POPULATION GENETIC AND GENOMIC ANALYSIS OF DEMOGRAPHY AND NATURAL SELECTION IN *ANOPHELES* MALARIA VECTORS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Jacob Crawford

August 2012

POPULATION GENETIC AND GENOMIC ANALYSIS OF DEMOGRAPHY AND NATURAL SELECTION IN *ANOPHELES* MALARIA VECTORS

Jacob Crawford, Ph.D.

Cornell University, 2012

Human malaria parasites are vectored primarily by three mosquito species of the genus *Anopheles*, and new technologies and strategies to control disease transmission by targeting the mosquito vector have been proposed or are in development. The success of these strategies depends on knowledge of genetic variation at relevant loci in targeted mosquito populations. However, we know very little about the selective forces shaping genetic variation of proteins involved in the mosquito-parasite interaction that could potentially be developed for intervention. Genetic variation in populations is largely shaped by natural selection, demography, and genetic drift. I used population genetic approaches to study the historical demographic and selective events of multiple populations of one of the primary vectors, *Anopheles gambiae*. Statistical inference through comparisons of population samples simulated under a variety of demographic models to genomically distributed empirical re-sequencing data revealed evidence for both historical population growth and migration in the M and S molecular forms, two insipient species of *A. gambiae.* Importantly, significantly different demographic histories were inferred for the two molecular forms. Both forms show evidence of population growth that predated the agricultural revolution, which has been suggested as a cause of population growth in this system. Novel vector-based disease intervention strategies are largely based on two types of mosquito proteins: non-immune proteins that physically interact with the malaria parasite during parasite development and immune genes involved in the anti-malaria immune response. To test whether two transmission-blocking vaccine candidate proteins saglin and laminin are adaptively evolving, I

sampled alleles from wild caught populations of the M and S molecular forms and analyzed both intraspecific and interspecific patterns of variation using population genetic tests. Neither protein showed significant evidence for positive selection in these populations. On the contrary, these proteins are evolving neutrally, one protein is evolving under particularly strong purifying selection, suggesting that it may be relatively reliable vaccine target. Immune genes show different patterns of evolution, however. I sampled alleles at 28 candidate immune genes in wild caught samples of three populations of *A. gambiae*: the M and S molecular forms and the recently discovered and genetically distinct GOUNDRY population. Population genetic neutrality tests revealed striking divisions of putative selection signals among these strata, with only 1 of the 11 loci that rejected the neutral model being shared among the populations. Interestingly, the S molecular form showed no evidence of positive selection at any loci. Putative positive selection was identified at loci that encode immune proteins from a variety of functional classes. When considered in the context of differences between the larval ecologies of these populations, these results point to a complex division of selection regimes among these strata of *A. gambiae*, probably related to larval pathogens encountered during niche expansion in the M molecular form and GOUNDRY.

Recent advancements in DNA sequencing technology make the prospects of whole-genome sequencing-based population genomic studies likely for *Anopheles* mosquitoes in the near future, but this so-called 'next-generation sequencing' (NGS) is complicated by relatively high sequencing error rates and subsequent uncertainty in genotype inference. To explore potential biases and statistical power of NGS-based population genomic studies, I used a simulation approach to identify biases introduced into demographic analysis, tests for positive natural selection, and analysis of genetic differentiation between populations. At relatively shallow sequencing depth (4x),

demographic inference and estimates of genetic differentiation are systematically biased, and positive selection can only be reliably detected if it is strong and recent. Many of the biases are mitigated and statistical power improved when sequencing depth is increased to even 8x, and 15x recovers the population genetic signals almost completely. This analysis provides insight into the biases that can be expected in NGS-based studies, and provides parameter values that can be used to inform the design of future NGS-based studies.

Anopheles funestus is a primary vector of malaria, but little is known about the basic biology and few genetic resources are available for this species. I used next-generation Illumina RNA-sequencing technology to sequence and assemble the transcriptome of *A. funestus de novo*, generating over 15,000 putative transcripts. I annotated the transcriptome through comparisons to the sequenced genomes of other Dipteran insects and functional domain databases, identified over 300 putative immune genes, and mapped the raw sequence reads back to the transcriptome and identified over 300,000 potential genetic variants. These data provide the largest and most exhaustive sequence and bioinformatic resource as well as putative genetic variants that can be developed for population genetic or mapping studies for this system

BIOGRAPHICAL SKETCH

I was born around Christmas to Leslie and Douglas Crawford in Leesburg, VA in 1980, joining my brother Jeffrey who had just turned two years old. We lived on a non-working farm in West Virginia till the age of 8 when we moved to Baltimore, MD so that my mother could pursue at Masters Degree in movement therapy at Goucher College. To avoid the pitfalls of the crumbling Baltimore City school system, my brother and I were switched from the local public school to the Waldorf School of Baltimore, where we enjoyed a wonderful blend of artistic and academic pursuits and spent lots of time outdoors, particularly in the woods of the Cylburn Arboretum next door. For high school, we returned to the public school system in Baltimore County to attend Towson High School. My interest in the natural world surely began early with my experiences as a young person, but several classes in high school with Mr. Gosnell and Mr. Lear provided my first taste of more formal biological education that I can credit as the beginning of a long pursuit of scientific knowledge. Despite a clear interest in biology, I enrolled in the Commercial Music degree program at The University of Memphis in Tennessee motivated to pursue my passion for music. This diversion lasted only a year before I transferred to Georgetown University in Washington, D.C. and switched my major to Biology. As a relatively small Biology department, my options for research were somewhat limited, but I found a home in the Plant-Insect Interactions lab where I conducted senior thesis research studying the role of olfaction and induced plant chemicals in predatory *Polistes* paper wasp learning under they supervision of Dr. Martha Weiss. While at Georgetown, I also had my first exposure to Cornell University through the Field Marine Science course I completed on Appledore Island in the Isle of the Shoals off the coast of Maine. Both my thesis research and my experience on Appledore Island cemented my interest in pursuing biological research in a more formal setting. After graduating, I remained at Georgetown and worked as a

research/laboratory technician with Dr. Peter Armbruster. Research in the Armbruster Lab focuses on the rapid ecological adaptation of invasive populations of the Asian Tiger mosquito *Aedes albopictus* using both ecological and molecular techniques, providing me the opportunity to develop molecular biology skills and think about ecology at the molecular level. Interested in learning more about mosquitoes, I accepted a one-year post-baccalaureate fellowship to work on ecological aspects of insipient speciation in *Anopheles gambiae* under the supervision of Dr. Tovi Lehmann at the National Institutes of Health. I found the *Anopheles* system (and mosquitoes in general) to be fascinating and a good system to study the process of adaptation, and liked that research involving this system could be medically relevant. In response, I joined the laboratory of Dr. Brian P. Lazzaro in the Entomology Department at Cornell University with the intention of studying life-history trade-offs at the organismal level in *A. gambiae* to identify trade-offs that limit the anti-malaria immune response in these mosquitoes. However, in my first semester at Cornell I completed Introduction to Population Genetics and shifted my focus to use population genetic tools to study natural variation in populations of *Anopheles* vectors to understand the role of natural selection in shaping the *Anopheles* immune response. In the process, I have learned a great deal about the biology of mosquitoes, theoretical and empirical population genetics, and the process of conducting sound scientific research. I hope to continue conducting population genetic research, in both mosquitoes and other systems, for many years to come to address fundamental questions in evolutionary biology in hopes of applying these findings in medically relevant settings.

ACKNOWLEDGEMENTS

I am deeply grateful to my thesis advisor Dr. Brian P. Lazzaro. From the moment I arrived at Cornell, Brian has been extremely generous with his time, meeting with me every week for many years where he allowed me to explore and stumble and find my way to a viable and valuable set of experiments. Brian has fostered a wonderful blend of molecular and organismal research in the lab that I believe keeps those in both camps thinking more holistically about their work and has shaped the way I approach my work. Outside of the lab, Brian was amazingly supportive when I decided to buy a house, or when I decided to join a band, but most importantly when I became a father. When my daughter Naomi was born, Brian allowed me the flexibility and support to spend invaluable time with her in her first months while we both found our way in this new reality and formed a bond that will shape our lifetime together. For this and all of his academic guidance, I am forever grateful.

I am very grateful to the members of my special committee, Drs. Laura Harrington and Chip Aquadro for providing valuable advice that improved my thesis and provided valuable guidance in making career decisions. I would also like to thank my collaborators that have provided extremely fruitful interactions and made possible several of the experiments described in this thesis. Dr. Kenneth Vernick at Institut Pasteur in Paris and his group have been wonderful to work with and made it possible for me to travel to Ouagadougou, Burkina Faso to conduct with Dr. N'fale Sagnon and his group at Centre National de Recherche et de Formation sur le Paludisme.

I am also very grateful to my family and friends for all of their support and for keeping my life interesting. My parents provided me with an enriching environment growing up that I credit for much of my success as well as unending support during all of my studies as an undergraduate and also during graduate school. They supported my decision first to pursue music in Memphis. Then they were on board with my decision to transfer, and

fully supported my decision to pursue a Ph.D. at Cornell. I could not have accomplished this without them. I would also like to thank my brother Jeffrey, my Baltimore friends, and all my friends from Ithaca who were always around when I needed their help or wanted a break and to have some fun.

Lastly, I am deeply grateful to my loving wife Deborah. I am forever in her debt for her willingness to turn down an opportunity in the Peace Corp and move to a tiny town she had no other reason to visit to be with me and try to make a life and be together. Over the years she has been so patient and flexible around my crazy grad school life, with its bouts of intense stress and work schedule, with grace and managing to keep us both smiling. She has been a light in my life, and sharing this experience with her has made a challenging time in my life wonderful.

TABLE OF CONTENTS

CHAPTER 1:

INTRODUCTION

The human malaria parasite, *Plasmodium falciparum*, is vectored predominantly and most efficiently by only three species of mosquitoes in sub-Saharan Africa: *Anopheles gambiae*, *Anopheles arabiensis*, and *Anopheles funestus (Collins and Paskewitz 1995)*. The parasite is remarkably specific and adapted to these vector species. Many other mosquito species are exposed to infectious human blood meals, sometimes on a daily basis in endemic areas, but the parasite fails to complete development and transmission is blocked (Billingsley and Sinden 1997; Sinden, Alavi, and Raine 2004). Even within the highly permissive species, a large proportion of parasites are killed and potential infections are limited or eliminated completely, owing in part to the deployment of a potent and efficient mosquito innate immune response that limits parasite development (Vaughan, Hensley, and Beier 1994; Luckhart et al. 1998; Gouagna et al. 1998; Han et al. 2000; Frolet et al. 2006; Hillyer, Barreau, and Vernick 2007). Collectively, these features of the malaria system mean that malaria is transmitted through a very small conduit that is dictated by both the parasite adaptation to a small number of mosquito species and the efficacy of the mosquito immune response to parasite infection. An important implication of this fact is that the mosquito immune response (and also the human immune response) is perhaps the most potent and effective mechanism of disease control available, and so should be carefully studied to identify opportunities for manipulation or augmentation that could further reduce disease transmission (Vernick and Waters 2004). Despite an impressive body of work functionally dissecting the mosquito-parasite interaction (reviewed in Yassine and Osta 2010; Cirimotich et al. 2010), less is known about the evolutionary and ecological contexts that have shaped

this interaction and influence the outcome. The research presented in this dissertation focuses on the evolutionary component, but the results also provide some insight into ecological factors affecting the mosquito-malaria system as well.

The host immune response is at the center of host-pathogen interactions, so understanding this system is fundamental. As a result of its crucial role in the conflict between hosts and pathogens, immune system molecules tend to be rapidly evolving, as was first postulated over a century ago (Biffen 1905; Haldane 1949). In *Drosophila*, molecules involved in the innate immune response as a class evolve faster than the genome on average (Sackton et al. 2007; Obbard et al. 2009). But rapid evolution is not a feature of all molecules and sub-classes of molecules involved in the immune response (Sackton et al. 2007; Lazzaro 2008; Waterhouse et al. 2007), making it important to identify which components of the immune system may be evolving in response to pathogenic pressure and to elucidate details and patterns of evolution among the molecules in play. In addition to the immune system, structural and receptor proteins are exploited by invading pathogens including *Plasmodium* and may also play a role in the host-pathogen conflict. Multiple proteins have been identified in *Anopheles* mosquitoes that physically interact with the malaria parasite during tissue invasion or recognition (Brennan et al. 2000; Arrighi et al. 2005; Rodrigues et al. 2012). These physical interactions have been proposed as potential opportunities for intervention and disease control, using transmission-blocking vaccines for example (Dinglasan and Jacobs-Lorena 2008), and understanding the evolutionary history of these host molecules is essential for development and success of such technologies.

One classical model for host-pathogen interactions is characterized by evolutionarily rapid and repeated reciprocal fixations of adaptive genetic variants in both the pathogen and the host in what is often called a host-pathogen co-evolutionary 'arms

race' (Dawkins and Krebs 1979). Over long time scales, rapidly evolving proteins can accumulate multiple amino substitutions, and statistical tests have been developed to detect this pattern through comparisons of the rate of nucleotide substitutions at non-synonymous, amino-acid changing positions relative to the rate of substitution at silent sites (McDonald and Kreitman 1991; R Nielsen and Yang 1998). During the time surrounding a selective event (i.e. shorter time-scale) when an evolutionarily favored genetic variant rises to high frequency in the population or species through the action of positive selection, sometimes even to fixation, patterns of intraspecific linked genetic variation are altered in the genomic region surrounding the selected site. This so-called 'genetic hitchhiking' effect leads to several hallmark signatures, including the local elimination of genetic variation, an enrichment of low and high frequency genetic variants in the derived state, as well as an increase in haplotype structure and homozygosity (Smith and Haigh 1974; Braverman et al. 1995; Przeworski 2002; Sabeti et al. 2002). Statistical tests that detect these signatures have been developed as a means to identify positively selected loci (Tajima 1989; Fay and Wu 2000; Kim and Stephan 2002; Sabeti et al. 2002). However, the signature of genetic hitchhiking is complex and fleeting, detectable most reliably in only certain phases of the selective event, and detecting this signature is further complicated by the confounding effects of demographic shifts, background selection, population structure and variations in mutation and recombination rates (reviewed in Nielsen 2005).

Prime among concerns in *Anopheles* mosquitoes is the highly complex demography and population structure in this system. Originally thought to be a single species based on morphology, *Anopheles gambiae sensu lato* has been taxonomically divided into a seven member species complex, including species that differ in remarkable ways such as blood-meal host preference and preference for fresh versus

salt water larval habitats (Krzywinski and Besansky 2003). Among these members, the type species *Anopheles gambiae sensu stricto* has been studied most thoroughly and has been found to be comprised of at least three insipient species, termed the M molecular form, the S molecular form, and GOUNDRY, with the M and S molecular forms being much more closely related to each other than either is to GOUNDRY (della Torre et al. 2001; Riehle et al. 2011). Genetic exchange among these insipient species is very limited, and population differentiation is genomically widespread between the M and S molecular forms (Lawniczak et al. 2010), although the genomic distribution of differentiation between GOUNDRY and the M and S molecular forms is not yet well understood. Both ecological and molecular data have led to suggestions in the literature that some compartments of this species are adapting to novel, perhaps human-derived, environments and may have experienced population shifts as a result (M Coluzzi et al. 1979; Donnelly, Licht, and Lehmann 2001; Mario Coluzzi et al. 2002). The M and S molecular forms of *A. gambiae* differ in a number of ecological and behavioral phenotypes (reviewed in Lehmann and Diabate 2008), although how these difference impact disease transmission is not yet clear. Chromosomal inversions also play a significant role in the ecology and evolution of this system; over 120 polymorphic inversions have been detected in natural populations of species in the complex, 10 of which are fixed between species, and multiple large inversions show strong correlations with ecological clines (M Coluzzi et al. 1979; Mario Coluzzi et al. 2002). Many decades of study have focused largely on *A. gambiae s.s.*, and some of the cryptic and complex details of this system are beginning to come into focus. Much less is known about the other primary vectors, *A. arabiensis* and *A. funestus*, although the limited data available suggests evidence of cryptic substructure and ecological heterogeneity within these species as well, highlighting the need for more attention to be focused on these systems

Overall, the complexities of this system continue to develop, and rigorous analysis of genetic variation, particularly with the goal of identifying the signatures of natural selection, can only move forward once these processes are better understood.

In this dissertation, I describe research that focuses on patterns of evolution at protein coding sequences in populations of two of the three primary vectors, *A. gambiae* and *A. funestus*, with the goal of understanding the contributions of natural selection and non-selective demographic processes in shaping genetic variation in these species. Chapter 2 is a statistical inference of the demographic history of the M and S molecular forms of *A. gambiae* that uses the statistical fit of data simulated under a variety of demographic models to empirical re-sequencing data from approximately 100 genes distributed around the genome. In addition to distinguishing between population growth, population bottleneck, and migration models, I use population genetic theory to estimate the timing of population shifts and demonstrate that the data are not consistent with the long-held hypothesis that demographic shifts in mosquito populations stemmed from the agricultural revolution in Africa. Chapter 3 is a population genetic analysis of genes encoding the salivary gland protein saglin and a basil lamina structural protein laminin, both of which have been shown to interact physically with malaria parasites. Analysis of intraspecific genetic variation data as well as interspecific comparative data reveals no evidence for non-neutral evolution at either protein in either the M and S molecular forms. Chapter 4 is an additional population genetic analysis of re-sequencing data at immune genes to test whether these immune genes bear the signature of positive selection in the M and S forms and GOUNDRY. I found signals of putative positive selection at 11 genes, but only one signal is shared among populations, and no signals are found in the S form population, leading to the interpretation that the signals identified in the M form and GOUNDRY may reflect ongoing niche specialization in these two

populations. Chapter 5 is a simulation study assessing the power and accuracy of population genomic analyses based on whole-genome re-sequencing data generated by new high-throughput DNA sequencing technologies (next-generation sequencing). This analysis illustrates that deeper sequencing is needed when the experimental goal is demographic inference or analysis of genetic differentiation among closely related populations, but that tests for positive selection retain significant power and accuracy with shallow read depths, particularly when positive selection is strong and recent. Chapter 6 presents the development of a novel approach to *de novo* transcriptome assembly from short-read sequence data in the absence of a sequenced genome. This approach is applied to *A. funestus*, which is an important vector but for which virtually no genomic resources existed. From this assembly approach, I obtain a final set of over 15,000 putative cDNA transcripts (contigs) that are annotated through comparisons to Dipteran species with sequenced genomes and functional domain databases. Many immune genes are identified in the contig set, and read mapping back to the contig set reveals over 300,000 putative single nucleotide polymorphisms, both representing significant contributions to the data-poor *A. funestus* system that can be used in future studies by the research community. Overall, the work I present in this thesis makes significant contributions to our understanding of the evolution of *A. gambiae*, particularly at immune genes, provides valuable insights and resources for future studies of *A. funestus*, and establishes informative parameters for design and implementation of population genomic studies in any system.

REFERENCES

Arrighi, Romanico B G, Gareth Lycett, Vassiliki Mahairaki, Inga Siden-Kiamos, and Christos Louis. 2005. “Laminin and the Malaria Parasite’s Journey Through the Mosquito Midgut.” *The Journal of Experimental Biology* 208 (Pt 13) (July): 2497–2502. doi:10.1242/jeb.01664.

Biffen, R. H. 1905. “Mendel’s Laws of Inheritance and Wheat Breeding.” *The Journal of Agricultural Science* 1 (01): 4–48. doi:10.1017/S0021859600000137.

Billingsley, P.F., and R.E. Sinden. 1997. “Determinants of Malaria-mosquito Specificity.” *Parasitology Today* 13 (8) (August): 297–301. doi:10.1016/S0169-4758(97)01094-6.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. “The Hitchhiking Effect on the Site Frequency Spectrum of DNA Polymorphisms.” *Genetics* 140 (2): 783–796.

Brennan, J D, M Kent, R Dhar, H Fujioka, and N Kumar. 2000. “Anopheles Gambiae Salivary Gland Proteins as Putative Targets for Blocking Transmission of Malaria Parasites.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (25) (December 5): 13859–13864. doi:10.1073/pnas.250472597.

Cirimotich, Chris M, Yuemei Dong, Lindsey S Garver, Shuzhen Sim, and George Dimopoulos. 2010. “Mosquito Immune Defenses Against Plasmodium Infection.” *Developmental and Comparative Immunology* 34 (4) (April): 387–395. doi:10.1016/j.dci.2009.12.005.

Collins, F H, and S M Paskewitz. 1995. “Malaria: Current and Future Prospects for Control.” *Annual Review of Entomology* 40: 195–219. doi:10.1146/annurev.en.40.010195.001211.

Coluzzi, M, A Sabatini, V Petrarca, and M A Di Deco. 1979. “Chromosomal Differentiation and Adaptation to Human Environments in the Anopheles Gambiae Complex.” *Transactions of the Royal Society of Tropical Medicine and Hygiene* 73 (5): 483–497.

Coluzzi, Mario, Adriana Sabatini, Alessandra della Torre, Maria Angela Di Deco, and Vincenzo Petrarca. 2002. “A Polytene Chromosome Analysis of the Anopheles Gambiae Species Complex.” *Science (New York, N.Y.)* 298 (5597) (November 15): 1415–1418. doi:10.1126/science.1077769.

Dawkins, R., and J. R. Krebs. 1979. “Arms Races Between and Within Species.” *Proceedings of the Royal Society of London. Series B. Biological Sciences* 205 (1161) (September 21): 489–511. doi:10.1098/rspb.1979.0081.

Dinglasan, Rhoel R, and Marcelo Jacobs-Lorena. 2008. “Flipping the Paradigm on Malaria Transmission-blocking Vaccines.” *Trends in Parasitology* 24 (8) (August): 364–370. doi:10.1016/j.pt.2008.05.002.

Donnelly, M J, M C Licht, and T Lehmann. 2001. “Evidence for Recent Population Expansion in the Evolutionary History of the Malaria Vectors Anopheles Arabiensis and Anopheles Gambiae.” *Molecular Biology and Evolution* 18 (7) (July): 1353–1364.

Fay, Justin C, and Chung-I Wu. 2000. “Hitchhiking Under Positive Darwinian Selection.” *Genetics* 155 (3) (July 1): 1405–1413.

Frolet, Cécile, Martine Thoma, Stéphanie Blandin, Jules A Hoffmann, and Elena A Levashina. 2006. “Boosting NF-kappaB-dependent Basal Immunity of Anopheles

Gambiae Aborts Development of Plasmodium Berghei." *Immunity* 25 (4) (October): 677–685. doi:10.1016/j.immuni.2006.08.019.
Gouagna, L C, B Mulder, E Noubissi, T Tchuinkam, J P Verhave, and C Boudin. 1998. "The Early Sporogonic Cycle of Plasmodium Falciparum in Laboratory-infected Anopheles Gambiae: An Estimation of Parasite Efficacy." *Tropical Medicine & International Health: TM & IH* 3 (1) (January): 21–28.
Haldane, J.B.S. 1949. "Disease and Evolution." *La Ricerca Scientifica Supplemento A* 19: 68–76.
Han, Y S, J Thompson, F C Kafatos, and C Barillas-Mury. 2000. "Molecular Interactions Between Anopheles Stephensi Midgut Cells and Plasmodium Berghei: The Time Bomb Theory of Ookinete Invasion of Mosquitoes." *The EMBO Journal* 19 (22) (November 15): 6030–6040. doi:10.1093/emboj/19.22.6030.
Hillyer, Julián F, Catherine Barreau, and Kenneth D Vernick. 2007. "Efficiency of Salivary Gland Invasion by Malaria Sporozoites Is Controlled by Rapid Sporozoite Destruction in the Mosquito Haemocoel." *International Journal for Parasitology* 37 (6) (May): 673–681. doi:10.1016/j.ijpara.2006.12.007.
Kim, Y., and W. Stephan. 2002. "Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome." *Genetics* 160 (2): 765.
Krzywinski, Jaroslaw, and Nora J Besansky. 2003. "Molecular Systematics of Anopheles: From Subgenera to Subpopulations." *Annual Review of Entomology* 48: 111–139. doi:10.1146/annurev.ento.48.091801.112647.
Lawniczak, M K N, S J Emrich, A K Holloway, A P Regier, M Olson, B White, S Redmond, et al. 2010. "Widespread Divergence Between Incipient Anopheles Gambiae Species Revealed by Whole Genome Sequences." *Science (New York, N.Y.)* 330 (6003) (October 22): 512–514. doi:10.1126/science.1195755.
Lazzaro, Brian P. 2008. "Natural Selection on the Drosophila Antimicrobial Immune System." *Current Opinion in Microbiology* 11 (3) (June): 284–289. doi:10.1016/j.mib.2008.05.001.
Lehmann, Tovi, and Abdoulaye Diabate. 2008. "The Molecular Forms of Anopheles Gambiae: a Phenotypic Perspective." *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 8 (5) (September): 737–746. doi:10.1016/j.meegid.2008.06.003.
Luckhart, S, Y Vodovotz, L Cui, and R Rosenberg. 1998. "The Mosquito Anopheles Stephensi Limits Malaria Parasite Development with Inducible Synthesis of Nitric Oxide." *Proceedings of the National Academy of Sciences of the United States of America* 95 (10) (May 12): 5700–5705.
McDonald, J H, and M Kreitman. 1991. "Adaptive Protein Evolution at the Adh Locus in Drosophila." *Nature* 351 (6328) (June 20): 652–654. doi:10.1038/351652a0.
Nielsen, R, and Z Yang. 1998. "Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene." *Genetics* 148 (3) (March): 929–936.
Nielsen, Rasmus. 2005. "Molecular Signatures of Natural Selection." *Annual Review of Genetics* 39: 197–218. doi:10.1146/annurev.genet.39.073003.112420.
Obbard, Darren J, John J Welch, Kang-Wook Kim, and Francis M Jiggins. 2009. "Quantifying Adaptive Evolution in the Drosophila Immune System." *PLoS Genetics* 5 (10) (October): e1000698. doi:10.1371/journal.pgen.1000698.
Przeworski, M. 2002. "The Signature of Positive Selection at Randomly Chosen Loci." *Genetics* 160 (3): 1179–1189.

Riehle, Michelle M, Wamdaogo M Guelbeogo, Awa Gneme, Karin Eiglmeier, Inge Holm, Emmanuel Bischoff, Thierry Garnier, et al. 2011. “A Cryptic Subgroup of Anopheles Gambiae Is Highly Susceptible to Human Malaria Parasites.” *Science (New York, N.Y.)* 331 (6017) (February 4): 596–598. doi:10.1126/science.1196759.

Rodrigues, Janneth, Giselle A Oliveira, Michalis Kotsyfakis, Rajnikant Dixit, Alvaro Molina-Cruz, Ryan Jochim, and Carolina Barillas-Mury. 2012. “An Epithelial Serine Protease, AgESP, Is Required for Plasmodium Invasion in the Mosquito Anopheles Gambiae.” *PloS One* 7 (4): e35210. doi:10.1371/journal.pone.0035210.

Sabeti, Pardis C, David E Reich, John M Higgins, Haninah Z P Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, et al. 2002. “Detecting Recent Positive Selection in the Human Genome from Haplotype Structure.” *Nature* 419 (6909) (October 24): 832–837. doi:10.1038/nature01140.

Sackton, Timothy B, Brian P Lazzaro, Todd A Schlenke, Jay D Evans, Dan Hultmark, and Andrew G Clark. 2007. “Dynamic Evolution of the Innate Immune System in Drosophila.” *Nature Genetics* 39 (12) (December): 1461–1468. doi:10.1038/ng.2007.60.

Sinden, R E, Yasmene Alavi, and J D Raine. 2004. “Mosquito--malaria Interactions: a Reappraisal of the Concepts of Susceptibility and Refractoriness.” *Insect Biochemistry and Molecular Biology* 34 (7) (July): 625–629. doi:10.1016/j.ibmb.2004.03.015.

Smith, J. M., and J. Haigh. 1974. “The Hitch-hiking Effect of a Favourable Gene.” *Genet Res* 23 (1): 23–35.

Tajima, F. 1989. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Genetics* 123 (3): 585.

della Torre, A, C Fanello, M Akogbeto, J Dossou-yovo, G Favia, V Petrarca, and M Coluzzi. 2001. “Molecular Evidence of Incipient Speciation Within Anopheles Gambiae S.s. in West Africa.” *Insect Molecular Biology* 10 (1) (February): 9–18.

Vaughan, J A, L Hensley, and J C Beier. 1994. “Sporogonic Development of Plasmodium Yoelii in Five Anopheline Species.” *The Journal of Parasitology* 80 (5) (October): 674–681.

Vernick, Kenneth D, and Andrew P Waters. 2004. “Genomics and Malaria Control.” *The New England Journal of Medicine* 351 (18) (October 28): 1901–1904. doi:10.1056/NEJMcibr042899.

Waterhouse, Robert M, Evgenia V Kriventseva, Stephan Meister, Zhiyong Xi, Kanwal S Alvarez, Lyric C Bartholomay, Carolina Barillas-Mury, et al. 2007. “Evolutionary Dynamics of Immune-related Genes and Pathways in Disease-vector Mosquitoes.” *Science (New York, N.Y.)* 316 (5832) (June 22): 1738–1743. doi:10.1126/science.1139862.

Yassine, Hassan, and Mike A Osta. 2010. “Anopheles Gambiae Innate Immunity.” *Cellular Microbiology* 12 (1) (January): 1–9. doi:10.1111/j.1462-5822.2009.01388.x.

CHAPTER 2:

The demographic histories of the M and S molecular forms of *Anopheles gambiae s.s.*

Published as a letter in Molecular Biology and Evolution *(2010) 27(8):1739–1744.*

Jacob E. Crawford, Department of Entomology, Cornell University

Brian P. Lazzaro, Department of Entomology, Cornell University

Research Institution:

Department of Entomology, Cornell University, Ithaca NY, 14850, USA

Submission type: Letter

Title length: characters with spaces: 82

Abstract length: # words:156

Total length of text: characters including spaces: 14,280

Total page requirements: 4.0 typeset pages (3.5 manuscript = 1 typeset MBE)

Number of references: 29

Corresponding author: Jacob E. Crawford, jc598@cornell.edu, Department of Entomology, Comstock Hall, Cornell University, Ithaca, NY 14851

Keywords: *Anopheles gambiae*, demographic history, population growth, maximum likelihood, population genetics

Running head: Demographic history of *Anopheles gambiae* mosquitoes

ABSTRACT

Anopheles gambiae is a primary vector of *Plasmodium falciparum*, a human malaria parasite that causes over a million deaths each year in sub-Saharan Africa. Population genetic tests have been employed to detect natural selection at suspected *An. gambiae* anti-malaria genes, but these tests have generally been compromised by the lack of demographically correct null models. Here, we used a coalescent simulation approach within a maximum likelihood framework to fit population growth, bottleneck and migration models to polymorphism data from Cameroonian *An. gambiae*. The best-fit models for both the 'M' and 'S' molecular forms of *An. gambiae* included ancient population growth and a high rate of migration from an unsampled subpopulation. After correcting for differences in effective population size, our models suggest that the molecular forms expanded at different times, and both expansions significantly pre-date the advent of agriculture. We show that correcting null models for demography increases the power to detect natural selection in *An. gambiae*.

INTRODUCTION

Anopheles gambiae sensu stricto (hereafter *An. gambiae*) is a primary vector of *Plasmodium falciparum* (Collins and Paskewitz 1995), a human malaria parasite responsible for the death of an estimated 1 million people each year in sub-Saharan Africa, most of whom are children under the age of 5 (WHO/UNICEF World Malaria Report 2008). Development of novel technologies for controlling disease transmission, including genetic engineering of *Plasmodium*-resistant transgenic mosquitoes (Alphey et al. 2002), depends on knowledge of the basic biology and evolution of the vector and the parasite. One approach to obtaining such information is to use population genetic data to identify *Anopheles* loci that evolve under pathogen-mediated natural selection, and a number of candidate loci have been tested for selection in *An. gambiae* (e.g. Cohuet et al. 2008; Obbard et al. 2009). Tests for selection in this system tend to rely on the site-frequency spectrum (SFS; the frequency distribution of polymorphic mutations in the population) due to the lack of a suitable outgroup for inter-species molecular evolutionary comparisons (Obbard et al. 2007). However, tests of the SFS are also sensitive to demographic processes such as population growth and bottlenecks (Tajima 1989a,b; Fu and Li 1993). One way to improve the power to distinguish patterns generated by selection from those generated by demography is to test selective hypotheses against a null model based on the demographic history of the species (e.g. Stajich and Hahn 2005; Haddrill et al. 2005), but the absence of genome-wide polymorphism data has prevented development of an adequate demographic null for *An. gambiae*. In this work, we use sequence polymorphism data from 109 *An. gambiae* genes recently published by Cohuet et al. (2008) to infer the demographic history of Cameroonian *An. gambiae*.

Several non-equilibrium demographic hypotheses have been previously proposed to describe *An. gambiae*. *An. gambiae* is highly anthropophilic and

ecologically dependent on humans, and has been hypothesized to have undergone a range and population expansion coincident with agriculture-related shifts in human populations (Coluzzi et al. 2002; Costantini et al. 2009). A study of microsatellite polymorphism from Kenyan *An. gambiae* found evidence for population growth (Donnelly et al. 2001), and the SFS in this system tends to be enriched with low-frequency alleles (e.g. Cohuet et al. 2008; Obbard et al. 2009) consistent with an historical population expansion. Such patterns of polymorphism could also, however, derive from population bottlenecks (Tajima 1989a,b). Additional evidence for a bottleneck stems from transposable element insertion site frequency data that are suggestive of population bottlenecks (e.g. Esnault et al. 2008), possibly related to founding events associated with the formation of incipient species or population fluctuations during the last glacial maximum (Weijers et al. 2007). Migration among sub-populations may also be an important demographic factor in this system. Extant *An. gambiae* are divided into two largely reproductively isolated units referred to as the 'M' and 'S' molecular forms (della Torre et al. 2001). Geographic and microecological sub-structure has been identified within both molecular forms as well (e.g. Lehmann et al. 2003; Slotman et al. 2007).

To distinguish among potential demographic hypotheses describing *An. gambiae*, we performed coalescent simulations under various parameterizations of the above demographic models (Supplementary fig S1, Supplemental Methods online) and employed a modified approximate-likelihood method (Weiss and von Haeseler 1998) to test the fit of simulated models to synonymous autosomal polymorphism data for each molecular form independently (Cohuet et al. 2008; Supplementary Material online). The Cohuet et al. (2008) data set consists of short coding fragments from 72 immune-related and 37 functionally random genes sequenced in M form (n=10-16 chromosomes) and S form (n = 10-18 chromosomes) mosquitoes collected in Cameroon. We treated the

demographic models in a hierarchy of increasing parameter number, such that the standard neutral equilibrium model (SNE) was the null hypothesis, population growth was the first alternative, and the bottleneck and migration models were alternatives to the growth model. We found that the population growth model fit the empirical data significantly better than the equilibrium model for both the M and the S forms (table 1; $p_M < 10^{-4}$, $p_S < 10^{-4}$). No support was found for a population bottleneck in either molecular form (table 1). However, models that included both population growth and migration fit the data significantly better than the simple growth model for both molecular forms (table 1; $p_M = 0.0019$, $p_S < 10^{-4}$). We confirmed that our best-fit models were able to adequately reproduce the empirical data by showing that the average number of pairwise differences and the number of segregating sites in samples simulated under the best fit migration models (Supplementary Material online) match those summary statistics from the empirical data very well (Supplementary fig. S2 and S3). Furthermore, approximately 10% of simulations were accepted for each model (Supplementary table 1), which implies a good fit considering that we used the 20% threshold approach within the approximate-likelihood method that should reject as high as 80% of simulations even when the model perfectly matches the evolutionary process underlying the data.

Although similar in structure, the most likely migration models for the M and S molecular forms differed in their timing of expansion. From profile likelihood curves, we obtained maximum likelihood estimates (MLE) and approximate 95% confidence regions for the growth parameters (fig. 1 and 2). To evaluate the potential impact selection may have on our demographic inference, we reanalyzed the likelihood surface after removing the 6 loci with the most extreme Tajima's *D* values and found that the migration models remained the best fit models and MLE parameter values were essentially unchanged from those inferred using whole datasets. From this, we conclude that it is unlikely that

any natural selection in the history of the empirical data is biasing our inference process. We estimate that both molecular forms underwent at least 13-fold growth (table 1), but that the M molecular form expanded more recently than the S molecular form (49,000-490,000 years before present (YBP) for M form versus 63,000-630,000 YBP for S form, assuming 10 generations per year and a reasonable mutation rate; table 2). Our estimated growth times likely pre-date the extant division between the two molecular forms (e.g. Mukabayire et al. 2001). One potential explanation for our estimate of differing times of expansion for the two forms is that the ancestral, pre-molecular form population underwent an expansion, and then the derived M molecular form underwent a second more recent expansion, that may have been associated with post-speciation niche specialization (Constantini et al. 2009), such that the M form genome bears a mixed demographic signal from the two expansions.

Models that include migrational exchange with an unspecified second population fit the data best for both molecular forms, but these models should be interpreted with caution. For both molecular forms, the profile likelihood curve for the rate of migration ($4Nm$) is bimodal with local maxima near $4Nm$ of 0 and 10 (fig. 1 and 2). Both of these maxima suggest near-panmixia, with little or no real migration component. We therefore next modeled each molecular form under the growth model, but manually adjusted the modeled effective population size to be larger (i.e., pooled the sampled and hypothetical unsampled 'populations' into a single panmictic unit). We found that both the MLE growth and MLE migration models fit the data significantly better than the N_e-adjusted growth model for both molecular forms (Supplementary Material online). In principle, the migration models might provide a statistically better fit to the data in the absence of true historical migration if they allow for greater variance in effective population size than the simple growth model does. The signal for ancient growth is strong and clear in both

forms regardless of whether historical migration is included in the model. Nonetheless, the best fitting models for both molecular forms are those in which the focal populations share migrants with an un-sampled population that is smaller than the sampled population. Since the effective population size of the S molecular form is thought to be significantly larger than that of the M molecular form (e.g. Cohuet et al. 2008), this is unlikely to reflect migration between progenitors of extant M and S form mosquitoes, at least when the M population is the focal population being modeled.

It has been hypothesized that the advent of agriculture played a major role in the history of *An. gambiae* populations (e.g. Coluzzi et al. 2002; Donnelly et al. 2001), but the empirical sequence data from Cohuet et al. (2008) do not support this hypothesis. Based on the MLE growth parameter values inferred in our study, one would have to assume a per-nucleotide mutation rate of 10^{-7} mutations per generation in order to reconcile the inferred timing of population expansion with the agricultural revolution (< 5,000 YBP; Phillipson 2004). Such a mutation rate is orders of magnitude higher than typical per nucleotide mutation rate estimates for *Drosophila* (e.g. Tamura et al. 2004; Keightley et al. 2009), which provides our best estimate of the *Anopheles* mutation rate. Calculations based on more plausible parameter values (table 2) suggest that earlier anthropogenic events such as the movement out of the ancestral East African range by early humans (ca. 130,000 YBP; Reed and Tishkoff 2006) or subsequent human population expansions (ca. 50,000 – 70,000 YBP; Rogers and Harpending 1992) may have been key factors allowing mosquito populations to grow.

Genetic sub-structure in *An. gambiae* has been associated with the incipient speciation between the M and S forms (della Torre et al. 2001) as well as with ecological factors and chromosomal inversions (e.g. Slotman et al. 2007), raising the possibility that the demographic signal inferred from any single population may not be universally

applicable. With specific respect to our study, the "Forest" M form population from Cameroon under analysis here is partially differentiated from the "Mopti" M form populations from West Africa (e.g., Slotman et al. 2007). However, the population size expansion we infer in this study surely pre-dates extant population structure between Forest M and Mopti M, and we are confident that the signature of this M form demographic history should be shared among extant un-inverted M form autosomes. The same logic can be applied to geographically distinct S-form populations. Sequences within polymorphic chromosomal inversions, particularly on the inversion-rich chromosome II, are likely to bear the signature of more recent demographic and selective events associated with the inversions themselves, which could confound model-based inference of demographic history. As our analysis was based entirely on autosomes with the standard (un-inverted) karyotype (Cohuet et al. 2008), thought to be the ancestral form of *An. gambiae* (Ayala and Coluzzi 2005), we believe our conclusions are insulated from this concern, and that they can be taken to provide a baseline, ancestral demographic history for the genomes of extant *An. gambiae*.

