and B, all located in the 5' 350 bp of the gene. This block of fixed differences, which spans part of the 5'UTR, the first exon, and the first intron, includes 4 SNPs that result in 3 amino acid changes near the predicted signal cleavage site (Figure 3). We considered that the divergent haplotype could have been introduced through paralogous gene conversion, but we were unable to find significant sequence matches to any CTL paralogs in either the *A. gambiae* genome or the entire NCBI nr database.

To determine if the presence of the two CTLMA2 haplotypes could be consistent with a partial selective sweep, a sweep with recombination, or a sweep from standing variation, we compared the patterns of genetic variation within each haplotype clade to genetic variation in *CTL4* and AGAP005327. If the presence of the haplotypes were due to a partial selective sweep, where the frequency of a selected allele increases in the population but does not or has not yet become fixed in the population, we would expect to see evidence of selection in one clade while variation in the other clade should be consistent with surrounding regions (Hudson *et al*. 1997). If the haplotypes are the result of a sweep from standing variation or a sweep with recombination, we would expect both clades to show signatures of selection (Pritchard *et al*. 2010). It is important to note that the expected values of neutrality statistics like those used here are not the same for both a random sample of chromosomes and a partitioned subset of

chromosomes. Most notably, the effective population size, which is a central parameter in these statistics, will be smaller in clade-wise estimates, since we are specifically conditioning on a specific subset of chromosomes in the population. In our sample, the clades are segregating at 50%, so we should expect an approximately 50% reduction in within-clade nucleotide diversity under neutral, equilibrium conditions, assuming these haplotypes have segregated at this frequency for many generations. Tajima's *D* is calculated as the normalized difference between two neutrality statistics that both depend on effective population size (Tajima, 1989), so we do not expect the reduction in effective population size to affect this statistic. We find that estimates of nucleotide diversity and the site frequency spectra (Tajima's *D*) of clade A are consistent with those observed in the adjacent *CTL4*, as well as the position matched control gene AGAP005327 (Table 4). Nucleotide diversity is slightly lower in clade B, which also shows a non-significant shift in the site frequency spectrum, but on the whole there is little evidence of positive selection on either clade, so we cannot attribute the presence of the two CTLMA2 haplotype clades to partial or soft selective sweep.

Alternatively, these two divergent haplotypes may be balanced polymorphisms maintained by either frequency dependent selection, spatially varying selection pressures, or overdominance. We might then predict that these haplotypes have been segregating for evolutionarily long time and thus might also be found in closely related species. We tested this hypothesis by comparing our sequences to a published dataset (Obbard *et al.* 2007) and found that both CTLMA2 haplotypes are also segregating in *A. gambiae* ("S form") as well as *A. arabiensis* (Figure 3, Figure 4). The presence of both haplotypes in all three species could be consistent with an origin that pre-dates the divergence of the species, or alternatively they could have been shared among species through adaptive introgression. Introgressed loci are expected to

be less diverged between species compared to adjacent chromosomal regions. Therefore, if the CTLMA2 haplotypes arose after the species diverged and were shared by introgression, we would expect the introgressed haplotype to show relatively low divergence between species, and divergence in the ancestral haplotype to be typical for the genomic region. If the CTLMA2 haplotypes pre-date the divergence of the species, divergence within each haplotype clade should be similar to the genome average, while divergence between haplotype clades should be elevated, reflecting a longer coalescence time. To determine which of these models best fit the data for CTLMA2, we calculated the average number of pairwise differences, D_{xy}, among species within and between the CTLMA2 haplotype clades, and also in two nearby loci from a published dataset (Obbard *et al.* 2007). Since D_{xy} depends on effective population size (Nei 1987), which is lower within clades as noted above, we would expect D_{xy} to be smaller relative to nearby regions without similar long-term haplotype structure. We found that D_{xy} among species ranged from 0.008 – 0.012 within each haplotype clade (Table 5), and from 0.030 – 0.035 between haplotype clades (Table 5). The within clade values are consistent with D_{xy} among species calculated for two nearby loci (D_{xy}=0.016 – 0.018, Table 5) and estimated genome-wide *gambiae-arabiensis* D_{xy} of 0.011 (O'Loughlin *et al.* 2014), whereas the between clade values are approximately three times higher, and even higher than the estimated genome-wide *gambiae-merus* D_{xy} of 0.0235, most likely reflecting an ancient origin. Although we can only compare our values to the mean and standard deviation among individual locus values of D_{xy} in the O'Loughlin *et al.* dataset, our between clade values are greater than the mean plus one standard deviation of even the *gambiae-merus* comparison, suggesting that the CTLMA2 values exceed most of the many independent genealogies in their study. Moreover, such high values of D_{xy} across clades are exceptional, given the expectation that clade-wise comparisons should be

scaled downwards. These results support the hypothesis that the two divergent CTLMA2 haplotypes arose before the species divergence and have been adaptively maintained as a balanced polymorphism in all three species.

DISCUSSION

We examined patterns of genetic variation and divergence in 20 immune related genes and 17 control genes. We found that on average, immune genes have elevated rates of amino acid divergence compared to non-immune genes, and this was particularly true for genes involved in the Imd pathway. Our findings are consistent with studies in other insects, which show that, as a group, immune genes are rapidly evolving (Schlenke and Begun 2003, Sackton *et al.*2007, Viljakainen *et al.* 2009). The rapid divergence of immune system genes is often hypothesized to be the result of positive selection driving adaptive evolution of the immune system in response to pathogen pressure (Sackton *et al.* 2007). However, when we examined patterns of nucleotide diversity in the genes in our dataset we found no evidence of recent positive selection driving the evolution of most of the immune genes. For example, although genes in the Imd pathway exhibit elevated amino acid divergence, they also tended to have higher levels of amino acid diversity than the control genes, which is parsimoniously consistent with relaxed purifying selection on this pathway.

Our data contrast to studies in *Drosophila* and termites that conclude adaptive evolution is particularly common in the Imd pathway. Rapid amino acid evolution has been detected in termites in the terminal NF-κB transcription factor of the Imd pathway (Bulmer and Croizer 2006). In *D. melanogaster*, positively selected sites appear to cluster in the interacting domains of Relish, $IKK_{\beta,}$ and Dredd, suggesting that the entire Relish cleavage complex is evolving by

positive selection (Sackton *et al.* 2007). Likewise, a survey of genes throughout the *D. melanogaster* immune system found that the trend of elevated rates of adaptive evolution observed in immune genes overall is primarily driven by genes in the Imd and RNAi pathways (Obbard *et al.* 2009b), although adaptation is not uniform throughout the pathways. It has been hypothesized that pathogen interaction with signaling molecules is driving the rapid evolution of the Imd pathway in *D. melanogaster* (Begun and Whitley 2000). Our findings of relaxed purifying selection in the Imd pathway may seem surprising, as signaling through the Imd transcription factor REL2 has been implicated in anti-bacterial and anti-plasmodium immunity in *A. gambiae*, and given that the Relish cleavage complex is thought to be a site of pathogen-driven evolution in *Drosophila*. However, the specific target or mode of pathogen-driven evolution is likely to be different among insects, shaped by the unique suite of pathogens to which they are exposed, and *A. gambiae* is exposed to a distinct set of pathogens due to its aquatic larval environment as well as its obligate blood feeding and potential exposure to human and animal pathogens. Furthermore, unlike *Drosophila* Relish, the NFkB transcription factor for the *A. gambiae* Imd pathway, REL2, exists in two isoforms (Meister *et al.* 2005). The full length isoform (REL2-F) contains an inhibitory domain which must be cleaved off prior to nuclear translocation, whereas the short isoform (REL2-S) which lacks the inhibitory domains is constitutively active (Meister *et al.* 2005). Additionally, some immune responses involving REL2 signaling have been shown to occur independent of the IMD protein (Garver *et al.* 2012), suggesting that novel mechanisms may play a role in REL2 signaling. Our observation of relaxed purifying selection on Imd pathway genes could be consistent with IMD-independent regulation of REL2 signaling in *A. gambiae*.

Evolution in lineage-specific genes or gene families may reflect adaptation to novel pathogens. We found that *STAT-B,* a retrotransposed duplicate of the STAT transcription factor *STAT-A,* which is specific to the Anopheles lineage, shows a strong signature of adaptive evolution, while the ancestral copy *STAT-A* is highly conserved. This pattern of selective constraint in *STAT-A* and rapid evolution in *STAT-B* is consistent with a model of neofunctionalization after duplication (Hahn 2009, Lynch and Force 2000). There are seven members of the STAT family in vertebrates, of which different copies have adopted specialized roles in the endocrine or immune system. Interestingly, the vertebrate STAT genes, which are predominantly involved in the endocrine system, evolve more slowly than those that are predominantly involved in immunity (Gorissen *et al.* 2011). The *A. gambiae* JAK-STAT pathway plays a role in regulating Nitric Oxide Synthase in response to bacterial and *Plasmodium* infections, and may additionally regulate TEP1 expression during late-phase *Plasmodium* infections (Gupta *et al.* 2009). A recent study reported that STAT-B, along with various other detoxification and immune genes, is expressed at higher levels in *A. coluzzii* larvae than in *A. gambiae* larvae (Cassone *et al.* 2014), presumably reflecting the shift to the more biotically and abiotically complex larval habitats preferred by *A. coluzzii,* and suggesting that adaptive evolution in *STAT-B* could be important to adaptation to novel ecological or pathogen environments.

Balancing selection can act to maintain functionally important polymorphisms in a population. Genes involved in host-pathogen interactions are considered likely targets of balancing selection, although instances of balancing selection are rarely observed (Leffler *et al.* 2013). We found two distinct haplotype clades in the 5' end of *CTLMA2*, with 11 fixed SNPs between the clades that result in three amino acid changes near the predicted signal cleavage site.

We identified both clades in previously published *A. gambiae* and *A. arabiensis* sequences. Patterns of divergence within and between the *CTLMA2* haplotype clades suggest that they arose before the species divergence and have been adaptively maintained as a balanced polymorphism in all three species, although further study will be required to determine what functional effect the amino acid variants between the two clades might have. CTLMA2 is primarily secreted into the hemolymph in the form of a heterodimer with CTL4 (Schnitiger *et al.* 2009), and since most of the amino acid substitutions are located in the signal peptide, the different haplotypes might lead to changes in peptide secretion. CTLs are a family of carbohydrate binding proteins that are involved in the immune response through pathogen recognition and as modulators of melanization. The CTL4/CTLMA2 heterodimer is required for successful melanization of Gram-negative bacteria (Schnitiger *et al.* 2009), so it is tempting to speculate that the amino acid variants in CTLMA2 might change the configuration of the heterodimer to create a unique carbohydrate recognition profile. In contrast to the protective role against bacterial pathogens, the CTL4/CTLMA2 heterodimer prevents melanization of rodent malaria (Osta *et al.* 2004), suggesting that this heterodimer has conflicting pleiotropic functions in the mosquito immune response that might contribute to the evolutionary pattern observed in *CTLMA2*.

SUMMARY

In this study, we used population genetic analysis to understand how natural selection operates on the *A. coluzzii* immune system. We found evidence of two distinct haplotypes in the *Anopheles*-specific C-type lectin CTLMA2 which have been adaptively maintained as a balanced polymorphism in *A. coluzzii* and the sister species *A. gambiae and A. coluzzii*. We also found strong evidence of adaptive evolution in the *Anopheles*-specific JAK-STAT transcription factor

STAT-B, consistent with neofunctionalization after duplication. In contrast to these *Anopheles*-specific immune genes, we found no evidence of adaptive evolution in genes involved in the canonical immune signal transduction pathways in *A. coluzzii*. As a group, genes in the Imd pathway exhibit patterns of elevated amino acid diversity and accelerated rates of protein evolution, consistent with relaxed purifying selection, relative to non-immune control loci. Taken together, these results suggest that host-pathogen interactions involving novel or lineage-specific molecular mechanisms likely play a larger role than canonical immune pathways in adaptively evolving resistance to infection in *A. coluzzii*.

Acknowledgements

We thank Mark Jandricic for his help collecting the data. We also thank Rob Unkless and Moria Chambers for helpful discussion and comments on the manuscript.

REFERENCES

Bahia, A. C., M. S. Kubota, A. J. Tempone, H. R. C. Arau´ jo, B. A. M. Guedes *et al.*, 2011 The JAK-STAT Pathway Controls *Plasmodium vivax* Load in Early Stages of *Anopheles aquasalis* Infection. PLoS Negl Trop Dis 5(11): e1317.

Barillas-Mury, C., A. Charlesworth, I. Gross, A. Richman, J. A. Hoffmann *et al*., 1996 Immune factor Gambif1, a new rel family member from the human malaria vector, *Anopheles gambiae*. EMBO J. 15: 4691–4701.

Bargielowski I. and J. C. Koella, 2009 A possible mechanism for the suppression of *Plasmodium berghei* development in the mosquito *Anopheles gambiae* by the microsporidian *Vavraia culicis*. PLoS ONE 2009, 4:e4676.