A primary motivation for establishing correct demographic models in *An. gambiae* and other systems is to accurately identify targets of natural selection. This is especially important in *Anopheles*, where sites of host-pathogen coevolution may serve as targets for malaria-control intervention. To show the effect of including demography in the null population genetic model on the inference of putatively non-neutral patterns of polymorphism, we re-evaluated the results from a frequency spectrum-based analysis of *An. gambiae* loci conducted by Obbard et al. (2009). These authors re-sequenced 16 serine protease inhibitor genes (serpins) and 16 control loci in a West African M form population from Burkina Faso (BK) and an East African S form population from Kenya (KY), although we will only consider loci on chromosome III (4 serpins and 4 control loci)

to avoid the potentially confounding effects of chromosome II inversions in Burkina Faso. Of the 8 chromosome III loci that had at least 4 segregating sites (4 serpins and 4 control loci), only control loci BK-5 and BK-6 departed significantly (5% threshold) from a null distribution simulated under the standard neutral equilibrium (SNE) model (Obbard et al. 2009). We compared Tajima's *D* values from all 8 loci from BK and KY first to null distributions simulated under SNE, then to null distributions simulated under the MLE migration models we developed here (Supplementary Material online; Supplementary table 2). We found that the negative values of *D* observed at control loci 5 and 6 remained significantly inconsistent with neutrality under the MLE migration model (locus BK-5: $p = 0.0160$; locus BK-6: $p = 0.0076$), and that the positive values of *D* observed at serpins 4C and 6 became significant when compared to the MLE models (KY-4C: $p = 0.0372$; KY-6: $p = 0.0069$; Supplementary table 2). Interestingly, while the mean *D* value under MLE migration models was consistently more negative than those under the SNE, the distributions showed less dispersion around the mean than distributions simulated under the SNE, resulting in a lower *p*-value for control loci BK-5 and BK-6 under the SNE model than under the MLE migration model (Supplementary table 2). These results highlight the increased power to detect putative signals of natural selection when using demographically corrected null models. The power gains associated with using correct null models should be even greater when sophisticated genome-scale methods such as the composite-likelihood ratio test of Kim and Stephan (2002) are employed.

ACKNOWLEDGEMENTS

We are grateful to two anonymous reviewers, Laura Harrington, Darren Obbard, Kirk Lohmueller and members of the Lazzaro Laboratory for helpful discussions and comments on earlier versions of the manuscript. This work was supported by NIH grant AI062995.

FIGURES

Figure 1: M molecular form genome-likelihood values relative to migration model parameters (A) Time of expansion in units of N_{curr} generations, (B) Size of population growth (N_{curr}/N_{anc}), (C) Rate of migration ($4Nm$) per generation and (D) Size of the unsampled subpopulation relative to the sampled subpopulation. For each parameter value, the highest genome-likelihood value from all models within the migration model family is plotted. Note log scale in panels (B) and (C). Horizontal dashed line indicates 95% threshold, such that all genome-likelihood values below this threshold are significantly different from the maximum-likelihood value. Vertical dashed line(s) indicate approximate boundaries of the 95% confidence region of the model parameter. For panels (A) and (D), all parameter values outside of the vertical dashed lines are significantly different from the MLE value. For panel (B), all parameter values to the left of the vertical line are significantly different from the MLE value. The shape of the curve in panel (C) did not allow determination of confidence region.

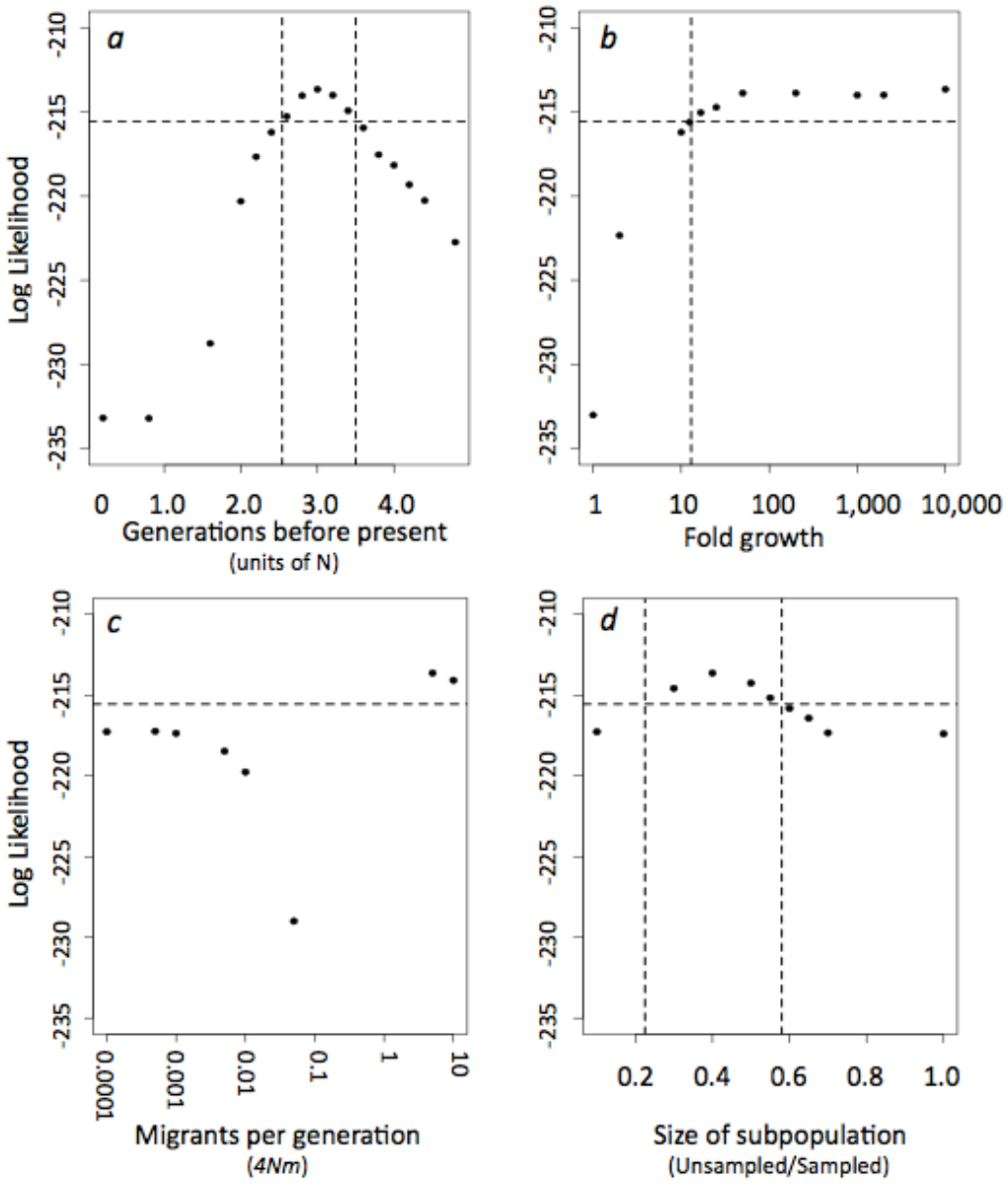
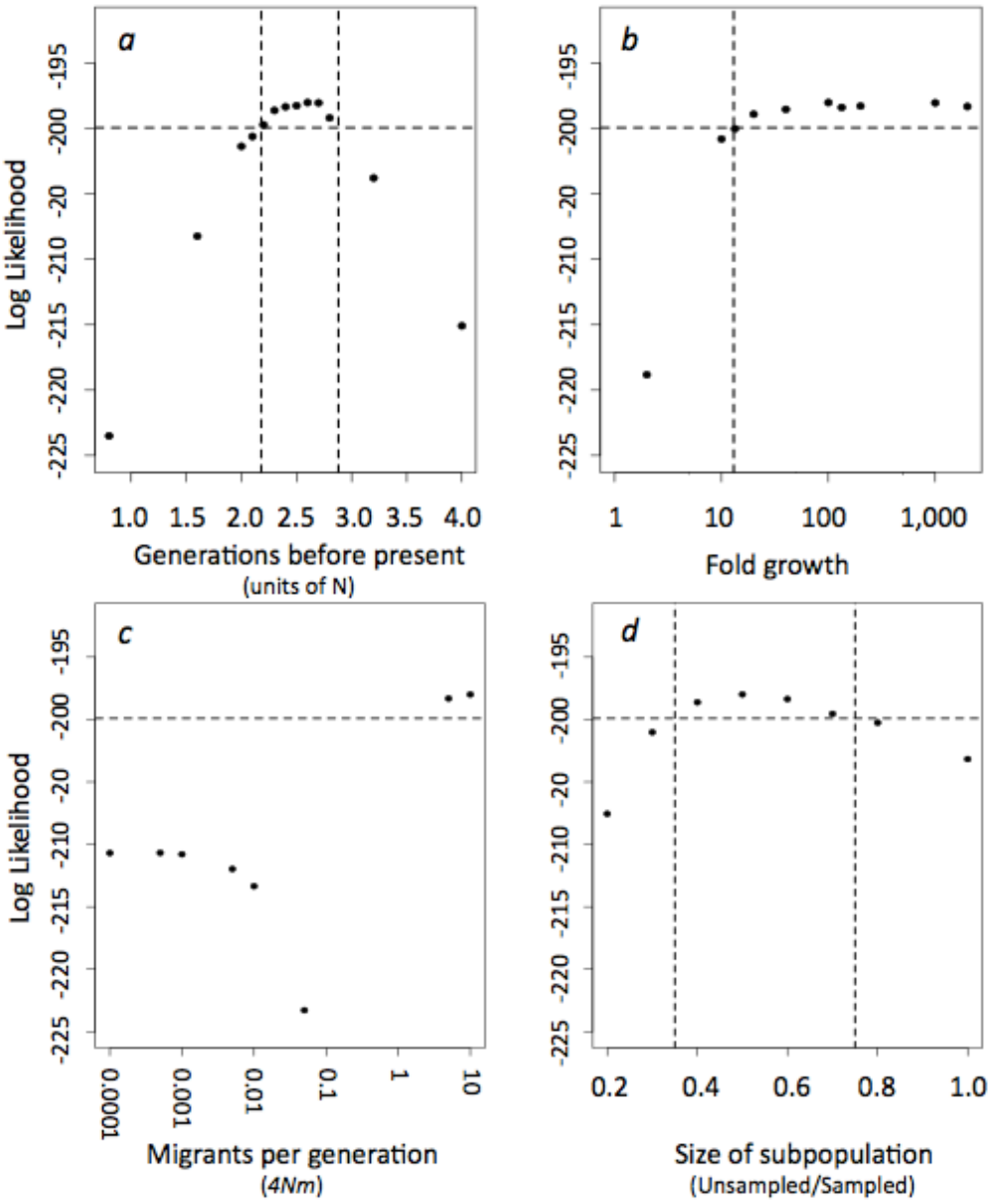

a
b
c
d
Log Likelihood
-210
-215
-220
-225
-230
-235
0
1.0
2.0
3.0
4.0
Generations before present
(units of N)
1
10
100
1,000
10,000
Fold growth
0.0001
0.001
0.01
0.1
1
10
Migrants per generation
(4Nm)
0.2
0.4
0.6
0.8
1.0
Size of subpopulation
(Unsampled/Sampled)

Figure 2: S molecular form genome-likelihood values relative to migration model parameters (A) Time of expansion in units of N_{curr} generations, (B) Size of population growth (N_{curr}/N_{anc}), (C) Rate of migration ($4Nm$) per generation and (D) Size of the unsampled subpopulation relative to the sampled subpopulation. For each parameter value, the highest genome-likelihood value from all models within the migration model family is plotted. Note log scale in panels (B) and (C). Horizontal dashed line indicates 95% threshold, such that all genome-likelihood values below this threshold are significantly different from the maximum-likelihood value. Vertical dashed line(s) indicate approximate boundaries of the 95% confidence region of the model parameter. For panels (A) and (D), all parameter values outside of the vertical dashed lines are significantly different from the MLE value. For panel (b), all parameter values to the left of the vertical line are significantly different from the MLE value. The shape of the curve in panel (C) did not allow determination of confidence region.


a
b
c
d
Log Likelihood
-195
-200
-20
-210
-215
-220
-225
1.0 1.5 2.0 2.5 3.0 3.5 4.0
Generations before present
(units of N)
1
10
100
1,000
Fold growth
0.0001
0.001
0.01
0.1
1
10
Migrants per generation
(4Nm)
0.2
0.4
0.6
0.8
1.0
Size of subpopulation
(Unsampled/Sampled)

REFERENCES

Alphey L, Beard CB, Billingsley P, Coetzee M, et al. 2002. Malaria control with genetically manipulated insect vectors. Science 298: 119-121.

Ayala FJ, Coluzzi M. 2005. Chromosome speciation: humans, Drosophila, and mosquitoes. Proc. Natl. Acad. Sci. 102 suppl 1:6535-6542.

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

Collins FH, Paskewitz SM. 1995. Malaria: current and future prospects for control. Annual Review of Entomology 40: 195-219.

Coluzzi M, Sabatini A, della Torre A, Di Deco MA, et al. 2002. A polytene chromosome analysis of the Anopheles gambiae species complex. Science 298: 1415-1418.

Costantini C, Ayala D, Guelbeogo WM, Pombi M, et al. 2009. Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in Anopheles gambiae. BMC Ecology 9:16.

della Torre A, Fanello C, Akogbeto M, Dossou-Yovo J, et al. 2001. Molecular evidence of incipient speciation within Anopheles gambiae ss in West Africa. Insect Molecular Biology 10: 9–18.

della Torre A, Tu Z, Petrarca V. 2005. On the distribution and genetic differentiation of Anopheles gambiae ss molecular forms. Insect biochemistry and molecular biology 35: 755–769.

Donnelly MJ, Licht MC, Lehmann T. 2001. Evidence for recent population expansion in the evolutionary history of the malaria vectors *Anopheles arabiensis* and *Anopheles gambiae*. Molecular Biology and Evolution 18: 1353–1364.

Esnault C, Boulesteix M, Duchemin JB, Koffi AA, et al. 2008. High genetic differentiation

between the M and S molecular forms of *Anopheles gambiae* in Africa. PloS One 3: e1968.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133: 693–709.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Research 15: 790-799.

Keightley PD, Trivedi U, Thomson M, Oliver F, et al. 2009. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. Genome Research 19: 1195-1201.

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765-777.

Lehmann T, Hawley WA, Grebert H, Collins FH. 1998. The effective population size of Anopheles gambiae in Kenya: implications for population structure. Mol. Biol. Evol. 15: 264-276.

Lehmann T, Licht M, Elissa N, Maega BTA, et al. 2003. Population Structure of Anopheles gambiae in Africa. The Journal of Heredity 94: 133-147.

Obbard DJ, Linton YM, Jiggins FM, Yan G, et al. 2007. Population genetics of Plasmodium resistance genes in Anopheles gambiae: no evidence for strong selection. Molecular ecology 16: 3497.

Obbard DJ, Welch JJ, Little TJ. 2009. Inferring selection in the Anopheles gambiae species complex: an example from immune-related serine protease inhibitors. Malaria Journal 8: 117.

Phillipson DW. 2005. African archaeology. p. 202. Cambridge: Cambridge University Press.

Reed F, Tishkoff S. 2006. African human diversity, origins and migrations. Current

Opinion in Genetics & Development 16: 597-605.

Rogers AR, H. Harpending H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. 9: 552-569.

Slotman MA, Tripet F, Cornel AJ, Meneses CR, et al. 2007. Evidence for subdivision within the M molecular form of Anopheles gambiae. Molecular Ecology 16: 639–649.

Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. Mol. Biol. Evol. 22: 63-73.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol. Biol. Evol. 21: 36-44.

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

Tajima F. 1989b. The effect of change in population size on DNA polymorphism. Genetics 123: 597-601.

Weijers JWH, Schefuss E, Schouten S, Damste JSS. 2007. Coupled Thermal and Hydrological Evolution of Tropical Africa over the Last Deglaciation. Science 315: 1701-1704.

Weiss G, von Haeseler A. 1998. Inference of population history using a likelihood approach. Genetics 149: 1539-1546.

WHO/UNICEF World Malaria Report. 2008. World Malaria Report. http://apps.who.int/malaria/wmr2008/.

Table 1: Maximum Likelihood Estimates (and 95% confidence intervals) for model families and model comparisons

	M form			S form		
Model	Growth	Bottleneck	Migration	Growth	Bottleneck	Migration
Generations since growth (N_{curr}) [a]	3.52 (2.88 - 4.12)	3.44 (2.92 - 4.16)	3.00 (2.54 – 3.50)	3.08 (2.58 – 3.50)	3.04 (2.60 – 3.52)	2.60 (2.18 – 2.88)
Fold growth (N_{curr} /N_{anc}) [a]	1,000 (4.65 - ND) [b]	10,000 (4.05 - ND) [b]	10,000 (13.0 - ND) [b]	2,000 (4.65 - ND) [b]	1,000 (5.10 - ND) [b]	100 (13.00 - ND) [b]
Reduction during bottleneck ($N_{pre\text{-}bottle}$/N_{anc}) [a]	---	10,000 (NA) [c]	---	---	667 (NA) [c]	---
Duration of bottleneck (T_{bot}) [a]	---	0.2 (NA) [c]	---	---	0.2 (NA) [c]	---
Migrants per generation ($4Nm$) [a]	---	---	5 (ND)	---	---	10 (ND)
Size of Unsampled subpopulation (relative to sampled)	---	---	0.40 (0.23 – 0.58)	---	---	0.50 (0.35 – 0.75)
Log Likelihood	-217.1336	-216.9759	-213.6302	-210.4305	-210.3666	-197.9891
AIC	438.2672	441.9518	435.2604	424.861	428.7332	403.9782
(k)	(2)	(4)	(4)	(2)	(4)	(4)
Λ_{SN} [d] (p value relative to equilibrium)	-28.7046 (<0.0001)	---	---	-44.7847 (<0.0001)	---	---
Λ_G [e] (p value relative to growth)	---	3.6846 (ns)	-3.0068 (0.0019)	---	3.8722 (ns)	-20.8828 (<0.0001)

AIC = -2(LogLikelihood -*k*) where *k* is the number of free parameters in model.

[a] Parameter units

[b] ND indicates cases where only one boundary of the confidence interval could be determined.

[c] Confidence intervals were not estimated for these parameters

[d] Λ_{SN} indicates comparisons made between the AIC under the MLE and the AIC under the standard-neutral equilibrium model.

[e] Λ_{G} indicates comparisons made between the AIC under the MLE and the AIC under the growth model.

1

Table 2: Calculations of the approximate timing of growth based on empirical parameter values

Form (θ_W[a])	μ[b]	N_e[c]	T_1[d]	Generations per year[e]	Years before present[f]
M (2.27%)	3.5×10^{-9}	1,623,571	3.0	10	487,071
	3.5×10^{-8}	162,357	3.0	10	48,707
	3.5×10^{-7}	16,236	3.0	10	4,871
	3.5×10^{-9}	1,623,571	3.0	20	243,536
	3.5×10^{-8}	162,357	3.0	20	24,353
	3.5×10^{-7}	16,236	3.0	20	2,435
S (3.4%)	3.5×10^{-9}	2,435,000	2.6	10	633,100
	3.5×10^{-8}	243,500	2.6	10	63,310
	3.5×10^{-7}	24,350	2.6	10	7,500
	3.5×10^{-9}	2,435,000	2.6	20	316,550
	3.5×10^{-8}	243,500	2.6	20	37,499
	3.5×10^{-7}	24,350	2.6	20	3,750

[a] θ_W was estimated from synonymous sites in the Cohuet et al. (2008) datasets (see Supplementary Material online) and is an estimator of $4N_e\mu$

[b] mutation rate per base pair per generation of 3.5×10^{-9} taken from Keightley et al. (2009)

[c] Effective population size calculated from θ_W using the stated mutation rate

[d] MLE time of growth in units of N_{curr}

[e] Estimates taken from Lehmann et al. 1998.

[f] Years before present calculated as (T_1 x N_e)/Generations per year

SUPPLEMENTAL MATERIAL

Molecular Form Datasets:

Our analysis is based on sequence polymorphism in coding fragments of 72 immune-related and 37 non-immune genes spread across all chromosome arms published by Cohuet et al. (2008). The mosquitoes sampled by Cohuet et al. (2008) are multiple *An. gambiae* individuals of the M (n=16 chromosomes) and S (n=18 chromosomes) molecular forms collected near Yaounde, Cameroon (03°51'N, 11°30'E). Mean nucleotide diversity was not significantly different between immune and non-immune loci and there was no evidence for strong selection in these data (Cohuet et al. 2008), so we consider all autosomal loci without regard to gene function in our analysis. We downloaded the heterozygous sequence fragments (accessions AM774672 – AM777160, AM900849 – AM900919), arbitrarily resolved the heterozygous sites to produce two hypothetical alleles for each individual and constructed alignments of each gene. Then, for each molecular form separately, the total number of segregating synonymous sites (S) was determined in each alignment and genetic diversity at synonymous sites was summarized for each molecular form as θ_W (Watterson 1975) and π (Tajima 1983) based on the total number of mutations using DnaSP (version 5.00.07, Librado and Rozas 2009). We used only synonymous sites to minimize any effects of natural selection in the dataset. θ_W and π are both estimators of the population parameter $4N_e\mu$ (Watterson 1975; Tajima 1983) where N_e is the effective population size and μ is the neutral substitution rate, but they are calculated from different features of the empirical data. Whereas π is the average number of differences between alleles and is thus sensitive to allele frequency, θ_W is calculated based on the number of segregating mutations regardless of their frequency in the sample. p and q respond differently to demographic shifts (Tajima 1989a,b). We used θ_W estimated from the empirical data to set the rate of mutation in the coalescent simulations, and we used π and S to

summarize diversity in the simulated and empirical samples. These latter two summary statistics are the main components of the frequently used Tajima's *D* statistic (Tajima 1989a) and provide information about the shape of the underlying genealogy. Their relationship can be used to detect demographic (or selective) perturbations reflected in a sample. We used the components of the *D* statistic instead of the statistic itself because the *D* statistic is a biased summary of the data when recombination rates are not correctly incorporated and can compromise approximate-likelihood inference of demographic parameters (Thornton 2005). However, the bias is minimized if *D* is decomposed and its components, π and *S,* are used in its place (Thornton 2005). Only autosomal loci from the Cohuet et al. (2008) data set that were represented by at least ten alleles (range of 10 to 16 alleles for the M form and 10 to 18 alleles per locus) and exhibited a value of θ_W greater than zero were included our analysis (92 and 95 loci for M and S form respectively). Although excluding polymorphism-free loci from the analysis may slightly bias the dataset, a non-zero value of θ_W is needed for simulations (see below). We excluded X-linked loci for this analysis because large regions of the X-chromosome lack polymorphism, possibly due to recent selective sweeps (Stump et al. 2005; Turner et al. 2007), making most of the chromosome difficult to simulate.

Coalescent Simulations and Demographic Models:

We were interested in identifying a demographic scenario that can explain observed patterns of polymorphism in each of the molecular forms of *An. gambiae.* Our approach was to simulate individual loci under specific population demographic scenarios, evaluate the fit of the simulated data to the empirical polymorphism data at individual loci, and combine these likelihood values into a 'genome likelihood.' We applied this approach to the M and S molecular forms independently. We modeled each gene individually by conducting 2×10^4 coalescent

simulations under varying demographic scenarios using the program *ms* (Hudson 2002) conditioned on the sample size and θ_W for each gene as reflected in the empirical data set. Based on the fact that the sequenced genes are physically dispersed but typically shorter than 700 base pairs, we assumed free recombination between genes, but no intragenic recombination. Underestimating recombination is a conservative approach and not likely to produce large biases in the inference process. We calculated π and S from the *ms* output, which were then used to evaluate the 'genome likelihood' fit to the empirical data (described below).

We considered three families of demographic models: population growth, population bottleneck, and migration between two growing populations (main text fig. 1). For each model family, we explored a wide range of parameter values, chosen to be comprehensive but biologically plausible. The first model, population growth, varied in two parameters: the timing of the expansion (T_1, in units of $4N$ generations) and the ratio of ancient to current effective population size (N_{anc}/N_{curr}). The population bottleneck model family included the growth parameters listed above with the addition of a pre-expansion bottleneck that varied in both the severity of size reduction ($N_{pre\text{-}bottle}/N_{anc}$) and the number of generations the population remained at the reduced size (T_{bot}). The last model, migration between expanding subpopulations, included the growth parameters as well as migration from a second, unsampled subpopulation with growth parameters identical to the sampled subpopulation. The relative size of the unsampled subpopulation ($N_{unsampled}/N_{sampled}$) and rate of migration ($4Nm$) was also allowed vary in the model. The standard neutral drift-equilibrium model was considered as a null hypothesis. All parameter values are listed in Table 1.

Approximate Likelihood Method:

To determine how well each demographic model fit the empirical data, the simulated population samples were evaluated using an adaptation of Weiss and von Haeseler's (1998) approximate likelihood method. An indicator variable I_δ was calculated as

$$I_\delta(j) = \begin{cases} 1, & if, |\pi_{data} - \pi_{sim}| \le \delta_\pi \; and \; |S_{data} - S_{sim}| \le \delta_S \\ 0, & othwerwise. \end{cases} \quad (1)$$

where π_{data} equals the average number of pairwise differences in the empirical sample and S_{data} equals the number of segregating sites in the empirical sample at locus *j*. π_{sim} and S_{sim} were summary statistics calculated from the simulation results for that locus under the given model and δ_π and δ_S were positive numbers that define an empirically determined interval (see below) for locus *j*. Our method differs slightly from that of Weiss and von Haeseler (1998) in that they required the number of segregating sites to exactly match the empirical data, which is a slightly more conservative method, but we used the threshold approach to accommodate uncertainty in empirical estimates of the true population θ_W. The numerical threshold was designed to capture 20% of stochastic variation natural to the coalescent process, such that simulated values of *p* or S falling more than 10% above or below the empirical value resulted in the assignment of zero to the indicator variable (Weiss and von Haeseler 1998). Both threshold values were determined for each gene and each molecular form by conducting 2×10^4 coalescent simulations conditioned on the empirical sample size and θ_W for each locus under the standard neutral model. The summary statistics, *p* and *S,* were calculated for all simulations, assembled into a distribution and the thresholds were determined as the values 10% greater and less than the median of the simulated distribution.

The likelihood of the model given the data for each gene was estimated as the proportion of simulations that were assigned an I_δ of 1. The approximate likelihood function can be written as

$$lik(\Phi \mid \pi_{data}, S_{data}) \approx \frac{1}{B}\sum_{j=1}^{B} I_{\delta}(j), \quad (2)$$

where Φ is the model, π and S are summary statistics from the empirical data at locus *j, B* is the number of simulations (2×10^4) and *I* is the indicator variable from above. To obtain a genome-wide likelihood value that reflects the likelihood of the model given genome-wide patterns of polymorphism, gene-specific likelihoods were then natural log transformed and summed.

To identify the most likely model within each model family, we evaluated a series of models organized across a grid of parameter values using the above likelihood function to obtain the genome-likelihood value for each model. First, we searched a coarse grid of parameter values for each family of models. Next, in order to improve the precision of our parameter estimates, we adjusted parameter scales to finer levels in regions of the parameter space that showed high likelihood values in the coarse grid search and searched our finer-scale grid using the same likelihood procedure. We identified the best-fit model within each model family as the combination of parameter values that maximized the genome-wide likelihood function, and these best-fit models were then compared to determine which model family is most likely given the data (discussed below). To visualize the likelihood surface, we generated profile-likelihood curves for each parameter by plotting the maximum likelihood value for each parameter value. We estimated approximate 95% confidence intervals for each parameter using asymptotic theory where all parameter values with a likelihood value within 1.92 likelihood units (i.e. $X^2_{df=1}$ and $\alpha = 0.05$) of the maximum likelihood value were considered not significantly different from the MLE. Linear interpolation of the profile-likelihood curves was used where points were not simulated directly.

Model Comparison:

After identifying the best-fit model within each model family, we compared models between families (e.g. growth vs. bottleneck) to identify the maximum likelihood estimate (MLE) for the demographic history of the *An. gambiae* population. We treated the models in a hierarchical fashion, with the standard-neutral model considered to be the primary null hypothesis. The simple growth model is an alternative to the standard neutral null. The more complex bottleneck and migration models both have the growth model nested within them, so the growth model is considered the null model for testing bottleneck and migration hypotheses. Thus, the standard-neutral was first compared to the growth model. If the standard neutral null was rejected in this first comparison, the growth model then became the secondary null model against which the bottleneck and migration models were compared. If neither the bottleneck nor migration models fit significantly better than the growth model, we concluded that simple growth was the most parsimonious and likely model.

We compared models using the Akaike Information Criterion (AIC; Akaike 1974). Our models were not nested in the fashion required for evaluation of likelihood ratios. We employed AIC values to compare the likelihoods of non-nested models by penalizing models according to the number of free parameters in the model. We calculated AIC values as $AIC_i = -2(lmax_i - k_i)$ where $lmax_i$ is the maximum likelihood value under model i and k is the number of free parameters in model i, such that a higher AIC value means a better fit to the data (Akaike 1974). Then we used the statistic $\Lambda = AIC_{alt} - AIC_{null}$ to compare AIC values between models (Caicedo et al. 2007). Negative values of this statistic indicate that the alternative model is a better fit. We established a null distribution by simulating 10^4 'genomes' comprised of the same number of loci as the empirical dataset under the null model, evaluating the maximum likelihood of each 'genome' under the null and alternative models and calculating Λ_{sim} as the difference between

the AIC statistics calculated under the null and alternative models. We calculated a *p*-value as the proportion of simulations with $\Lambda_{sim} < \Lambda_{obs}$.

Model performance:

Although the approximate likelihood method used here explicitly evaluates the fit of the model to the entire dataset, we wanted to confirm that our best-fit model is able to adequately reproduce the empirical data for each molecular form. To this end, we simulated all chromosome III loci under the best-fit migration model for each molecular form and plotted the median value of π and S from 10^4 coalescent simulations next to the empirical data (Supplementary figs. 4 and 5). Simulations were conducted as above where each locus was simulated using *ms* conditioned on the empirical sample size and θ_W. The distributions of the summary statistics are often skewed so we compared the median value to the data in order to minimize biases associated with mean values of skewed distributions. We considered the comparison of loci on only one chromosome sufficient to demonstrate the adequacy of model performance and arbitrarily chose chromosome III.

Comparison of the timing of expansion between Molecular forms:

To determine whether the MLE timings of expansion were significantly different between the molecular forms, we asked whether the timing inferred for one molecular form was within the confidence region of the timing or growth parameter (T_1) estimated for the other molecular form. For example, the inferred timing of growth for the M form is $3.0N_{curr}$, which corresponds to $2.1N_{curr}$ for the S form after calibration for the relative effective population sizes (we estimate that the M form is 0.7 times the S form). This value is outside of the confidence interval

estimated for the timing of growth for the S form (95% C.I. 2.18 - 2.88), suggesting that this more recent timing of the M form expansion did not overlap in time with the S form expansion.

Population genetic re-analysis of Obbard et al. (2009) data:

To test the effects of applying the demographic correction to the null model on population genetic analyses, we compared Tajima's *D* values from 4 serpin loci and 4 control loci obtained by Obbard et al. (2009) first to null distributions simulated under the standard-neutral equilibrium (SNE) model then to null distributions simulated under the MLE migration models inferred here. We simulated each sample and locus individually using the same simulation framework described above. 10^4 coalescent simulations were conducted using the coalescent simulation program *ms* (Hudson 2002) for each locus-population combination conditioned on the number of chromosomes sampled for that locus-population combination and θ_W estimated from the empirical data. Tajima's *D* was calculated from each simulated sample and assembled into null distributions for a given locus-population combination. Null distributions were generated both under the standard-neutral equilibrium as well as under the MLE migration models for each form. Empirical *D* values were then compared to null distributions in a one-tailed test, the polarity of which depended on whether *D* was positive or negative. No correction for multiple testing was made, so *D* was considered significantly unlikely under a given null model if the empirically observed value fell into the 5% tail of the null distribution. The simulations assumed no recombination, which is a reasonable approximation given the short sequences (range of 354 to 783 basepairs), and so are conservative with regard to testing hypotheses of selection.

Simulations of N_e-adjusted migration models:

One possible explanation for the better fit of migration models is that the effective population size is increased through migration, and thus no migration is actually necessary in the models. To determine whether manually adjusting the effective population size can account for the increased likelihood of the migration models over the growth models, we simulated each locus under the MLE growth model, but we adjusted θ_W to reflect the larger effective population size. For example, the MLE migration model for the M molecular form includes migration between the sampled population and an unsampled population that is 0.4 times the size of the sampled population, so we multiplied the empirical θ_W for each locus by 1.4 so that the adjusted simulated population is one panmictic unit 1.4 times as large as its unadjusted counterpart. These adjusted models were compared to the unadjusted MLE growth and MLE migration models using the model comparison framework described above.

SUPPLEMENTARY MATERIAL REFERENCES

Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716-723.

Caicedo AL, Williamson SH, Hernandez RD, Boyko A, et al. 2007. Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice. PLoS Genetics 3: e163.

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

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

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

Obbard DJ, Welch JJ, Little TJ. 2009. Inferring selection in the Anopheles gambiae species complex: an example from immune-related serine protease inhibitors. Malaria Journal 8: 117.

Stump AD, Fitzpatrick MC, Lobo NF, Traoré S, et al. 2005. Centromere-proximal differentiation and speciation in Anopheles gambiae. Proc. Natl. Acad. Sci. 102: 15930–15935.

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

Tajima F. 1989b. The effect of change in population size on DNA polymorphism. Genetics 123: 597-601.

Thornton K. 2005. Recombination and the properties of Tajima's D in the context of approximate-likelihood calculation. Genetics 171: 2143-2148.

Turner TL, Hahn MW. 2007. Locus- and population-specific selection and differentiation between incipient species of Anopheles gambiae. Molecular Biology and Evolution 24: 2132-2138.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7: 256-276.

Supplementary Table 1: Demographic model parameter ranges, sampling density and rejection statistics

Parameter	Range[a] (units)	Growth		Bottleneck		Migration	
		M	S	M	S	M	S
Fold population expansion (N_{anc}/N_{curr})	0 – 10,000	42	33	22	19	11	11
Generations since growth (T_1)	0.05 – 1.2 ($4N_{curr}$ generations)	64	55	29	29	17	13
Fold population reduction during bottleneck ($N_{pre\text{-}bottle}/N_{anc}$)	1.25 – 10,000	---	---	4	4	---	---
Duration of bottleneck (T_{bot})	0.01 – 0.5 ($4N_{curr}$ generations)	---	---	4	5	---	---
Subpopulation size ($N_{unsampled}/N_{sampled}$)	0.1 – 1.0	---	---	---	---	9	8
Rate of migration ($4Nm$)	10^{-4} – 10 (migrants per generation)	---	---	---	---	10	10
Total number parameter combinations[b]		2,688	1,815	7,888	7,556	16,830	11,440
Percentage of simulations accepted[c]		9.74	10.66	9.42	10.17	8.28	9.47
Total number parameter combinations accepted[d]		432	267	1,269	1,062	88	149
Percentage of simulations accepted within accepted models[e]		10.39	11.64	10.40	11.66	11.29	13.63

[a] For each demographic model and molecular form, we searched the parameter space uniformly over coarse intervals. We then adjusted the parameter space to include a higher density of grid points (parameter combinations) in the region with the highest likelihood values in the first search and evaluated the grid a second time.

[b] Total number of parameter combinations searched in grid after increasing density of parameter values sampled in the second grid search.

[c] Percentage of all simulations that was not rejected within likelihood framework. Each locus was simulated 20,000 times for each parameter combination.

[d] Total number of parameter combinations that received likelihood value within 1.92 likelihood units of the maximum.

[e] Percentage of simulations within accepted models (see [d]) that were not rejected within the likelihood framework.

Supplementary Table 2: Population genetic re-analysis of Obbard et al. (2009) data under SNE and MLE migration models

			Tajima's *D*	
Locus	Population[a]	*D*	*SNE*[b]	*MLE*[c]
Control 4	Burkina Faso	0.16	0.3715	0.1894
	Kenya	0.34	0.3066	0.0622
Serpin 4C	Burkina Faso	-0.11	0.5093	0.6333
	Kenya	0.90	0.1575	**0.0372**
Control 5	Burkina Faso	-1.74	**0.0235**	**0.0160**
	Kenya	0.36	0.3123	0.1784
Serpin 5	Burkina Faso	-1.33	0.0737	0.0609
	Kenya	0.07	0.4158	0.2077
Control 6	Burkina Faso	-1.85	**0.0113**	**0.0076**
	Kenya	---	---	---
Serpin 6	Burkina Faso	-0.63	0.2943	0.3361
	Kenya	1.25	0.0836	**0.0069**
Control 16	Burkina Faso	-0.75	0.2569	0.2850
	Kenya	---	---	---
Serpin 16	Burkina Faso	-0.29	0.4284	0.5539
	Kenya	-0.77	0.2507	0.3504

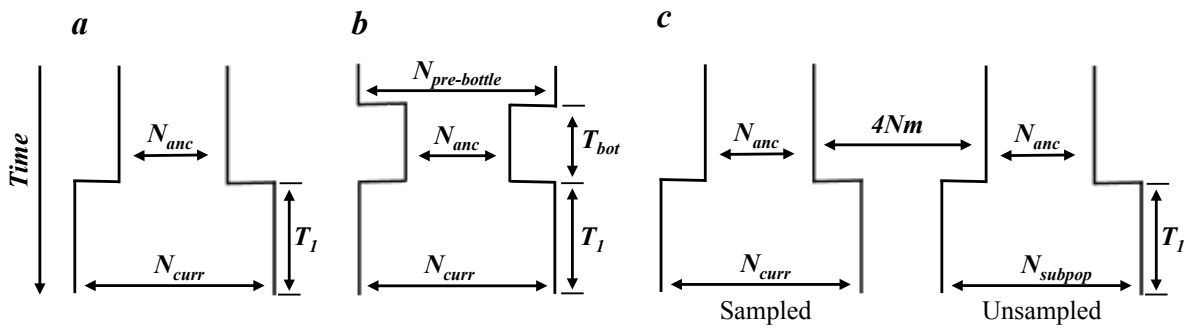
[a] location where *An. gambiae* were sampled

[b] *P* values indicating probability of statistic when compared to null distribution simulated under standard-neutral equilibrium

[c] *P* values indicating probability of statistic when compared to null distribution simulated under MLE migration model

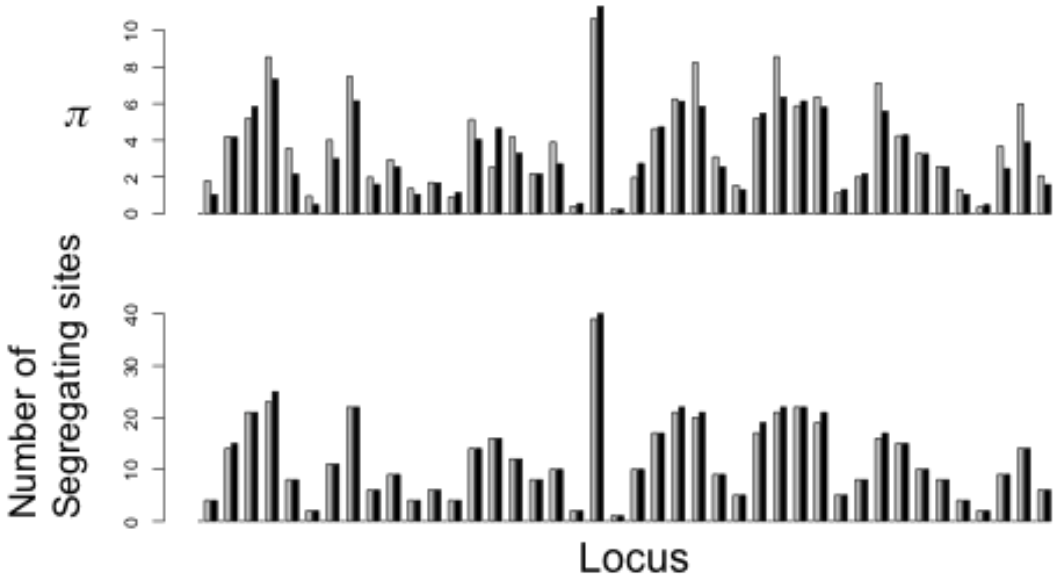
- Values in bold font were significantly inconsistent with the simulated null model at the nominal 5% threshold (no correction for multiple testing)

Supplementary Figure S1: Demographic models and their varied parameters *(a)* Population growth included time of expansion (T_1) and size of expansion (N_{curr}/N_{anc}) variables *(b)* Population bottleneck included growth parameters (T_1 and N_{curr}/N_{anc}), the size of population reduction and duration of bottleneck (T_{bot}) *(c)* Migration between growing subpopulations including growth parameters (T_1 and N_{curr}/N_{anc}), the rate of migration ($4Nm$) and the size of the unsampled subpopulation relative to the sampled subpopulation.