Barillas-Mury, C., Y. S. Han, D. Seeley, and F. C. Kafatos, 1999 *Anopheles gambiae* Ag-STAT, a new insect member of the STAT family, is activated in response to bacterial infection. EMBO J. 18, 959–967.

Begun, D. J., and P. Whitley, 2000 Adaptive evolution of *Relish*, a Drosophila NF-kB/IkB protein. Genetics 154**:** 1231–1238.

Bulmer, M. S. and R. H. Crozier, 2006 Variation in positive selection in termite GNBPs and Relish. Mol. Biol. Evol. 23, 317–326.

Cassone, B. J., C. Kamdem, C. Cheng, J. C. Tan, M. W. Hahn, *et al.*, 2014 Gene expression divergence between malaria vector sibling species *Anopheles gambiae* and *An. coluzzii* from rural and urban Yaounde Cameroon. *Molecular Ecology*. 23:2242-2259.

Coetzee, M., R. H. Hunt, R. Wilkerson, A. Della Torre, M. B. Coulibaly, and N.J. Besansky, 2013 *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. Zootaxa, 3619, 246–274.

Cohuet, A., S. Krishnakumar, F. Simard, I. Morlais, A. Koutsos *et al*., 2008 SNP discovery and molecular evolution in *Anopheles gambiae*, with special emphasis on innate immune system. BMC Genomics 9: 227.

Collins, F. H., R. K. Sakai, K. D. Vernick, S. Paskewitz, D. C. Seeley *et al*., 1986 Genetic selection of a *Plasmodium*-refractory strain of the malaria vector *Anopheles gambiae*. Science 234: 607–610.

Favia, G., A. Lanfrancotti, L. Spanos, I. Siden-Kiamos, and C. Louis, 2001 Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of *Anopheles gambiae* s.s. Insect Mol. Biol. 10: 19–23.

Frolet C., M. Thoma, S. Blandin, J. A. Hoffmann, and E. A. Levashina, 2006 Boosting NF-kappaB-dependent basal immunity of *Anopheles gambiae* aborts development of *Plasmodium berghei*. Immunity 25: 677–685.

Garver L. S., Y. Dong Y, and G. Dimopoulos, 2009 Caspar controls resistance to *Plasmodium falciparum* in diverse anopheline species. PLoS Pathog 5:e1000335.

Garver L. S., A. C. Bahia, S. Das, J. A. Souza-Neto, J. Shiao *et al*., 2012 *Anopheles* Imd pathway factors and effectors in infection intensity-dependent anti-Plasmodium action. PLoS Pathog 8: e1002737.

Gorissen M., E. de Vrieze, G. Flik, and M. O. Huising, 2011 STAT genes display differential evolutionary rates that correlate with their roles in the endocrine and immune system. J Endocrinol 209: 175–184.

Gupta L., A. Molina-Cruz, S. Kumar, J. Rodrigues, R. Dixit *et al.*, 2009 The STAT pathway mediates late-phase immunity against *Plasmodium* in the mosquito *Anopheles gambiae*. Cell Host Microbe 5: 498–507.

Hahn, M.W. , 2009 Distinguishing among evolutionary models for the maintenance of gene duplicates. Journal of Heredity. 100:605-617.

Hudson, R. R., M. Kreitman, and M. Aguade, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–159.

Hudson, R. R., A. G. Sáez, and F. J. Ayala, 1997 DNA variation at the Sod locus of Drosophila melanogaster: an unfolding story of natural selection. Proc. Natl. Acad. Sci. 94: 7725–7729.

Krzywinski, J., O. G. Grushko, and N. J. Besansky, 2006 **Analysis of the complete mitochondrial DNA from** *Anopheles funestus***: an improved dipteran mitochondrial genome annotation and a temporal dimension of mosquito evolution.** *Molecular Phylogenetics and Evolution*, **39:**417-423.

Leffler, E. M., Z. Gao, S. Pfeifer, L. Segurel, A. Auton, *et al*., 2013 Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 339: 1578-1582.

Librado, P. and J. Rozas, 2009 DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452

McTaggart, S. J., D. J. Obbard, C. Conlon, and T. J. Little, 2012 Immune genes undergo more adaptive evolution than non-immune system genes in *Daphnia pulex*. BMC Evolutionary Biology 12:63.

McDonald, J. H. and M. Kreitman, 1991 Adaptive evolution at the Adh locus in Drosophila. Nature 351: 652-654.

Meister, S., S. M. Kanzok, X. L. Zheng, C. Luna, T. R. Li *et al*., 2005 Immune signaling pathways regulating bacterial and malaria parasite infection of the mosquito *Anopheles gambiae*. Proc. Natl. Acad. Sci. USA 102: 11420–11425.

Meister, S., B. Agianian, F. Turlure, A. Relogio, I. Morlais *et al*., 2009 *Anopheles gambiae* PGRPLC-mediated defense against bacteria modulates infections with malaria parasites. PLoS Pathog. 5: e1000542.

Mitri, C., J. C. Jacques, I. Thiery,M.M. Riehle, J. Xu *et al.,* 2009 Fine pathogen discrimination within the APL1 gene family protects *Anopheles gambiae* against human and rodent malaria species. PLoS Pathog. 5: e1000576.

Muspratt J., 1946 On Coelomomyces fungi causing high mortality of *Anopheles gambiae* larvae in Rhodesia. Ann Trop Med Parasitol 40:10–17.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton *et al*., 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3: e170.

Nei, M. 1987 Molecular evolutionary genetics. New York, NY: Columbia University Press.

Obbard, D. J., F. M. Jiggins, D. L. Halligan, and T. J. Little, 2006 Natural Selection Drives Extremely Rapid Evolution in Antiviral RNAi Genes. Current Biology 16:580.

Obbard, D. J., Y. M. Linton, F. M. Jiggins, G. Yan, and T. J. Little, 2007 Population genetics of Plasmodium resistance genes in Anopheles gambiae: no evidence for strong selection. Mol. Ecol. 16: 3497–3510.

Obbard, D. J., D. M. Callister, F. M. Jiggins, D. C. Soares, G. Yan *et al*., 2008 The evolution of TEP1, an exceptionally polymorphic immunity gene in *Anopheles gambiae*. BMC Evol. Biol. 8: 274.

Obbard, D. J., J. J. Welch, and T. J. Little, 2009a Inferring selection in the *Anopheles gambiae* species complex: an example from immune-related serine protease inhibitors. Malar. J. 8: 117.