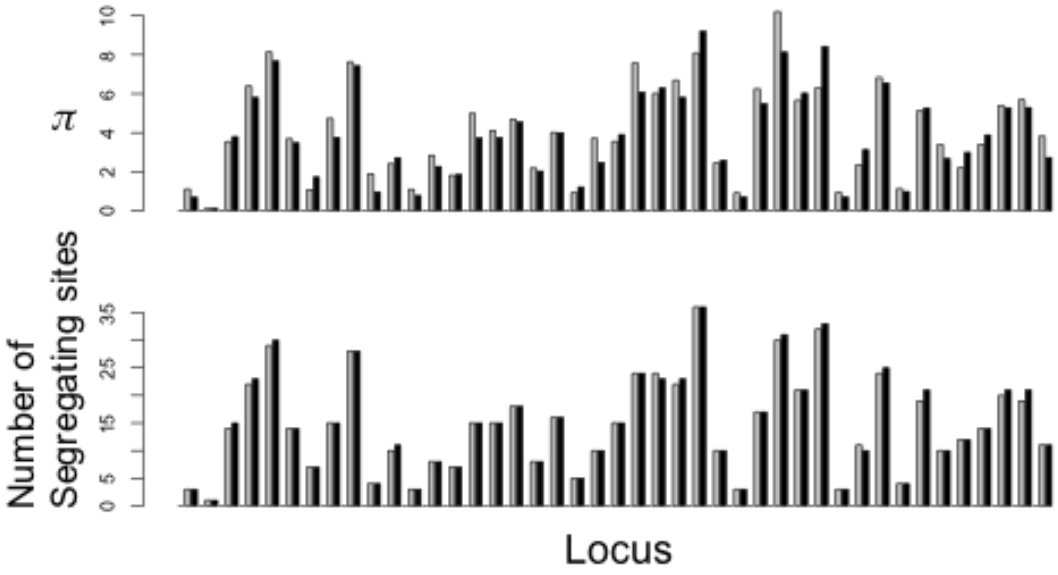


Supplementary Figure S2: Comparison of M form loci modeled under the MLE to M form empirical data. We conducted 10^4 simulations using the program *ms* under the MLE migration model for each 3rd chromosome locus and plotted the median value (gray bars) of the average number of pairwise differences (π) and the number of segregating sites (S) next the empirical value (black bars) of each statistic for that locus. No intralocus recombination was included in the simulations. Loci are ordered according to their

relative positions on the 3rd chromosome.



Supplementary Figure S3: Comparison of S form loci modeled under the MLE to S form empirical data. We conducted 10^4 simulations using the program *ms* under the MLE migration model for each 3rd chromosome locus and plotted the median value (gray bars) of the average number of pairwise differences (π) and the number of segregating sites (S) next the empirical value (black bars) of each statistic for that locus. No intralocus recombination was included in the simulations. Loci are ordered according to their relative positions on the 3rd chromosome.


π
0 2 4 6 8 10
Number of
Segregating sites
0 5 15 25 35
Locus

CHAPTER 3:

No evidence for positive selection at two malaria Transmission-Blocking Vaccine target molecules in *Anopheles gambiae s.s.*

Jacob E. Crawford, Susan Rottschaefer, Brian P. Lazzaro

ABSTRACT:

Human malaria causes nearly a million deaths in sub-Saharan Africa each year, and the development of parasitic drug-resistance and mosquito vector insecticide resistance has complicated control measures and made the need for new control strategies more urgent. *Anopheles gambiae s.s.* is one of the primary vectors of human malaria in Africa, and parasite-transmission-blocking vaccines targeting *Anopheles* proteins have been proposed as a possible strategy to control the spread of the disease. However, the success of these hypothetical technologies would depend on the ability to successfully target potentially heterogeneous mosquito populations. Understanding the evolutionary pressures shaping genetic variation among candidate target molecules offers a first step towards evaluating the prospects of successfully deploying such technologies. We studied the population genetics of two candidate target molecules, the salivary gland protein saglin and the basil lamina structural protein laminin, in wild populations of the M and S molecular forms of *A. gambiae* in Mali. Through analysis of intraspecific genetic variation and interspecific comparisons, we found no evidence of positive natural selection at the genes encoding these proteins. On the contrary, we found evidence for particularly strong purifying selection at one protein, and we discuss these findings in relation to the potential development of these molecules as vaccine targets.

INTRODUCTION:

Mosquito-stage, transmission blocking vaccines have been proposed as an alternative novel technology for blocking transmission of human malaria parasites (Gwadz 1976; Carter and Chen 1976; Barreau et al. 1995; Brennan et al. 2000). The logic of this technology is that mosquito proteins essential for parasite development could be targeted, for example by human derived anti-bodies, and blocked such that transmission is halted within the mosquito. At the heart of this hypothetical technology is the assumption that the vaccine target (i.e. mosquito proteins) can be easily and exhaustively identified and blocked by anti-bodies, but genetic variation segregating in natural populations could result in a heterogeneous molecule population that is difficult to target. Such genetic variation could be neutral with respect to natural selection or it could evolve adaptively if some functional alleles are favored over others in resisting parasite establishment. Substantial traction has been made in identifying potential target molecules within the mosquito, but mostly through the use of genetically inbred lab strains of both the mosquito and the parasite (Brennan et al. 2000; Arrighi et al. 2005; Saul 2007; Dinglasan and Jacobs-Lorena 2008), providing little information on potential genetic heterogeneity and its consequences in nature. Understanding the evolutionary pressures shaping genetic variation among candidate target molecules offers a first step towards evaluating the prospects of successfully deploying such technologies.

The suggestions that pathogen-related selection pressures are likely to drive host evolution date back 100 years (Biffen 1905; Haldane 1949; Lederberg 1999). As one of only a very few permissive and common vectors of the human malaria parasite *Plasmodium falciparum*, *Anopheles gambiae sensu stricto* is a good candidate for experiencing such selection. Several non-immune mosquito proteins directly interact with the developing malaria parasite and are in some cases required for successful

parasite tissue invasion and development. One of the first *Anopheles* proteins shown interact directly with both ookinete and oocyst stages of *Plasmodium* parasites is laminin, a component of basil laminae in the mosquito, including the one surrounding the midgut (Adini and Warburg 1999; D Vlachou et al. 2001; Arrighi and Hurd 2002; Dessens et al. 2003). Subsequent studies of laminin suggests that this protein may act as a trigger for the transition from ookinete to oocyst development or even as a protective coating that masks the parasite from immune detection (Arrighi et al. 2005; Warburg et al. 2007). Further evidence for the intimate and perhaps protective role of laminin was provided by the observations that laminin becomes localized within the oocysts and sporozoites, and it is incorporated into the oocyst capsule (Nacer, Walker, and Hurd 2008).

A second host protein, saglin, plays a crucial role in parasite localization and invasion of the salivary glands through a receptor-ligand interaction with the *Plasmodium* TRAP protein (Brennan et al. 2000; Korochkina et al. 2006; Okulate et al. 2007; Anil K. Ghosh et al. 2009). Blocking this interaction with either antibody interference or receptor saturation by SM1, a short peptide whose physical conformation resembles TRAP, inhibits parasitic salivary gland invasion (Brennan et al. 2000; A K Ghosh, Ribolla, and Jacobs-Lorena 2001; Anil K. Ghosh et al. 2009). Moreover, point mutations in TRAP completely abrogate gland invasion (Matuschewski et al. 2002). Population genetic analysis of *Plasmodium falciparum* and *Plasmodium vivax* suggests adaptive maintenance of variation, especially in the A-domain that binds to saglin (Weedall et al. 2007; Barry et al. 2009).

Based on the evidence that laminin and saglin mediate *Plasmodium* infection in *Anopheles*, we hypothesized that these proteins may be under pathogen-related selection pressure. To address this hypothesis, we sequenced alleles of the genes

coding for these proteins in wild populations of the two incipient species of *Anopheles gambiae*, the M and S molecular forms from Mali. We analyzed patterns of intraspecific polymorphism and divergence at these loci but found no significant evidence for non-neutral evolution at these loci in either population.

MATERIALS AND METHODS:

Mosquito Samples

Anopheles gambiae individuals were collected inside dwellings from the villages of Bancoumana and N'gabakoro Droit outside the Malian capital city, Bamako (12°39′N 8°0′W), and an additional collection was drawn from Toumani-Oulena, Mali (10°83′N 7°81′W). The M/S molecular form of each individual mosquito was determined using the PCR diagnostic developed by Favia et al. (2001). Of the mosquitoes sampled from Bancoumana, four were M form and eleven were S form. All mosquitoes sampled from N'gabakoro Droit were M form (n = 10), and all Toumani-Oulena individuals were S form (n = 7). *Anopheles merus* DNA from mosquitoes of the OPHANSI colony was obtained from MR4.

DNA extraction, PCR and sequencing

DNA was extracted from the mosquitoes using DNeasy kits (Qiagen) under slight modifications to the manufacturers' suggested protocols. PCR primers were designed based on the published *A. gambiae* genome sequence (Holt et al. 2002). Each gene was amplified from genomic 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 the PureLink gel extraction kit (Invitrogen). Adenosine tails

were added to the purified products by incubating for 20 minutes at 72°C with PCR buffer, dATP and Taq polymerase. Products were then cloned using the TOPO XL cloning kit (Invitrogen). Colonies to be sequenced were grown overnight at 37°C 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 then sequenced directly from the plasmids using the BigDye Terminator Cycle Sequencing Kit v3.1(ABI). The sequences were assembled using Sequencher (Gene Codes Corp.) and CodonCode Aligner (CodonCode Corporation). Only one of the two alleles at each gene was sequenced from any given mosquito in the study. All sequences have been deposited into Genbank under accession numbers XXXXXX – XXXXXX.

To control for sequencing error, all singleton polymorphisms were verified by re-amplification and direct sequencing of heterozygous PCR products. The entire gene was amplified from genomic DNA using iProof high fidelity DNA Polymerase (BioRad) and this full-length amplicon was then used as template in a secondary PCR using 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 then sequenced using the BigDye Terminator Cycle Sequencing Kit v3.1(ABI). To further avoid errors stemming from homopolymers, we deleted all homopolymer sequences from the alignment.

The two loci were sometimes sequenced from different individuals within the Bancoumana and N'gabakoro Droit populations, and only *SAG* sequences were generated from the Toumani-Oulena population.

Loci Analyzed

We analyzed the *SAG* locus (AGAP000610), which is X-linked, and *LANB2* (AGAP007629), which is found on the distal tip of chromosome 2L.. The original predicted exon structure of *SAG* in Ensemble was changed to remove a predicted intron leaving a single coding sequence (Brennan et al. 2000). However, to be certain, we obtained cDNA from live *A. gambiae* females and directly sequenced *SAG* transcripts to identify intronic boundaries if they exist. We used the same PCR, cloning, and sequencing conditions and reagents described above, only first-strand cDNA from whole mosquitoes was used as template in the initial PCR reaction. After aligning the cDNA sequence to the Agam PEST reference sequence, we identified a single 175 basepair (bp) intron beginning at position 1007 of the coding sequence, consistent with the original Ensemble annotation prior to the re-annotation based on short peptide mapping (Brennan et al. 2000). According to our sequencing results, the final saglin protein is predicted to be 374 amino acids in length, and for the analyses here, we assumed that the putative 175 bp intron is non-coding and analyzed the sequence accordingly. For *LANB2*, we analyzed the genomic region according to the exon structure annotated in Vectorbase (www.vectorbase.org)

Population genetic analysis

Measures of nucleotide diversity estimated from the average number of differences between haplotypes (π) and the number of polymorphic sites (θ_W) were calculated on synonymous and non-synonymous sites alone as well as on all sites combined using DnaSP version 5 (Librado and Rozas 2009). Three neutrality tests that emphasize different features of the data including Tajima's *D* (Tajima 1989), a normalized version of Fay and Wu's *H* (Fay and Wu 2000; Zeng et al. 2006), and Ewens-Watterson's

haplotype homozygosity statistic (*EW*) were calculated using a program kindly provided by K. Zeng. Tajima's *D* detects significant excesses of rare SNPs or intermediate frequency SNPs, normalized *H* detects excesses of high-frequency derived SNPs, and *EW* detects excess homozygosity between haplotypes, all of which are deviations from the neutral equilibrium model expected under a hitchhiking model of positive selection. Three compound statistics (*DH*, *HEW*, *DHEW*) that combine the *p*-values of these neutrality tests were also calculated. These compound statistics were developed to take advantage of each of the features from the different neutrality tests and are more robust to confounding factors such as demography and background selection (Zeng et al. 2006; Zeng, Shi, and Wu 2007; Zeng et al. 2007). The statistical significance of the neutrality tests and compound statistics was evaluated through comparisons to 10^5 neutral coalescent simulations without recombination conditioned on the sample size and θ_W estimated from the data conducted using the program from K. Zeng.

We also used tests to identify patterns of potentially adaptive rates of evolution at non-synonymous sites as would be expected under a model of repeated episodes of historical positive selection at these genes. First, we applied the McDonald-Kreitman test (McDonald and Kreitman 1991) within DnaSP using *A. merus* as an outgroup and evaluating significance using Fisher's exact test. For *SAG*, we applied this test first to *SAG* alone then to all three genes in the sequenced region (*SG1-2*, *SAG*, *gSG1a*). We also calculated the ratio of evolutionary rates at non-synonymous to synonymous sites (K_a/K_s) within DnaSP using *A. merus* for estimates of divergence. This ratio is expected to equal one under a completely neutral model and to be greater than 1 when positive selection has caused repeated adaptive evolution. However, purifying selection often acts to conserve amino acid state such that the rate of substitutions at non-synonymous sites tends to be much lower than that of synonymous sites. As such, K_a/K_s is often less

than 1 when considered across an entire gene or gene region due to the widespread effects of purifying selection at most positions. Thus, the K_a/K_s test is conservative. To circumvent the obscuring effects of purifying selection, we used a sliding window approach to localize any potential cluster of rapidly evolving sites. For both datasets, we used a window size of 1200 basepairs (bp) with a step size of 500bp.

RESULTS:

Saglin

To test the hypothesis that saglin (*SAG*) has experienced recent positive selection, we re-sequenced this gene in the M and S molecular forms and used population genetic analyses to probe patterns of inter- and intra-species genetic variation at these genes. We sequenced a 3.7 kilobase (kb) region that includes the *SAG* gene as well as 1.6kb of sequence upstream and 0.8kb of sequence downstream of the *SAG* coding region. Saglin is one of several members of the SG1 protein family of salivary gland proteins arrayed on the X chromosome (Arcá et al. 1999; Lanfrancotti et al. 2002; Arcà et al. 2005), and in addition to *SAG*, we captured the complete coding sequence of *SG1-2* (upstream of *SAG*) and partial coding sequence of the downstream paralog *gSG1a* within our 3.7kb sequenced region. Across the entire region, nucleotide diversity is low in both populations, consistent with the expectations of low diversity on the X chromosome in *A. gambiae* (e.g. Cohuet et al. 2008). Analysis of each paralog and the intergenic regions separately revealed that diversity is approximately equal across the sequence except at *SAG* and the upstream intergenic region (data not shown). Non-synonymous diversity is particularly rare at the *SAG* gene in these populations. The ratio of non-synonymous to synonymous nucleotide diversity (π_{NS}/π_S) is 0.059 in the M form population and 0.096 in the S form population. In contrast, this ratio equals

approximately 0.5 in both populations for SG1-2, the paralog only 300bp upstream. If we assume that the mutation rate is not especially low at *SAG* since diversity at synonymous sites is comparable across the region, the relatively low π_{NS}/π_S at this locus implies particularly strong effects of purifying selection acting on this protein.

The process of genetic hitchhiking that affects genetic variation linked to a positively selected site in a population modifies linked nucleotides in a number of hallmark ways, including increase in the physical scale of observed linkage disequilibrium (LD) and shifts in the site-frequency spectrum of allele frequencies at polymorphic sites (Smith and Haigh 1974; Braverman et al. 1995; Przeworski 2002). Statistical tests have been developed to detect such genomic footprints (e.g. Tajima 1989; Fay and Wu 2000; Kim and Stephan 2002). To test for the presence of hitchhiking at *SAG*, we used a compound statistic, *HEW* (Zeng, Shi, and Wu 2007), that combines a site-frequency spectrum based test, Fay and Wu's *H* (Fay and Wu 2000), and a haplotype-based test, *EW* (Watterson 1978). We are specifically aiming to detect scenarios where genetic hitchhiking has resulted in an excess of high-frequency derived variants and an increase in haplotype structure. When we applied this test to the entire sequenced region, we found that patterns of genetic variation in the M form were consistent with neutrality (*HEW*: $p = 1.0$), but the S form harbored patterns of variation that were nearly significant (*HEW*: $p = 0.0507$) and may reflect the effects of recent positive selection. This S form signal appears to be driven largely by an increase in haplotype structure that is unexpected under a neutral model, although not significantly so (*EW* = 0.2089, $p = 0.0712$; Table 1). Scanning this region using a sliding-window approach to localize the signal indicated that genetic variation downstream of *SAG*, including part of *gSG1a*, harbors the most significant deviations from neutrality (*HEW*:

uncorrected $p = 0.0267$), while the windows that include *SAG* and *SG1-2* do not reject the neutral model (*HEW*: $p > 0.05$; Figure 1).

We applied the McDonald-Kreitman test to these data for the signature of historical selection using *A. merus* as the outgroup species and found no evidence for an excess of non-synonymous fixed differences for either population (M form $p = 0.4892$; S form $p = 0.2757$; Table 2) as would be expected under a model of recurrent direction selection at this locus. Moreover, a comparison of the rates of substitution at synonymous and non-synonymous sites (K_a/K_s) confirms purifying selection as the predominant mode of evolution at *SAG*. If multiple episodes of positive selection has fixed non-synonymous sites at this locus, the K_a/K_s ratio would be expected to exceed one (Hughes and Nei 1988). However, $K_a/K_s = 0.294$ for the S form and $K_a/K_s = 0.303$ for the M form, reflecting fewer fixations at non-synonymous sites relative to synonymous sites in both subpopulations, consistent with the operation of purifying selection on the locus. Taken together, these data provide no evidence for recent or more ancient natural selection at the *SAG* gene. We sequenced only a small portion of the coding sequence for *gSG1a*, so more data will be required to determine whether this gene could be under positive selection, perhaps associated with a role in blood-feeding (Arcá et al. 1999; Lanfrancotti et al. 2002; Arcà et al. 2005).

LANB2

We tested the hypothesis that *LANB2* is under positive selection by analyzing both intraspecific and interspecific genetic variation. *LANB2* is a relatively large gene composed of nine exons that span approximately 8kb of the distal tip of the left arm of chromosome 2. We sequenced an approximately 10kb region that includes *LANB2* as well as 497bp of sequence upstream and 789bp of sequence downstream of *LANB2*.

Consistent with other estimates of nucleotide diversity at *A. gambiae* autosomal loci (Cohuet et al. 2008), *LANB2* harbors substantial genetic diversity in both populations with per site estimates of diversity (θ_W) equaling 0.0180 and 0.0175 for the M and S molecular forms, respectively, for all functional classes of nucleotide sites (Table 1). The vast majority of this variation maps to synonymous sites, however, and non-synonymous variation is nearly absent (Table 1).

We tested *LANB2* for evidence of recent positive selection. As for *SAG*, we applied the compound *HEW* test statistic (Zeng, Shi, and Wu 2007) using *A. merus* as the outgroup species. When applied to the entire region, *HEW* is not significant for either population (p_M = 1.0; p_S = 1.0; Table 1). To rule out the possibility that a smaller region within the 10kb sequenced region could have experienced positive selection and is being masked by neutral patterns in the larger surrounding sequence, we calculated *HEW* using a sliding-window of 1200 bp with a step size of 500 bases, but this approach failed to reveal any significant windows (all $p > 0.05$). Our inability to reject a neutral equilibrium model at this locus suggests that the protein coding sequence of *LANB2* has been evolving neutrally or under purifying selection in the recent past.

We also used interspecies comparative analyses to determine whether this gene has experienced repeated selective sweeps in the more distant past. First, we applied the McDonald-Kreitman test to these data for each population independently using *A. merus* as the outgroup but found no evidence of deviations from neutral expectations (M form p = 0.1367; S form p = 0.1846; Table 2). Second, we used K_a/K_s to test for accelerating rates of evolution at non-synonymous sites. Since the K_a/K_s test is conservative across large spans of sequence, we used a sliding window approach as described above. We found that the maximum K_a/K_s value never exceeded 0.1 for either molecular form in any window (1200bp window with maximum K_a/K_s = 0.046 for S and

0.053 for M). Coupled with the results from intra-specific analyses, these data provide no evidence for positive selection on *LANB2*.

DISCUSSION

Structural and receptor host proteins can also be intimately involved in host-pathogen interactions, for example related to their crucial role in tissue recognition and invasion (Brennan et al. 2000; Arrighi et al. 2005). Like immune proteins, non-immune proteins with a direct role in pathogen development within the host could be subject to pathogen-related selection pressure. One such case of non-immune pathogen-related selection has been documented in humans at the Duffy erythrocyte receptor protein, which is exploited by the malaria parasite for merozoite invasion (Hamblin and Di Rienzo 2000), and HIV-related selection is thought to be currently driving the evolution of the chemokine CCR5 in Africa (Schliekelman, Garner, and Slatkin 2001). Several *A. gambiae* proteins have been identified as playing roles in the establishment or development of malaria parasites within the mosquito host (Brennan et al. 2000; Arrighi et al. 2005; Rodrigues et al. 2012). We tested the genes that code for two of these proteins, *saglin* and *laminin*, for the signatures of *Plasmodium*-related selection pressure. To address hypotheses of selection at these genes, we re-sequenced the genomic regions harboring these genes in a small population sample of both the M and S molecular forms of *A. gambiae* and applied population genetic tests to identify significant deviations from expectations under a neutral model. Neither intra-specific analyses based on the site-frequency spectrum of genetic variation, nor inter-specific comparisons designed to detecte accelerated rates of evolution at non-synonymous sites revealed any evidence for non-neutral evolution at these loci. The *HEW* neutrality test failed to reject the neutral model, confirming that the number of high-frequency

derived alleles and the degree of haplotype homozygosity are both consistent with expectations under a neutral model. Inter-specific comparisons with *A. merus* also revealed patterns expected under models of purifying selection and failed to provide any evidence for adaptive evolution in the protein coding sequences.

It is possible that the population genetic tests may be underpowered in this study because of the limited number of chromosomes sampled and the limited divergence between *A. gambiae* and *A. merus*. Small sample sizes lead to increased sampling variance in population genetic analyses and also limit the amount of information available for analysis (i.e. number of segregating sites), in turn potentially resulting in false negative test results (Simonsen, Churchill, and Aquadro 1995). This may be of particular concern for *SAG* since levels of diversity are particularly low in this gene region, partly owing to its location on the X chromosome, which is known to be depauperate of diversity in *A. gambiae* (Cohuet et al. 2008; Lawniczak et al. 2010; Neafsey et al. 2010). This issue was partially mitigated by sequencing a larger physical region to capture both additional segregating sites as well as the spatial distribution of diversity. Moreover, statistical tests based on divergence from an outgroup such as the McDonald-Kreitman test may be generally underpowered because within the *Anopheles gambiae* species complex because of the very recent divergence of the species (Besansky et al. 2003; Darren J Obbard, Welch, and Little 2009), violating assumptions of the model upon which these tests are based (McDonald and Kreitman 1991). These violations are of most concern when comparing the sister species *A. gambiae* and *Anopheles arabiensis*, however, and previous studies have suggested that divergence between *A. merus* and *A. gambiae* may be sufficient (synonymous divergence ranging K_s = 4 - 11%) to allow the proper application of divergence based tests (Wang-Sattler et al. 2007; D J Obbard et al. 2007; Parmakelis et al. 2008; Darren J Obbard, Welch, and

Little 2009; Rottschaefer et al. 2011). Nonetheless, some caution is warranted when interpreting the results of these tests since low statistical power could have led to false positives and/or false negatives.

Our results lead us to accept the null hypothesis of neutral evolution at these loci, suggesting that the proteins coded by these loci are not involved in the evolutionary host-parasite conflict, or at least are not evolving in response. One alternate hypothesis to explain these data could be that both saglin and laminin serve a critical purpose, structural or otherwise, that requires the protein structure to remain conserved, consistent with the evidence of particularly strong purifying selection at *LANB2* (Table 1). Overall, these results suggest that both proteins would be reliable candidates for interventions such as transmission blocking vaccines (Brennan et al. 2000; Dinglasan and Jacobs-Lorena 2008). If vaccines are to be developed to directly target either of these proteins, it would be detrimental to the technology if the proteins were adaptively evolving or were adaptively maintaining functionally variable alleles in natural populations, neither of which is the case here. Although *SAG* harbors a substantial number of fixed differences from the outgroup, the level of non-synonymous variations currently segregating in the population is very low for both proteins. This suggests that technologies could easily be developed to target existing alleles, and the technologies could be expected to remain effective in at least the natural populations sampled here.

TABLES

Table 1: Nucleotide diversity and neutrality tests for *SAG* and *LANB2*

	n	**S**	**θ_W**	**π**	**π_S**	**π_{NS}**	***H***	***EW***	***HEW***
SAG									
M form	14	59	0.0051	0.0044	0.0092	0.0005	-1.0300	0.0769	1.0
S form	15	58	0.0048	0.0049	0.0081	0.0008	-0.3680	0.2089	0.0507
LANB2									
M form	14	480	0.0180	0.0157	0.0369	0.0002	0.4972	0.0714	1.0
S form	11	412	0.0175	0.0151	0.0358	0.0002	0.2954	0.0909	1.0

n – Number of chromosomes sampled per population.
S – Total number of segregating sites detected in each sample.
θ_W – Per site Watterson's Theta calculated on all segregating sites.
π – Average number of pairwise differences (nucleotide diversity) calculated on all segregating sites.
π_S – nucleotide diversity at synonymous sites only.
π_{NS} – nucleotide diversity at non-synonymous sites only.

Table 2: McDonald-Kreitman test results for both *SAG* and *LANB2*

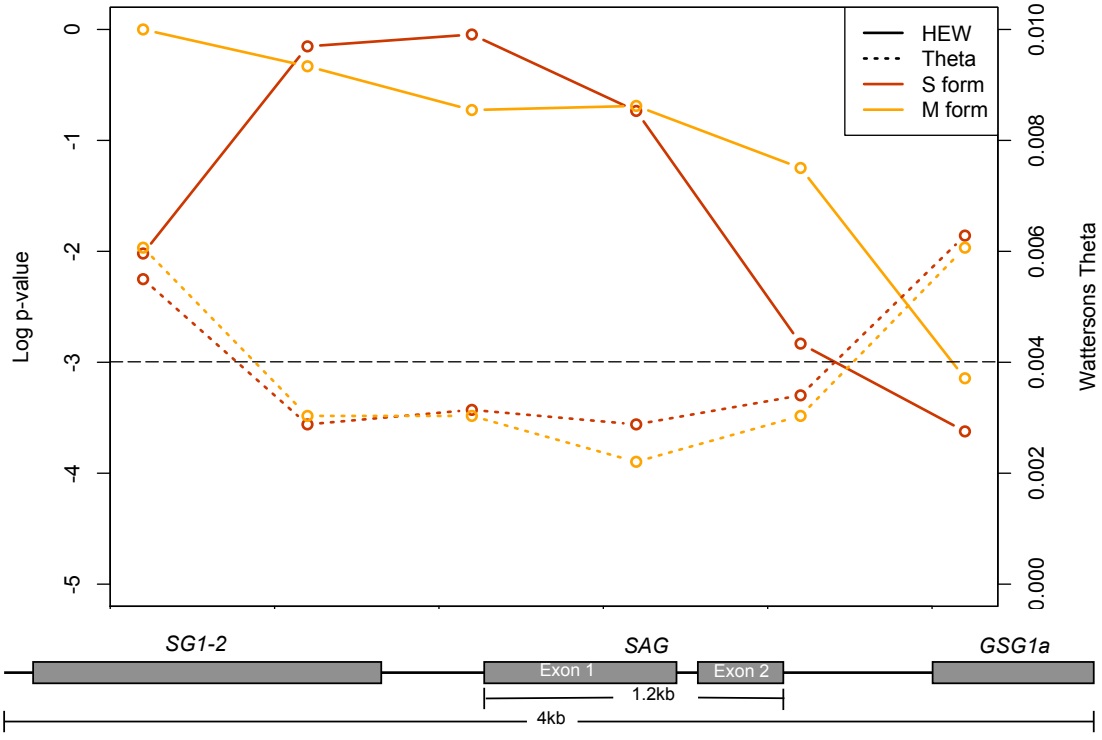
	Polymorphic sites		Fixed sites		
	Synonymous	Non-Synonymous	Synonymous	Non-Synonymous	*p*-value
S form					
SAG[a]	7	2	27	25	0.2757
All[b]	19	17	64	77	0.4587
M form					
SAG[a]	7	3	27	24	0.4892
All[b]	17	20	64	77	1.0
S form					
LANB2	151	4	22	2	0.1846
M form					
LANB2	162	4	19	2	0.1367

a – Test applied to all coding regions in the sequenced region (*SG1-2*, *SAG*, *gSG1a*)
b – Test applied to *SAG* alone.
Significance of contingency table evaluated using Fisher's exact test, no corrections for multiple testing applied.

FIGURES:

Figure 1: Sliding window analysis of *SAG* using *HEW* neutrality test.

The compound neutrality test statistic *HEW* plotted as a function of position across sequenced region containing three genes. *HEW* was calculated in a sliding window analysis using a 1200bp window with a 500bp step size. The schematic below the plot indicates the location of each gene in the region. Both the uncorrected log transformed *HEW p*-values as well as corresponding per site Watterson's Theta value for each window is plotted and the dotted line indicates 0.05 threshold such that any point below the line indicates a value that satisfies the criteria for statistical significance.



REFERENCES:

Adini, A, and A Warburg. 1999. “Interaction of Plasmodium Gallinaceum Ookinetes and Oocysts with Extracellular Matrix Proteins.” *Parasitology* 119 (Pt 4) (October): 331–336.

Aly, Ahmed S I, Ashley M Vaughan, and Stefan H I Kappe. 2009. “Malaria Parasite Development in the Mosquito and Infection of the Mammalian Host.” *Annual Review of Microbiology* 63: 195–221. doi:10.1146/annurev.micro.091208.073403.

Arcá, B, F Lombardo, M de Lara Capurro, A della Torre, G Dimopoulos, A A James, and M Coluzzi. 1999. “Trapping cDNAs Encoding Secreted Proteins from the Salivary Glands of the Malaria Vector Anopheles Gambiae.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (4) (February 16): 1516–1521.

Arcà, Bruno, Fabrizio Lombardo, Jesus G Valenzuela, Ivo M B Francischetti, Osvaldo Marinotti, Mario Coluzzi, and José M C Ribeiro. 2005. “An Updated Catalogue of Salivary Gland Transcripts in the Adult Female Mosquito, Anopheles Gambiae.” *The Journal of Experimental Biology* 208 (Pt 20) (October): 3971–3986. doi:10.1242/jeb.01849.

Arrighi, Romanico B G, and Hilary Hurd. 2002. “The Role of Plasmodium Berghei Ookinete Proteins in Binding to Basal Lamina Components and Transformation into Oocysts.” *International Journal for Parasitology* 32 (1) (January): 91–98.

Arrighi, Romanico B G, Gareth Lycett, Vassiliki Mahairaki, Inga Siden-Kiamos, and Christos Louis. 2005. “Laminin and the Malaria Parasite’s Journey Through the Mosquito Midgut.” *The Journal of Experimental Biology* 208 (Pt 13) (July): 2497–2502. doi:10.1242/jeb.01664.

Barreau, C, M Touray, P F Pimenta, L H Miller, and K D Vernick. 1995. “Plasmodium Gallinaceum: Sporozoite Invasion of Aedes Aegypti Salivary Glands Is Inhibited by Anti-gland Antibodies and by Lectins.” *Experimental Parasitology* 81 (3) (November): 332–343. doi:10.1006/expr.1995.1124.

Barry, Alyssa E, Lee Schultz, Caroline O Buckee, and John C Reeder. 2009. “Contrasting Population Structures of the Genes Encoding Ten Leading Vaccine-candidate Antigens of the Human Malaria Parasite, Plasmodium Falciparum.” *PloS One* 4 (12): e8497. doi:10.1371/journal.pone.0008497.

Besansky, N J, J Krzywinski, T Lehmann, F Simard, M Kern, O Mukabayire, D Fontenille, Y Touré, and N’F Sagnon. 2003. “Semipermeable Species Boundaries Between Anopheles Gambiae and Anopheles Arabiensis: Evidence from Multilocus DNA Sequence Variation.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (19) (September 16): 10818–10823. doi:10.1073/pnas.1434337100.

Biffen, R. H. 1905. “Mendel’s Laws of Inheritance and Wheat Breeding.” *The Journal of Agricultural Science* 1 (01): 4–48. doi:10.1017/S0021859600000137.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. “The Hitchhiking Effect on the Site Frequency Spectrum of DNA Polymorphisms.” *Genetics* 140 (2): 783–796.

Brennan, J D, M Kent, R Dhar, H Fujioka, and N Kumar. 2000. “Anopheles Gambiae Salivary Gland Proteins as Putative Targets for Blocking Transmission of Malaria Parasites.” *Proceedings of the National Academy of Sciences of the United*

States of America 97 (25) (December 5): 13859–13864. doi:10.1073/pnas.250472597.

Carter, Richard, and David H. Chen. 1976. “Malaria Transmission Blocked by Immunisation with Gametes of the Malaria Parasite.” *, Published Online: 02 September 1976; | Doi:10.1038/263057a0* 263 (5572) (September 2): 57–60. doi:10.1038/263057a0.

Cohuet, Anna, Sujatha Krishnakumar, Frédéric Simard, Isabelle Morlais, Anastasios Koutsos, Didier Fontenille, Michael Mindrinos, and Fotis C Kafatos. 2008. “SNP Discovery and Molecular Evolution in Anopheles Gambiae, with Special Emphasis on Innate Immune System.” *BMC Genomics* 9: 227. doi:10.1186/1471-2164-9-227.

Dessens, Johannes T, Inga Sidén-Kiamos, Jacqui Mendoza, Vassiliki Mahairaki, Emad Khater, Dina Vlachou, Xiao-Jin Xu, et al. 2003. “SOAP, a Novel Malaria Ookinete Protein Involved in Mosquito Midgut Invasion and Oocyst Development.” *Molecular Microbiology* 49 (2) (July): 319–329.

Dinglasan, Rhoel R, and Marcelo Jacobs-Lorena. 2008. “Flipping the Paradigm on Malaria Transmission-blocking Vaccines.” *Trends in Parasitology* 24 (8) (August): 364–370. doi:10.1016/j.pt.2008.05.002.

Favia, G, A Lanfrancotti, L Spanos, I Sidén-Kiamos, and C Louis. 2001. “Molecular Characterization of Ribosomal DNA Polymorphisms Discriminating Among Chromosomal Forms of Anopheles Gambiae S.s.” *Insect Molecular Biology* 10 (1) (February): 19–23.

Fay, Justin C, and Chung-I Wu. 2000. “Hitchhiking Under Positive Darwinian Selection.” *Genetics* 155 (3) (July 1): 1405–1413.

Ghosh, A K, P E Ribolla, and M Jacobs-Lorena. 2001. “Targeting Plasmodium Ligands on Mosquito Salivary Glands and Midgut with a Phage Display Peptide Library.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (23) (November 6): 13278–13281. doi:10.1073/pnas.241491198.

Ghosh, Anil K., Martin Devenport, Deepa Jethwaney, Dario E. Kalume, Akhilesh Pandey, Vernon E. Anderson, Ali A. Sultan, Nirbhay Kumar, and Marcelo Jacobs-Lorena. 2009. “Malaria Parasite Invasion of the Mosquito Salivary Gland Requires Interaction Between the Plasmodium TRAP and the Anopheles Saglin Proteins.” Ed. David S. Schneider. *PLoS Pathogens* 5 (1) (January 16): e1000265. doi:10.1371/journal.ppat.1000265.

Gwadz, R. W. 1976. “Successful Immunization Against the Sexual Stages of Plasmodium Gallinaceum.” *Science* 193 (4258) (September 17): 1150–1151. doi:10.1126/science.959832.

Haldane, J.B.S. 1949. “Disease and Evolution.” *La Ricerca Scientifica Supplemento A* 19: 68–76.

Hamblin, M T, and A Di Rienzo. 2000. “Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus.” *American Journal of Human Genetics* 66 (5) (May): 1669–1679.

Holt, Robert A, G Mani Subramanian, Aaron Halpern, Granger G Sutton, Rosane Charlab, Deborah R Nusskern, Patrick Wincker, et al. 2002. “The Genome Sequence of the Malaria Mosquito Anopheles Gambiae.” *Science (New York, N.Y.)* 298 (5591) (October 4): 129–149. doi:10.1126/science.1076181.

Hughes, A L, and M Nei. 1988. “Pattern of Nucleotide Substitution at Major Histocompatibility Complex Class I Loci Reveals Overdominant Selection.” *Nature* 335 (6186) (September 8): 167–170. doi:10.1038/335167a0.

Kim, Y., and W. Stephan. 2002. “Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome.” *Genetics* 160 (2): 765.

Korochkina, S., C. Barreau, G. Pradel, E. Jeffery, J. Li, R. Natarajan, J. Shabanowitz, D. Hunt, U. Frevert, and K. D Vernick. 2006. “A Mosquito-specific Protein Family Includes Candidate Receptors for Malaria Sporozoite Invasion of Salivary Glands.” *Cellular Microbiology* 8 (1): 163–175.

Lanfrancotti, Alessandra, Fabrizio Lombardo, Federica Santolamazza, Massimiliano Veneri, Tiziana Castrignanò, Mario Coluzzi, and Bruno Arcà. 2002. “Novel cDNAs Encoding Salivary Proteins from the Malaria Vector Anopheles Gambiae.” *FEBS Letters* 517 (1-3) (April 24): 67–71.

Lawniczak, M K N, S J Emrich, A K Holloway, A P Regier, M Olson, B White, S Redmond, et al. 2010. “Widespread Divergence Between Incipient Anopheles Gambiae Species Revealed by Whole Genome Sequences.” *Science (New York, N.Y.)* 330 (6003) (October 22): 512–514. doi:10.1126/science.1195755.

Lederberg, Joshua. 1999. “J. B. S. Haldane (1949) on Infectious Disease and Evolution.” *Genetics* 153 (1) (September 1): 1–3.

Lehmann, Tovi, Jen C C Hume, Monica Licht, Christopher S Burns, Kurt Wollenberg, Fred Simard, and Jose’ M C Ribeiro. 2009. “Molecular Evolution of Immune Genes in the Malaria Mosquito Anopheles Gambiae.” *PloS One* 4 (2): e4549. doi:10.1371/journal.pone.0004549.

Librado, P, and J Rozas. 2009. “DnaSP V5: a Software for Comprehensive Analysis of DNA Polymorphism Data.” *Bioinformatics (Oxford, England)* 25 (11) (June 1): 1451–1452. doi:10.1093/bioinformatics/btp187.

Matuschewski, Kai, Alvaro C Nunes, Victor Nussenzweig, and Robert Ménard. 2002. “Plasmodium Sporozoite Invasion into Insect and Mammalian Cells Is Directed by the Same Dual Binding System.” *The EMBO Journal* 21 (7) (April 2): 1597–1606. doi:10.1093/emboj/21.7.1597.

McDonald, J H, and M Kreitman. 1991. “Adaptive Protein Evolution at the Adh Locus in Drosophila.” *Nature* 351 (6328) (June 20): 652–654. doi:10.1038/351652a0.

Nacer, Adéla, Karen Walker, and Hilary Hurd. 2008. “Localisation of Laminin Within Plasmodium Berghei Oocysts and the Midgut Epithelial Cells of Anopheles Stephensi.” *Parasites & Vectors* 1 (1): 33. doi:10.1186/1756-3305-1-33.

Neafsey, D E, M K N Lawniczak, D J Park, S N Redmond, M B Coulibaly, S F Traoré, N Sagnon, et al. 2010. “SNP Genotyping Defines Complex Gene-flow Boundaries Among African Malaria Vector Mosquitoes.” *Science (New York, N.Y.)* 330 (6003) (October 22): 514–517. doi:10.1126/science.1193036.