Obbard, D. J., J. J. Welch, K. W. Kim, and F. M. Jiggins, 2009b Quantifying Adaptive Evolution in the Drosophila Immune System. PLoS Genet 5(10): e1000698.

O'Loughlin, S. M., S. Magesa, C. Mbogo, F. Mosha, J. Midega *et al.*, 2014 Genomic analyses of three malaria vectors reveals extensive shared polymorphism but contrasting population histories. Mol Biol Evol. 2014;31(4):889–902

Osta, M. A., G. K. Christophides, and F. C. Kafatos, 2004 Effects of mosquito genes on Plasmodium development. Science 303: 2030–2032.

Parmakelis, A., M. A. Slotman, J. C. Marshall, P. H. Awono-Ambene, C. Antonio-Nkondjio *et al.*, 2008 The molecular evolution of four anti-malarial immune genes in the *Anopheles gambiae* species complex. BMC Evol Biol 8: 79.

Povelones, M., R. M. Waterhouse, F. C. Kafatos, and G. K. Christophides, 2009 Leucine-rich repeat protein complex activates mosquito complement in defense against Plasmodium parasites. Science 324: 258–261.

Pritchard, J. K., J. K. Pickrell, and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. 20: R208–R215.

Riehle, M. M., J. Xu, B. P. Lazzaro, S. M. Rottschaefer, B. Coulibaly *et al*., 2008 *Anopheles gambiae* APL1 is a family of variable LRR proteins required for Rel1-mediated protection from the malaria parasite, Plasmodium berghei. PLoS ONE 3:e3672.

Rottschaefer, S. M., M. M. Riehle, B. Coulibaly, M. Sacko, O. Niaré *et al*., 2011 Exceptional diversity, maintenance of polymorphism, and recent directional selection on the APL1 malaria resistance genes of *Anopheles gambiae*. PLoS Biol. 9: e1000600.

Sackton, T. B., B. P. Lazzaro, T. A. Schlenke, J. D. Evans, D. Hultmark *et al*., 2007 Dynamic evolution of the innate immune system in Drosophila. Nat. Genet. 39: 1461–1468.

Smith, N. G. C., and A. Eyre-Walker, 2002 Adaptive protein evolution in Drosophila. Nature 415:1022-1024.

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

Schnitger, A.K., H. Yassine, F. C. Kafatos, and M. A. Osta, 2009 Two C-type lectins cooperate to defend *Anopheles gambiae* against gram-negative bacteria. *J Biol Chem* **284:** 17616–17624.

Scott J. A., W. G. Brogdon, and F. H. Collins, 1993 Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. *Am J Trop Med Hyg* 1993, 49:520-529.

Slotman, M. A., A. Parmakelis, J. C. Marshall, P. H. Awono-Ambene, C. Antonio-Nkondjo *et a*l., 2007 Patterns of selection in anti-malarial immune genes in malaria vectors: evidence for adaptive evolution in LRIM1 in Anopheles arabiensis. PLoS ONE 2: e793.

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

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei *et al*., 2011 MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28: 2731–2739.

Viljakainen, L., J. D. Evans, M. Hasselmann, O. Rueppell, S. Tingek *et al.*, 2009 Rapid evolution of immune proteins in social insects. Mol. Biol. Evol. 26, 1791–1801.

Wang, Y., T. M. Gilbreath III, P. Kukutla, G. Yan, and J. Xu, 2011 Dynamic gut microbiome across life history of the malaria mosquito Anopheles gambiae in Kenya. PLoS ONE 2011, 6:e24767.

Waterhouse, R. M., E. V. Kriventseva, S. Meister, Z. Xi, K. S. Alvarez *et al*., 2007 Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. Science 316: 1738–1743.

Welch, J.J., 2006 Estimating the genomewide rate of adaptive protein evolution in Drosophila. Genetics 173: 821-837.

White, B. J., M. K. Lawniczak, C. Cheng, M. B. Coulibaly, M. D. Wilson *et al*., 2011 Adaptive divergence between incipient species of Anopheles gambiae increases resistance to Plasmodium. Proc. Natl. Acad. Sci. USA 108: 244–249.

Wright S.I. and B. Charlesworth, 2004 The HKA test revisited: a maximum likelihood-ratio test of the standard neutral model. Genetics 168: 1071–1076.

Table 1. Average pairwise divergence in immune and control groups

	n[a]	K_{total}[b]	K_S[c]	K_A[d]	K_A/K_S[e]
All Data					
immune	20	0.0335	0.0652	0.0120	0.1747
control	17	0.0319	0.0545	0.0043	0.0772
		U = 149, p = 0.537	U =117.5, p =0.113	U = 64.5, p = 0.001	U = 62.5, p = 0.001
IMD pathway					
immune	8	0.0336	0.0613	0.0121	0.1968
control	8	0.0302	0.0526	0.0041	0.0675
		U = 24, p = 0.442	U = 21, p = 0.279	U = 8, p = 0.010	U = 5, p = 0.003
TOLL pathway					
immune	7	0.0332	0.0713	0.0087	0.1090
control	6	0.0335	0.0565	0.0032	0.0555
		U = 22, p = 0.945	U = 14, p = 0.366	U = 8, p = 0.073	U = 7.5, p = 0.063
Exclude IMD pathway					
immune	12	0.0335	0.0679	0.0119	0.1600
control	10	0.0322	0.0557	0.0043	0.0835
		U = 58, p = 0.923	U = 44, p = 0.314	U = 28.5, p = 0.041	U = 29.5, p = 0.048

[a]number of genes considered

[b]Average per gene divergence at all sites between *A. coluzzii* and *A. merus*

[c]Average per gene silent divergence between *A. coluzzii* and *A. merus*

[d]Average per gene replacement divergence between *A. coluzzii* and *A. merus*

[e]Average per gene K_A/K_S ratio between *A. coluzzii* and *A. merus*

Table 2. Multilocus MK-test model comparison

Model	Description	Par	log($\mathcal{L}$)	2Δlog($\mathcal{L}$)	χ2 p-value	AICc[a]	Akaike weight[b]	α_a[c]	α_b[d]
All loci									
M0	α = 0	38	-615.83			1335.9	0.806	[0]	[0]
M1	α ~ (all loci)	39	-615.51	0.64	0.4237	1339.0	0.169	[0.07]	[0.07]
M2	α ~ (control, immune)	40	-615.48	0.06	0.8065	1342.8	0.025	0.04	0.09
Imd Pathway									
M0	α = 0	18	-235.44			522.1	0.003	[0]	[0]
M1	α ~ (all loci)	19	-228.01	14.86	0.0001	511.3	0.695	[-0.64]	[-0.64]
M2	α ~ (control, immune)	20	-226.72	2.58	0.1069	513.0	0.302	-0.20	-0.91
Toll Pathway									
M0	α = 0	14	-181.22			403.2	0.305	[0]	[0]
M1	α ~ (all loci)	15	-178.55	5.34	0.0208	402.1	0.520	[-0.5]	[-0.5]
M2	α ~ (control, immune)	16	-177.37	2.36	0.1245	404.3	0.175	-1.09	-0.24

[a]The Akaike information criterion corrected for sample size

[b]The likelihood of the model, given the relative support for each of the models tested.