Niaré, Oumou, Kyriacos Markianos, Jennifer Volz, Frederick Oduol, Abdoulaye Touré, Magaran Bagayoko, Djibril Sangaré, et al. 2002. “Genetic Loci Affecting Resistance to Human Malaria Parasites in a West African Mosquito Vector Population.” *Science (New York, N.Y.)* 298 (5591) (October 4): 213–216. doi:10.1126/science.1073420.

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.” *Molecular Ecology* 16 (16) (August): 3497–3510. doi:10.1111/j.1365-294X.2007.03395.x.

Obbard, Darren J, John J Welch, and Tom J Little. 2009. “Inferring Selection in the Anopheles Gambiae Species Complex: An Example from Immune-related Serine Protease Inhibitors.” *Malaria Journal* 8: 117. doi:10.1186/1475-2875-8-117.

Okulate, M. A., D. E. Kalume, R. Reddy, T. Kristiansen, M. Bhattacharyya, R. Chaerkady, A. Pandey, and N. Kumar. 2007. “Identification and Molecular Characterization of a Novel Protein Saglin as a Target of Monoclonal Antibodies Affecting Salivary Gland Infectivity of Plasmodium Sporozoites.” *Insect Molecular Biology* 16 (6): 711–722.

Parmakelis, Aristeidis, Michel A Slotman, Jonathon C Marshall, Parfait H Awono-Ambene, Christophe Antonio-Nkondjio, Frederic Simard, Adalgisa Caccone, and Jeffrey R Powell. 2008. “The Molecular Evolution of Four Anti-malarial Immune Genes in the Anopheles Gambiae Species Complex.” *BMC Evolutionary Biology* 8: 79. doi:10.1186/1471-2148-8-79.

Przeworski, M. 2002. “The Signature of Positive Selection at Randomly Chosen Loci.” *Genetics* 160 (3): 1179–1189.

Riehle, Michelle M, Kyriacos Markianos, Oumou Niaré, Jiannong Xu, Jun Li, Abdoulaye M Touré, Belco Podiougou, et al. 2006. “Natural Malaria Infection in Anopheles Gambiae Is Regulated by a Single Genomic Control Region.” *Science (New York, N.Y.)* 312 (5773) (April 28): 577–579. doi:10.1126/science.1124153.

Rodrigues, Janneth, Giselle A Oliveira, Michalis Kotsyfakis, Rajnikant Dixit, Alvaro Molina-Cruz, Ryan Jochim, and Carolina Barillas-Mury. 2012. “An Epithelial Serine Protease, AgESP, Is Required for Plasmodium Invasion in the Mosquito Anopheles Gambiae.” *PloS One* 7 (4): e35210. doi:10.1371/journal.pone.0035210.

Rottschaefer, Susan M., Michelle M. Riehle, Boubacar Coulibaly, Madjou Sacko, Oumou Niaré, Isabelle Morlais, Sekou F. Traoré, Kenneth D. Vernick, and Brian P. Lazzaro. 2011. “Exceptional Diversity, Maintenance of Polymorphism, and Recent Directional Selection on the APL1 Malaria Resistance Genes of Anopheles Gambiae.” *PLoS Biol* 9 (3) (March 8): e1000600. doi:10.1371/journal.pbio.1000600.

Saul, Allan. 2007. “Mosquito Stage, Transmission Blocking Vaccines for Malaria.” *Current Opinion in Infectious Diseases* 20 (5) (October): 476–481. doi:10.1097/QCO.0b013e3282a95e12.

Schliekelman, Paul, Chad Garner, and Montgomery Slatkin. 2001. “Natural Selection and Resistance to HIV.” *Nature* 411 (6837) (May 31): 545–546. doi:10.1038/35079176.

Simonsen, K L, G A Churchill, and C F Aquadro. 1995. “Properties of Statistical Tests of Neutrality for DNA Polymorphism Data.” *Genetics* 141 (1) (September): 413–429.

Sinden, Robert E, and Peter F Billingsley. 2001. “Plasmodium Invasion of Mosquito Cells: Hawk or Dove?” *Trends in Parasitology* 17 (5) (May 1): 209–211. doi:10.1016/S1471-4922(01)01928-6.

Smith, J. M., and J. Haigh. 1974. “The Hitch-hiking Effect of a Favourable Gene.” *Genet Res* 23 (1): 23–35.

Tajima, F. 1989. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Genetics* 123 (3): 585.

Vlachou, D, G Lycett, I Sidén-Kiamos, C Blass, R E Sinden, and C Louis. 2001. “Anopheles Gambiae Laminin Interacts with the P25 Surface Protein of Plasmodium Berghei Ookinetes.” *Molecular and Biochemical Parasitology* 112 (2) (February): 229–237.

Vlachou, Dina, Timm Schlegelmilch, Ellen Runn, Antonio Mendes, and Fotis C. Kafatos. 2006. “The Developmental Migration of Plasmodium in Mosquitoes.” *Current*

Opinion in Genetics & Development 16 (4) (August): 384–391. doi:10.1016/j.gde.2006.06.012.

Wang-Sattler, Rui, Stephanie Blandin, Ye Ning, Claudia Blass, Guimogo Dolo, Yeya T Touré, Alessandra delle Torre, et al. 2007. "Mosaic Genome Architecture of the Anopheles Gambiae Species Complex." *PloS One* 2 (11): e1249. doi:10.1371/journal.pone.0001249.

Warburg, Alon, Alex Shtern, Noa Cohen, and Noa Dahan. 2007. "Laminin and a Plasmodium Ookinete Surface Protein Inhibit Melanotic Encapsulation of Sephadex Beads in the Hemocoel of Mosquitoes." *Microbes and Infection / Institut Pasteur* 9 (2) (February): 192–199. doi:10.1016/j.micinf.2006.11.006.

Watterson, G. A. 1978. "The Homozygosity Test of Neutrality." *Genetics* 88 (2) (February 1): 405–417.

Weedall, Gareth D, Benjamin M J Preston, Alan W Thomas, Colin J Sutherland, and David J Conway. 2007. "Differential Evidence of Natural Selection on Two Leading Sporozoite Stage Malaria Vaccine Candidate Antigens." *International Journal for Parasitology* 37 (1) (January): 77–85. doi:10.1016/j.ijpara.2006.09.001.

White, Bradley J, Mara K N Lawniczak, Changde Cheng, Mamadou B Coulibaly, Michael D Wilson, N'Fale Sagnon, Carlo Costantini, Frederic Simard, George K Christophides, and Nora J Besansky. 2011. "Adaptive Divergence Between Incipient Species of Anopheles Gambiae Increases Resistance to Plasmodium." *Proceedings of the National Academy of Sciences of the United States of America* 108 (1) (January 4): 244–249. doi:10.1073/pnas.1013648108.

Zeng, Kai, Yun-Xin Fu, Suhua Shi, and Chung-I Wu. 2006. "Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants." *Genetics* 174 (3) (November 1): 1431–1439. doi:10.1534/genetics.106.061432.

Zeng, Kai, Shuhei Mano, Suhua Shi, and Chung-I Wu. 2007. "Comparisons of Site- and Haplotype-frequency Methods for Detecting Positive Selection." *Molecular Biology and Evolution* 24 (7) (July): 1562–1574. doi:10.1093/molbev/msm078.

Zeng, Kai, Suhua Shi, and Chung-I Wu. 2007. "Compound Tests for the Detection of Hitchhiking Under Positive Selection." *Molecular Biology and Evolution* 24 (8) (August): 1898–1908. doi:10.1093/molbev/msm119.

CHAPTER 4:

Evidence for population-specific positive selection on immune genes of *Anopheles gambiae.*

Submitted to Molecular Biology and Evolution

Jacob E. Crawford, Emmanuel Bischoff, Thierry Garnier, Awa Gneme, Karin Eiglmeier, Inge Holm, Michelle Riehle, Wamdaogo M. Guelbeogo, N'fale Sagnon, Brian P. Lazzaro, Kenneth D. Vernick

ABSTRACT:

Host-pathogen interactions can be powerful drivers of adaptive evolution, shaping the patterns of molecular variation at the genes involved. In this study, we sequenced alleles from 28 immune-related loci in wild samples of multiple genetic subpopulations of the African malaria mosquito *Anopheles gambiae*, obtaining unprecedented sample sizes and providing the first opportunity to contrast patterns of molecular evolution at immune-related loci in the recently discovered GOUNDRY population to those of the indoor-collected M and S molecular forms. In contrast to previous studies that focused on immune genes identified in laboratory studies, we centered our analysis on genes that fall within a quantitative trait locus associated with resistance to *P. falciparum* in natural populations of *A. gambiae.* Analyses of haplotypic and genetic diversity at these 28 loci revealed striking differences among populations in levels of genetic diversity and allele frequencies in coding sequence. Moreover, putative signals of positive selection were identified at 11 loci, but only one was shared among subgroups of *A. gambiae*. We discuss these results with respect to ecological differences among these strata as well as potential implications for disease transmission.

INTRODUCTION:

Over six decades ago, J.B.S. Haldane inferred from simple ecological observations that host-pathogen interactions must be a unique and powerful driving agent of adaptive evolution (Haldane 1949). Such evolutionary dynamics are expected to leave traces in genomes of both host and pathogen; especially in immunity-related genes for the former and in virulence genes for the latter. Consistent with this expectation, molecular evolutionary studies of genes in primates, *Drosophila*, and plants have shown that immune-related genes tend to evolve adaptively and are among the most rapidly evolving genes in the genome (Clark et al. 2003; Schlenke and Begun 2003; Nielsen et al. 2005; Tiffin and Moeller 2006; Sackton et al. 2007). Here, we examine patterns of genetic variation in multiple natural populations of the African malaria mosquito, *A. gambiae s.s.* to identify possible evidence of pathogen-driven molecular evolution.

One form of pathogen-driven evolution is positive selection acting on a beneficial allele or alleles. In the classical selective sweep model, positive selection acts on a new allele that arises by mutation, driving it to fixation within a population or species (Maynard Smith and Haigh 1974). Partial selective sweeps are also possible, in which the beneficial mutation increases in frequency within the population but does not or has not yet reached fixation (Hudson, Sáez, and Ayala 1997). Partial sweeps may be expected when the selective event is either very recent or occurs in a heterogeneous environment. In models of positive directional selection, the chromosomal background carrying the beneficial mutation also rises in frequency through so-called genetic hitchhiking, leaving a distinct footprint on linked neutral variation in the genomic region surrounding the mutation (Smith and Haigh 1974; Kaplan, Darden, and Hudson 1988; Kaplan, Hudson, and Langley 1989; Braverman et al. 1995).

Positive selection can also act on polymorphisms segregating at low or intermediate frequency in the population prior to the selective event. Such selection on standing genetic variation can yield a rapid response to a change in selective pressure, such as a switch of ecological niche (Przeworski, Coop, and Wall 2005; Pritchard, Pickrell, and Coop 2010). This could explain the fixation of the Duffy null allele in sub-Saharan African human populations, where two major haplotypes carrying the null allele were driven to fixation presumably in response to selection pressure from the human malaria parasite *Plasmodium vivax* (Hamblin and Di Rienzo 2000). Statistical population genetic tests have been developed to detect the footprint left by positive selection on natural variation in a population (Tajima 1989; Fay and Wu 2000; Kim and Stephan 2002; Sabeti et al. 2002), although the signature of selection on standing variation is more complex and more difficult to identify reliably, especially when the adaptive allele is relatively common at the beginning of the selective event (Przeworski, Coop, and Wall 2005).

Identifying adaptive evolution in *A. gambiae s.s.* may also have implications for human health, since this species is a primary vector of the human malaria parasite *Plasmodium falciparum* in sub-Saharan Africa. Malaria remains the most deadly vector borne disease and a major public health concern in tropical Africa. Control of vector-borne diseases such as malaria is difficult due to their complex mode of transmission, and development of novel control strategies depends on a clear understanding of host-pathogen dynamics. Natural mosquito phenotypic variation for susceptibility to malaria parasite infection has a large genetic component, and has been mapped to loci on all three chromosomes in multiple studies in both West and East Africa (Niaré et al. 2002; Menge et al. 2006; Riehle et al. 2006; Riehle et al. 2007). In particular, the left arm of chromosome 2 carries a genomic region containing a cluster of significantly associated

genetic markers that has been termed a "*Plasmodium*-Resistance Island" (PRI; Menge et al. 2006; Riehle et al. 2006; Riehle et al. 2007). The PRI is 15 megabases (Mb) in size and encompasses approximately 1000 coding genes (Riehle et al. 2006), and the genetic variants responsible for the resistance trait have not been identified. Nonetheless, a set of plausible candidate genes has been proposed that may harbor the causative variation, and we analyzed sequence diversity in a large sample of wild-caught mosquitoes at a subset of the candidate genes to analyze the selective history and genetic structure at these loci.

Anopheles gambiae s.s. is comprised of genetically differentiated but morphologically identical subgroups that can be distinguished only on the basis of molecular diagnostic assays, termed 'molecular forms'. The "S" molecular form has the largest range, and is widespread throughout sub-Saharan Africa. The "M" molecular form has arisen only in West Africa (della Torre et al. 2001), likely as a population derived from the S form, and the two are broadly sympatric over the M form range. The M and S forms are reproductively isolated at the prezygotic level (Diabaté et al. 2007) and display restricted gene flow (Wondji et al. 2005; Neafsey et al. 2010). However, the evolutionary status of the subgroups appears to be dynamic, because some sympatric populations of M and S forms display elevated levels of hybridization, with complex patterns of directional introgression (Oliveira et al. 2008; Marsden et al. 2011).

We identified an additional population form in the Sudan Savanna zone of Burkina Faso, the GOUNDRY subgroup, which is genetically distinct from both M and S forms, and in which the markers typically diagnostic for the M and S forms segregate freely (Riehle et al 2011). In this sampling location, the three subgroups are sympatric and larval collections yield all three forms. Phenotypic differences among the M and S molecular forms have been described at the aquatic larval stage, but the larval ecology

of GOUNDRY is not well understood. M and S can differ by the ecotype of their larval sites (Costantini et al. 2009), where the S form favors temporary breeding sites and the M form prefers permanent freshwater pools, although larval breeding sites are also often shared. The larval habitats of these mosquitoes harbor a diverse community of invertebrates, microbes, fungi and protozoa, and there is some evidence that the M and S molecular forms may differ in their interactions with this community, particularly with respect to their ability to avoid predators that are more often found in habitats preferred by the M form population (Diabaté et al. 2008; Fillinger et al. 2009; Gimonneau et al. 2010; Gimonneau et al. 2012). At the adult stage, the M and S molecular forms are highly endophilic and tend to rest indoors after feeding, and are thus sometimes referred to as ENDO forms (Riehle et al. 2011). In contrast, GOUNDRY adults presumably exploit yet undiscovered outdoor resting sites, although occasional indoor-resting GOUNDRY adults have been captured (Riehle MM, Vernick KD, Sagnon N, Guelbeogo WM, unpublished observation).

The genomic region containing the *Plasmodium*-Resistance Island partially overlaps the large 2La paracentric chromosomal inversion. 2La polymorphism is widespread in Africa and was originally noted for the allele frequency correlation with degree of environmental aridity (Coluzzi et al. 1979; Powell et al. 1999). In the Burkina Faso study area, the M and S form populations are nearly fixed for the inverted $2La^{a}$ arrangement, while the GOUNDRY subgroup segregates both the standard $2La^{+}$ and inverted $2La^{a}$ arrangements in Hardy-Weinberg equilibrium (Riehle et al. 2011).

In this study, we used a population genetic approach to test hypotheses of adaptive evolution at candidate immune genes within the PRI infection control locus. The candidate gene set is comprised of genes with a diverse array of putative immune functions including immune effector molecules, pathogen recognition, signal transduction

and modulation (Riehle et al. 2006). We hypothesized that the differences in ecology and adaptation of the M and S molecular forms, the 2La inversion arrangements, and the GOUNDRY subgroup might result in exposure to distinct suites of pathogens in the environment, and that these differences in exposure could result in different host-pathogen evolutionary dynamics that may have important implications for malaria transmission. We analyzed the patterns of genetic variation in 28 immunity-related genes in four population strata of *A. gambiae* collected from the village of Goundry, Burkina Faso. We find that signals of putative positive selection vary among the genetic subpopulations and 2La inversion types. We discuss these results in the context of *Plasmodium* selection pressure and transmission, ecological differences among these populations, and the ongoing incipient speciation process.

MATERIALS AND METHODS

Mosquito collection and sample sets

The mosquitoes used in this study are part of the sample set described by Riehle and colleagues (2011), with full details on the origin of mosquito specimen, sample composition and genotyping methodologies given in that publication. Briefly, *A. gambiae s.s.* were collected as larvae in 2007 and 2008 from larval habitats in and around Goundry, Burkina Faso (coordinates 12°30'N, 1°20'W), about 30 km north of the capital city Ouagadougou. The collection area is situated within the Sudan-Savanna (Sudano-Sahelian) ecological zone of tropical shrubland and dry forest. Mosquitoes from 56 larval collections were reared to adults in an insectary. Sixteen distinct larval collections from the years 2007 and 2008 were used to compose the samples.

DNA isolation

DNA was extracted from individual adult female mosquito carcasses in 100 μl DNAzol (Invitrogen) according to the manufacturer's recommendations. The genomic DNA from each mosquito was resuspended in 500 μl H_2O.

Genotyping and population assignment:

Species, molecular form and 2La inversion karyotype were genotyped as described previously in Riehle et al. (Riehle et al. 2011). Briefly, species and molecular form were typed using the SINE200 X6.1 assay (Santolamazza et al. 2008). The 2La inversion was typed using the published molecular assay (White et al. 2007). Fluorescent primers were used in both assays, and PCR fragments were sized using an ABI Genetic Analyzer 3730 as previously described (Riehle et al. 2011).

Genotyping of microsatellites on the third chromosome was carried out as described previously (Riehle et al. 2011). The genotyped markers were: 3R.H59, 3R.H93, 3R.H249, 3R.H119, 3R.H555, 3L.H242, 3L.H758, and 3L.H817. These markers are regularly spaced on the two chromosome arms, present no detectable null alleles and segregate at HWE. Microsatellites were used to assign the samples to either ENDO M/S or GOUNDRY subpopulations using the program STRUCTURE (Pritchard, Stephens, and Donnelly 2000), then the standard molecular diagnostic (Favia et al. 2001) was used to assign M versus S molecualar form as described in (Riehle et al. 2011). Mosquitoes not assigned to a single cluster with greater than 80% probability by STRUCTURE were removed from the analysis. Heterokaryotypic $2La^a/2La^+$ mosquitoes were not included in this study. Mosquito identifiers, sampling year, molecular form status, and 2La karyotype for each mosquito is provided in Supplementary Table 1.

PCR amplification of target gene fragments

For 28 selected candidate genes, two alternative primer pairs were designed manually based on the *A. gambiae* genome (Vectorbase, *A. gambiae* genome, version AgamP3). Primers were designed in coding exons to generate a PCR product of ~500 bp. If possible, the PCR amplicon was designed to span an intron to increase the number of sequence variants. The two alternative primer pairs were first tested for efficient amplification with DNAs from three unrelated *A. gambiae* s.s. of different 2L inversion karyotypes ($2La^a/2La^a$, $2La^+/2La^+$, $2La^+/2La^a$) and one mosquito from the Ngousso colony of *A. gambiae*. The generated PCR amplicons were analyzed on a 1% agarose gel, and single band PCR products of the correct size were sequenced in both directions to confirm the amplification of the correct target gene. If needed, additional primer pairs were designed to optimize the amplification. The primer pair that performed best with all four mosquito DNAs was retained for the experiment with the field-collected specimens. The sequences of the primers are given in Supplementary Table 2. In the sample set from 2007, PCR fragments were generated from all 28 target genes. In the 2008 samples set PCR fragments were produced from 7 of the 28 target genes (Supplementary Table 1). To facilitate the direct sequencing of the PCR products, all retained primer pairs were synthesized with 5' extensions corresponding to the universal forward and reverse M13 primers, respectively.

PCR reactions were performed in 20 μl with 2 μl of genomic DNA using the AccuPrime SuperMixII (Invitrogen) according to the supplier's recommendations. The amplification conditions included an initial denaturation step of 94°C for 3 min, followed by 40 cycles at 94°C for 30 sec, 55°C for 45 sec and 72°C for 1 min. A final extension step of 72°C lasted 10 min.

To obtain PCR amplicons suitable for sequencing, unincorporated primer molecules and nucleotides were removed from the PCR product by centrifugation over

Sephadex P100 columns in Multi Screen filter plates (Millipore). To obtain an efficient purification, a mixture of 6 μl PCR amplification product and 26 μl of H_2O were carefully overlaid on a 300 μl P100 column and centrifuged for 4 min at 1800 rpm.

Sequencing

Sequencing reactions were conducted using 2 to 4 μl of the P100 purified PCR product. All amplicons were sequenced in both directions using the universal forward and reverse M13 primers, the ABI Big Dye Terminator v.3.1 Cycle Sequencing kit (LifeTechnologie) and an ABI Prism 3730 DNA Analyzer (Applied Biosystems). Genotypes were called for each heterozygous sequence individually automatically using the internal software of the DNA Analyzer based on intensity ratios of the sequence chromatograms. The sequences were assembled using CodonCode Aligner (CodonCode Corporation). To assemble a contig, the two sequences of a PCR fragment from a single mosquito were aligned using the reference genome of *A. gambiae* as a guide. Then a multiple sequence alignment of all consensus sequences was constructed using ClustalW. Gene names, genomic locations, and AGAP identifiers are provided in Supplementary Tables 2 and 3. The number of chromosomes sequenced from each population for each fragment is provided in Supplementary Table 4 and in the Results section.

Haplotype phasing

Haplotypes were inferred from the aligned sequences using PHASE 2.1.1 (Stephens, Smith, and Donnelly 2001) for each population independently using default options. The FASTA sequence alignment obtained for each population was converted into the input file format requested for phase inference using the program Seqphase

(Flot 2010). The same software was also used to transform the PHASE output file back into FASTA. Within the GOUNDRY subgroup, for sequenced amplicons located within the 2La inversion, the phase reconstruction was done independently for $2La^a/2La^a$ and $2La^+/2La^+$ mosquitoes.

Outgroup sequence

For inference of ancestral versus derived allele states, we used sequence from *Anopheles arabiensis*, *Anopheles quadriannulatus*, and *Anopheles merus*, all of which are members of the *Anopheles gambiae* species complex. In order to collect sequence for each gene from these species, we downloaded paired-end lanes of Illumina short read sequence data from the NCBI Short Read Archive (*A. merus*: SRR314654 and SRR314646; *A. arabiensis*: SRR314650; *A. quadriannulatus*: SRR314661), deposited by a public sequencing initiative (Besansky and Anopheles Genomes Cluster Commitee 2008). These short reads were generated from whole genome sequencing of a pool of two individuals from the *A. merus* OPHANSI strain, a pool of two individuals from the DONGOLA strain of *A. arabiensis*, and a pool of two individuals from the SKUQUA strain of *A. quadriannulatus*. Each paired-end lane was mapped to the *Anopheles gambiae* PEST genome sequence (AgamP3, August 2011 release from VectorBase.org) using BWA (Li and Durbin 2009) with default parameter settings except for the edit distance, which was set to 8 to accommodate the relatively high expected genetic distance between the reads and the reference. Read mapping resulted in median alignment depths of 20, 20, and 23 for *A. merus, A. arabiensis*, and *A. quadriannulatus*, respectively. We used the mpileup function in SAMtools (Li et al. 2009) to generate pileups and call variants. We extracted an outgroup sequence for each species by substituting the alternative nucleotide into the *A. gambiae* reference sequence whenever

the short-read data from the outgroups differed from the *A. gambiae* reference. The inferred sequence for each species was aligned with the *A. gambiae* sequences and used for subsequent analyses.

Genetic differentiation

To estimate levels of genetic differentiation between the strata at the immune genes studied here, we calculated F_{ST} using Weir and Cockerham's unbiased estimator (Weir and Cockerham 1984) as implemented in an R script written by Eva Chan (www.evachan.org). F_{ST} was calculated for each gene separately between all pairs of population strata. To determine whether each estimate was significantly greater than zero, we randomly permuted population assignments 10^5 times, recalculated F_{ST}, and asked how many of the randomly permuted data sets exhibited F_{ST} greater than the observed value. Values were considered significantly greater than zero if fewer than 5% of simulations resulted in F_{ST} greater than the observed value. We were also interested in determining whether each gene-wise value of F_{ST} is greater than genomic levels of F_{ST}, so we compared gene-wise values to values of F_{ST} that had been calculated previously using microsatellite loci on the 3rd chromosome in a previous study of these populations (Riehle et al. 2011).

Neutrality tests and population genetic statistics

We calculated population genetic statistics for each gene within each population. To identify patterns of genetic variation that are consistent with positive selection, we used a compound statistic called *HEW* (Zeng, Shi, and Wu 2007). This approach combines Fay and Wu's *H* statistic (Fay and Wu 2000), which is based on site-frequency estimators of the population parameter $4N_e\mu$ while giving extra weight to high frequency

derived alleles, with the Ewens-Watterson (*EW*) haplotype based statistic (Watterson 1978) that measures haplotype homozyosity. *HEW* is implemented by calculating *H* and *EW* separately, comparing each test statistic to a null distribution to obtain *p*-values, combining the *p*-values into a vector, and then comparing this vector against empirically determined thresholds to determine whether the vector is consistent with neutrality (Zeng, Shi, and Wu 2007). Empirical thresholds for statistical significance for *HEW* were established by comparing the distribution of *p*-values for the component statistics and finding the threshold that provided the desired statistical cutoff (0.05) for the vector combining the two component *p*-values (Zeng, Shi, and Wu 2007). Null distributions of all test statistics were generated using coalescent simulations conditioned on the number of haplotypes in the sample with the mutation rate set to the population parameter θ ($4N_e\mu$) estimated from the data using Watterson's estimator (Watterson 1975), as described in Zeng et al. (Zeng, Shi, and Wu 2007). All coalescent simulations were conservatively conducted with no recombination. The assumption of no recombination is justified because the gene fragments sequenced in the present study are short enough that recombination is not likely to be pervasive in the samples. Accurate estimates of recombination are not available in *A. gambiae*, especially in and around the 2La inversion, and estimating recombination rates from population sequence data can produce inaccurate results (Wall 2000). It should be noted that we used a normalized version of the *H* statistic that was derived by Zeng et al. (Zeng et al. 2006) and was shown to consistently have slightly more power than the original un-normalized version. Since *H* requires an outgroup and *EW* is based on haplotype diversity, the *A. merus* outgroup sequence and *A. gambiae* haplotypes consisting only of silent (synonymous and non-coding) sites were used to calculate the component statistics of *HEW* for each gene and for each population. Calculations of the component statistics of

HEW and all coalescent simulations were carried out using a program kindly provided by K. Zeng. The resulting *p*-values were corrected for multiple testing using the Benjamini and Hochberg (Benjamini and Hochberg 1995) correction as implemented in the p.adjust function in R (R Development Core Team 2011).

Additional population statistics including nucleotide diversity, the number of haplotypes, and Tajima's *D* (Tajima 1989) were calculated. Nucleotide diversity (π) was measured as the average number of nucleotide differences per site using DnaSP v.5.10 (Librado and Rozas 2009). The number of haplotypes, *h*, was also calculated using DnaSP. To test the significance of haplotype and nucleotide diversity for genes with structured genealogies, we simulated neutral genealogies using *ms* (Hudson 2002) conditioned on the empirical clade structure and number of segregating sites in the empirical sample. For each simulated genealogy that satisfied these criteria, we calculated nucleotide and haplotype diversity for each sub-clade and counted how many simulated genealogies showed values equal to or more extreme than the empirically observed values. Tajima's *D* was calculated using the software package provided by K. Zeng mentioned above and evaluated for statistical significance in the same simulation framework as for *H* and *EW*.

We also used a multilocus version of the Hudson-Kreitman-Aguade (HKA) Test (Hudson, Kreitman, and Aguadé 1987) to test for deviations from neutral expectations. We employed a multi-locus version of the HKA test implemented in the program *hka* written by J. Hey (http://genfaculty.rutgers.edu/hey/software#HKA). For this test, we used only variation at synonymous or non-coding sites and *A. merus* as the outgroup (species 2). To determine significance of the observed values and sum of deviations, 10^5 neutral coalescent simulations were conducted modeled on parameters inferred from the data within the program to establish an empirical distribution of the χ^2 distribution.

Since the comparison is designed to be between a locus of interest and a 'neutral' locus and we didn't sequence any functionally random control loci, we instead compared the focal locus to the nearest upstream and downstream neighbors that did not show a significant *HEW* test statistic.

Linkage disequilibrium and haplogroup analysis

We estimated genetic correlation, r^2, between all variant sites within each gene for each population and used this statistic to identify blocks of high linkage disequilibrium (LD) in genes that rejected neutrality based on the *HEW* statistic above. r^2 was calculated using an R script written by Eva Chan (www.evachan.org) and plotted using the R package LDheatmap (Shin et al. 2006). LD plots were visually inspected for all loci with a significant *HEW* result in the homokaryotype groups of GOUNDRY ($2La^a/2La^a$, $2La^+/2La^+$), and the gene with the most striking LD block was chosen for further analysis. Where we hypothesized that incomplete sweeps or sweeps from standing variation were plausible models in this data, MEGA5 (Tamura et al. 2011) was used to calculate and draw neighbor-joining gene trees using the Maximum Composite Likelihood method with uniform substitution rates among sites, and trees were inspected for evidence of distinct clades of genetically similar haplotypes. Alignments were inspected for distinct haplotypes using CodonCode (v3.7), and evidence for increased linkage within and differentiation between groups of haplotypes was used to delineate distinct haplotype groups within the sample. These haplogroups were then designated A and B, and each haplogroup was further analyzed for evidence of positive selection based on the same summary statistics as above. To test whether the distribution of segregating sites and patterns of nucleotide diversity across the two haplogroups were consistent with the neutral equilibrium model, we conducted an additional 10^5 coalescent simulations for

each gene conditioned on the observed clade structure and the observed number of segregating sites. We then asked how often the simulated data showed values equal or more extreme than those observed in our empirical datasets.

We also conducted additional analyses of specific clades within certain datasets. To identify specific regions of high divergence, we calculated Jukes-Cantor corrected divergence (K_{JC}) using a sliding window analysis in DnaSP with a physical window size of 50bp and a shift size of 10bp. After identifying a region of high divergence, we extracted the sequence from both the low and high divergence clades and searched for transcription factor binding sites by comparing the sequences to insect matrices within the TRANSFAC database using the Match™ 1.0 webserver (BioBase). We set the selection cutoff to minimize false positives and only searched high quality matrices within the insect group.

Results

Population differentiation and genetic variation

We examined genetic variation at 28 loci, comprised mostly of genes selected by a filtering process designed to enrich for immune-related genes inside the *Plasmodium*-Resistance Island (PRI) on the *A. gambiae* second chromosome near the proximal boundary of the 2La inversion (Niaré et al. 2002; Riehle et al. 2006; Riehle et al. 2007). Population samples were drawn from three strata of *A. gambiae s.s.*: ENDO M and ENDO S molecular forms of *A. gambiae* and the recently discovered cryptic sub-group GOUNDRY (Riehle et al. 2011). The genes analyzed here can be grouped by genomic context since many of the genes (18/28) are located inside the large 2La inversion on the left arm of the second chromosome, although many of the genes are located near the proximal breakpoint of the inversion (Figure 1). Half of the remainder of the genes lie

on 2L outside the inversion and the rest are on other chromosomal arms. Twenty of the 28 sequenced loci fell within the PRI (Figure 1). The 2La inversion is nearly fixed for the inverted form in the molecular forms in Burkina Faso, but is segregating at Hardy-Weinberg equilibrium frequencies in GOUNDRY (Riehle et al. 2011). When GOUNDRY $2La^{+}/2La^{+}$ and $2La^{a}/2La^{a}$ individuals are contrasted for genetic differentiation, genes in collinear regions of the genome that are physically distant from the inversion show no differentiation among homozygous groups (mean F_{ST} = -0.0005), while those inside or near (<2.2MB) the inversion show extremely high levels of differentiation (mean F_{ST} = 0.52), indicating strong reductions of recombination and independent evolutionary trajectories within and surrounding the 2La inversion (Figure 1). On the other hand, comparisons among the M and S molecular forms and GOUNDRY individuals homozygous for the inverted form revealed relatively constant genetic differentiation across all loci, irrespective of distance from the inversion (Figure 1). Both the M and S molecular forms were compared to GOUNDRY separately and exhibited qualitatively similar levels and patterns of differentiation, so only the S form comparison is presented in Figure 1. These results highlight reductions of interbreeding between the M and S molecular forms and GOUNDRY as well as reductions in recombination between the two forms of the 2La inversion, especially at the breakpoints, and lead us to delineate four groups (M, S, GOUNDRY $2La^{+}/2La^{+}$, and GOUNDRY $2La^{a}/2La^{a}$) for subsequent analysis. The contrast between the molecular forms and the GOUNDRY $2La^{a}/2La^{a}$ group also indicates that the origin of the 2La inversion predates the split between all of these groups, since differentiation is similar between loci inside and outside the inversion. This pattern is contrary to that which would be expected if the inversion had been introgressed after the subgroups diverged, in which case the inversion might show less differentiation than loci outside the 2La region.

Patterns of genetic diversity also differentiate these population strata (Table 1). Levels of synonymous coding variation are 35% lower in GOUNDRY 2La^a homokaryotypes (average θ_W = 1.64%) compared to M form (θ_W = 2.53%) and S form (θ_W = 3.66%), indicating that the effective population size of GOUNDRY is substantially smaller than that of the M and S form populations, which are distributed across most of West Africa and sub-Saharan Africa, respectively (Lehmann and Diabate 2008). Furthermore, the distributions of allele frequencies differ between GOUNDRY and the molecular forms, possibly indicating distinct demographic histories. While genes of M and S molecular form mosquitoes generally have negative values of Tajima's *D* consistent with recent population growth previously inferred for these populations (Crawford and Lazzaro 2010), GOUNDRY exhibits a distribution of *D* approximately centered on *D* = 0 with a substantially increased variance. The reduced nucleotide diversity and non-negative values of *D* may indicate either a recent bottleneck of at least moderate size and duration in GOUNDRY or that this population has maintained a relatively small and consistent effective population size in the recent past. We also compared levels of variation and distributions of allele frequencies between genes inside and outside of the 2La inversion and found that, while loci associated with the inversion may be more differentiated among inversion forms, levels of diversity and *D* do not differ among genes inside and out of the inversion (Table 1).

Evidence for positive selection

To determine whether genetic variation at the immune-related genes under study exhibit patterns that are consistent with positive selection, we used summary statistics that measure enrichment of high frequency derived alleles (Fay and Wu 2000) and levels of haplotype homozygosity (Watterson 1978) to ask whether the observed patterns are

consistent with neutral evolution. A recent series of papers introduced the *HEW* test that combines Fay and Wu's site-frequency spectrum based *H* statistic (Fay and Wu 2000) with Ewens-Watterson test for haplotype homozygosity (Watterson 1978) into a compound statistic. The compound statistic *HEW* provides a sensitive and specific approach for distinguishing the footprint of positive selection from both stochastic neutral evolution as well as demographic effects, particularly for short sequence fragments as analyzed here (Zeng et al. 2006; Zeng et al. 2007; Zeng, Shi, and Wu 2007). We tested all genes in each of the four groups using *HEW* and, despite the relatively conservative nature of the compound *HEW* statistic, we found putative evidence for positive selection at 11 immune-related genes after correcting for multiple testing (Table 2). Interestingly, of the 11 genes that exhibit evidence for positive selection, only one (*APL1B*) is shared across population strata (Figure 3), implying that the strata reflect ecologically distinct subpopulations whose immune genes are under substantially different selection regimes. The contrast between the molecular forms is remarkable in that evidence for selection was identified at six loci in the M form, but not one gene showed significant departure from neutrality in the S form. Furthermore, there is no overlap among adaptive signals between the two GOUNDRY classes of 2La homokaryotypes (Figure 3), suggesting that the two inversion states may experience distinct selective pressure.

The *HEW* statistic is relatively robust to demographic effects, but false positive results can occur if haplotype reconstruction is incorrect (Zeng et al. 2007). To rule out this potential source of false positive test results, we evaluated the phase inference results based on several criteria to determine whether genes with significant *HEW* statistics also showed relatively low confidence phase reconstruction. When the statistical software PHASE assigns heterozygous sites to haplotypes, confidence probabilities are calculated for each site that reflect the degree of statistical confidence in

the assignment, where a probability of one reflects no ambiguity and 0.5 indicates complete ambiguity (Stephens, Smith, and Donnelly 2001). In our dataset, more than half of inferred sites were assigned to a haplotype with a confidence probability of one, thanks in part to the power of haplotype inference achieved with large numbers of individuals sampled and small sequence windows with little recombination. There is some uncertainty in haplotype reconstruction based on the remaining sites with probabilities less than one, and we sought to use the information contained in these probabilities to evaluate the possible effects of this uncertainty on our *HEW* results. We evaluated each run of PHASE (see Methods) based on the proportion of sites that were inferred or imputed as well as the distribution of confidence probabilities, reasoning that haplotype reconstruction might be most problematic in genes with relatively more missing and heterozygous sites and, therefore, more low confidence probabilities. To determine whether this was a concern with respect to our inferences of positive selection, we first evaluated genes based on the proportion of the total sequence that was phased or imputed. None of the genes that rejected neutrality based on *HEW* were in the top 5% for proportion of either phased or imputed sites (Supplementary Table 5). We then ranked the genes by proportion of variant sites at which the confidence probability estimated by PHASE was less than one. One of the genes that rejected neutrality, *LRR(7030)* (for simplicity of presentation, the unnamed LRR genes are labeled according to shortened forms of their AGAP identifiers) in GOUNDRY $2La^a/2La^a$, was in the 5% tail (Supplementary Table 5). This suggests that the results from this gene should be interpreted with caution. Generally speaking, however, the mean confidence probability among sites with probabilities less than one was 0.78 at this locus. Although it is difficult to fully evaluate the success of the phase inference process in the absence of experimental validation, since the genes that show evidence of positive

selection are not among those with the lowest confidence or even those that required the most phasing, we do not believe that phasing errors are likely to be causing false positives in our tests for positive selection.

TEP1, LRIM1, and APL1

An important validation of our analysis was the recovery of signals of positive selection in two genes previously indicated as evolving adaptively through analysis of independent datasets and analytical approaches. Positive selection has been identified at both *TEP1* (White et al. 2011) and the *APL1* gene cluster (Rottschaefer et al. 2011) in the M form population, and our new analysis indicated adaptive evolution at both of these loci (Figure 3). In addition, it has been speculated that, since a physical complex is formed between the proteins encoded by *TEP1* and *APL1C* as well as a third protein (LRIM1), the patterns of variation at *TEP1* and the *APL1* locus may reflect coordinate adaptive evolution (Fraiture et al. 2009; Povelones et al. 2009; Rottschaefer et al. 2011). Thus far, no signals of adaptive evolution have been identified at *LRIM1* in *A. gambiae* (Obbard et al. 2007; Slotman et al. 2007; Cohuet et al. 2008), but our new analysis points to an enrichment of high frequency derived alleles (H_{norm} = -2.49; uncorrected p = 0.0222) and an increase in haplotype homozygosity (*EW* = 0.12; uncorrected p = 0.0513) at *LRIM1* in the M form population that is inconsistent with neutral evolution (*HEW* corrected p = 0.0336). We also find evidence for positive selection at the *TEP1* locus (*HEW* corrected p = 0.0233), consistent with coordinate adaptive evolution among the proteins making up the complex. When we analyze the *APL1* paralogs separately, we find that *APL1C*, the only *APL1* paralog involved in the described protein complex, is a clear outlier from the majority of loci in this population (Figure 3), although its *HEW* statistic is marginally non-significant after multiple testing (*HEW* corrected p = 0.084). If

selection is acting on the three members of the complex in a coordinate fashion, *TEP1* is an outlier (Figure 4) suggesting that selection is stronger on this locus than on *LRIM1* and *APL1C*.

FBN32

Of all genes that showed a significant departure from the neutrality by the *HEW* statistic, *FBN32* (AGAP007041; *HEW* corrected $p = 0.0482$) in the GOUNDRY $2La^a/2La^a$ subgroup showed the most striking pattern of linkage disequilibrium (Figure 4). Inspection of the sequence data revealed three derived SNPs in perfect linkage disequilibrium at the boundaries of the sequenced fragment. These SNPs mark two distinct major haplotype clades, hereafter referred to as haplogroups A and B, the larger of the two also harboring two possible recombinant haplotypes (clade A*; Figure 4). This genealogical structure is significantly unlikely under a neutral model ($p = 0.0062$), and we hypothesized that the presence of two sharply defined clades could be consistent with either a partial selective sweep (Hudson, Sáez, and Ayala 1997), a sweep from standing genetic variation (Przeworski, Coop, and Wall 2005), or a classical selective sweep with at least one recombination event occurring during the sweep. Several lines of evidence suggest that an incomplete sweep cannot explain the data. First, if we assume that the less variable B haplotype is the selected haplotype, it would be segregating at a relatively low frequency in the population (27% or 30 out of 110 chromosomes) that the *HEW* test statistic has very low power to detect (Zeng, Shi, and Wu 2007), implying that the deviation from neutrality detected by the *HEW* test stems from the entire sample instead of just one clade. Second, under a partial selective sweep model, we would expect the selected clade to lack variation while the other clade harbors pre-sweep variation, but this is not what we find. Comparison to simulated neutral genealogies

indicates that the low average number of pairwise differences (π_A = 0.0021; π_B = 0.0002) is significantly unlikely under the neutral model in both clades (clade A $p < 0.0005$; clade B $p < 0.05$; Table 3). Moreover, when applied to the larger (n = 80 chromosomes) and more diverse A clade, neutrality tests (*D, H*, *EW, HEW*) reject the neutral model (all $p <$ 0.05; Table 3). Collectively, these results confirm that the A and B clades both harbor patterns of genetic variation that are inconsistent with neutrality, thus the data are more consistent with a complete sweep with recombination rather than an incomplete sweep. To rule out the possibility that this locus could have an unusually low mutation rate that could be driving these results, we compared patterns at *FBN32* to its nearest neighbors in the dataset using an HKA test (Hudson, Kreitman, and Aguadé 1987), and found that *FBN32* harbors significantly fewer polymorphisms and is significantly more diverged from *A. merus* than expected under neutral model (uncorrected $p = 0.0176$).

The division of the haplotypes into two large clades could have arisen due to either the presence of the selected site on two chromosomal backgrounds prior to selection or a recombination event could have occurred during the selective event. It is difficult to distinguish between these two models. Overall, the data are consistent with a model of positive selection at *FBN32* in this population that may have involved a sweep with recombination or selection on standing variation.

Toll9

In the GOUNDRY $2La^+/2La^+$ group, the only gene to show significant evidence for non-neutral evolution based on the *HEW* statistic was *Toll9* (Figure 3). Despite showing a significant departure from neutrality, this gene harbors substantial variation in this subpopulation (θ_W = 0.0204), signaling that these data are not consistent with a recent and strong selective sweep. Similarly to the pattern observed in *FBN32* in GOUNDRY

$2La^a/2La^a$, two distinct clades are present in the *Toll9* data, a genealogical structure that is significantly unlikely under a neutral model ($p < 10^5$; Figure 5). Analysis of LD in this region reveals an LD block consisting of 18 sites in linkage disequilibrium separating the two clades, three of which are nonsynonymous substitutions in the fourth exon (Figure 5). Of these sites, 17 are fixed between the two clades. A neighbor-joining tree reveals an interesting and unexpected topology where the A clade shares a common ancestor with the outgroups before coalescing with the B clade sequences (Figure 5). Plotting divergence across the sequence for each clade separately reveals a large spike in divergence between clade B and the outgroups (K_{JC} = 0.495) restricted to the intronic sequence (Figure 6). We found that divergence from the outgroups (K_{JC}) never exceeded 0.155 in similarly sized sequence windows from other genes in our data set, confirming that the *Toll9* sequence is an outlier. We also examined divergence at *Toll9* in other subpopulations and found similar, albeit smaller spikes in the intronic sequence (maximum 50bp window K_{JC} = 0.322 in M form population, mean across all populations K_{JC} = 0.252), suggesting that both haplotype groups existed and were segregating at intermediate frequency prior to the split of GOUNDRY from the M and S molecular forms. We considered that the unusual B clade could have arisen through a paralogous gene conversion event, for example with another member of the *Toll* family. To test this hypothesis, we used BLAST to search the clade B sequence against the *A. gambiae* genome and the NCBI nr sequence database, but we found no significant matches to any other available sequence other than existing *A. gambiae Toll9* sequences. It is possible that the sequence may have been introgressed from a species not sampled in this study, but we have no data to support or refute that hypothesis.

The divergent sequences defining clades A and B may represent a functional balanced polymorphism. Among other possible functions, introns can harbor

transcription-binding sites that affect expression patterns of the surrounding gene, or even a different gene in *trans*. We compared the intronic sequence from both haplotype groups to known insect transcription factor binding site motifs in the TRANSFAC database and found that both haplotypes showed approximately equal matches to three binding motifs (BR-C z1, Hairy, Elf-1). Each haplotype also showed matches to at least one unique motif. The divergent B haplotype showed a match (matrix match = 0.852) to a motif that recruits the NF-kappaB transcription factor Dorsal that has been shown in *Drosophila* to function both in dorsal-ventral patterning during embryogenesis as well as in activating an immune response as a component of the Toll signaling pathway (Lemaitre et al. 1995; Ghosh, May, and Kopp 1998; De Gregorio et al. 2002). The homolog of Dorsal in *Anopheles*, Rel1, has been shown to play a role in driving immune responses against a variety of pathogens, perhaps in part through the action of the APL1 proteins (Barillas-Mury et al. 1996; Frolet et al. 2006; Mitri et al. 2009), so it is tempting to speculate that this motif may serve a role in immunity, but further experimental analysis is required to determine any functional differences between the divergent haplotypes.

We propose similar population genetic models to explain the *Toll9* and *FBN32* data: a partial selective sweep or a selective sweep with recombination either from standing genetic variation or with subsequent recombination during the sweep. Under the partial selective sweep model, the selected haplogroup would be expected to show a departure from neutrality while the alternative haplogroup would show patterns of genetic variation consistent with neutral expectations. To determine whether the *Toll9* data fit these expectations, we analyzed the haplogroups separately for evidence of non-neutral evolution. Population genetic analysis of the B haplogroup indicates a significant departure from neutrality in this clade, reflecting a scooped shape in the site-frequency

spectrum enriched in both rare and high frequency derived sites (H = -3.0625; $p < 0.005$; Table 2). The A haplogroup, however, also exhibited a significant paucity of genetic variation compared to neutral expectations ($p < 0.005$), suggesting the possibility of recent positive selection acting to remove linked variation in this clade, but the very low level of polymorphism (3 segregating sites) precludes further analysis of this clade with neutrality tests (Table 2). The alternative hypothesis of selection on standing variation predicts that the selected allele was segregating at appreciable frequencies in the population on multiple backgrounds at the time of the selective event. Under this model, we would expect to find private fixations in this population as well as chromosomes bearing both genetic backgrounds segregating in other populations. Although the degree of linkage disequilibrium is much lower in other populations, the SNPs that delineate the two clades in GOUNDRY $2La^{+}/2La^{+}$ are segregating at intermediate and even high frequencies in the M and S molecular forms as well as GOUNDRY $2La^{a}/2La^{a}$ (data not shown). Furthermore, comparison to the outgroup species *A. merus* reveals 4 derived fixations that are unique to GOUNDRY $2La^{+}/2La^{+}$, three of which are synonymous substitutions and the fourth of which falls within an intron (Figure 5). Although these fixations and deficits of diversity could reflect sites linked to an adaptive fixation as expected under a sweep model, they could also be a feature of *Toll9* residing inside the polymorphic and highly diverged 2La inversion. We tested this hypothesis by comparing *Toll9* to its neighbors (*LRR(7030)* and *IRSP1*) using a multi-locus HKA test (Hudson, Kreitman, and Aguadé 1987) and found that, although the data were significantly inconsistent with equivalent evolutionary rates across the genes ($p < 0.0125$), the deviation is largely driven by excess divergence and lower than expected polymorphism at *IRSP1*, while the *Toll9* data more closely fit expected levels of divergence and polymorphism. The results of this test support the neutral model for

Toll9, contradicting the *HEW* test result. However, this test may be inappropriate for these data since the diverged intronic sequence may have been a single mutational event and reflect a balanced polymorphism that could reduce the rate of fixation at this locus, in turn downward biasing the estimate of the mutation parameter 4Nμ in the HKA model reducing the power to detect reductions in polymorphism. In both intraspecific and interspecific comparisons, the presence of the highly diverged intronic haplotypes makes these data somewhat difficult to interpret. Nonetheless, both global and clade-specific summary statistics point to positive selection at this locus (Table 3), and functional studies of *Toll9* are needed to identify both the selective agent as well as the functional role, if any, of the diverged intronic haplotypes.

DISCUSSION

We sequenced alleles from 28 immune-related loci in wild samples of multiple genetic subpopulations of *A. gambiae*, obtaining unprecedented sample sizes and providing the first opportunity to contrast patterns of molecular evolution at immune-related loci in the recently discovered GOUNDRY subgroup (Riehle et al. 2011) with those in the indoor-collected M and S molecular forms. Analyses of haplotypic and genetic diversity revealed sharp differences among these strata in levels of genetic diversity and allele frequencies in coding sequence, as well as evidence for significant deviations from neutrality at 11 loci among these populations. Further experimentation will be necessary to determine the nature of the selective pressures behind the signals observed here. Our results do, however, allow some speculation on the distribution and nature of the selective events affecting these loci.

Selection across functional classes

Our results reveal possible evidence for positive selection at genes coding for proteins with a broad range of functions. In one case, we found evidence for positive selection at the developmental morphogen *Distal-less (DLL)*. Our previous studies showed a highly significant association between microsatellite (H603) alleles inside an intron of *DLL* and infection by *P. falciparum* (Niaré et al. 2002; Riehle et al. 2006), and the signal identified here may reflect linked selection on the functional site driving these signals, although the genetic mapping was conducted in the M and S molecular forms while the signal detected here came from GOUNDRY, where such mapping has not yet been done. However, we also sequenced an intronic fragment flanking H603 and did not find evidence for selection, highlighting the need for further analysis of this genomic

region to identify the linked functional site(s) identified in our genetic mapping studies. In another case, the gene encoding FBN32 (also known as FREP39), a member of the pathogen recognition receptor family in invertebrates (Gokudan et al. 1999; Dong and Dimopoulos 2009), also showed possible evidence for selection. Expression analyses of this gene revealed expression patterns restricted to the abdomen and midgut, the larval salivary gland, and male accessory glands (Dong and Dimopoulos 2009; Neira Oviedo et al. 2009; Baker et al. 2011). With respect to immune function, *FBN32* was up-regulated in response to immune challenge with a gram-negative bacteria, a fungal pathogen, and *P. falciparum*, but not the rodent malaria parasite *P. berghei* (Dong and Dimopoulos 2009), indicating some degree of generality in its immune function.

Leucine-rich repeat-containing (LRR) proteins, a superfamily composed of 180 proteins in *A. gambiae* (Waterhouse, Povelones, and Christophides 2010), featured prominently among the original filtered gene set within the PRI and also among the genes that showed possible evidence for selection. In fact, the PRI locus as a window contains the largest number of LRR genes in the *A. gambiae* genome (Riehle et al. 2006). One class of LRRs, Toll-like receptors (TLRs), are trans-membrane proteins known to act as pathogen recognition molecules in mammals (Means, Golenbock, and Fenton 2000). But strong functional evidence is available for only a small number of TLRs in insects, and these TLRs have roles in development and immune-related signal transduction (Imler and Hoffmann 2001; Imler and Zheng 2004). We found evidence for positive selection acting on the TLR *Toll9* in the GOUNDRY subpopulation. The exact functional role of *Toll9* is unknown, but several studies have shown that *Toll9* is slightly up-regulated following a bacterial immune challenge in larvae and expression is concentrated in the midgut of adults, particularly after a blood-meal (Luna et al. 2002; Marinotti et al. 2005; Baker et al. 2011). Phylogenetic comparisons of the *Toll* genes

may provide insight into protein function in that *A. gambiae Toll9* clusters with mammalian TLRs based on its ectodomain structure and sequence similarity of the intracytoplasmic domain, suggesting that it may be an ancestral TLR in insects (Du et al. 2000; Imler and Zheng 2004; Waterhouse et al. 2007).

The other LRRs that showed evidence for positive selection included several more characterized LRRs (*APL1B*, *APL1A* and *LRIM1*) as well as three uncharacterized LRRs (*LRR(7030), LRR(7059), LRR(7060)*). Interestingly, structural similarities between APL1 and LRIM1 proteins prompted a bioinformatic search through the *A. gambiae* genome for genes that code for other LRIM-like proteins that turned up 24 candidates, but the three uncharacterized LRRs studied here were not among them (Waterhouse, Povelones, and Christophides 2010), implying yet another sub-class of LRRs in mosquitoes. Collectively, our results point to selection acting on proteins with a broad range of putative immune function, most of which show little specificity with respect to the pathogen classes to which they respond. This observation is consistent with gathering evidence supporting a model wherein selection pressures derive from a diverse suite of pathogens, including those that infect larvae, driving the evolution of a generalized immune response (Rottschaefer et al. 2011; Mitri and Vernick 2012).

No evidence for Plasmodium*-driven selection*

Our focus on candidate genes within the PRI provides the potential opportunity to identify *Plasmodium*-driven selection pressure on *Anopheles* immune genes, since our candidate ascertainment process intentionally enriched for immune-related genes that may play a role in resisting *Plasmodium* infection based on experimental and bioinformatic evidence (Riehle et al. 2006). The epidemiological importance of GOUNDRY is not currently known, although this population is physiologically more

permissive than M or S form *A. gambiae* to infection with *P. falciparum* (Riehle et al. 2011). Since the M and S molecular forms are both primary malaria vectors in sub-Saharan Africa and rates of natural *P. falciparum* infection in wild M and S mosquitoes are equivalent (Wondji et al. 2005; Ndiath et al. 2008; Trout Fryxell et al. 2012), *Plasmodium*-driven selective pressure should be shared among these subgroups, if it exists. However, in our data, only 1 of the 11 signals of positive selection (that at *APL1B*) is shared among subpopulations (Figure 3), and we found no evidence of non-neutral evolution at *APL1B* or any other genes in the S form population. Differences in sample sizes or the number of variant sites among populations could lead to reductions in statistical power that could generate false negatives, but this is not likely to be the explanation here because the populations with the fewest signals of possible selection (S form and GOUNDRY $2La^+/2La^+$) had larger sample sizes in most cases than both of the populations that did show signals of putative natural selection (Supplementary Table 4).

Alternatively, the selection pressures driving non-neutral evolution at these loci could be related to differential pathogen exposure associated with ecological differences among these populations. Although some of the immune genes studied here affect susceptibility to infection by *P. falciparum* in laboratory gene silencing experiments, most of these loci appear to also play a role in resistance to bacteria and non-human malaria parasites. Thus while the polymorphism in these genes might contribute to variation in susceptibility to *Plasmodium* infection, it is unlikely that *Plasmodium* is itself the agent driving selection. Nevertheless, selective pressure driven by other pathogens, acting on immunity related genes, could modify the mosquito response to the parasite, in turn enhancing or decreasing malaria transmission. In fact, the increased susceptibility of GOUNDRY to malaria parasite infection (Riehle et al. 2011) could be explained by such

a mechanism, under the hypothesis that a change of ecological niche in this population has shaped a new pattern of selection in the immune genes.

Niche specialization

The striking division of selective pressures among populations points to a model of recent ecological niche specialization. Specifically, our results are consistent with a model in which the M form and GOUNDRY are moving into novel ecological niches exposing genetic variation to novel selection pressures in the new environments. Under this model, the genetic variation segregating in GOUNDRY and the M form would be expected to be largely a subset of S form variation. As expected under this model, the majority of segregating sites in both the M form and GOUNDRY $2La^a/2La^a$ are shared with the S form (mean proportion of M and GOUNDRY $2La^a/2La^a$ sites shared with S form = 0.5892 and 0.6309, respectively, after correction for sample size), but the opposite is not the case (proportion of S form sites shared with M form = 0.4729 and with GOUNDRY $2La^a/2La^a$ = 0.3564, after correction for sample size). Indeed, this model has been proposed previously to explain the relationship between the M and S forms based on the ecological observation that the M form exploits human-derived marginal habitats (della Torre et al. 2002; Simard et al. 2009; Costantini et al. 2009), and to explain molecular data indicating lower levels of genetic diversity (Cohuet et al. 2008) as well as more recent population growth in the M form relative to the S form (Crawford and Lazzaro 2010).

Although adaptive divergence at immune genes is not likely to drive the speciation and niche specialization process, exploitation of novel environments is likely to be accompanied by novel pathogen pressures, potentially leading to adaptation of immune factors as observed here (Lee 2002). For example, the more permanent,

disturbed, human-derived larval habitats preferred by the M form population harbor more abundant and complex insect communities (Diabaté et al. 2008; Gimonneau et al. 2010; Gimonneau et al. 2012), reflecting the more permanent, ecologically viable nature of these habitats that may also harbor a greater diversity of pathogens. The larval habitat ecology of GOUNDRY has not been well characterized, but the lower genetic diversity relative to the M and S molecular forms and the fact that genetic variation in this group is largely a subset of that in the S molecular form suggests the possibility that it is also a derived population that could be moving into novel environments.

Differences between the two 2La homokaryotype groups of GOUNDRY in the genes that show signals of putative positive selection is not likely to be explained by ecology, but rather by the lack of recombination between the two forms of the inversion, particularly near the breakpoints. Under this model, the two forms of the inversion represent distinct gene pools in this region of the genome that harbor distinct sets of genetic variants that respond differentially to selection. Consistent with this observation and the model that the $2La^{+}$ form is the derived form, we find that only 35.6% of segregating sites among $2La^{a}/2La^{a}$ individuals are shared with $2La^{+}/2La^{+}$ individuals, while 63.1% of sites segregating among $2La^{+}/2La^{+}$ individuals are shared with $2La^{a}/2La^{a}$ individuals, after correcting for sample size. Taken together, these lines of evidence support a model of ongoing insipient speciation and niche specialization in this system that has led to adaptive evolution at immune genes.

As in any molecular population genetic analysis of natural selection, it is impossible to state conclusively the selective agent responsible for the observed patterns. However, it seems reasonable in this case to exclude the human malaria parasite *P. falciparum* as the driving force behind the signals detected in our data, considering the striking division of selective signals among population strata. A

compelling alternative explanation is that pathogens in the larval habitats may be driving evolution of these immune genes, and this may have implications for malaria transmission. If broad spectrum immune factors involved in responding to multiple pathogen classes, one of which including *P. falciparum* (Meister et al. 2005; Dong et al. 2006; Dong, Manfredini, and Dimopoulos 2009; Meister et al. 2009), are evolving in response to non-*Plasmodium* pathogens, susceptibility to the malaria parasite could be affected. Many of the immune proteins studied here (e.g. FBN32, LRR(7060)) have not been thoroughly tested functionally for anti-*Plasmodium* activity. But almost all of the loci that show putative signals of selection are within the PRI region that showed significant association to *Plasmodium* resistance phenotypes, and these proteins warrant further genetic and functional analysis.

CONCLUSION

In 1949, Haldane hypothesized, with very little knowledge of the underlying molecular mechanisms, that host-pathogen interactions must be an important factor in shaping the ecological patterns observed in nature (1949). In this study, we tested candidate immune-related genes for evidence of this evolutionary conflict and found a striking pattern that provides grounds to make the reverse inference from molecular evolution to ecology. Namely, the distribution of putative pathogen-related signals of selection among populations of *A. gambiae* implies that these populations may occupy distinct ecological niches and correspondingly experience disparate host-pathogen interactions.

TABLES

Table 1: Summary of nucleotide diversity and the site-frequency spectrum among populations of *A. gambiae*.

Population	**n**[a]	**θ**[b]		**Tajima's *D***[c]	
		All	2La[d]	All	2La[d]
M form	94	0.0253	0.0278	-1.28	-1.24
S Form	136	0.0366	0.0412	-1.41	-1.39
GOUNDRY $2La^a/2La^a$	56	0.0164	0.0186	-0.07	-0.13
GOUNDRY $2La^+/2La^+$	170	0.0114	0.0125	0.22	0.17

[a] average number of chromosomes sequenced per gene fragment.
[b] average θ calculated for each gene fragment using all sites.
[c] average *D* calculated for each gene fragment using only synonymous sites.
[d] statistic calculated using only genes inside 2La inversion.

Table 2: Population genetic summary statistics and test results for loci with a significant *HEW* test *p*-value (statistics for all loci are presented in Supplementary Table 4).

Locus	n[a]	S_{syn}[b]	*D*[c]	*H*[d]	*EW*[e]	*HEW p*-value[f]
M form						
LRIM1	64	25	-0.1096	-2.4951	0.1221	0.0336
TEP1	64	10	-2.1817	-2.0996	0.7979	0.0233
APL1A	62	110	-1.1500	-2.5176	0.3002	0.0140
APL1B	100	88	-0.5362	-1.7003	0.0898	0.0294
LRR (7059)	100	50	-1.8131	-1.6246	0.1480	0.0302
IRSP1	64	55	-1.4194	-2.4365	0.0557	0.0140
GOUNDRY $2La^a/2La^a$						
APL1B	100	70	0.2394	-2.4022	0.2408	0.0252
LRR (7030)[g]	34	22	-1.3252	-1.5011	0.1211	0.0482
LRR (7060)	100	38	-2.0531	-2.2083	0.3992	0.0280
FBN32	100	15	-1.2392	-1.6624	0.5080	0.0482
DLL	100	72	-2.3283	-2.5965	0.2512	0.0252
GOUNDRY $2La^+/2La^+$						
Toll9	100	50	-0.1716	-3.0625	0.0868	0.0224

[a] Number of chromosomes in the sample. Loci with n=100 had more than 100 in the original sample, but were down-sampled to 100 for this analysis.
[b] Number of synonymous segregating sites.
[c] Tajima's *D* calculated using only synonymous sites.
[d] Normalized Fay and Wu's *H* calculated using only synonymous sites.
[e] Ewens-Watterson's haplotype homozygosity statistic calculated using only synonymous sites.
[f] *HEW p*-value with Benjamini and Hochberg correction for multiple tests. Statistical significance of *HEW* was evaluated by comparison to 10^5 neutral coalescent simulations of each sample (see Methods).
[g] For simplicity of presentation, the unnamed LRR genes are labeled according to a shortened form of their AGAP identifier

Table 3: Population genetic summary statistics for haplogroups of *FBN32* in GOUNDRY 2La[a]/2La[a] and *Toll9* in GOUNDRY 2La[+]/2La[+].

Clade	*n*[a]	*h*[b, i]	S[c, i]	π[d, i]	*D*[e]	*H*[f, j]	*EW*[g, j]	*HEW*[h, j]
FBN32								
All	110	10	18	0.0046*	-1.2392	-1.6624	0.508	**
A	80	8 [NS]	17	0.0021***	-2.16**	-2.60*	0.81**	**
B	30	2 [NS]	1**	0.0002*	NA	NA	NA	NA
Toll9								
All	122	42	42	0.0183[NS]	-0.1716	-3.0625**	0.0868*	***
A	20	3 [NS]	3**	0.0013*	NA	NA	NA	NA
B	102	39 [NS]	39	0.0001***	-1.4	-1.79*	0.09**	**

[a] Number of chromosomes in each clade.
[b] Number of haplotypes in each clade.
[c] Number of segregating sites in each clade.
[d] Per site nucleotide diversity calculated on all segregating sites.
[e] Tajima's *D* calculated on synonymous sites.
[f] Normalized Fay and Wu's *H* calculated on synonymous sites.
[g] Ewens-Watterson's haplotype homozygosity statistic calculated on synonymous sites.
[h] *P*-value of the *HEW* test corrected for multiple tests.
[i] For *h,* S, and π, statistical significance was evaluated by comparison to 10^5 coalescent simulations conditioned on clade structure (see Methods).
[j] For the neutrality tests, statistical significance was evaluated by comparison to 10^5 neutral coalescent simulations of each sample sub-set/clade (see Methods). Statistical significance is indicated as * < 0.05, ** < 0.005, *** < 0.0005.

FIGURES

Figure 1: Barplot distributions of genetic differentiation among groups of *A. gambiae* at immune genes.

Genetic differentiation was estimated using Weir and Cockerham's unbiased estimator of F_{ST} between groups of *A. gambiae* at each gene separately. Loci are arranged according to their genomic coordinates with the 2La inversion presented in the inverted arrangement. The vertical bar colors specify genomic region according to the legend. The schematic below indicates the physical distribution of the genes on 2L with 'C' and 'T' representing the centromere and telomere respectively. The PRI region and 2La are also indicated on the chromosome schematic. Dashed lines indicate levels of differentiation estimated from 3rd chromosome microsatellites in corresponding population comparisons in Riehle et al. (Riehle et al. 2011). Asterisks indicate F_{ST} values significantly greater than zero as determined by permutation tests (see Methods). Both the M and S molecular forms were compared to GOUNDRY separately and exhibited qualitatively similar levels and patterns of differentiation, so only the S form comparison. is presented here.

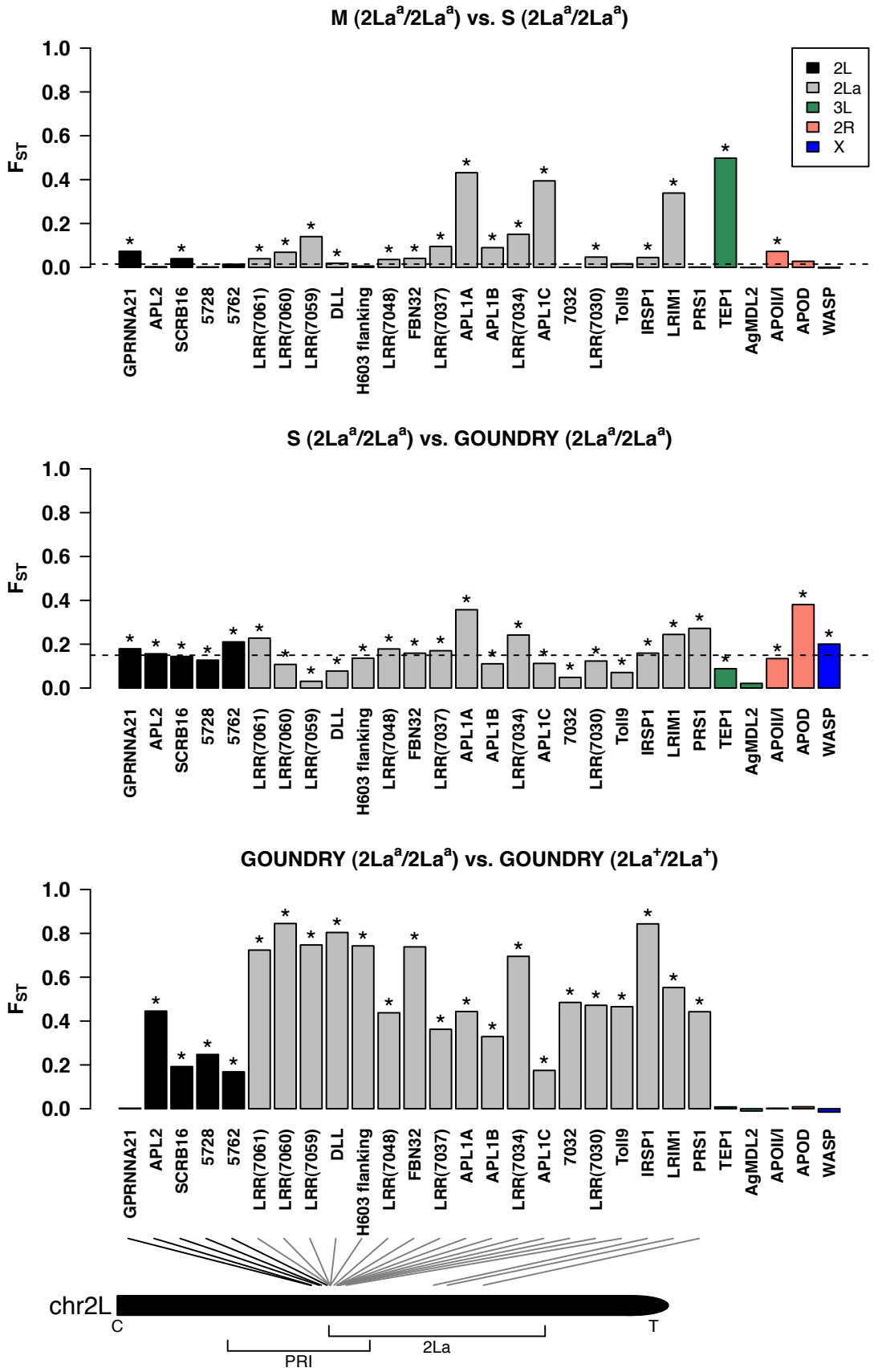

M (2La^a/2La^a) vs. S (2La^a/2La^a)
S (2La^a/2La^a) vs. GOUNDRY (2La^a/2La^a)
GOUNDRY (2La^a/2La^a) vs. GOUNDRY (2La^+/2La^+)
F_ST
1.0
0.8
0.6
0.4
0.2
0.0
2L
2La
3L
2R
X
GPRNNA21
APL2
SCRB16
5728
5762
LRR(7061)
LRR(7060)
LRR(7059)
DLL
H603 flanking
LRR(7048)
FBN32
LRR(7037)
APL1A
APL1B
LRR(7034)
APL1C
7032
LRR(7030)
Toll9
IRSP1
LRIM1
PRS1
TEP1
AgMDL2
APOII/I
APOD
WASP
chr2L
C
T
PRI
2La

Figure 2: Distribution of Tajima's *D*.

Tajima's *D* was calculated for all genes using synonymous sites and both boxplots and data points are presented. The dotted line indicates the expected value of *D* under neutral equilibrium population models. 'GNDRY a/a' and 'GNDRY +/+' refer to GOUNDRY $2La^a/2La^a$ and $2La^+/2La^+$, respectively.

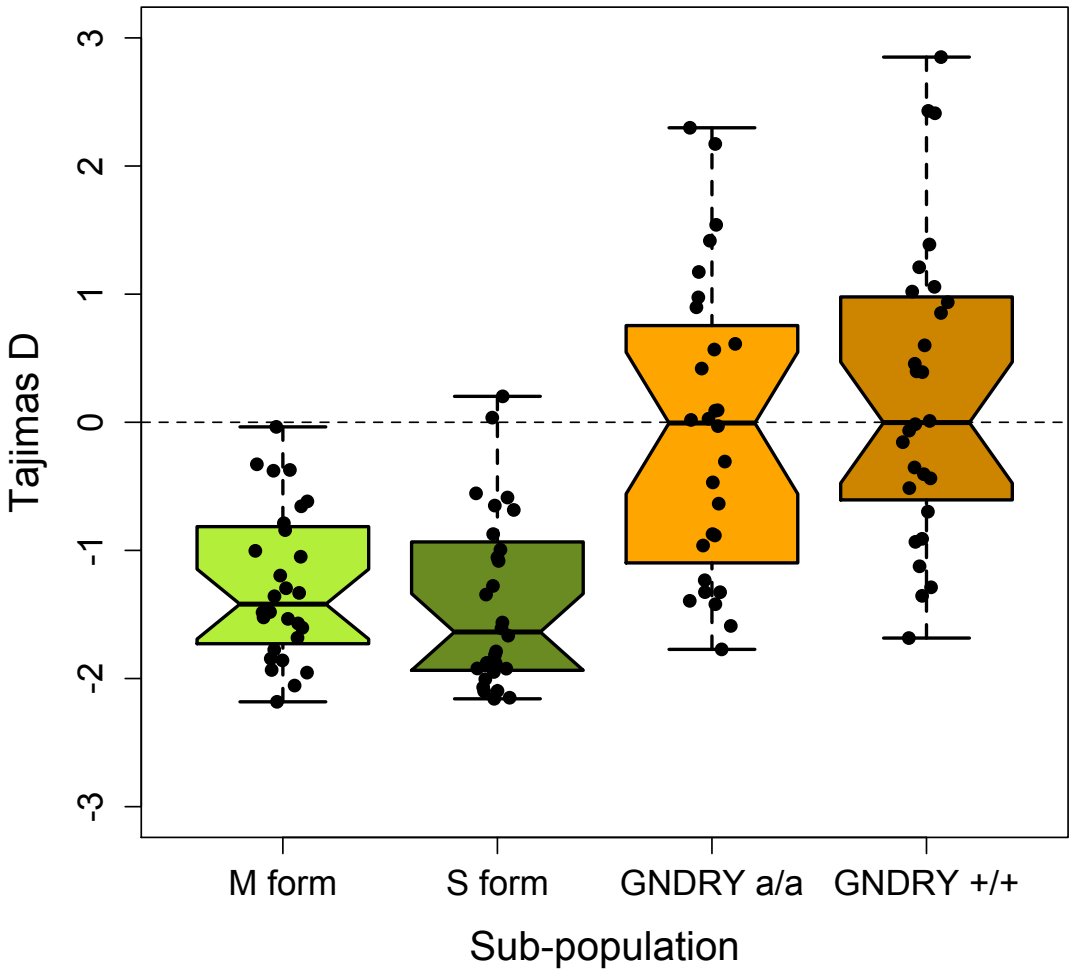


Figure 3: The distribution of Fay and Wu's *H* and the Ewens-Watterson statistic for all loci in *A. gambiae* populations.

Normalized Fay and Wu's *H* and Ewens-Watterson statistics were calculated on synonymous variation at each gene and evaluated using coalescent simulations (see Methods). Red dots indicate that the *HEW* statistic is significant at a 5% threshold level after correcting for multiple tests. Each panel presents the results for a different group and the gene names are presented next to the corresponding data point.

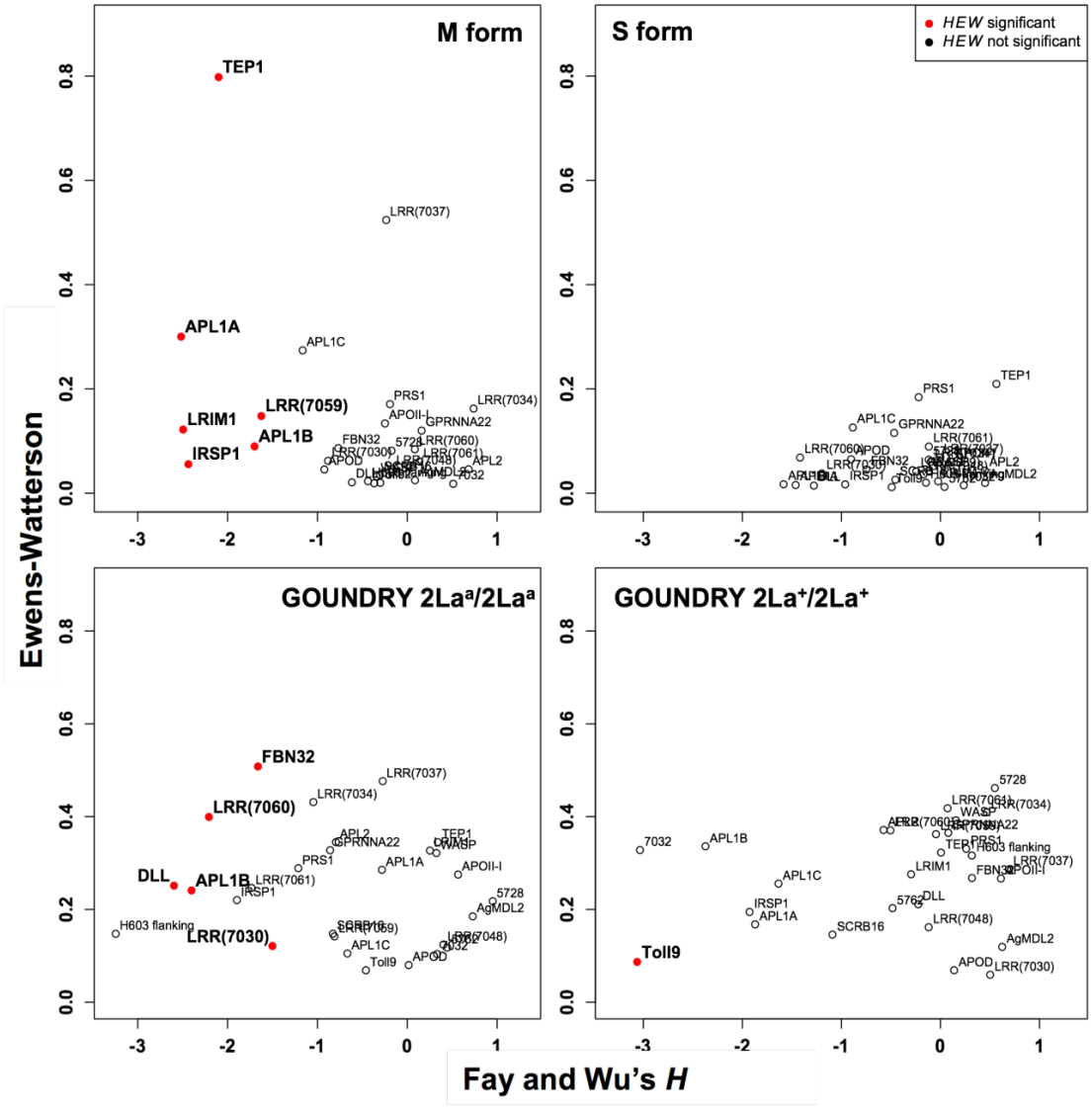


Figure 4: LD plot and Neighbor-joining tree of *FBN32* in GOUNDRY 2La^{a}/2La^{a}

A) Linkage disequilibrium (r^2) plotted among variant sites in the sequenced fragment of *FBN32*. Each pixel represents an r^2 value according to the shade of grey, as indicated in the scale. Higher r^2 values indicate increased linkage among those sites. The exon structure of *FBN32* is placed on the diagonal of the plot to indicate the physical location of each variant site. The frequency of the derived allele frequency (DAF) relative to *A. merus* is plotted in the barplots above and to the side of the LD plot. B) Neighbor-joining tree of all *FBN32* sequences from GOUNDRY 2La^{a}/2La^{a} as well as three outgroups: *Anopheles merus, Anopheles arabiensis*, and *Anopheles quadriannulatus*. The scale bar indicates genetic distance. Large clades of genetically similar taxa were collapsed for presentation.

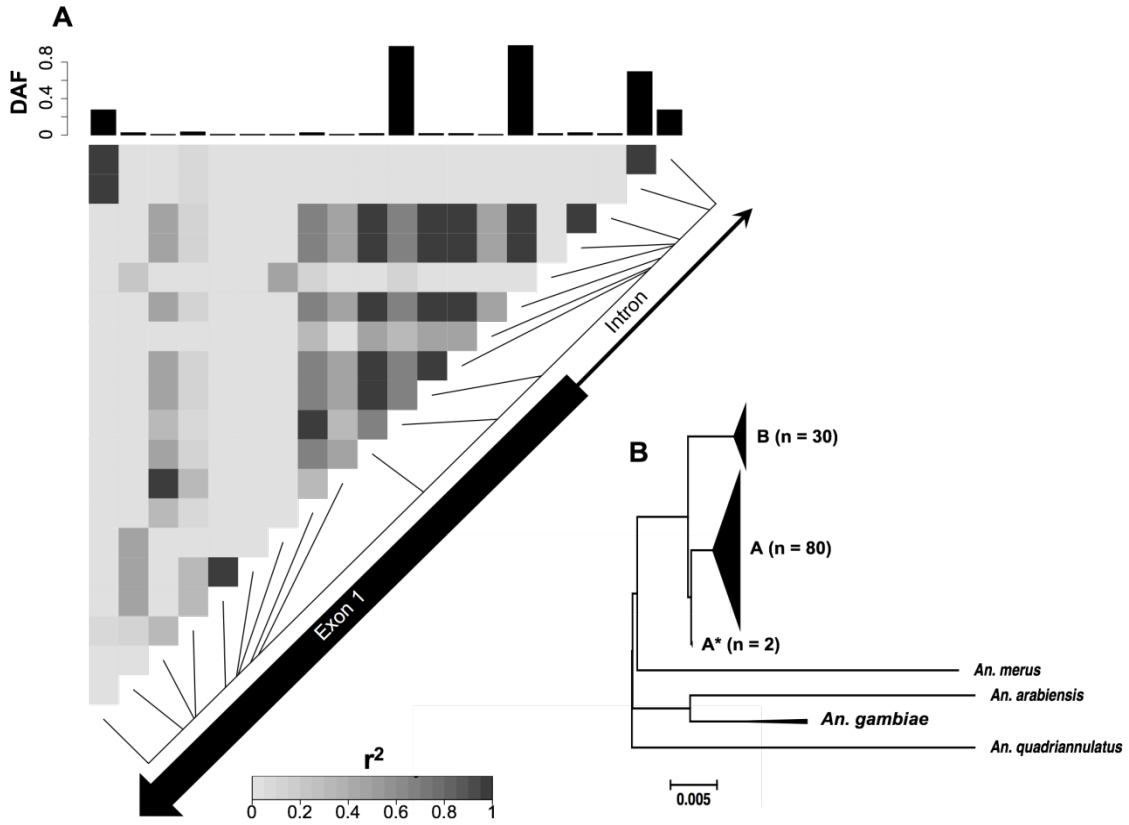


Figure 5: LD plot and Neighbor-joining tree of *Toll9* in GOUNDRY $2La^+/2La^+$.