[c]Estimate of the proportion of adaptive substitutions in the control genes. Square brackets indicate where α is constrained by the model.

[d]Estimate of the proportion of adaptive substitutions in the immune genes. Square brackets indicate where α is constrained by the model.

Table 3. Average pairwise genetic diversity in immune and control groups

	n[a]	π_{total}[b]	π_s[c]	π_a[d]	*D*[e]
All data					
immune	19	0.0140	0.0283	0.0048	-0.434
control	19	0.0127	0.0252	0.0016	-0.764
		V = 102, $p = 0.798$	V = 128, $p = 0.196$	V = 172, $p = 0.001$	V = 141, $p = 0.066$
IMD pathway					
immune	8	0.0182	0.0385	0.0066	-0.512
control	8	0.0134	0.0297	0.0018	-0.601
		V = 27, $p = 0.25$	V = 34, $p = 0.023$	V = 36, $p = 0.008$	V = 20, $p = 0.844$
TOLL pathway					
immune	6	0.0116	0.0231	0.0031	-0.353
control	6	0.0141	0.0239	0.0014	-0.834
		V = 4, $p = 0.219$	V = 9, $p = 0.844$	V = 17, $p = 0.219$	V = 19, $p = 0.094$

[a]number of genes considered, 18-20 alleles sampled per gene

[b]Average per gene genetic diversity calculated for all sites

[c]Average per gene genetic diversity calculated for synonymous sites

[d]Average per gene genetic diversity calculated for nonsynonymous sites

[e]Average per gene Tajima's D calculated using silent sites.

Table 4. Population genetic statistics for CTLMA2 haplotype clades and surrounding loci

	n[a]	S[b]	π[c]	D[d]
CTLMA2 all	20	55	0.022	1.254
clade A	10	30	0.010	-0.465
clade B	10	25	0.007	-1.737
CTL4	20	44	0.013	-0.263
AGAP005327	19	54	0.010	-1.107

[a]number of alleles sampled

[b]number of segregating sites

[c]Average number of pairwise differences per site

[d]Tajima's D calculated using silent sites

Table 5. D_{xy} among species within and between CTLMA2 haplotype clades and in two nearby loci.

a.

A. coluzzii	*A. arabiensis*[a]	D_{xy}
clade A	clade A	0.009
clade B	clade B	0.011
clade A	clade B	0.034
clade B	clade A	0.031
AGAP005540[a]	AGAP005540[a]	0.016
APL2[a]	APL2[a]	0.017

b.

A. gambiae[a]	*A. arabiensis*[a]	D_{xy}
clade A	clade A	0.008
clade B	clade B	0.012
clade A	clade B	0.035
clade B	clade A	0.030
AGAP005540[a]	AGAP005540[a]	0.017
APL2[a]	APL2[a]	0.018

[a]Sequences from published data (Obbard *et al.* 2007)

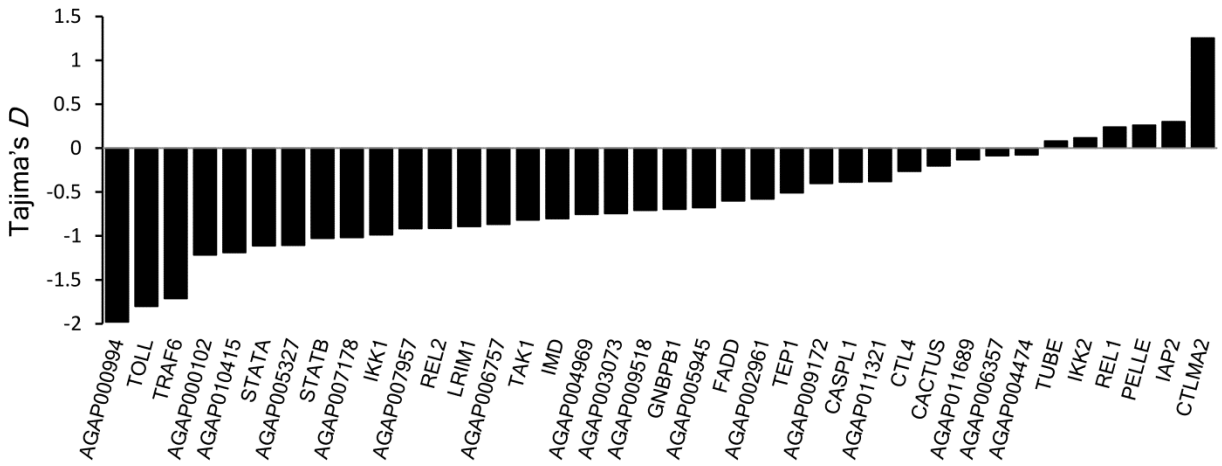


Figure1. Distribution of Tajima's *D* across all 37 genes. Tajima's *D* was calculated for each gene using silent sites and plotted as a histogram showing deviation from neutral expectations. Loci are ordered based on the value of *D* in order to draw attention to the contrast between CTLMA2 and the rest of the dataset.