Linkage disequilibrium (r^2) plotted among variant sites in the sequenced fragment of *Toll9*. Each pixel represents an r^2 value according to the shade of grey, as indicated in the scale. Higher r^2 values indicate increased linkage among those sites. The exon structure of *Toll9* is placed above and beside the plot to indicate the structural location of each variant site. The frequency of the derived allele relative to *A. merus* is plotted in the barplots above and to the side of the LD plot. The red bars indicate the three non-synonymous sites and the triangle delineates the block of linked sites. B) Neighbor-joining tree of all *Toll9* sequences from GOUNDRY $2La^+/2La^+$ as well as three outgroups: *Anopheles merus, Anopheles arabiensis*, and *Anopheles quadriannulatus*. The scale bar indicates genetic distance. Large clades of genetically similar taxa were collapsed for presentation.

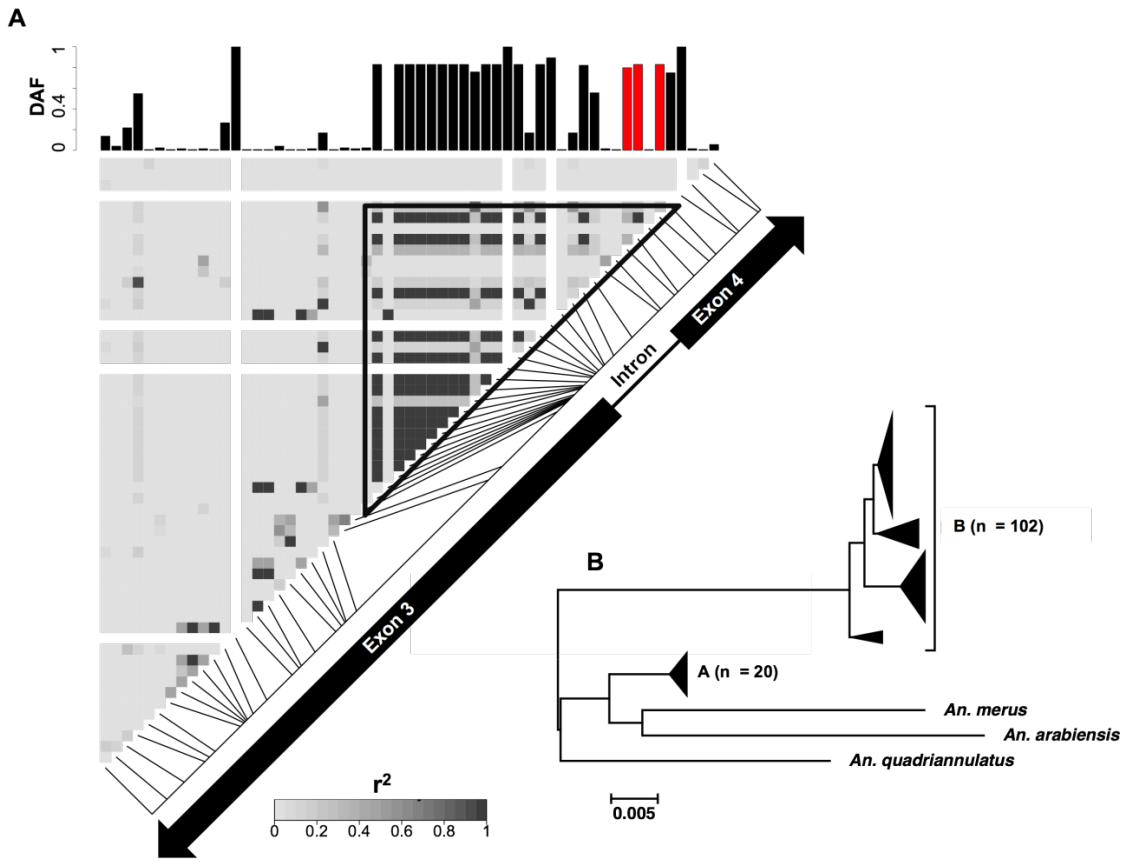
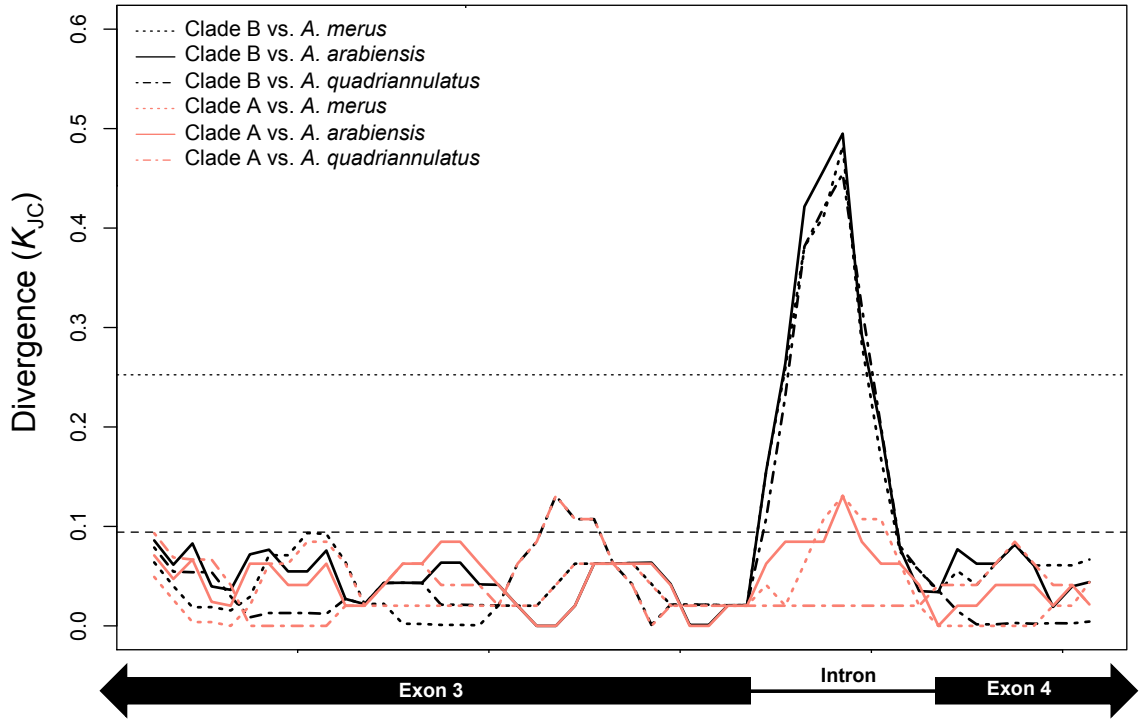


Figure 6: Clade specific patterns of divergence at *Toll9* intron.

Sliding window analysis of Jukes-Cantor corrected divergence (K_{JC}, range 0 to 1) at all sites relating the two haplotype groups (A and B) to three outgroup species. Divergence was calculated for 50bp physical windows shifting 10bp for every consecutive window. The top horizontal dotted line indicates the average maximum divergence per window for *Toll9* in the M and S molecular forms as well as GOUNDRY $2La^a/2La^a$. The lower horizontal dashed line indicates the average maximum divergence per window for all other sequenced loci that included an intron in the GOUNDRY $2La^+/2La^+$ population. The legend indicates the color and line style for each clade/outgroup comparison. The schematic under the plot depicts the exon structure in the sequenced region of *Toll9*.



Supplementary Table 1: Mosquito population assignment, molecular form, and 2La karyotype.
(Microsoft Excel spreadsheet attached)

Each row in the spreadsheet corresponds to an individual mosquito. Columns indicate mosquito identifier, the malaria transmission season in which the mosquito was collected, the molecular form (see Methods), and karyotype of the 2La chromosomal inversion (see Methods). This spreadsheet is provided on www.jacobecrawford.com.

Supplementary Table 2: Fragment and PCR Oligo information.

Identifier	Gene	PCR fragment	Forward oligo	Reverse oligo	Fragment Length	Year Sampled
AGAP005681	GPRNNA21	AmS010b	5'-GCATCATCATCGGTCACCG-3'	5'-CTGAGTCACCTGCAAACCG-3'	528	2007
AGAP005693	APL2	AmS011b	5'-CTATCCACCGTCCAGTTTG-3'	5'-GGTTCGGTGGAATTCTAACC-3'	561	2007
AGAP005716	SCRB16	AmS013a	5'-TGCCGAAGATGAAACGTACG-3'	5'-CGTGCTAAAGATTGTCATCCG-3'	534	2007
AGAP005728	5728	AmS048b	5'-GAATTGCGCAAACAGTCCAG-3'	5'-CACGTTCGATATCTCGCTGA-3'	650	2007
AGAP005762	5762	AmS049b	5'-ATCGATGTCCTCGGCACTAC-3'	5'-GATGGTCAAAGCCAACGAAC-3'	525	2007
AGAP006102	PRS1	AmS052b	5'-CGAAAGTGATTCCGGACAAG-3'	5'-TTATCGCTCGCACAGCAC-3'	304	2007
AGAP006348	LRIM1	AmS059e	5'-CCTCGTACCGCTTGACGAT-3'	5'-GTGACCTGGATCAGTCTGC-3'	576	2007
AGAP006421	IRSP1	AmS053a	5'-ATGGCCATCTGGATAGCTTG-3'	5'-GATATTCGCTCCACCAGCTC-3'	563	2007
AGAP006974	TOLL9	AmS001b	5'-GCATCTCGAACTGACACCAG-3'	5'-TTCGGATATTCCGGAGGAG-3'	536	2007
AGAP007030	LRR(7030)	AmS002a	5'-AGAAACAACAGTCGCAAGCTC-3'	5'-ACATGCTGTGCACCATAAAGAA-3'	501	2007
AGAP007032	7032	AmS046a	5'-CGAAAGCAGCAGAAGAATCG-3'	5'-TGCTGCATCGTTGTGCACG-3'	601	2007
AGAP007033	APL1C	AmS003a	5'-CTTCTGAATAGTGTGCGCGTAA-3'	5'-TGAGACAAACTTTGGAGGTCAG-3'	371	2007
AGAP007034	LRR(7034)	AmS047b	5'-CACAGATGCTCCAGCTTCG-3'	5'-CGTACTTGGTGGACCAACG-3'	738	2007
AGAP007035	APL1B	AmS036a	5'-AGATGGGTCTGTGTTTGCTG-3'	5'-CGCACAACCATTTGATGTGGG-3'	833	2007-2008
AGAP007036	APL1A	AmS037d	5'-TGTGATTTAYMCAACTCATGC-3'	5'-TCAAAGTGCTCGATYTGTCG-3'	802	2007
AGAP007037	LRR(7037)	AmS038a	5'-CGTTGTCCAGTATCCACAG-3'	5'-CAACAACACGATCAAGCAGC-3'	580	2007
AGAP007041	FBN32	AmS004b	5'-GTACGATGGTACGGTCGATTTC-3'	5'-GGTAGGAATGCTTTCCGATTAG-3'	486	2007-2008

Supplementary Table 2: Continued

Identifier	Gene	PCR fragment	Forward oligo	Reverse oligo	Fragment Length	Year Sampled
AGAP007048	LRR(7048)	AmS005b	5'-TTTTTAAGCCTAGCCCGTCTG-3'	5'-CAGCTCGGTAAGCCGATTG-3'	491	2007-2008
AGAP007058	DLL	AmS007b	5'-GTACGTAGCCACCCATCTG-3'	5'-GTTAAGATCTGGTTTCAAAATCG-3'	641	2007-2008
AGAP007059	LRR(7059)	AmS051b	5'-ACCAGGCGCTAGTTCTTTGA-3'	5'-TACCGGCAACGGTCTTTAAC-3'	641	2007-2008
AGAP007060	LRR(7060)	AmS006b	5'-AGTAGCAGGCTCGTGAGTGAG-3'	5'-GAAGCACTTCCACTGGTGCT-3'	658	2007-2008
AGAP007061	LRR(7061)	AmS045b	5'-AGGAAAGATCAAGCAGCTCG-3'	5'-CTGGCGATCGTCAACAACG-3'	532	2007
H603flank	intergenic region	AmS050a	5'-CAAGGCAGCTTCTTCGTTCT-3'	5'-GTTCACGAGTTTGGTCTTGC-3'	522	2007-2008
AGAP001826	APOII/I	AmS056a	5'-CCGTTGACGTGGTACTTGG-3'	5'-ATGTGGCTGCCGATTTCTAC-3'	566	2007
AGAP002593	APOD	AmS057c	5'-GGTACGATCAACACTTCGAG-3'	5'-TGATGCGCATATCCTGTCG-3'	490	2007
AGAP010815	TEP1	AmS054c	5'-CGTATTTGGACGTCCGACG-3'	5'-CCATGCAATCAATGAGAACG-3'	579	2007
AGAP012352	AgMDL1	AmS055b	5'-CAGCAGGATTCACTGTTCTC-3'	5'-ATCCATGAGGTTTCGATCTC-3'	486	2007
AGAP001081	WASP	AmS058b	5'-TTCGTCCTCGGTAGCAAAG-3'	5'-TGGTGCAGCTGTACACGAC-3'	440	2007

Supplementary Table 3: Sequenced fragment physical genomic locations

Identifier	Gene Name	Chromosome	Chromosomal Location[a]	Inside PRI?	Chromosomal Strand
AGAP005681	GPRNNA21	2L	18693376	Yes	-
AGAP005693	APL2	2L	18785249	Yes	+
AGAP005716	SCRB16	2L	19542880	Yes	+
AGAP005728	5728	2L	19769635	Yes	-
AGAP005762	5762	2L	20339460	Yes	-
AGAP006102	PRS1	2L	26685850	No	+
AGAP006348	LRIM1	2L	30329656	No	-
AGAP006421	IRSP1	2L	31693742	No	+
AGAP006974	TOLL9	2L	40434581	Yes	-
AGAP007030	LRR(7030)	2L	41057548	Yes	-
AGAP007032	7032	2L	41245076	Yes	+
AGAP007033	APL1C	2L	41257877	Yes	-
AGAP007034	LRR(7034)	2L	41262272	Yes	-
AGAP007035	APL1B	2L	41266619	Yes	-
AGAP007036	APL1A	2L	41271509	Yes	-
AGAP007037	LRR(7037)	2L	41274607	Yes	-
AGAP007041	FBN32	2L	41381834	Yes	+
AGAP007048	LRR(7048)	2L	41648960	Yes	+
AGAP007058	Distalless	2L	42005592	Yes	-
AGAP007059	LRR(7059)	2L	42005592	Yes	+
AGAP007060	LRR(7060)	2L	42062331	Yes	-
AGAP007061	LRR(7061)	2L	42067616	Yes	-
H603flank	intergenic region	2L	42071847	Yes	-
AGAP001826	APOII/I	2R	11201968	No	-
AGAP002593	APOD	2R	40909880	No	-
AGAP010815	TEP1	3L	11117128	No	-
AGAP012352	AgMDL1	3L	23677600	No	+
AGAP001081	WASP	X	23372448	No	-

[a] Gene starting positions on chromosome 2 given according to locations in *A. gambiae* PEST genome sequence, which corresponds to the inverted 2La+ form of the 2La inversion.

Supplementary Table 4: Population genetic summary statistics for all genes in each population.

Locus	Identifier	n[a]	S_{syn}[b]	*D*[c]	*H*[d]	*EW*[e]	*HEW*[f]
M form							
Toll9	AGAP006974	64	79	-0.6612	-0.3681	0.019	1
LRR(7030)	AGAP007030	64	34	-0.7755	-0.8827	0.062	0.0921
APL1C	AGAP007033	64	21	-0.546	-1.1648	0.2739	0.084
FBN32	AGAP007041	100	33	-1.0852	-0.7677	0.0864	0.0992
LRR(7048)	AGAP007048	100	68	-2.0115	-0.1663	0.046	0.1441
LRR(7060)	AGAP007060	100	42	-1.6659	0.084	0.0842	0.2152
DLL	AGAP007058	100	87	-2.0234	-0.6157	0.0208	0.1036
GPRNNA21	AGAP005681	64	16	-0.7143	0.1597	0.1201	0.2411
APL2	AGAP005693	64	27	-1.0359	0.6826	0.0459	0.7155
SCRB16	AGAP005716	64	47	-0.3479	-0.3037	0.0347	0.1326
APL1B	AGAP007035	100	88	-0.5362	-1.7003	0.0898	0.0294
APL1A	AGAP007036	62	110	-1.15	-2.5176	0.3002	0.014
LRR(7037)	AGAP007037	64	13	-1.682	-0.2348	0.5239	0.1441
LRR(7061)	AGAP007061	64	37	-0.8415	0.1359	0.0605	0.2283
7032	AGAP007032	64	73	-1.4061	0.5115	0.0181	1
LRR(7034)	AGAP007034	62	20	0.174	0.7383	0.1623	0.779
5758	AGAP005758	64	29	-1.5625	-0.1749	0.0815	0.1441
5762	AGAP005762	64	73	-1.6686	-0.2999	0.02	0.1326
H603 flanking	AGAP007058	100	54	-1.866	-0.4372	0.023	0.1326
LRR(7059)	AGAP007059	100	50	-1.8131	-1.6246	0.148	0.0302
PRS1	AGAP006102	64	16	-1.6028	-0.1924	0.1709	0.1441
IRSP1	AGAP006421	64	55	-1.4194	-2.4365	0.0557	0.014
TEP1	AGAP010815	64	10	-2.1817	-2.0996	0.7979	0.0233
AgMDL2	AGAP012352	64	66	-1.7729	0.0863	0.0249	0.2667
APOII-I	AGAP001826	64	23	-1.5496	-0.2477	0.1338	0.1441
APOD	AGAP002593	64	37	-0.8402	-0.9231	0.0454	0.4199
WASP	AGAP001081	64	42	-1.6375	-0.3393	0.0322	0.1326

LRIM1	AGAP006348	64	25	-0.1096	-2.4951	0.1221	0.0336
S form							
Toll9	AGAP006974	100	81	-0.6536	-0.4918	0.0118	0.1352
LRR(7030)	AGAP007030	100	59	-1.1477	-1.1961	0.0378	0.091
APL1C	AGAP007033	100	72	-1.3893	-0.885	0.126	0.0928
FBN32	AGAP007041	100	47	-1.7597	-0.7475	0.0454	0.1117
LRR(7048)	AGAP007048	100	79	-2.2608	-0.1704	0.0364	0.1473
LRR(7060)	AGAP007060	100	50	-2.0471	-1.4173	0.0684	0.091
DLL	AGAP007058	100	96	-1.9502	-1.2775	0.0146	0.091
GPRNNA21	AGAP005681	100	28	-1.7265	-0.4675	0.1156	0.1473
APL2	AGAP005693	100	43	-1.7926	0.4499	0.0434	0.3721
SCRB16	AGAP005716	100	58	-0.7949	-0.4554	0.026	0.1352
APL1B	AGAP007035	100	120	-0.7912	-1.5837	0.0172	0.091
APL1A	AGAP007036	100	147	-0.061	-1.4618	0.0158	0.1352
LRR(7037)	AGAP007037	100	49	-2.0094	-0.0114	0.0688	0.1535
LRR(7061)	AGAP007061	100	33	-0.9212	-0.1171	0.0894	0.1473
7032	AGAP007032	100	83	-1.7822	0.2356	0.0154	0.2135
LRR(7034)	AGAP007034	100	42	-1.9241	-0.099	0.0598	0.1473
5758	AGAP005758	100	37	-1.7718	-0.1192	0.0642	0.1473
5762	AGAP005762	100	100	-1.564	0.0407	0.0122	0.1535
H603 flanking	AGAP007058	100	83	-2.2543	-0.1461	0.0202	0.1473
LRR(7059)	AGAP007059	100	51	-0.986	-0.2439	0.042	0.1473
PRS1	AGAP006102	100	17	-1.4331	-0.2205	0.1842	0.1473
IRSP1	AGAP006421	100	61	-0.8321	-0.9617	0.017	0.0928
TEP1	AGAP010815	100	15	0.317	0.565	0.2096	0.5221
AgMDL2	AGAP012352	100	68	-1.4512	0.4507	0.0198	0.3682
APOII-I	AGAP001826	100	44	-2.0695	0.1091	0.0598	0.1926
APOD	AGAP002593	100	53	-1.819	-0.9022	0.065	0.0928
WASP	AGAP001081	100	77	-2.2284	-0.1177	0.0432	0.1473

LRIM1	AGAP006348	100	47	-0.5435	-0.0234	0.0228	0.1535
GOUNDRY $2La^{+}/2La^{+}$							
Toll9	AGAP006974	100	50	-0.1716	-3.0625	0.0868	0.0224
LRR(7030)	AGAP007030	100	23	-0.1144	0.5018	0.059	0.5892
APL1C	AGAP007033	100	22	-0.1767	-1.6356	0.2556	0.0999
FBN32	AGAP007041	100	23	0.2029	0.3183	0.2676	0.425
LRR(7048)	AGAP007048	100	14	1.5159	-0.1189	0.1616	0.3041
LRR(7060)	AGAP007060	100	10	0.0003	-0.5047	0.3704	0.425
DLL	AGAP007058	100	16	0.9977	-0.2242	0.211	0.3027
GPRNNA21	AGAP005681	100	11	-0.4812	0.0792	0.365	0.3683
APL2	AGAP005693	100	13	-1.0815	-0.5745	0.3712	0.2448
SCRB16	AGAP005716	100	21	1.0104	-1.0902	0.1456	0.2402
APL1B	AGAP007035	100	70	0.7992	-2.3727	0.336	0.0999
APL1A	AGAP007036	100	63	2.4401	-1.8712	0.1678	0.7745
LRR(7037)	AGAP007037	100	8	0.3655	0.6926	0.2866	0.7745
LRR(7061)	AGAP007061	100	8	2.4058	0.072	0.418	0.425
7032	AGAP007032	100	19	1.1648	-3.0339	0.328	0.2402
LRR(7034)	AGAP007034	100	6	0.8999	0.4665	0.4092	0.7579
5758	AGAP005758	100	2	2.3379	0.5497	0.4616	0.7745
5762	AGAP005762	100	20	1.0203	-0.4835	0.2026	0.2448
H603 flanking	AGAP007058	100	16	-1.469	0.3172	0.3162	0.425
LRR(7059)	AGAP007059	100	9	1.6054	-0.0451	0.3622	0.425
PRS1	AGAP006102	100	14	-1.6311	0.2602	0.3306	0.499
IRSP1	AGAP006421	100	10	-0.5248	-1.9277	0.1946	0.425
TEP1	AGAP010815	100	11	0.1881	0.0061	0.3224	0.7133
AgMDL2	AGAP012352	100	48	-0.423	0.6236	0.1192	0.7126
APOII-I	AGAP001826	100	8	1.1057	0.6108	0.2666	0.7186
APOD	AGAP002593	100	38	-0.57	0.1383	0.0688	0.4118

WASP	AGAP001081	100	13	-0.8877	0.1567	0.392	0.4439
LRIM1	AGAP006348	100	22	0.8458	-0.2961	0.2756	0.2863
GOUNDRY 2La[a]/2La[a]							
Toll9	AGAP006974	40	59	0.7813	-0.4603	0.0688	0.1255
LRR(7030)	AGAP007030	34	22	-1.3252	-1.5011	0.1211	0.0482
APL1C	AGAP007033	40	29	-0.1794	-0.6664	0.105	0.16
FBN32	AGAP007041	100	15	-1.2392	-1.6624	0.508	0.0482
LRR(7048)	AGAP007048	100	73	-1.6192	0.4025	0.1242	0.4164
LRR(7060)	AGAP007060	100	38	-2.0531	-2.2083	0.3992	0.028
DLL	AGAP007058	100	72	-2.3283	-2.5965	0.2512	0.0252
GPRNNA21	AGAP005681	40	9	-0.0237	-0.8588	0.3275	0.2329
APL2	AGAP005693	40	17	-1.4316	-0.7944	0.345	0.1101
SCRB16	AGAP005716	40	28	0.3575	-0.8282	0.1475	0.1033
APL1B	AGAP007035	100	70	0.2394	-2.4022	0.2408	0.0252
APL1A	AGAP007036	38	78	2.4679	-0.2821	0.2853	0.5291
LRR(7037)	AGAP007037	38	5	0.0936	-0.2734	0.4765	0.16
LRR(7061)	AGAP007061	40	19	-0.8833	-1.7333	0.2463	0.4164
7032	AGAP007032	40	36	-0.3302	0.3314	0.1038	0.3951
LRR(7034)	AGAP007034	40	10	1.2186	-1.045	0.4313	0.5291
5758	AGAP005758	38	9	1.173	0.9511	0.2175	0.931
5762	AGAP005762	38	25	-0.1673	0.4451	0.1177	0.4444
H603 flanking	AGAP007058	100	24	-1.312	-3.2452	0.1474	0.0994
LRR(7059)	AGAP007059	100	44	-0.924	-0.8104	0.1418	0.1033
PRS1	AGAP006102	40	8	0.5677	-1.2126	0.2888	0.1255
IRSP1	AGAP006421	40	34	-1.2448	-1.8962	0.22	0.1255
TEP1	AGAP010815	40	11	0.2615	0.3394	0.345	0.5291
AgMDL2	AGAP012352	40	44	-0.7708	0.7287	0.185	0.6782
APOII-I	AGAP001826	40	8	0.6999	0.5657	0.275	0.5291

APOD	AGAP002593	40	38	-1.0434	0.0154	0.08	0.2317
WASP	AGAP001081	38	16	-1.3667	0.3246	0.3213	0.4164
LRIM1	AGAP006348	38	17	2.1208	0.2522	0.3269	0.3668

a- Number of chromosomes in the sample. Loci with n=100 had more than 100 in the original sample, but were down-sampled to 100 for this analysis.
b- Number of synonymous segregating sites.
c- Tajima's *D* calculated using only synonymous sites.
d- Normalized Fay and Wu's *H* calculated using only synonymous sites.
e- Ewens-Watterson's haplotype homozygosity statistic calculated using only synonymous sites.
f- *HEW p*-value with Benjamini and Hochberg correction for multiple tests.

Statistical significance of *HEW* was evaluated by comparison to 10^5 neutral coalescent simulations of each sample (see Methods).

Supplementary Table 5: Results from *post-hoc* evaluation of haploype reconstruction for genes with significant *HEW* result.

Gene	Confidence Rank[a]	Mean Probability (<1)[b]	Proportion Imputed Rank[c]	Proportion Phased Rank[d]
M form				
LRIM1	50	0.82	55	26
TEP1	20	0.76	90	100
APL1A	86	0.74	33	12
APL1B	76	0.81	36	19
LRR (7059)	40	0.78	70	72
IRSP1	24	0.73	35	32
GOUNDRY $2La^a/2La^a$				
DLL	53	0.73	27	70
APL1B	94	0.86	39	7
LRR (7030)	2	0.78	24	58
LRR (7060)	64	0.76	37	83
FBN32	74	0.77	62	88
GOUNDRY $2La^+/2La^+$				
TOLL9	63	0.78	29	24

Haplotype reconstruction and imputation was conducted for each gene in each population separately. For GOUNDRY, genes inside the 2La inversion were treated separately according to 2La homokaryotype, while genes outside of the inversion were treated as one group. In total, this resulted in 102 runs of the program PHASE (Stephens, Smith, and Donnelly 2001). When a heterozygous site is phased, it is given a statistical confidence probability ranging from 0.5 to 1, where 0.5 is low confidence or complete ambiguity and 1 being the highest level of confidence.

[a] Genes ranked based on the proportion of phased sites that were given a confidence probability less than 1, with 1 indicating the gene with the most

[b] Mean probability calculated for each gene for sites with probabilities less than 1.

[c] Genes ranked based on the proportion of sequenced sites that were imputed, with 1 representing gene with the most imputed sites.

[d] Genes ranked based on the proportion of sequenced sites that were phased, with 1 representing the gene with the most statistically phased sites.

REFERENCES

Baker DA, Nolan T, Fischer B, Pinder A, Crisanti A, Russell S. 2011. A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector, Anopheles gambiae. BMC Genomics 12:296.

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

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57:289–300.

Besansky N, Anopheles Genomes Cluster Commitee. 2008. Genome analysis of vectorial capacity in major Anopheles vectors of malaria parasites.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140:783–796.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963.

Cohuet A, Krishnakumar S, Simard F, Morlais I, Koutsos A, Fontenille D, Mindrinos M, Kafatos FC. 2008. SNP discovery and molecular evolution in Anopheles gambiae, with special emphasis on innate immune system. BMC Genomics 9:227.

Coluzzi M, Sabatini A, Petrarca V, Di Deco MA. 1979. Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. Trans. R. Soc. Trop. Med. Hyg. 73:483–497.

Costantini C, Ayala D, Guelbeogo WM, Pombi M, Some CY, Bassole IH, Ose K, Fotsing J-M, Sagnon N, Fontenille D, et al. 2009. Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in Anopheles gambiae. BMC Ecol. 9:16.

Crawford JE, Lazzaro BP. 2010. The demographic histories of the M and S molecular forms of Anopheles gambiae s.s. Mol. Biol. Evol. 27:1739–1744.

Diabaté A, Dabiré RK, Heidenberger K, Crawford J, Lamp WO, Culler LE, Lehmann T. 2008. Evidence for divergent selection between the molecular forms of Anopheles gambiae: role of predation. BMC Evol. Biol. 8:5.

Diabaté A, Dabire RK, Millogo N, Lehmann T. 2007. Evaluating the effect of postmating isolation between molecular forms of Anopheles gambiae (Diptera: Culicidae). J. Med. Entomol. 44:60–64.

Dong Y, Aguilar R, Xi Z, Warr E, Mongin E, Dimopoulos G. 2006. Anopheles gambiae Immune Responses to Human and Rodent Plasmodium Parasite Species. PLoS Pathog 2:e52.

Dong Y, Dimopoulos G. 2009. Anopheles Fibrinogen-Related Proteins Provide Expanded Pattern Recognition Capacity Against Bacteria and Malaria Parasites. J. Biol. Chem. 284:9835–9844.

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

Du X, Poltorak A, Wei Y, Beutler B. 2000. Three novel mammalian toll-like receptors: gene structure, expression, and evolution. Eur. Cytokine Netw. 11:362–371.

Favia G, Lanfrancotti A, Spanos L, Sidén-Kiamos I, Louis C. 2001. Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of Anopheles gambiae s.s. Insect Mol. Biol. 10:19–23.
Fay JC, Wu C-I. 2000. Hitchhiking Under Positive Darwinian Selection. Genetics 155:1405–1413.
Fillinger U, Sombroek H, Majambere S, van Loon E, Takken W, Lindsay SW. 2009. Identifying the most productive breeding sites for malaria mosquitoes in The Gambia. Malar. J. 8:62.
Flot J-F. 2010. seqphase: a web tool for interconverting phase input/output files and fasta sequence alignments. Mol Ecol Resour 10:162–166.
Fraiture M, Baxter RHG, Steinert S, Chelliah Y, Frolet C, Quispe-Tintaya W, Hoffmann JA, Blandin SA, Levashina EA. 2009. Two mosquito LRR proteins function as complement control factors in the TEP1-mediated killing of Plasmodium. Cell Host Microbe 5:273–284.
Frolet C, Thoma M, Blandin S, Hoffmann JA, Levashina EA. 2006. Boosting NF-kappaB-dependent basal immunity of Anopheles gambiae aborts development of Plasmodium berghei. Immunity 25:677–685.
Ghosh S, May MJ, Kopp EB. 1998. NF-kappa B and Rel proteins: evolutionarily conserved mediators of immune responses. Annu. Rev. Immunol. 16:225–260.
Gimonneau G, Bouyer J, Morand S, Besansky NJ, Diabate A, Simard F. 2010. A behavioral mechanism underlying ecological divergence in the malaria mosquito Anopheles gambiae. Behavioral Ecology: Official Journal of the International Society for Behavioral Ecology 21:1087–1092.
Gimonneau G, Pombi M, Dabire RK, Diabate A, Morand S, Simard F. 2012. Behavioural responses of Anopheles gambiae sensu stricto M and S molecular form larvae to an aquatic predator in Burkina Faso. Parasites & Vectors 5:65.
Gokudan S, Muta T, Tsuda R, Koori K, Kawahara T, Seki N, Mizunoe Y, Wai SN, Iwanaga S, Kawabata S-I. 1999. Horseshoe Crab Acetyl Group-Recognizing Lectins Involved in Innate Immunity Are Structurally Related to Fibrinogen. PNAS 96:10086–10091.
De Gregorio E, Spellman PT, Tzou P, Rubin GM, Lemaitre B. 2002. The Toll and Imd pathways are the major regulators of the immune response in Drosophila. EMBO J. 21:2568–2579.
Haldane JBS. 1949. Disease and Evolution. La Ricerca Scientifica Supplemento A 19:68–76.
Hamblin MT, Di Rienzo A. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66:1669–1679.
Hudson R, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.
Hudson RR, Sáez AG, Ayala FJ. 1997. DNA Variation at the Sod Locus of Drosophila Melanogaster: An Unfolding Story of Natural Selection. PNAS 94:7725–7729.
Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.
Imler J-L, Hoffmann JA. 2001. Toll receptors in innate immunity. Trends in Cell Biology 11:304–311.
Imler J-L, Zheng L. 2004. Biology of Toll receptors: lessons from insects and mammals. J. Leukoc. Biol. 75:18–26.

Kaplan NL, Darden T, Hudson RR. 1988. The coalescent process in models with selection. Genetics 120:819–829.
Kaplan NL, Hudson RR, Langley CH. 1989. The “hitchhiking effect” revisited. Genetics 123:887–899.
Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765.
Lee CE. 2002. Evolutionary genetics of invasive species. Trends in Ecology & Evolution 17:386–391.
Lehmann T, Diabate A. 2008. The molecular forms of Anopheles gambiae: a phenotypic perspective. Infect. Genet. Evol. 8:737–746.
Lemaitre B, Meister M, Govind S, Georgel P, Steward R, Reichhart JM, Hoffmann JA. 1995. Functional analysis and regulation of nuclear import of dorsal during the immune response in Drosophila. EMBO J. 14:536–545.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.
Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452.
Luna C, Wang X, Huang Y, Zhang J, Zheng L. 2002. Characterization of four Toll related genes during development and immune responses in Anopheles gambiae. Insect Biochem. Mol. Biol. 32:1171–1179.
Marinotti O, Nguyen QK, Calvo E, James AA, Ribeiro JMC. 2005. Microarray analysis of genes showing variable expression following a blood meal in Anopheles gambiae. Insect Mol. Biol. 14:365–373.
Marsden CD, Lee Y, Nieman CC, Sanford MR, Dinis J, Martins C, Rodrigues A, Cornel AJ, Lanzaro GC. 2011. Asymmetric introgression between the M and S forms of the malaria vector, Anopheles gambiae, maintains divergence despite extensive hybridization. Mol. Ecol. 20:4983–4994.
Means TK, Golenbock DT, Fenton MJ. 2000. The biology of Toll-like receptors. Cytokine Growth Factor Rev. 11:219–232.
Meister S, Agianian B, Turlure F, Relógio A, Morlais I, Kafatos FC, Christophides GK. 2009. Anopheles gambiae PGRPLC-mediated defense against bacteria modulates infections with malaria parasites. PLoS Pathog. 5:e1000542.
Meister S, Kanzok SM, Zheng X-L, Luna C, Li T-R, Hoa NT, Clayton JR, White KP, Kafatos FC, Christophides GK, et al. 2005. Immune signaling pathways regulating bacterial and malaria parasite infection of the mosquito Anopheles gambiae. Proc. Natl. Acad. Sci. U.S.A. 102:11420–11425.
Menge DM, Zhong D, Guda T, Gouagna L, Githure J, Beier J, Yan G. 2006. Quantitative trait loci controlling refractoriness to Plasmodium falciparum in natural Anopheles gambiae mosquitoes from a malaria-endemic region in western Kenya. Genetics 173:235–241.
Mitri C, Jacques J-C, Thiery I, Riehle MM, Jiannong Xu, Bischoff E, Morlais I, Nsango SE, Vernick KD, Bourgouin C. 2009. Fine Pathogen Discrimination within the APL1 Gene Family Protects Anopheles gambiae against Human and Rodent Malaria Species. PLoS Pathogens 5:1–10.

Mitri C, Vernick KD. 2012. Anopheles gambiae pathogen susceptibility: the intersection of genetics, immunity and ecology. Current Opinion in Microbiology [Internet]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22538050

Ndiath MO, Brengues C, Konate L, Sokhna C, Boudin C, Trape JF, Fontenille D. 2008. Dynamics of transmission of Plasmodium falciparum by Anopheles arabiensis and the molecular forms M and S of Anopheles gambiae in Dielmo, Senegal. Malar. J. 7:136.

Neafsey DE, Lawniczak MKN, Park DJ, Redmond SN, Coulibaly MB, Traoré SF, Sagnon N, Costantini C, Johnson C, Wiegand RC, et al. 2010. SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. Science 330:514–517.

Neira Oviedo M, Ribeiro JMC, Heyland A, VanEkeris L, Moroz T, Linser PJ. 2009. The salivary transcriptome of Anopheles gambiae (Diptera: Culicidae) larvae: A microarray-based analysis. Insect Biochem. Mol. Biol. 39:382–394.

Niaré O, Markianos K, Volz J, Oduol F, Touré A, Bagayoko M, Sangaré D, Traoré SF, Wang R, Blass C, et al. 2002. Genetic loci affecting resistance to human malaria parasites in a West African mosquito vector population. Science 298:213–216.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. PLoS Biol 3:e170.

Obbard DJ, Linton Y-M, Jiggins FM, Yan G, Little TJ. 2007. Population genetics of Plasmodium resistance genes in Anopheles gambiae: no evidence for strong selection. Mol. Ecol. 16:3497–3510.

Oliveira E, Salgueiro P, Palsson K, Vicente JL, Arez AP, Jaenson TG, Caccone A, Pinto J. 2008. High levels of hybridization between molecular forms of Anopheles gambiae from Guinea Bissau. J. Med. Entomol. 45:1057–1063.

Povelones M, Waterhouse RM, Kafatos FC, Christophides GK. 2009. Leucine-rich repeat protein complex activates mosquito complement in defense against Plasmodium parasites. Science 324:258–261.

Powell JR, Petrarca V, della Torre A, Caccone A, Coluzzi M. 1999. Population structure, speciation, and introgression in the Anopheles gambiae complex. Parassitologia 41:101–113.

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. 20:R208–215.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. Evolution 59:2312–2323.

R Development Core Team. 2011. R: A language and Environment for Statistical Computing. Available from: http://www.R-project.org/

Riehle MM, Guelbeogo WM, Gneme A, Eiglmeier K, Holm I, Bischoff E, Garnier T, Snyder GM, Li X, Markianos K, et al. 2011. A cryptic subgroup of Anopheles gambiae is highly susceptible to human malaria parasites. Science 331:596–598.

Riehle MM, Markianos K, Lambrechts L, Xia A, Sharakhov I, Koella JC, Vernick KD. 2007. A major genetic locus controlling natural Plasmodium falciparum infection is shared by East and West African Anopheles gambiae. Malar. J. 6:87.

Riehle MM, Markianos K, Niaré O, Xu J, Li J, Touré AM, Podiougou B, Oduol F, Diawara S, Diallo M, et al. 2006. Natural malaria infection in Anopheles gambiae is regulated by a single genomic control region. Science 312:577–579.
Rottschaefer SM, Riehle MM, Coulibaly B, Sacko M, Niaré O, Morlais I, Traoré SF, Vernick KD, Lazzaro BP. 2011. Exceptional Diversity, Maintenance of Polymorphism, and Recent Directional Selection on the APL1 Malaria Resistance Genes of Anopheles gambiae. PLoS Biol 9:e1000600.
Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.
Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in Drosophila. Nat. Genet. 39:1461–1468.
Santolamazza F, Mancini E, Simard F, Qi Y, Tu Z, della Torre A. 2008. Insertion polymorphisms of SINE200 retrotransposons within speciation islands of Anopheles gambiae molecular forms. Malar. J. 7:163.
Schlenke TA, Begun DJ. 2003. Natural selection drives Drosophila immune system evolution. Genetics 164:1471–1480.
Shin J-H, Blay S, Mcneney B, Graham J. 2006. LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms. Journal Of Statistical Software 16:Code Snippet 3.
Simard F, Ayala D, Kamdem GC, Pombi M, Etouna J, Ose K, Fotsing JM, Fontenille D, Besansky NJ, Costantini C. 2009. Ecological niche partitioning between Anopheles gambiae molecular forms in Cameroon: the ecological side of speciation. BMC ecology 9:17.
Slotman MA, Parmakelis A, Marshall JC, Awono-Ambene PH, Antonio-Nkondjo C, Simard F, Caccone A, Powell JR. 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.
Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res 23:23–35.
Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68:978–989.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585.
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28:2731–2739.
Tiffin P, Moeller DA. 2006. Molecular evolution of plant immune system genes. Trends in Genetics 22:662–670.
della Torre A, Costantini C, Besansky NJ, Caccone A, Petrarca V, Powell JR, Coluzzi M. 2002. Speciation Within Anopheles Gambiae-- the Glass Is Half Full. Science 298:115–117.
della Torre A, Fanello C, Akogbeto M, Dossou-yovo J, Favia G, Petrarca V, Coluzzi M. 2001. Molecular evidence of incipient speciation within Anopheles gambiae s.s. in West Africa. Insect Mol. Biol. 10:9–18.
Trout Fryxell RT, Nieman CC, Fofana A, Lee Y, Traoré SF, Cornel AJ, Luckhart S, Lanzaro GC. 2012. Differential Plasmodium falciparum infection of Anopheles gambiae s.s. molecular and chromosomal forms in Mali. Malaria journal 11:133.

Wall JD. 2000. A Comparison of Estimators of the Population Recombination Rate. Mol Biol Evol 17:156–163.
Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, Barillas-Mury C, Bian G, Blandin S, Christensen BM, et al. 2007. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. Science 316:1738–1743.
Waterhouse RM, Povelones M, Christophides GK. 2010. Sequence-structure-function relations of the mosquito leucine-rich repeat immune proteins. BMC Genomics 11:531.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7:256–276.
Watterson GA. 1978. The Homozygosity Test of Neutrality. Genetics 88:405–417.
Weir B, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution:1358–1370.
White BJ, Lawniczak MKN, Cheng C, Coulibaly MB, Wilson MD, Sagnon N, Costantini C, Simard F, Christophides GK, Besansky NJ. 2011. Adaptive divergence between incipient species of Anopheles gambiae increases resistance to Plasmodium. Proc. Natl. Acad. Sci. U.S.A. 108:244–249.
White BJ, Santolamazza F, Kamau L, Pombi M, Grushko O, Mouline K, Brengues C, Guelbeogo W, Coulibaly M, Kayondo JK, et al. 2007. Molecular karyotyping of the 2La inversion in Anopheles gambiae. The American journal of tropical medicine and hygiene 76:334–339.
Wondji C, Frédéric S, Petrarca V, Etang J, Santolamazza F, Della Torre A, Fontenille D. 2005. Species and populations of the Anopheles gambiae complex in Cameroon with special emphasis on chromosomal and molecular forms of Anopheles gambiae s.s. J. Med. Entomol. 42:998–1005.
Zeng K, Fu Y-X, Shi S, Wu C-I. 2006. Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. Genetics 174:1431–1439.
Zeng K, Mano S, Shi S, Wu C-I. 2007. Comparisons of site- and haplotype-frequency methods for detecting positive selection. Mol. Biol. Evol. 24:1562–1574.
Zeng K, Shi S, Wu C-I. 2007. Compound tests for the detection of hitchhiking under positive selection. Mol. Biol. Evol. 24:1898–1908.

CHAPTER 5:

Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data

Published in Frontiers in Evolutionary and Population Genetics

AUTHORS:

Jacob E. Crawford and Brian P. Lazzaro

Department of Entomology, Cornell University, New York, USA.

CORRESPONDENCE:
Brian P. Lazzaro
Department of Entomology
Cornell University
Ithaca, NY, 14853, USA
bplazzaro@cornell.edu

RUNNING TITLE: Assessing next-generation population genomics

KEYWORDS: Next-generation sequencing, population genetics, population genomics, natural selection, demography, population structure

ABSTRACT

Next-generation sequencing technologies have made it possible to address population genetic questions in almost any system, but high error rates associated with such data can introduce significant biases into downstream analyses, necessitating careful experimental design and interpretation in studies based on short-read sequencing. Exploration of population genetic analyses based on next-generation sequencing has revealed some of the potential biases, but previous work has emphasized parameters relevant to human population genetics and further examination of parameters relevant to other systems is necessary, including situations where sample sizes are small and genetic variation is high. To assess experimental power to address several principal objectives of population genetic studies under these conditions, we simulated population samples under selective sweep, population growth, and population subdivision models and tested the power to accurately infer population genetic parameters from sequence polymorphism data obtained through simulated 4x, 8x, and 15x coverage short-read sequence data. We found that estimates of population genetic differentiation and population growth parameters were systematically biased when inference was based on 4x sequencing, but biases were markedly reduced at even 8x read depth. We also found that the power to identify footprints of positive selection depends on an interaction between read depth and the strength of selection, with strong selection being recovered consistently at all read depths, but weak selection requiring deeper read depths for reliable detection. Although we have explored only a small subset of the many possible experimental designs and population genetic models, using only one SNP calling approach, our results reveal some general patterns and provide some assessment of what biases could be expected under similar experimental structures.

INTRODUCTION

Principal objectives in population genetics are to identify targets of natural selection, infer historical shifts in demography, and define genetic differentiation among groups. Over the past four decades, the power to address these questions has improved markedly with the increase in scale and availability of genetic markers. The recent arrival of next-generation sequencing (NGS) marks another shift on multiple scales (Pool et al. 2010). The relative low cost and high throughput nature of next-generation sequencing technologies has made it possible to collect full genome sequence data on population samples, providing the opportunity to address population genetic questions at the genomic scale, sometimes across multiple populations (e.g. Xia et al. 2009; Durbin et al. 2010; Magwene et al. 2011). For example, NGS makes possible unbiased scans of the genome for signatures of positive selection (Durbin et al. 2010), tests of demography and population structure that include rare (<5%) variants (Henn et al. 2010; Gravel et al. 2011), as well as genomic mapping of population parameters such as nucleotide diversity or fine-scale linkage disequilibrium (e.g. Branca et al. 2011; e.g. Magwene et al. 2011).

While NGS has expanded the realm of possible experiments, NGS-based population genomic analyses and experimental designs are not yet standard and free of complications. The main challenges to population genomic analysis using NGS data stem from the substantially higher error rates in NGS relative to traditional Sanger sequencing, which complicates identification of low-frequency variants in populations (Johnson & Slatkin 2006, 2008; Hellmann et al. 2008; Lynch 2008, 2009; Jiang et al. 2009), uneven sequencing of the homologous chromosomes in a diploid individual, which may compromise accuracy in detecting heterozygotes (Hellmann et al. 2008; Johnson & Slatkin 2008; Lynch 2008, 2009; Jiang et al. 2009), and a higher false

negative SNP detection rate due to the Poisson read sampling, which can result in some regions not being sequenced at all (Durbin et al. 2010). One approach to mitigating these challenges is to sequence each sampled individual to substantially greater coverage depth or to obtain larger sample sizes of individuals. However, current experimental designs typically consist of either small, deeply sequenced samples (Xia et al. 2009; Branca et al. 2011; Magwene et al. 2011) or large samples sequenced to low read-depths (Durbin et al. 2010), reflecting a common and practical trade-off between sample size and sequencing depth. In general practice, population genomic experiments in ecological and other non-model systems will likely have to compromise on both sample size and read depth, possibly resulting in losses in power and biases not incurred with larger experimental designs.

In addition to modifying experimental design to mitigate challenges related to NGS data analysis, statistical corrections may also provide a means for accommodating uncertainty in the data. Most current methods for conducting population genetic analysis are based on allele frequencies (reviewed in Nielsen 2005) or a summary of allele frequencies (e.g. Gutenkunst et al. 2009; but see Yi et al. 2010), and, broadly speaking, two statistical approaches have been proposed to estimate this information from NGS data. The first approach entails calling genotypes of each individual using either a Bayesian or Likelihood framework (Hoberman et al. 2009; Li et al. 2009b; Bansal et al. 2010; DePristo et al. 2011). The other approach attempts to estimate allele frequencies directly from the data without first inferring individual genotypes (Lynch 2009; Kim et al. 2010, 2011; Martin et al. 2010). In some cases, a posterior probability is generated that provides a quantification of the uncertainty of each genotype call (e.g. Martin et al. 2010; DePristo et al. 2011) that could be directly incorporated into population genetic analyses (e.g. Yi et al. 2010). However, until population genetic analyses are further adapted to

incorporate posterior probabilities, standard population genetic analyses must be applied directly to genotype calls. The number of applications of such statistical approaches to empirical data is thus far relatively small with a bias towards human-based studies (e.g. Hellmann et al. 2008; Durbin et al. 2010; Yi et al. 2010), but some examples in non-human systems exist as well (e.g. Williams et al. 2010; Ahmad et al. 2011). More importantly, the biases introduced when population genetic analyses are applied to genotypes inferred from NGS data have not been well characterized, particularly in systems other than humans.

The aim of the present study is to determine how variation in the structure of NGS experiments and inaccuracies inherent to NGS-based genotype calling impact the ability to address several common population genetic questions in non-model or ecological systems. In particular, we sought to provide some assessment of what can be accomplished with NGS data when genetic variation is high, sample sizes and sequencing budgets are small and independent datasets are not available for calibration. We simulated population genetic samples under Wright-Fisher equilibrium, selective sweep, population growth, and population sub-division models. Short read datasets were generated *in silico* and processed through a read-mapping and multi-sample-genotyping-based SNP calling pipeline similar to that used by the human 1000 Genomes Project (Durbin et al. 2010). We determined the power to infer population genetic parameters and conduct populatin genetic tests using NGS data of varying depths. Our results demonstrate that very low sequencing depth introduces systematic biases under some, but not all, inference frameworks, yet significant power and accuracy is recovered with as little as 8x sequencing depth.

METHODS

A graphical flowchart presentation of our analysis pipeline can be found in Supplementary Figure 1.

Coalescent Simulations

We conducted coalescent simulations to generate population samples under a variety of equilibrium and non-equilibrium population models. Our null model is at Wright-Fisher equilibrium with no natural selection, constant population size, and complete random mating. Our alterative models included selective sweeps, exponential population growth, and sub-divided populations. The general structure of our simulation approach was to simulate 100 population samples comprised of 30 haplotypes that were 30-kilobases (kb) in length per sample under each set of measured parameters. The departures from this structure were that we conducted 500 simulations under the growth model, and for the sub-divided population, we simulated two subpopulations with 30 haplotypes each (total of 60 haplotypes per iteration). Because we wanted to consider levels of genetic variation seen in many organisms with naturally large population sizes, we modeled a population with an effective population size (N) of 10^6, a per base mutation of $3.5x10^{-9}$ (Keightley et al. 2009), and a recombination rate of 10^{-8} per base per generation. These parameters correspond to levels of genetic variation of $\theta = 0.011$ per site for the Wright-Fisher model and $\theta = 0.027$ per site for population structure model, and are similar to what one might expect from abundant insects with large geographic ranges such as *Drosophila* (e.g. Charlesworth 2009) or *Anopheles* mosquitoes (e.g. Michel et al. 2006).

We used the coalescent simulation program *ssw* to simulate population samples under the selective sweep model of Kim and Stephan (Kim & Stephan 2002). We conducted two rounds of simulations for each parameterization of the sweep model. In

the first round, we conducted simulations under a rejection framework in which we kept the simulation result only if the likelihood ratio obtained using the program *SweepFinder* (Nielsen et al. 2005) was deemed significant (described below). As such, this round of simulations contains only datasets that contain patterns of polymorphism that reject in favor of selection with true genotypes, thereby providing a direct contrast when the simulations are processed through the sequencing pipeline and re-tested for selection. In a second round of simulations, we conducted a set of sweep simulations that were retained without regard to whether the null hypothesis of no selection could be rejected with the complete data set. This round of simulations provides a more complete power curve reflecting both the inherent power of the test implemented in *Sweepfinder* as well as the loss of power due to sequencing. For both rounds, we generated population samples under 6 parameterizations of the sweep model, including three strengths of selection ($\alpha = 2Ns$) varying from weak ($\alpha = 50$) to moderate ($\alpha = 200$) to strong ($\alpha = 1000$). For each value of selection strength, we also varied the time since completion of the sweep (τ [in units of $2N_{CURR}$ generations] = 0.01 and 0.005) to reflect a variety of plausible recent selective sweep events. In addition, we simulated population samples under the null Wright-Fisher equilibrium model with *ssw*. *ssw* allows for both coding and non-coding sequences to be modeled, and we included six coding regions (covering ~21% of the 30kb simulated) where the mutation rate was reduced by a factor of 0.3, the default value in *ssw*.

To simulate population samples under growth and structure models, we used the coalescent simulation program *ms* (Hudson 2002). The growth model included exponential growth resulting in a doubling of the population size ($N_{ANC}/N_{CURR} = 0.5$) over the last $1N_{CURR}$ generations. We also simulated paired samples from diverging populations. These models were based on the island model in which, going backwards

in time, two sub-populations exchanged migrants at a rate of $4N_{CURR}m = 0.05$, and at either $2N_{CURR}$, $1N_{CURR}$, $0.025N_{CURR}$ generations ago began exchanging migrants at a much greater rate ($4N_{CURR}m = 10$ or 100) meant to reflect near panmixia. These models generated paired sub-populations with current F_{ST} values of approximately 0.54, 0.37, 0.15, and 0.01. Chromosomes were sampled evenly from each current sub-population and population assignment was maintained throughout the analysis.

Short-read generation and mapping

Short-read sequence libraries were computationally generated and mapped to a reference sequence. A reference sequence of randomly chosen nucleotides was generated and used as the starting material for each simulated chromosome. To generate chromosomes, simulated polymorphism data from above was used as a guide for applying nucleotide changes resulting in a population sample of nucleotide sequences that reflects the simulated sample. Nucleotide changes, or SNPs, were applied to reflect the presence of derived alleles in the simulated polymorphism data. Diploid "individuals" were generated by randomly pairing simulated chromosomes. These diploid sequences were then computationally fragmented and sampled to generate short-reads that emulate Illumina's HiSeq 2000 platform using the short-read simulation program SimSeq (Earl et al. 2011). 100 base-pair (bp) paired-end reads were sampled with an average insert size of 500bp (std dev = 50) and a duplicate probability of 0.01. SimSeq adds a fixed rate and distribution of 'sequencing' errors using an error profile trained on alignments of human-derived HiSeq 2000 reads. Indels were not included in this model. While human data was used to train the error model, patterns of sequencing errors are largely based on sequencing platform and are likely to be similar between experimental systems. We converted BAM alignment files generated by

SimSeq into SAM format using SAMtools (Li et al. 2009a), and ultimately into FastQ short-read libraries using the BAMtoFastQ program in the Picard Tools (v1.48) package (http://picard.sourceforge.net/). Paired-end short-read libraries were aligned to the reference sequence generated above using BWA (Li & Durbin 2009) allowing for an edit distance of 4 between each read and the reference, except for samples from the population structure models which we mapped with an edit distance of 5 to accommodate the higher number of SNPs in these samples. Reads with duplicate mapping positions were removed with the *rmdup* function in SAMtools (Li et al. 2009a). Cleaned BAM files from each of 15 diploids per population were used in subsequent steps for SNP calling. To determine how the power of inference is affected by read depth, we generated short-read libraries for each diploid such that the resulting alignments would achieve an average of 4, 8 and 15X read depth. Thus, for each individual diploid, we generated BAM alignments at each of the three read depths resulting in a total of 45 BAM alignment files per population sample.

SNP calling

We used the Genome Analysis Toolkit (GATK v. 1.1-30) to recalibrate FastQ quality scores and call candidate SNPs. The GATK implements a FastQ quality score re-calibration step that is designed to recalibrate reported quality scores by accounting for technology and sequence features that are known to co-vary with the reported quality score (DePristo et al. 2011). The GATK builds a recalibration model by ignoring all sites in a dbSNP database file provided by the user, correlating sequence and technology features with reported quality scores at remaining sites that differ from the reference, and calculating a recalibrated score based on residuals from this model (DePristo et al. 2011). This approach is designed for human genetic analysis, and relies heavily on the

well- populated human dbSNP database. It is less ideal for systems with fewer independent SNP datasets. Although, it may be possible in a full genome re-sequencing study to use high-confidence SNPs as the "known" set, this step is likely to be project-specific so we opted not to model SNP ascertainment in this way. Instead, we circumvented the issue by training the re-calibration model on a sample of 15 diploid alignments that are 20 Mb in length and entirely lack "true" SNPs, but that have been processed using SimSeq to introduce 'sequencing' error. We confirmed the validity of the re-calibration model and its effectiveness by analyzing the covariates and quality scores before and after re-calibration using the GATK. All BAM alignments in the primary study were re-calibrated with this model, and re-calibrated BAMs were used for subsequent SNP calling.

We tested the GATK's Unified Genotyper (UG) for calling SNPs in our simulated population samples. The UG considers all individuals simultaneously to make genotype calls in a technology-aware fashion and uses a Bayesian genotype likelihood model to calculate genotypes for each individual and estimate the allele frequency at each variant site. We submitted each batch of 15 BAMs to the UG with default settings except the expected heterozygosity was set equal to the scaled mutation rate used in the simulations. The UG generates a Phred-scaled Quality score (Q) for each variant indicating the probability that a SNP exists at each site, where Q of 20 indicates a 1 in 100 chance that the call is incorrect. These Q scores are calculated without regard to the surrounding sequence context so the GATK implements a sophisticated variant quality score re-calibration method that has been shown to be more effective at sorting true from false positives than hard filtering based on un-calibrated Q scores or other parameters (DePristo et al. 2011). However, despite efforts to simulate larger SNPs datasets for training, we did not find that re-calibration led to better distinction between

true and false positives under our simulation framework (data not shown). Therefore, we opted to use hard filtering based on the Q score and use all SNPs with a score of at least 5. Although this is a quite liberal threshold compared to the standard of Q20, preliminary analyses indicated that, even at 4X read depth, many true positives had Q values on this scale (data not shown). We found a threshold of 5 struck a good balance of minimizing false negatives while permitting a small number of false positives (8.4 false negatives for each false positive for 4x read depth). SNP calls were made for each population sample, converted into appropriate input formats and used in subsequent population genetic analyses.

Population genetic analysis

The goal of this study was to determine the impact of the short-read sequencing, alignment and SNP calling process on the inference of population genetic patterns. Therefore, for each simulation under a specific model and sequence read depth, we quantified the difference between the population genetic model inferred using complete, pre-sequencing data and the model inferred from post-sequencing data to directly measure the effect of sequencing.

To infer selective sweeps, we scanned SNP sets from the sweep simulations as well as the Wright-Fisher simulations using the Parametric version of the Composite Likelihood Ratio Test implemented in the program *SweepFinder* (Nielsen et al. 2005), which compares the likelihood of a selective sweep under the model of Kim and Stephan (Kim & Stephan 2002) to the likelihood of a model without selection based on background allele frequencies. We used a grid size of 25, which corresponds to 1250bp. For this part of the study, we aimed to compare the power to infer selection before and after sequencing as well as to specifically quantify the false-negative rate after

sequencing. To quantify the difference in power to infer selection before and after sequencing, we searched for selective sweeps in the pre-sequencing data and then searched the SNP set inferred from the post-sequencing alignments. To determine significance, we established a null distribution of the likelihood ratio by conducting 10^4 simulations under the neutral Wright-Fisher equilibrium model, and collecting the maximum likelihood ratio from each simulation. The simulated 'experimental' sweep datasets were considered significant if their likelihood ratio was greater than the 95% threshold from this null distribution. Simulations from both the significance-naïve set and the significance-enriched set (see section on *Coalescent Simulations*) were analyzed in this way. The proportions of significant CLRT results were compared between read-depths and sweep models.

To infer the population growth parameters, we searched a grid of population growth models using a Poisson log-likelihood approach to determine the fit of the complete and inferred data to each simulated dataset. We used the program PRFREQ (Boyko et al. 2008) to calculate the expected site-frequency spectrum (SFS) across the grid and find the model within the grid with the maximum likelihood for each SNP set. We used the Poisson likelihood function in PRFREQ to determine the best fit between the data and the models, first using the full pre-sequencing data and then with the SFS inferred from each post-sequencing SNP set. We used the scaled mutation rate and effective population size used in the simulations above, the instantaneous population growth model (two epochs), and the neutral distribution of selective effects ($2Ns = 0$). The grid consisted of 16 grid points for the timing of growth (τ), varying from 0.05 to 0.9, in units of $2N_{CURR}$ and 16 points for the ratio of ancestral to current population size (ω), again varying from 0.05 to 0.9, for a total of 256 models. Since SNPs from the same population sample were used for inference before and after sequencing, we compared

the pre-sequencing model to the post-sequencing model by subtracting the post-sequencing parameter values from the pre-sequencing values and plotting the difference.

To determine the effect of short-read sequencing on inference of genetic divergence between populations, we calculated F_{ST} between the two simulated sub-populations before and after short-read sequencing. We estimated global genetic differentiation between the two sub-populations across the 30kb fragment using Weir and Cockerham's unbiased estimator (Weir & Cockerham 1984) implemented in an R script written by Eva Chan (www.evachan.org). We compared the differences between pre- and post-sequencing F_{ST} values between read depths using a Paired Student's t-test [t.test function in R (R Development Core Team 2011)] to determine whether increasing read depth resulted in significantly better inference of population differentiation. We also estimated a line of best-fit for each read depth using the *lm* function in R (R Development Core Team 2011).

RESULTS

SNP Recovery

To quantify the effect of short-read sequencing on the power to infer population genetic models, we simulated a typical empirical re-sequencing pipeline including sequencing and SNP-calling errors inherent to such experimental frameworks. For all population genetic models, short-read datasets were generated at three read depths, aligned to a simulated reference and queried for SNPs. We found that, under the parameterization of this simulation pipeline, read depth had a significant effect on the rate of true SNP recovery as well as on the rate of false-positive SNP ascertainment. The rate of true SNP recovery was high, increasing as a function of read depth, with

average recovery rates across population genetic models of 86.7% at 4x read depth (standard deviation, σ = 0.0134), 95.7% at 8x (σ = 0.0067), and 99.2% at 15x (σ = 0.0025). Furthermore, even without using the false positive SNP culling steps in the GATK, the false positive rates across all models were reasonably low with 4.1% (σ = 0.0206) of called SNPs being spurious at 4x, 0.95% (σ = 0.0047) at 8x, and 0.39% (σ = 0.0025) at 15x, highlighting the effect of read depth on the false positive rate. Importantly, false negative and false positive rates differed among population genetic models and disproportionately affected low (<0.1) frequency SNPs (Table 1). For example, at 4x read depth, the rates of true SNP recovery (or 1 minus the false negative rates) differed among population genetic models, with the rate under the selective sweep model being significantly lower than that under the Wright-Fisher equilibrium model ($t_{df=170} = 2.17$, $p = 0.0314$; Figure 1), and the rate under the growth model being significantly lower than the sweep model ($t_{df=165} = 8.02$, $p = 1.85 \times 10^{-13}$; Figure 1). On the other hand, the rate of false positives was highest under the selection model at 6.1%, which was significantly greater than the rate under the growth model ($t_{df=195} = 5.67$, $p = 5.09 \times 10^{-8}$), and the rate under the growth model (5.2%) was significantly greater than that under the equilibrium model (4.4%; $t_{df=195} = 7.93$, $p = 1.7 \times 10^{-13}$). Since rare variants are disproportionately missed at low read depths (Figure 2; Table 1), the lower rate of recovery under the selective sweep and growth models can likely be attributed to the proportionally greater number of low-frequency variants under these models relative to the Wright-Fisher model (Supp Fig 2). The structure model showed a lower true SNP recover rate relative to Wright-Fisher (Figure 1) and a substantially lower rate of false positive SNPs (Table 1), but these are not a fair comparison since the structure model is different genotyping environment as it is comprised of twice as many chromosomes and approximately three times as many true SNPs as the other models.

Inferring selective sweeps

We assessed the effect of short read sequencing on the power to infer a selective sweep by testing for selection in patterns of variation in simulated population samples before and after simulated next-generation sequencing. When we tested a set of random sweep simulations under each sweep model parameterization, we found that power curves are quite comparable before and after simulated sequencing, with only minor loss in power to identify the signature of a selective sweep at lower read depths (Supplementary Figure 3), although further losses might be incurred under different parameterizations of the selection model, including modeling of older sweeps. To more directly quantify the loss in power of inference at lower read depths, we conducted a second set of genealogical simulations in which we required that the simulation give a significant result prior to simulated sequencing. In this case, we found that the power to infer selection depended on both the strength of selection and the depth of sequencing (Figure 3). Depth of sequencing did not have a strong effect on the power to identify the signature of selection after sequencing when selection was strong ($2Ns = 1000$; Figure 3). But, when the strength of selection is weak ($2Ns = 50$), we found a 29.9% reduction in power to infer selection with 8x sequencing relative to 15x, and an additional 29.1% reduction with 4x relative to 8x read depth (Figure 3). Interestingly, a small but noteworthy number of simulated samples that did not give a significant test based on full sequence data yielded a significant result after simulated next-generation sequencing (data not shown). Inspection of these test results indicated that many of the pre-sequencing likelihood ratios were nearly significant in their rejection of the null hypothesis. We suspect that the higher rate of undetected SNPs in low pass sequencing datasets altered inference of the background (no-selection) model just enough to result

in a significant likelihood ratio supporting selection. Collectively, our results indicate that strong and recent selective sweeps can be detected reliably even with low read depths, but that deeper sequencing will be required for consistent detection of weak selective sweeps and, by extrapolation, older sweeps. Although not directly assessed here, we suspect that the shift in power we observe would not apply to the detection of incomplete sweeps since the such sweeps are most reliably detected with statistical tests based on haplotype structure (Sabeti et al. 2002), a feature of the data not likely to be greatly affected by the SNP recovery patterns observed here.

Inferring demography

We simulated population samples under a simplistic model of population size expansion in which the population doubled in effective size $1N_{CURR}$ generations ago, where N_{CURR} equals the effective population size at the time of sampling. To determine the effect of short-read sequencing on the accuracy of demographic inference, we used a Poisson log-likelihood approach to infer growth parameters from each simulation dataset before and after sequencing. Comparison of inferred values before and after simulated sequencing showed that SNP data inferred from 15x sequencing recovers the demographic signal extremely well (Figure 4). At 8x sequencing depth, a large proportion of simulations returned post-sequencing parameter values that differed slightly from the pre-sequencing values. Sequencing depths of only 4x introduce a systematic downward bias in the inferred timing of expansion, resulting in conclusion of more recent growth (by an approximate difference of $0.17*N_{CURR}$ generations, under our simulation framework).

Inference of genetic differentiation

We simulated structured population samples with three levels of current genetic differentiation between the two sub-populations. The three groups of simulated subdivided populations have mean F_{ST} values of 0.54, 0.37, 0.15, and 0.01 as estimated from the simulated samples. Analysis of these same samples after they had been processed through the simulated NGS pipeline revealed a systematic downward bias in F_{ST} values (Figure 5). Although, the bias was most severe at 4x read depth with a mean reduction of 0.0147, higher read depths also suffer the downward bias (mean diff at 8x = 0.0069, and 0.0018 at 15x), albeit significantly less so (4x vs. 8x $t_{df=299} = 39.34$, $p < 2.2\times10^{-16}$; 8x vs. 15x $t_{df=299} = 46.41$, $p < 2.2\times10^{-16}$). Interestingly, the bias increases with the degree of differentiation (Figure 5). While this suggests that the bias is minimized when differentiation is low, it is in systems with low differentiation that such a bias would have the greatest effect on biological interpretation. Therefore, the significant improvements in precision achieved with greater read depths would prove particularly valuable when differentiation is being estimated between closely related sub-populations.

DISCUSSION

NGS technologies hold promise for expanding the field of population genomics into a diverse array of biological and ecological systems (Pool et al. 2010). However, careful consideration of experimental structure and statistical analysis is essential to avoid compounding data-related uncertainty and biases in downstream analyses. Multiple statistical approaches have been proposed to accommodate the limitations of NGS, many specifically designed to handle low (<5x) read depth data (e.g. Lynch 2009; Martin et al. 2010; DePristo et al. 2011; Kim et al. 2011). A prefered standard approach that performs well in population genetic analyses under a variety of experimental

structures has not surfaced, perhaps due to the limited number of applications to empirical data. While these approaches result in improved ability to call SNPs and estimate their population frequencies, particularly for human data or data simulated to resemble human data, the effect of sampling error in low-coverage sequencing data on the capacity to address population genetic questions has not been previously addressed in a broad sense. It is important to note that our study is not meant to be an evaluation of any particular SNP calling approach, but instead is intended to provide an evaluation of how using NGS data processed through a typical SNP calling pipeline can affect the power to address population genomic questions, recognizing that both the sequencing technologies and related statistical approaches are likely to change and improve over time. Moreover, we chose to address population genetic models that are most vulnerable to NGS related errors, but other models such as population bottlenecks and incomplete sweeps are also of interest and should be explored.

We evaluated the effect of sequence coverage on the ability to detect three common population genetic scenarios: directional selection, population growth, and partial subdivision. The primary source of variation among read depths is the rate of true SNP recovery, or the false negative rate. Consistent with previous observations (e.g. Jiang et al. 2009; Lynch 2009), rare variants are disproportionately missed (Figure 2) due to sparse read sampling, and the rate of SNP recovery increased substantially with read depth (Table 1). Contrary to previous reports that showed an excess of rare variants when the SFS is inferred from short-read data (e.g. Kim et al. 2011), we inferred a deficit of rare variants after computational elimination of putative sequencing errors (Figure 2), underlining how the ability to distinguish between errors and true SNPs in a system with high genetic variation differs from the ability in systems where variation is rare. Another possible explanation for this discrepency is the fact that in our study the

sequencing quality scores were recalibrated in a way that may have lead to an unrealistically accurate estimate of error rates, perhaps resulting in false negative and false positive SNP calling rates unachievable in empirical studies. Comparison of recalibration performance in our study to that achieved using human data from the 1000 genomes (DePristo et al. 2011) suggests that our approach resulted in comparable improvements in base quality distributions with this empirical example. Interestingly, we also observed significantly lower rates of SNP recovery from simulations under the growth and sweep models compared to those under the equilibrium model (Figure 1), a difference we attribute to the rare-skewed SFS under the growth and sweep models (Supplementary Figure 2). While this result may be specific to the models and simulation framework used here, the broader implications is that the dependence of the SNP recovery rate on the SFS itself could lead to heterogeneous error in SFS inference, even among regions of the same genome. Genomic regions of low recombination exhibiting low diversity may experience further complications in SNP recovery since SNP detection is also sensitive to the diversity-to-error ratio (Lynch 2009). However, in humans and possibly other systems with sufficient external data, improvements in rare-variant recovery can be made through imputation from haplotype information or statistical tuning modeled on independent deep sequencing data from the same diploid individuals have been employed (Durbin et al. 2010; Gravel et al. 2011).

How do the differences in SNP recovery across depths of sequencing affect the ability to address population genetic questions? We compared the power to detect selective sweeps, infer demographic shifts, and estimate genetic differentiation among populations from “true” complete sequence information to the power when sequences inferred from NGS data at 4x, 8x and 15x read depths are used. Interpretation of our findings follows below, but it is important to recall throughout that our results are specific

in their detail to our particular simulated experimental structure. General conclusions can be drawn from our results, but numerical details, such as exact power curves, will depend on experimental parameters such as sample size and levels of genetic variation.

Detecting Positive Selection

The rapid fixation of a newly arising beneficial mutation leaves a distinct pattern diversity in flanking chromosomal segments, including an excess of rare variants and high frequency derived alleles. Multiple statistical tests have been developed to detect such selective sweeps (reviewed in Nielsen 2005). We used the composite likelihood ratio test (Nielsen et al. 2005) to detect sweeps among population samples simulated under a selective sweep model (Kim & Stephan 2002). We found that, overall, the strength of selection had a larger effect on the power to detect selective sweeps than that of the sequencing process and changes in sequencing coverage (Supplemental Figure 3). Since the effects of NGS on genotyping accuracy are physically diffuse but the genomic footprint of positive selection is genomically local (reviewed in Nielsen 2005), it stands to reason that the selection footprint can be inferred with relative accuracy provided that the data is not riddled with false positive SNPs that will both obscure the selection footprint and alter the neutral background model based on data genomic patterns of variation. However, when we addressed the reduction in power related to read depth with greater resolution, we found an interaction between the strength of selection and read depth (Figure 3), suggesting that strong selective sweeps, leaving large and dramatic selection footprints, can be detected with very low read depths, but weaker selective events will only be detected with greater genotyping and allele frequency accuracy. It should be noted, however, that weak selective events are difficult to detect even with complete true data (Supplemental Figure 3; Nielsen et al.

2005). Incomplete sweeps (e.g., Sabeti et al. 2002; Juneja & Lazzaro 2010) and sweeps from standing genetic variation (Przeworski et al. 2005) are also likely to be difficult to detect with low read depth sequencing.

Inferring Demography

Many systems show genomic patterns of genetic variation that are inconsistent with expectations under canonical equilibrium models, making inference of demography a standard component of genomic analyses, both for its own sake and to inform accompanying tests of other hypotheses (Boyko et al. 2008; e.g. Crawford & Lazzaro 2010; Gravel et al. 2011; Locke et al. 2011). Demographic inference is typically accomplished by testing the fit of one or several summaries of polymorphism data that include information about both the number of SNPs and their frequency in the sample (e.g. Crawford & Lazzaro 2010; Gravel et al. 2011; Locke et al. 2011). Thus, accurate inference of polymorphism from NGS is essential for avoiding biases in demographic inference. We simulated a population growth model and quantified the difference in parameter estimates between models inferred from complete sequence data and models inferred from simulated short-read sequence data. Population growth has been shown to result in a negative skew in the SFS owing to an enrichment of external branches in geneological structures of populations that have experienced growth (Tajima 1989; Slatkin & Hudson 1991; Rogers & Harpending 1992), suggesting that the lower recovery rate for rare SNPs in low pass sequencing will obscure the signal of growth. We found that the high genotyping accuracy at 15x read depth results in near perfect recovery of the demographic signal (Figure 4). However, at lower depths, we found a systematic bias towards inference that growth was more recent that it truly was, without any bias in the inferred magnitude of growth. These results suggest that accurately inferring

demographic parameters will hinge on full recovery across the SFS, most likely via sequencing depths of at least 8x. This need may be somewhat mitigated in systems that allow alternative approaches for recovery of rare variants such as haplotype imputation (Durbin et al. 2010) or statistical tuning based on reduced-representation deep sequencing data (Gravel et al. 2011). It should be noted that we have tested only one, argruably simplistic, population growth model here. Further study will be required to extend these results to more complex models.

Inferring Genetic Differentiation

When a panmictic ancestral population is divided into two predominantly reproductively isolated populations, allele frequencies of shared polymorphisms diverge over time via neutral genetic drift at a rate that depends on the amount of gene flow between the populations and the effective population size of the nascent populations (reviewed in Holsinger & Weir 2009). The signature of this process can be summarized using, among other statistics, F_{ST}, which directly compares the partitioning of genetic variance among populations (Weir & Cockerham 1984; Holsinger & Weir 2009). Rare variants contribute less to estimates of F_{ST} than do intermediate frequency variants (Weir & Cockerham 1984), suggesting that the missing rare-variant issue inherent to low pass sequencing may not have a large impact on estimates of genetic differentiation. We compared the accuracy of F_{ST} estimates of genetic differentiation between two partially isolated populations inferred from NGS data of various depths and found a systematic underestimation of F_{ST}, even at 15x read depth (Figure 5). Inspection of Figure 2 suggests that underestimation of allele frequency is more common than overestimation. A systematic reduction in perceived diversity as well as a tendency to underestimate allele frequencies both result in reduced estimates of differentiation. Interestingly, the

bias we inferred here did not vary substantially across a range of F_{ST} values (Figure 5), although we explored only highly differented samples and this bias may differ at lower levels of differentiation. When population differentiation is substantial, even short read data as shallow as 4x is sufficient to detect substantial differences in allele frequencies, suggesting significant progress can be made towards measuring genetic differentiation with minimal investment in sequencing.

In summary, we assessed the power to address population genetic questions using NGS, providing quantification of both the power and accuracy of population inference under experimental conditions typical of many ecological systems with large population sizes. We found that the prospect of identifying strong selective sweeps is good even at low sequencing depths, while inferring weak selection, non-equilibrium population demographics and population structure may suffer significant biases without higher coverage. While our results improve our understanding of the dependencies between read depth, SNP calling and allele frequency estimates, and population genetic inference using NGS, further investigation is warranted to explore how biases and power-loss changes across a broader set of population genetic models and experimental parameterizations.

ACKNOWLEDGEMENTS

We are grateful to Matteo Fumagalli and Zhen Wang for helpful discussions and comments on earlier versions of the manuscript. We are also thankful to two anonymous reviewers for their helpful comments. This work was supported by NIH grant AI062995. JEC is supported by a Cornell Center for Comparative and Population Genomics fellowship.

Table 1: Effect of read depth and population genetic model on false negative and false positive SNP rates.

Model	Proportion False Negative SNPs[a]			Proportion False Positive SNPs[b]		
	Total	Low Freq[c]	High Freq[d]	Total	Low Freq[c]	High Freq[d]
4x Read Depth[e]						
Equilibrium[f]	0.1268	0.2761	0.0026	0.0444	0.1187	0
Sweep[g]	0.1295	0.2822	0.0024	0.0608	0.1672	0
Growth	0.1463	0.2653	0.0021	0.0516	0.1154	0
Structure[h]	0.1340	0.2583	0.0011	0.0084	0.0201	0
8x Read Depth						
Equilibrium	0.0400	0.0875	0.0005	0.0099	0.0235	0
Sweep	0.0417	0.0910	0.0006	0.0134	0.0318	0
Growth	0.0464	0.0843	0.0003	0.0119	0.0234	0
Structure	0.0449	0.0862	0.0007	0.0031	0.0065	0
15x Read Depth						
Equilibrium	0.0062	0.0133	0.0003	0.0041	0.0091	0
Sweep	0.0064	0.0137	0.0004	0.0053	0.0118	0
Growth	0.0067	0.0121	0.0001	0.0048	0.0088	0
Structure	0.0095	0.0177	0.0006	0.0013	0.0026	0

a – Proportion of all true SNPs that were not called after sequencing.
b – Proportion of all called SNPs that were not present in true data.
c – SNPs with true frequency less than or equal to 0.1 in sample.
d – SNPs with true frequency greater than 0.1 in sample.
e – Simulated read depth for each diploid individual.
f – Equilibrium refers to Wright-Fisher equilibrium model.
g – Only rates for sweep model with $\tau = 0.005$ and $\alpha = 1000$ are presented.
h – Only rates for structure model with $F_{ST} = 0.37$ are presented.

FIGURES

Figure 1: Proportion of true SNPs recovered with 4x sequencing. Data from 100 simulations is presented for each population model (only the selection model with τ = 0.005 and α = 1000 and structure model with $F_{ST} \approx$ 0.37 is presented in each case, see Methods). Models were compared with paired t-test with significance threshold of 5%. Note that the y-axis scale is limited from 0.84 to 0.92. The population structure model was not included in this particular statistical contrast (see Methods).