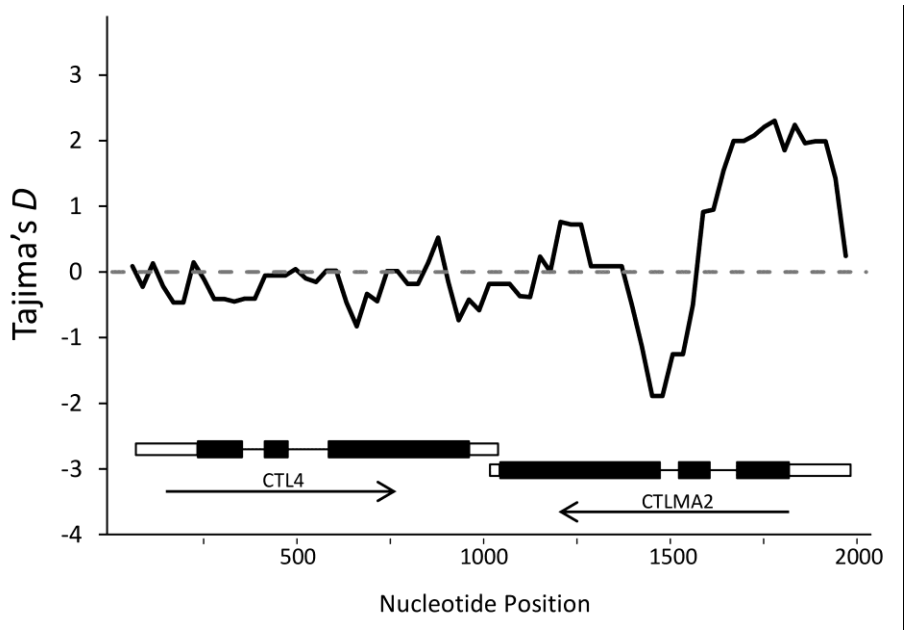


Figure 2. Sliding window analysis of Tajima's *D* in CTL4 and CTLMA2**.** CTL4 and CTLMA2 are located directly adjacent on the chromosome and were PCR amplified and cloned as a single fragment. Tajima's *D* was calculated using silent sites in a sliding window along the entire sequenced region, using a 200 bp window with a 25 bp step size. The dashed line indicates the expected value of *D* under a neutral equilibrium model. The schematic below the plot shows the exon structure and direction of transcription for CTL4 and CTLMA2. Positive values of *D* in the 5' region of CTLMA2 indicate an excess of intermediate frequency polymorphism in this region.

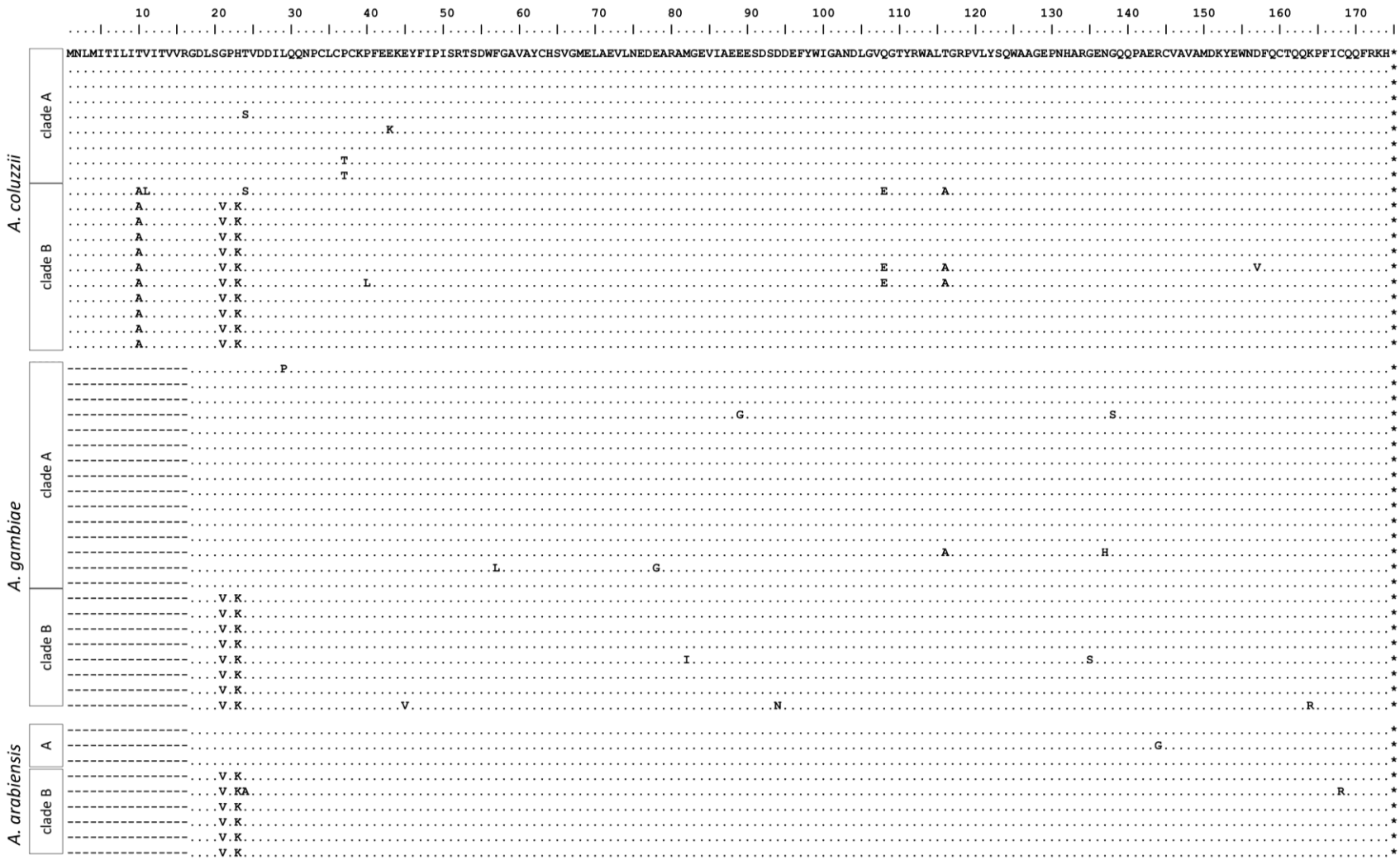


Figure 3. Amino acid alignment of CTLMA2 sequences of *A. coluzzii* collected in Burkina Faso, along with previously published sequences from *A. gambiae* and *A. arabiensis* collected in Kenya. Dashes note unavailable sequence as we sequenced a different amplicon than was available for *A. gambiae* or *A. arabiensis*.

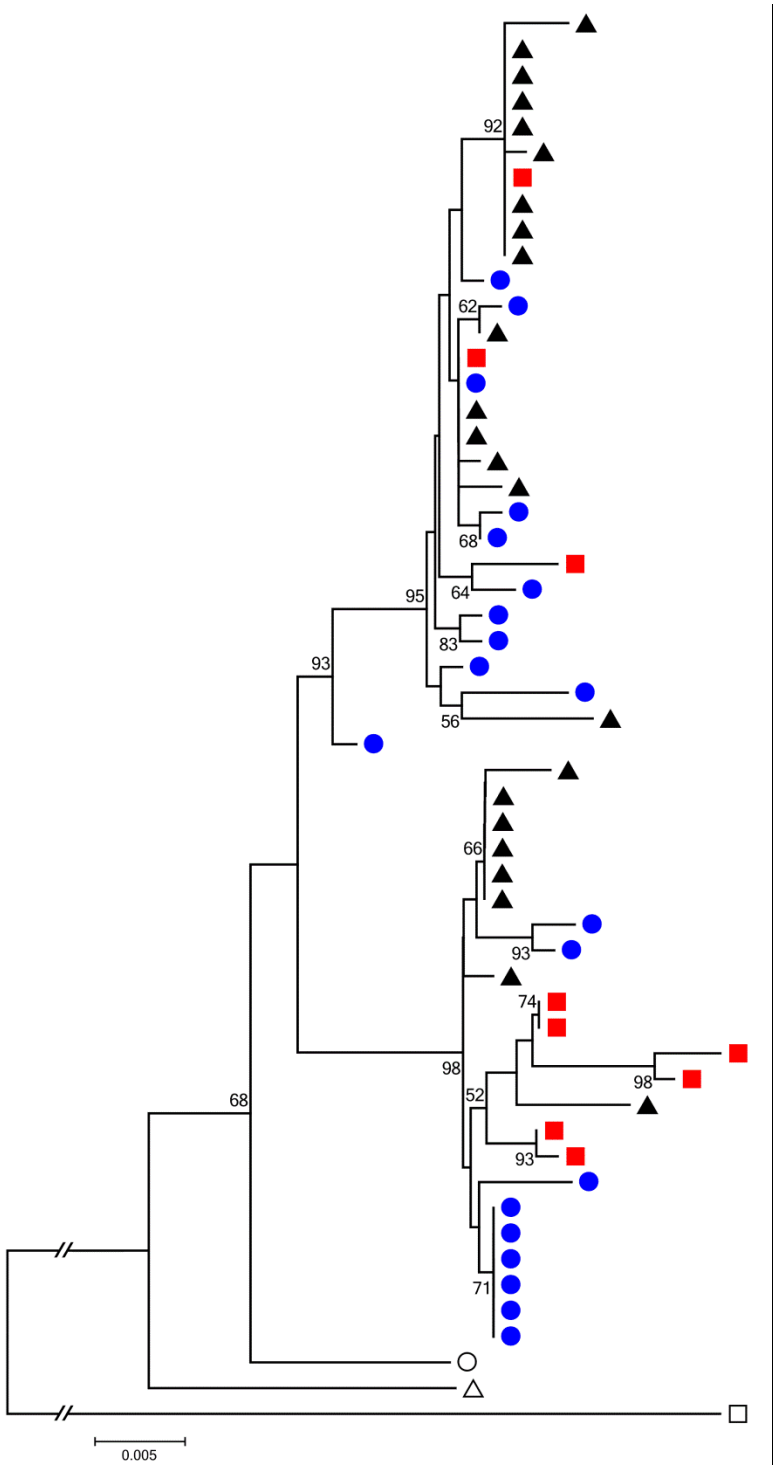


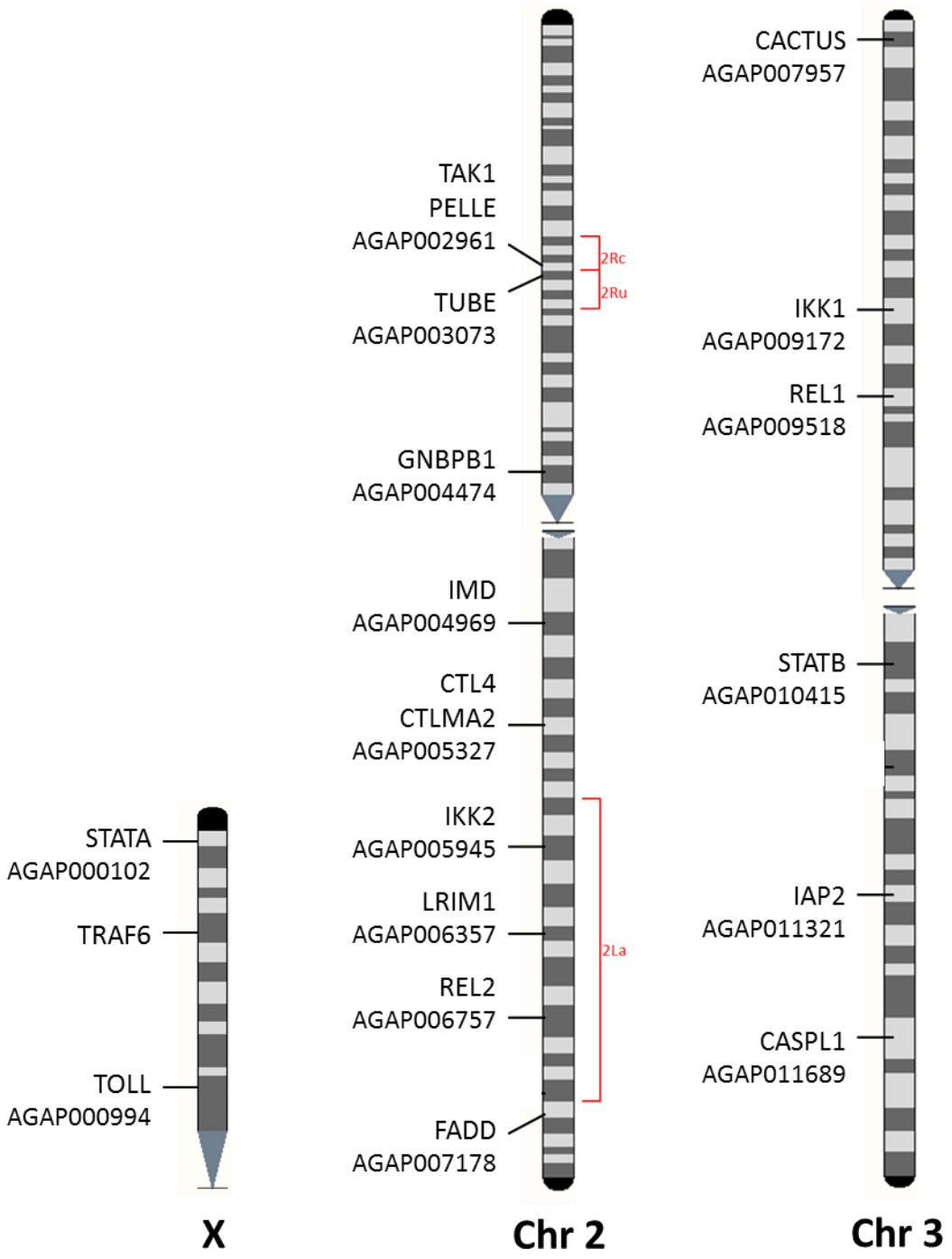
Figure 4. Neighbor-joining tree of *CTLMA2* sequences from *A. coluzzii* (blue circles), *A. gambiae* (black triangles), *A. arabiensis* (red squares), *A. quadriannulatus* (open circle), *A. merus* (open triangle), and *A. christyi* (open square). Bootstrap support was determined using 1000 bootstrap replicates, and nodes with >50% bootstrap support are shown.

Supplementary Table S1. Population genetic statistics for all sampled loci

locus	N^a	Sites^b	S^c	$\pi_{total}{}^d$	$\pi_s{}^e$	$\pi_a{}^f$	D^g	$K_{total}{}^h$	$K_S{}^i$	$K_A{}^j$	$K_A/K_S{}^k$
CACTUS	20	2660	133	0.0132	0.0215	0.0021	-0.201	0.0247	0.047	0.0016	0.033
CASPL1	19	1424	133	0.024	0.0608	0.0116	-0.385	0.0269	0.061	0.0134	0.212
CTL4	20	882	44	0.0125	0.0311	0.0039	-0.263	0.0326	0.06	0.0147	0.237
CTLMA2	20	891	55	0.0217	0.0258	0.0084	1.254	0.047	0.1021	0.0204	0.188
FADD	20	980	124	0.0294	0.0458	0.0149	-0.601	0.0447	0.0663	0.0248	0.363
GNBPB1	20	1782	79	0.0102	0.0202	0.0013	-0.697	0.0334	0.0773	0.0064	0.078
IAP2	20	873	78	0.0253	0.0451	0.006	0.3	0.0467	0.0783	0.0131	0.16
IKK1	20	2681	170	0.0141	0.0336	0.0051	-0.987	0.0294	0.0728	0.0077	0.102
IKK2	20	2035	98	0.0119	0.0373	0.0043	0.116	0.037	0.0795	0.0082	0.098
IMD	20	1270	65	0.0116	0.0195	0.005	-0.804	0.0328	0.0481	0.0215	0.439
LRIM1	20	1636	76	0.0106	0.0265	0.0055	-0.892	0.0302	0.0638	0.0179	0.272
PELLE	20	1986	87	0.013	0.0325	0.0039	0.261	0.0342	0.0751	0.0096	0.125
REL1	19	6802	363	0.016	0.0226	0.0024	0.238	0.03	0.0493	0.0055	0.108
REL2	19	5075	311	0.013	0.0288	0.0038	-0.912	0.0151	0.0338	0.004	0.116
STAT-A	20	2740	55	0.0042	0.0064	0	-1.111	0.0221	0.0268	0	0
STAT-B	19	2123	15	0.0018	0.0021	0.0015	-1.025	0.0377	0.0627	0.0295	0.46
TAK1	20	5409	395	0.0163	0.037	0.0023	-0.819	0.0359	0.0506	0.0044	0.084
TOLL	20	3595	40	0.0017	0.0054	0.0004	-1.802	0.0225	0.066	0.0065	0.094
TRAF6	20	1814	53	0.0048	0.0079	0.0023	-1.713	0.0574	0.1132	0.0181	0.149
TUBE	18	2285	110	0.0156	0.0365	0.0087	0.081	0.0305	0.0711	0.0131	0.176
AGAP000102	20	3500	131	0.0078	0.0104	0.0034	-1.218	0.0498	0.0388	0.0145	0.368

AGAP000994	20	2519	47	0.0027	0.0049	0.0001	-1.98	0.0288	0.0477	0.0019	0.039
AGAP002961	20	4700	237	0.0116	0.0315	0.0017	-0.58	0.0208	0.0504	0.0032	0.062
AGAP003073	20	1768	115	0.0152	0.0241	0.0003	-0.743	0.0393	0.0895	0.0033	0.035
AGAP004474	19	2363	150	0.0181	0.0282	0.0013	-0.077	0.0442	0.0579	0.0031	0.052
AGAP004969	20	1653	64	0.009	0.0209	0.0034	-0.753	0.022	0.0594	0.0053	0.086
AGAP005327	19	1088	54	0.0102	0.0208	0.0007	-1.107	0.0099	0.0117	0.0003	0.029
AGAP005945	19	2553	183	0.0169	0.0367	0.0016	-0.678	0.0487	0.0562	0.0027	0.046
AGAP006357	20	3166	213	0.0183	0.0336	0.0013	-0.086	0.034	0.0728	0.0065	0.085
AGAP006757	20	5014	409	0.0183	0.0204	0.0004	-0.868	0.0489	0.0406	0.0002	0.005
AGAP007178	20	1699	84	0.0105	0.0253	0.0017	-1.017	0.0203	0.0424	0.0032	0.074
AGAP007957	20	4364	317	0.0158	0.0293	0.0029	-0.915	0.0287	0.0666	0.0065	0.094
AGAP009172	19	2826	125	0.0115	0.0242	0.0006	-0.4	0.0269	0.0492	0.0021	0.042
AGAP009518	19	4695	411	0.021	0.0253	0.0021	-0.709	0.039	0.0267	0.0014	0.051
AGAP010415	20	1275	25	0.0035	0.0117	0.0011	-1.188	0.0271	0.0947	0.0021	0.02
AGAP011321	20	1566	89	0.0142	0.0376	0.005	-0.381	0.0318	0.0809	0.0134	0.158
AGAP011689	20	1684	96	0.0153	0.041	0.0003	-0.132	0.022	0.0414	0.0029	0.067

[a]Number of *A. coluzzii* alleles sampled
[b]Number of base pairs sequenced per allele
[c]Number of segregating sites
[d]Pairwise genetic diversity calculated for all sites
[e]Pairwise genetic diversity calculated for synonymous sites
[f]Pairwise genetic diversity calculated for nonsynonymous sites
[g]Tajima's D calculated using silent sites.
[h]Pairwise divergence at all sites between *A. coluzzii* and *A. merus*
[i]Pairwise silent divergence between *A. coluzzii* and *A. merus*
[j]Pairwise replacement divergence between *A. coluzzii* and *A. merus*
[k]K_A/K_S ratio between *A. coluzzii* and *A. merus*



Supplementary Figure S1. Approximate chromosomal locations of loci sampled. AGAP identifiers are provided for all non-immune control loci. Control loci are located within 40-100KB of their "matched" controls. Positions of chromosomal inversions are shown in red.