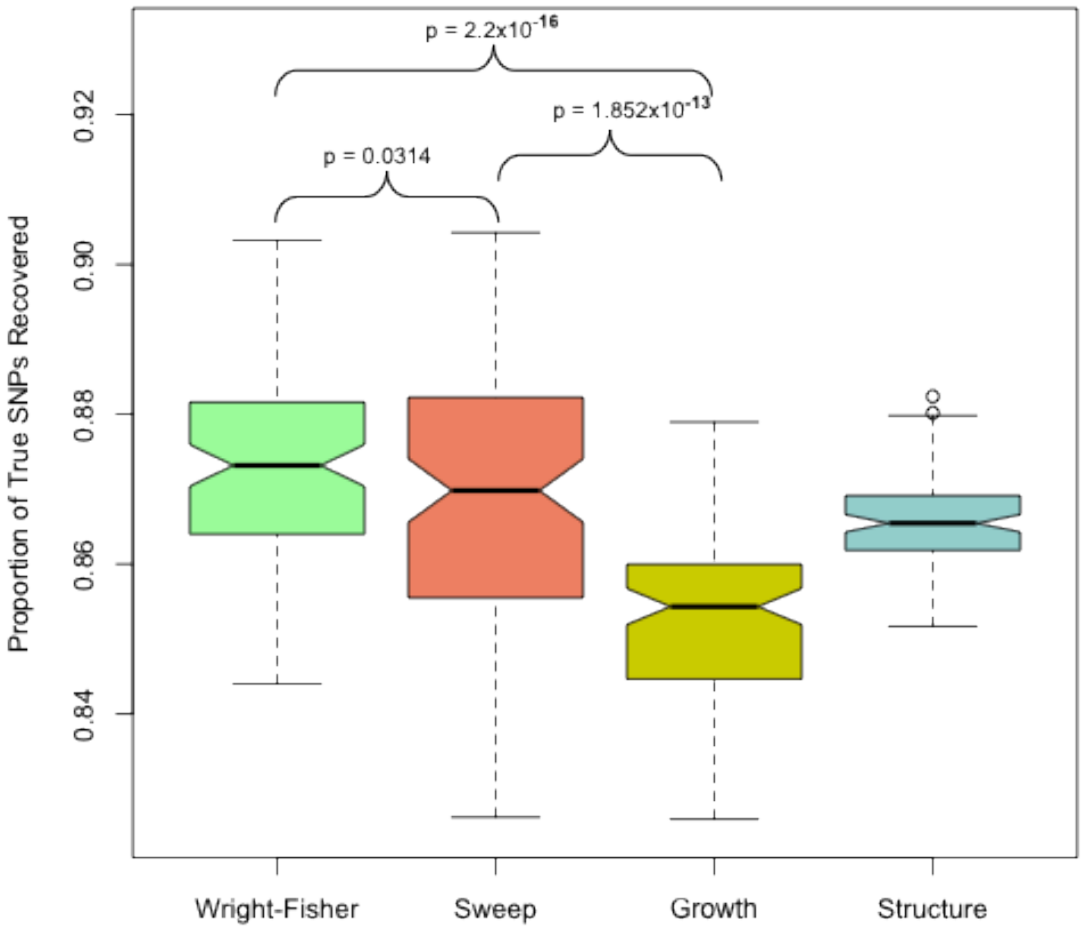


Figure 2: Comparison of allele frequencies before and after sequencing at 4x. Frequencies for SNPs, including false positives and false negatives, from population samples simulated under a) Wright-Fisher model, b) Population Growth model, c) Selective Sweep model ($\alpha = 1000$, $\tau = 0.005$), and d) Population Structure ($F_{ST} \approx 0.37$). The frequencies of false positive SNPs are found in the left most column of each plot. The frequencies of true SNPs that were missed after sequencing (false negatives) are plotted in the bottom row of each plot. For the population structure model, frequency was calculated as average frequency across both subpopulations. Colors indicate the total number of SNPs (on a log scale) in each bin from 100 simulations for each model.

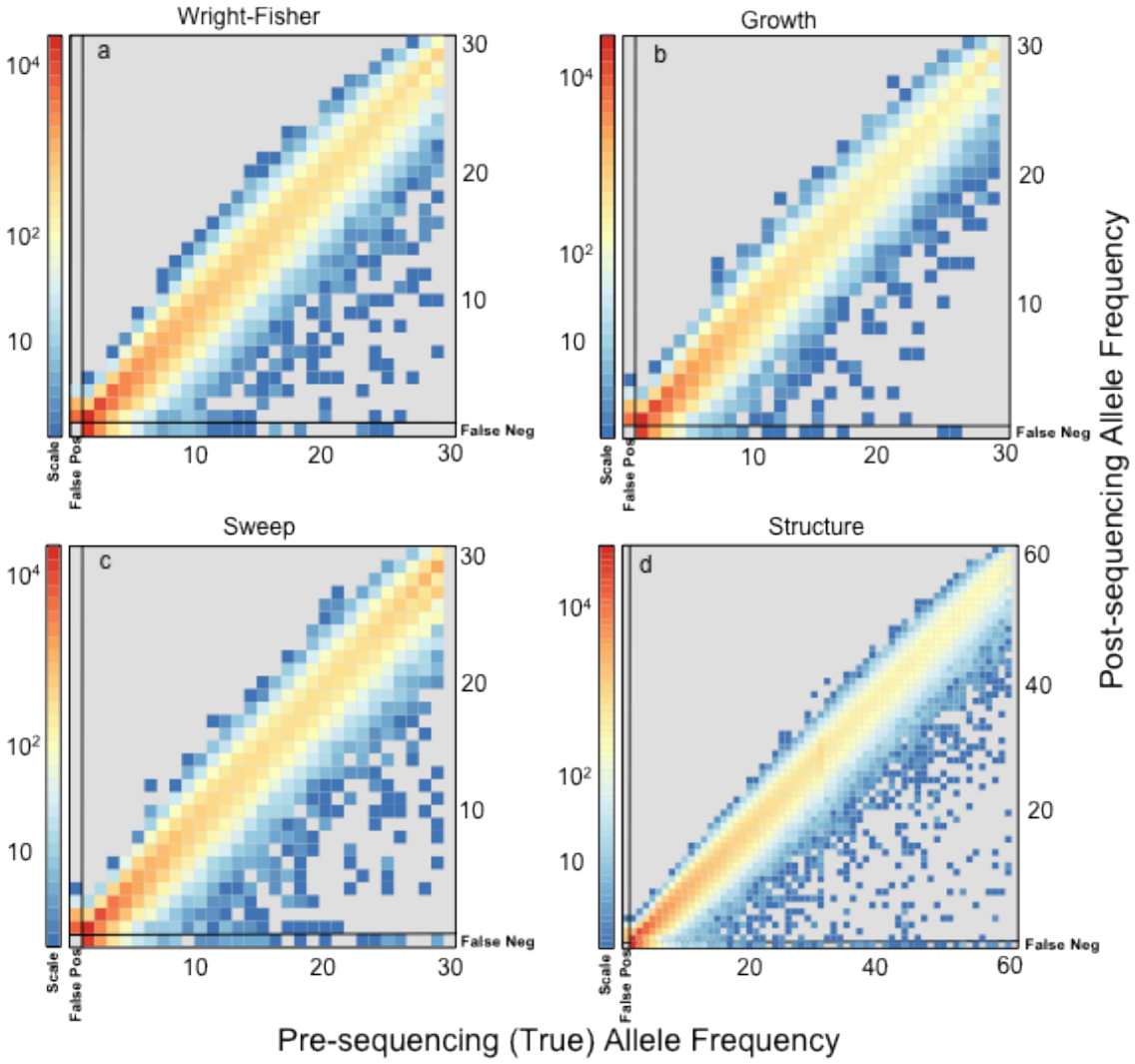

Wright-Fisher
a
Growth
b
Sweep
c
Structure
d
Scale
False Pos
False Neg
Post-sequencing Allele Frequency
Pre-sequencing (True) Allele Frequency

Figure 3: The effect of next-generation sequencing on the power to infer selective sweeps. The proportion of simulated samples that rejected the null hypothesis with complete, pre-sequencing data and also rejected after sequencing is plotted as a function of the strength of selection. 30kb regions of 30 chromosomes with one selective sweep (n = 100 for each α – τ) were simulated. Simulations were compared to a null distribution generated by neutral simulations and considered significant at 0.05 cutoff.

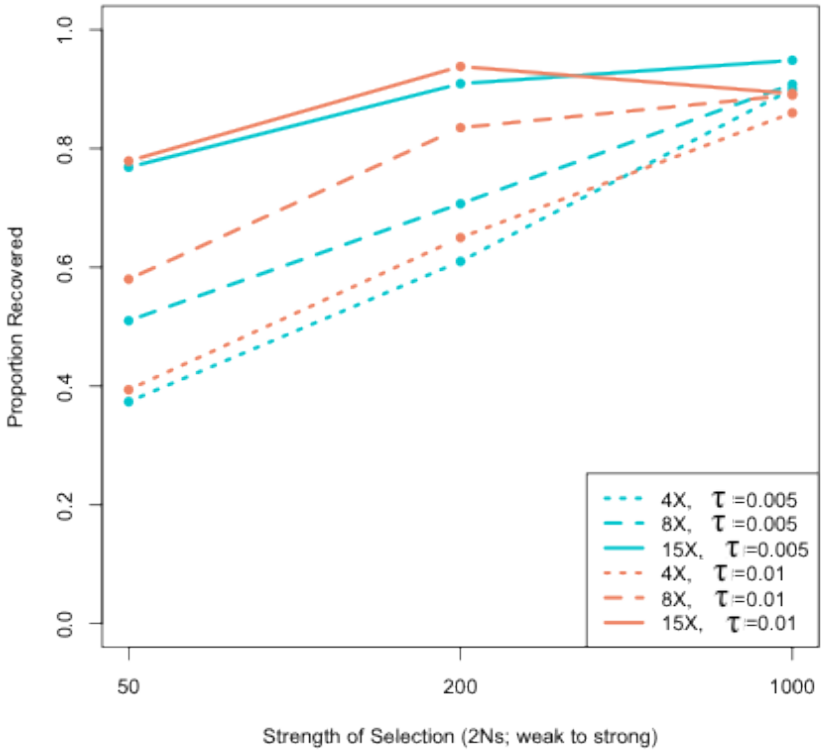


Figure 4: The effect of 4x, 8x, and 15x sequencing on demographic inference. Population samples of 30 chromosomes were simulated under a population expansion model (n = 500), and the timing and magnitude of growth were inferred using a likelihood approach both from full sequence data and after simulated a) 4x, b) 8x, and c) 15x next-generation sequencing. The difference in the timing of growth, or bias, was calculated by subtracting the parameter value inferred post-sequencing from the pre-sequencing value. The same calculation was used for the magnitude of growth. Colors indicate the proportion of simulations in each region of the parameter space.

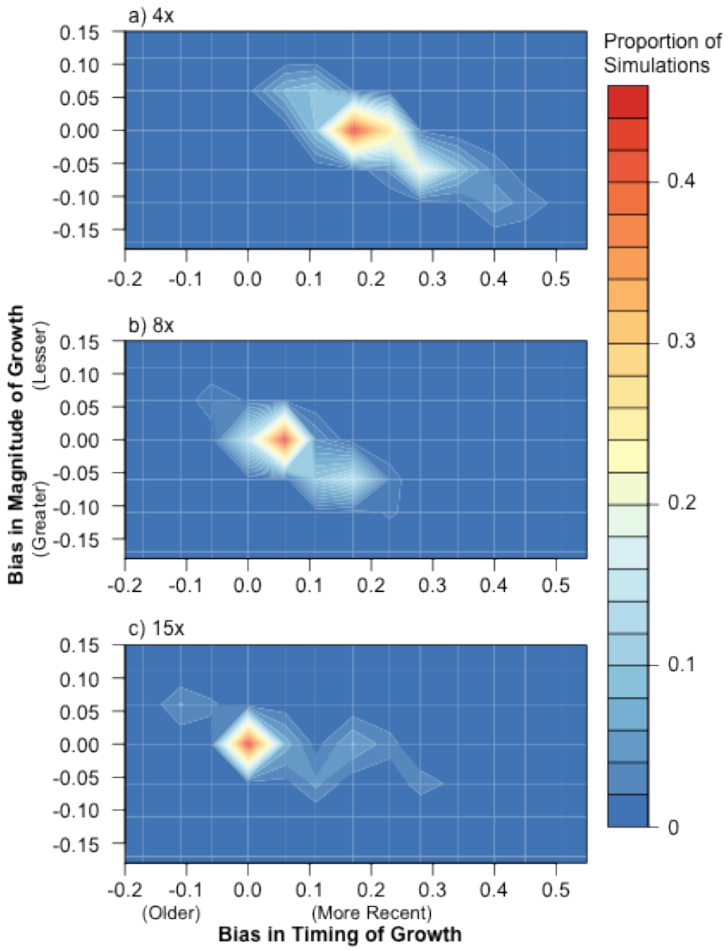
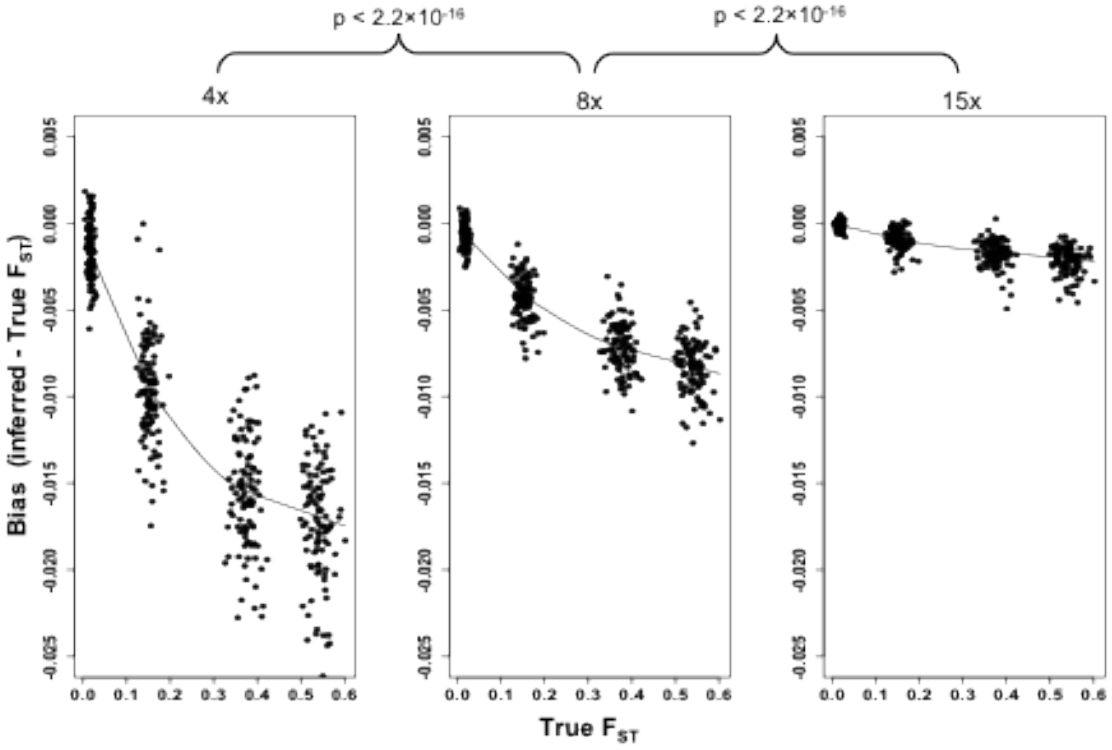


Figure 5: The effect of 4x, 8x, and 15x sequencing on inference of genetic differentiation. F_{ST} was calculated with full sequence data ('True F_{ST}') and after

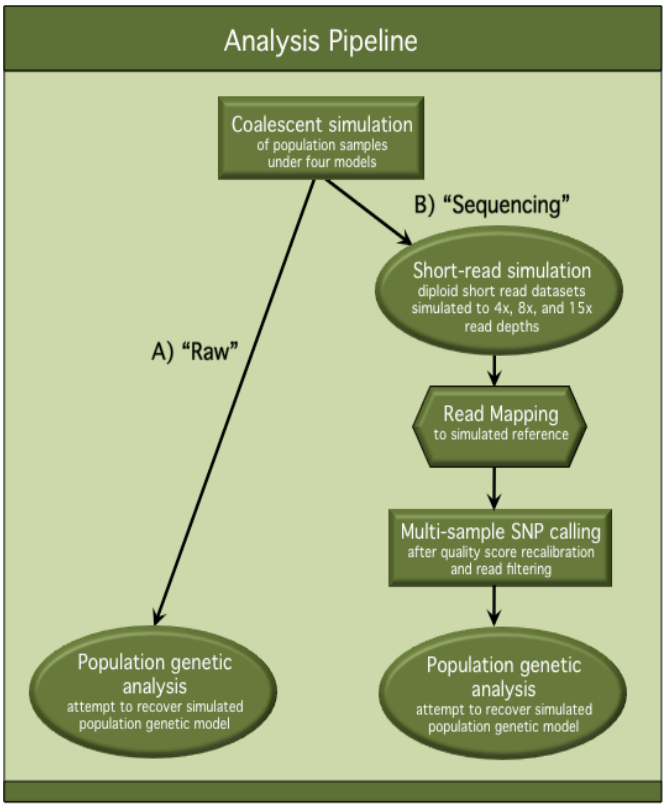
simulated next-generation sequencing using Weir and Cockerham's unbiased estimator (Weir & Cockerham 1984), and the post-sequencing bias is plotted. A loess curve was fit to the data for each sequencing depth to illustrate both the effect of increasing F_{ST} as well as read depth. The difference between pre- and post-sequencing F_{ST} was calculated for all simulations (n = 100 for each value of F_{ST}) and these differences were compared between sequencing depths using a paired t-test.



SUPPLEMENTARY FIGURES

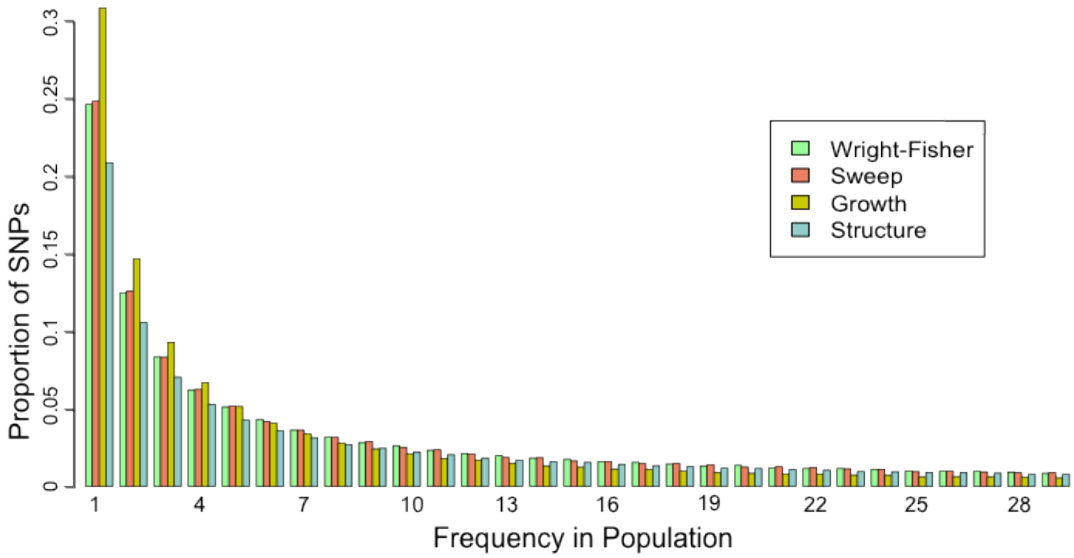
Supplemental Figure 1: Flowchart of analysis pipeline.

This flowchart describes the analysis pipeline used to assess the effects of NGS on population genetic inference and hypothesis testing. Population samples were simulated under four population genetic model and processed through both A) the "Raw" track of the pipeline and B) the "Sequencing" track of the pipeline. In the "Raw" track, unmodified, simulated polymorphism datasets were used for population genetic analysis. In the "Sequencing" track, simulated polymorphism datasets were processed through an *in silico* sequencing pipeline, and polymorphisms inferred from the 'sequence' data were used for population genetic analysis. Comparisons were made between population genetic analysis results from the "Raw" track and the "Sequencing" track to quantify the differences in accuracy and power of inference after "Sequencing".

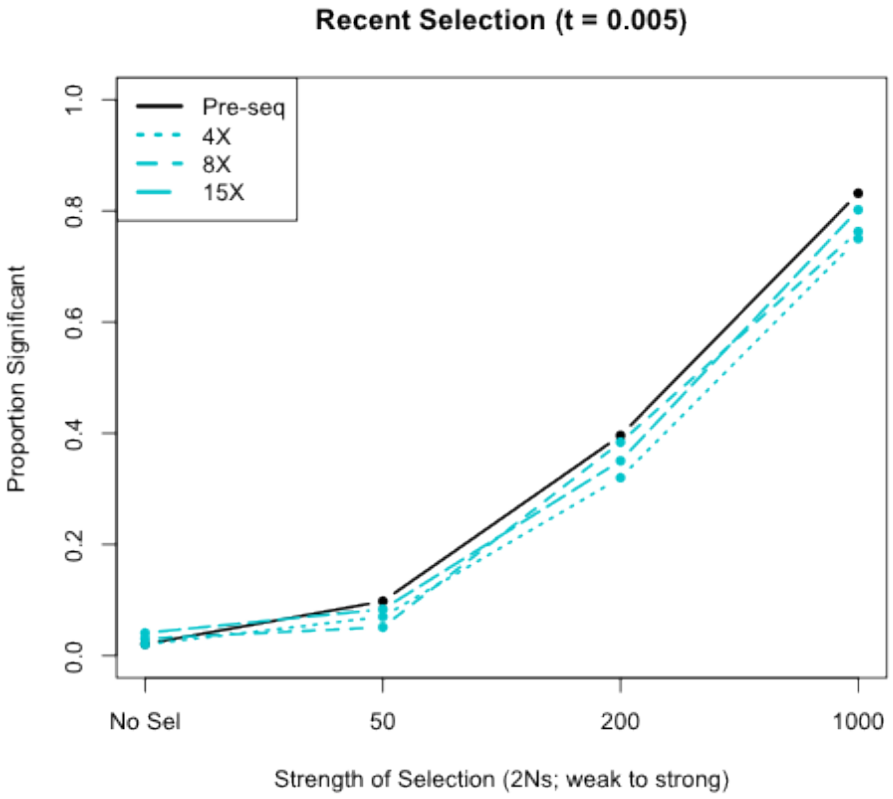


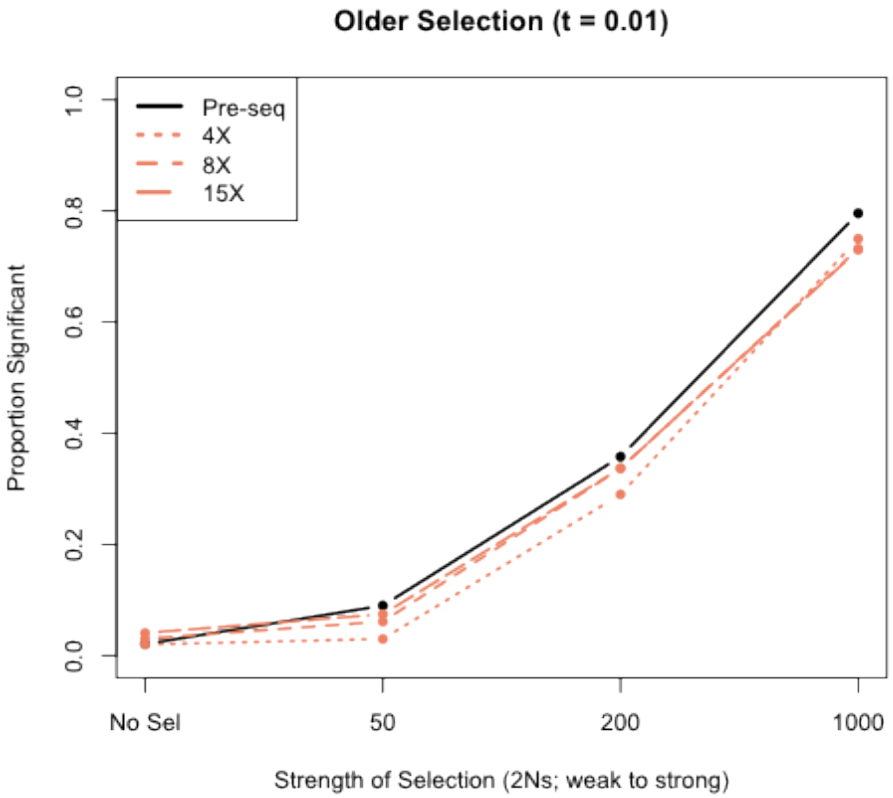
Supplemental Figure 2: Site-frequency spectrum from 'complete' data for each population genetic model.

The proportion of all true SNPs at various frequencies in the population is presented. For the structure model, frequency was calculated across both sub-populations (60 chromosomes) and proportions calculated according to that distribution, but only SNPs with frequencies less than 30 are presented here. Data from only one selective sweep model (alpha = 1000, tau = 0.005) and one structure model (F_{ST} = 0.38) is presented here.



Supplemental Figure 3: The power to detect selective sweeps before and after simulated next-generation sequencing. The proportion of statistically significant detections of positive selection was calculated from simulations of a 30kb region of 30 chromosomes with one selective sweep. Simulations were compared to a null distribution generated by neutral simulations and considered significant at 0.05 cutoff. The top panel shows the results for recent selective sweep models ($\tau = 0.005$) and the bottom panel shows results for models with relatively older selective sweeps ($\tau = 0.01$). "Pre-seq" refers to the complete sequence information, without simulated next-generation sequencing and SNP calling.


Recent Selection (t = 0.005)
Pre-seq
4X
8X
15X
Proportion Significant
0.0
0.2
0.4
0.6
0.8
1.0
No Sel
50
200
1000
Strength of Selection (2Ns; weak to strong)


Older Selection (t = 0.01)
Pre-seq
4X
8X
15X
Proportion Significant
0.0
0.2
0.4
0.6
0.8
1.0
No Sel
50
200
1000
Strength of Selection (2Ns; weak to strong)

REFERENCES

Ahmad, R., Parfitt, D.E., Fass, J., Ogundiwin, E., Dhingra, A., Gradziel, T.M., et al. (2011). Whole genome sequencing of peach (Prunus persica L.) for SNP identification and selection. *BMC Genomics*, 12, 569.

Bansal, V., Harismendy, O., Tewhey, R., Murray, S.S., Schork, N.J., Topol, E.J., et al. (2010). Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res.*, 20, 537–545.

Boyko, A.R., Williamson, S.H., Indap, A.R., Degenhardt, J.D., Hernandez, R.D., Lohmueller, K.E., et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.*, 4, e1000083.

Branca, A., Paape, T.D., Zhou, P., Briskine, R., Farmer, A.D., Mudge, J., et al. (2011). Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume Medicago truncatula. *Proceedings of the National Academy of Sciences*, 108, E864 –E870.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10, 195–205.

Crawford, J.E. & Lazzaro, B.P. (2010). The demographic histories of the M and S molecular forms of Anopheles gambiae s.s. *Mol. Biol. Evol.*, 27, 1739–1744.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43, 491–498.

Durbin, R., Abecasis, G., Altshuler, D., Auton, A., Brooks, L. & Gibbs, R. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467, 1061–1073.

Earl, D.A., Bradnam, K., St John, J., Darling, A., Lin, D., Faas, J., et al. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*.

Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., et al. (2011). Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.*, 108, 11983–11988.

Hellmann, I., Mang, Y., Gu, Z., Li, P., de la Vega, F.M., Clark, A.G., et al. (2008). Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res.*, 18, 1020–1029.

Henn, B.M., Gravel, S., Moreno-Estrada, A., Acevedo-Acevedo, S. & Bustamante, C.D. (2010). Fine-scale population structure and the era of next-generation sequencing. *Hum. Mol. Genet.*, 19, R221–226.

Hoberman, R., Dias, J., Ge, B., Harmsen, E., Mayhew, M., Verlaan, D.J., et al. (2009). A probabilistic approach for SNP discovery in high-throughput human resequencing data. *Genome Res.*, 19, 1542–1552.

Holsinger, K.E. & Weir, B.S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nat Rev Genet*, 10, 639–650.

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

Jiang, R., Tavaré, S. & Marjoram, P. (2009). Population genetic inference from resequencing data. *Genetics*, 181, 187–197.

Johnson, P.L.F. & Slatkin, M. (2006). Inference of population genetic parameters in metagenomics: a clean look at messy data. *Genome Res.*, 16, 1320–1327.

Johnson, P.L.F. & Slatkin, M. (2008). Accounting for bias from sequencing error in population genetic estimates. *Mol. Biol. Evol.*, 25, 199–206.
Juneja, P. & Lazzaro, B.P. (2010). Haplotype structure and expression divergence at the Drosophila cellular immune gene eater. *Mol. Biol. Evol.*, 27, 2284–2299.
Keightley, P.D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S. & Blaxter, M.L. (2009). Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Research*, 19, 1195–1201.
Kim, S.Y., Li, Y., Guo, Y., Li, R., Holmkvist, J., Hansen, T., et al. (2010). Design of association studies with pooled or un-pooled next-generation sequencing data. *Genet. Epidemiol.*, 34, 479–491.
Kim, S.Y., Lohmueller, K.E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., et al. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, 12, 231.
Kim, Y. & Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160, 765.
Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754–1760.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.
Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., et al. (2009b). SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, 19, 1124–1132.
Locke, D.P., Hillier, L.W., Warren, W.C., Worley, K.C., Nazareth, L.V., Muzny, D.M., et al. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, 469, 529–533.
Lynch, M. (2008). Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol. Biol. Evol.*, 25, 2409–2419.
Lynch, M. (2009). Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*, 182, 295–301.
Magwene, P.M., Kayıkçı, Ö., Granek, J.A., Reininga, J.M., Scholl, Z. & Murray, D. (2011). Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 108, 1987 –1992.
Martin, E.R., Kinnamon, D.D., Schmidt, M.A., Powell, E.H., Zuchner, S. & Morris, R.W. (2010). SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics*, 26, 2803–2810.
Michel, A.P., Grushko, O., Guelbeogo, W.M., Sagnon, N., Costantini, C. & Besansky, N.J. (2006). Effective population size of Anopheles funestus chromosomal forms in Burkina Faso. *Malaria Journal*, 5, 115.
Nielsen, R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet.*, 39, 197–218.
Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G. & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research*, 15, 1566.
Pool, J.E., Hellmann, I., Jensen, J.D. & Nielsen, R. (2010). Population genetic inference from genomic sequence variation. *Genome Res.*, 20, 291–300.
Przeworski, M., Coop, G. & Wall, J.D. (2005). The signature of positive selection on standing genetic variation. *Evolution*, 59, 2312–2323.

R Development Core Team. (2011). R: A language and Environment for Statistical Computing.

Rogers, A.R. & Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, 9, 552.

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

Slatkin, M. & Hudson, R.R. (1991). Pairwise Comparisons of Mitochondrial DNA Sequences in Stable and Exponentially Growing Populations. *Genetics*, 129, 555–562.

Tajima, F. (1989). The effect of change in population size on DNA polymorphism. *Genetics*, 123, 597.

Weir, B. & Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 1358–1370.

Williams, L.M., Ma, X., Boyko, A.R., Bustamante, C.D. & Oleksiak, M.F. (2010). SNP identification, verification, and utility for population genetics in a non-model genus. *BMC Genetics*, 11, 32.

Xia, Q., Guo, Y., Zhang, Z., Li, D., Xuan, Z., Li, Z., et al. (2009). Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). *Science*, 326, 433–436.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329, 75–78.

CHAPTER 6:

De novo transcriptome sequencing in *Anopheles funestus* using Illumina RNA-seq technology.

Published in PLoS One (2010) 5(12): e14202. doi:10.1371/journal.pone.0014202.

Authors: Jacob E. Crawford[1], Wamdaogo M. Guelbeogo[2], Antoine Sanou[2], Alphonse Traoré[2], Kenneth D. Vernick[3,4], N'Fale Sagnon[2], Brian P. Lazzaro[1]

Corresponding Author: Jacob E. Crawford, jc598@cornell.edu

Affiliations:
[1] Department of Entomology, Cornell University, Ithaca, New York, United States of America
[2] Centre National de Recherche et de Formation sur le Paludisme, Ouagadougou, Burkina Faso
[3] Department of Parasitology and Mycology, Centre National de la Recherche Scientifique Unit URA3012: Hosts, Vectors and Infectious Agents, Institut Pasteur, Paris, France
[4] Department of Microbiology, University of Minnesota, Saint Paul, Minnesota, United States of America

ABSTRACT

Background: *Anopheles funestus* is one of the primary vectors of human malaria, which causes a million deaths each year in sub-Saharan Africa. Few scientific resources are available to facilitate studies of this mosquito species and relatively little is known about its basic biology and evolution, making development and implementation of novel disease control efforts more difficult. The *An. funestus* genome has not been sequenced, so in order to facilitate genome-scale experimental biology, we have sequenced the adult female transcriptome of *An. funestus* from a newly founded colony in Burkina Faso, West Africa, using the Illumina GAIIx next generation sequencing platform.

Methodology/Principal Findings: We assembled short Illumina reads *de novo* using a novel approach involving iterative *de novo* assemblies and 'target-based' contig clustering. We then selected a conservative set of 15,527 contigs through comparisons to four Dipteran transcriptomes as well as multiple functional and conserved protein domain databases. Comparison to the *Anopheles gambiae* immune system identified 339 contigs as putative immune genes, thus identifying a large portion of the immune system that can form the basis for subsequent studies of this important malaria vector. We identified 5,434 1:1 orthologues between *An. funestus* and *An. gambiae* and found that among these 1:1 orthologues, the protein sequence of those with putative immune function were significantly more diverged than the transcriptome as a whole. Short read alignments to the contig set revealed almost 367,000 genetic polymorphisms segregating in the *An. funestus* colony and demonstrated the utility of the assembled transcriptome for use in RNA-seq based measurements of gene expression.

Conclusions/Significance: We developed a pipeline that makes *de novo* transcriptome sequencing possible in virtually any organism at a very reasonable cost

($6,300 in sequencing costs in our case). We anticipate that our approach could be used to develop genomic resources in a diversity of systems for which full genome sequence is currently unavailable. Our *An. funestus* contig set and analytical results provide a valuable resource for future studies in this non-model, but epidemiologically critical, vector insect.

INTRODUCTION

Anopheles funestus is a primary vector of human malaria parasites, which cause almost a million deaths of children under the age of 5 annually in sub-Saharan Africa [1]. The genome of *An. funestus* has not been sequenced, although it is expected to be within the next couple of years [2]. The current absence of a sequenced genome prevents many valuable experimental approaches from being applied to *An. funestus*, including determination of gene expression patterns after exposure to malaria parasites, comparison of genome content to other *Anopheles* and insect species, and reverse genetic manipulation to determine gene function. Despite the absence of a fully sequenced and assembled genome, however, many of these experiments could be pursued after sequencing of the transcriptome, the complete set of expressed genes.

Short read sequencing technologies such as the Solexa/Illumina (Illumina), 454 (Roche) and SOLiD (ABI) platforms have made it increasingly possible to perform *de novo* transcriptome sequencing [3][4]. For example, a single experiment on the instrument used in the present study (Illumina Genome Analyzer IIx, Illumina) can sequence 225-250 million nucleic acid molecules, generating 45-50 Gigabases of 100 base pair (bp) paired-end sequence in roughly 9.5 days, where “paired-end” refers to sequences obtained from the respective opposite ends of a single DNA molecule. As we will show, this volume of sequencing provides ample read coverage for *de novo*

transcriptome assembly as well as for gene expression analyses and polymorphism discovery. The challenge with *de novo* transcriptome sequencing using data from short read technology lies in the difficulty of assembling the reads into contigs reflecting transcriptional units [3]. Short read sequence assembly is an active area of research, and has produced an array of assembly options (e.g. Velvet [5]; ALLPATHS2 [6]; ABySS [7]; Oases, Schulz and Zerbino, unpublished). The Roche 454 platform produces longer reads (~450 bases) than the Illumina platform (<120 bp), helping to overcome the difficulties of *de novo* assembly, but Illumina produces orders of magnitude more sequence at a fraction of the cost, making it an attractive option for researchers with limited budgets. Despite this, *de novo* short read assembly of eukaryotic transcriptome sequence has been largely confined to 454-based sequencing efforts e.g. [8-10], with only a very few examples of *de novo* transcriptome sequencing using the Illumina platform occurring in the literature (e.g. *Pachycladon* [11]; Chinese Hamster Ovary Cell [12]; Whitefly [13]). Illumina-based transcriptome sequencing has been hampered in part by the absence of simple and effective assembly workflows capable of handling Illumina RNA-seq, or mRNA derived, datasets.

Understanding the basic biology of mosquito disease vectors such as *An. funestus* is essential for disease control efforts and development of new control technologies to be effective [14]. Valuable insights have been gained through studies of *An. gambiae* [15][16] and the sequencing of its genome [17], but *An. gambiae* is just one of several potent vectors of human malaria in Africa, and many open questions remain, including those regarding the genetic similarities and differences between the three most important and congeneric vectors: *An. gambiae, An. arabiensis* and *An. funestus*. *An. funestus* is estimated to have shared a common ancestor with the closely related sibling species pair *An. gambiae* and *An. arabiensis* 30 – 80 million years ago [18], and

previous studies found the degree of genetic differentiation between *An. funestus* and *An. gambiae* to be high (substitutions per synonymous site (K_s) = 0.612 ± 0.392 [19]), prohibiting simple use of the *An. gambiae* genome for gene discovery in *An. funestus* (such as through specific PCR in *An. funestus* with primers designed to the *An. gambiae* genome). Furthermore, *An. funestus* exhibits many epidemiologically important ecological differences from *An. gambiae*, including its ability to thrive in arid conditions unsuitable to many other vectors [20][21]. Disease control efforts will have to be tailored specifically to *An. funestus* in order to be fully effective. To date, there have been no efforts to sequence the complete transcriptome of *An. funestus*, although approximately 2,800 Expressed Sequence Tags (ESTs) have been obtained from traditional sequencing efforts aimed at genetic mapping [19], salivary gland protein discovery [22] and general transcript discovery [23].

We used the Illumina Genome Analyzer IIx platform (Illumina) coupled with a novel assembly approach to sequence the transcriptome of *An. funestus*. Historically, the generation of scientific data, genetic and otherwise, from *An. funestus* has been limited by the difficulty in rearing *An. funestus* in colony. We have recently established a new colony from specimens caught in Burkina Faso [Materials and Methods], bringing the number of *An. funestus* colonies worldwide from two [24] to three. We sequenced mRNA deriving from this colony using the Illumina sequencing platform and assembled the adult transcriptome of this species *de novo* using a hybrid assembly approach. Through bioinformatic analyses we identified ~15,500 largely novel, high confidence transcription units. Short read alignments revealed almost 367,000 single nucleotide polymorphisms (SNPs) and insertion/deletion polymorphisms (indels), as well as substantial variation in expression levels among contigs. We confirmed homology of a large majority of our *An. funestus* contigs to several Dipteran transcriptomes, and

identified 5,434 transcripts that could be paired to an *An. gambiae* gene as 1:1 orthologues. Using bioinformatics, we putatively assigned contigs to broad functional categories and found that protein divergence was not evenly distributed among functional categories. Contigs that do not contain ambiguous bases or previously published *An. funestus* EST sequence have been deposited into NCBI and can be downloaded through the NCBI Sequence Read Archive website. The final catalog of our inferred transcriptional units is publicly available at www.jacobecrawford.com and at www.lazzaro.entomology.cornell.edu. We expect that this *An. funestus* transcriptome will provide a valuable genomic resource for future studies, including facilitating experimental genetic experiments and providing empirical support for gene models of the *An. funestus* genome when it is eventually sequenced.

MATERIALS AND METHODS

An. funestus colony:

We collected fed and gravid female *An. funestus* mosquitoes from the village of Koubri (12°11'54"N; 1°23'43"W) 35 kilometers South East of Ouagadougou, Burkina Faso, in February of 2007. Approximately 50 females were used to establish the colony, and the colony is maintained at a large size with overlapping generations in the insectary of Centre National de Recherche et de Formation sur le Paludisme in Ouagadougou, Burkina Faso. The females used to establish the colony were monomorphic with respect to chromosomal rearrangements and thus of the Kiribina chromosomal form as defined by [25]. We sampled females from the colony for mRNA extraction in November of 2008, corresponding to the 17^{th}-19^{th} generation since the colony was established.

RNA extraction and sequencing:

3-5 day old female mosquitoes (n = 30) were removed from the colony, knocked down at -20°C, washed in ice-cold 95% ethanol to overcome the hydrophobic properties of mosquito cuticles and rinsed in ice-cold water, and then submerged in RNAlater (Qiagen, USA) and frozen at -80°C. Carcasses frozen in RNAlater were transported from Burkina Faso to the US where they were stored at -80°C. Total RNA was extracted using standard protocols (Trizol; Invitrogen, USA) from all 30 carcasses after grinding them under liquid nitrogen. mRNA selection, library preparation and sequencing was performed by the Cornell University Life Sciences Core Facilities on an Illumina GAIIx sequencer according to manufacturer specifications. Briefly, mRNA was selected using oligo(dT) probes and then fragmented using divalent cations. cDNA was synthesized using random primers, modified and enriched for attachment to the Illumina flowcell. We sequenced one 60-cycle paired-end lane and two 87-cycle paired-end lanes, generating ~102.6 million reads for a total of 8,150 MB of sequence. All three un-filtered paired-end lanes of sequence have been deposited as a series with the accession number GSE21977 at NCBI's GEO database or at the NCBI Short Read Archive under submission number SRA020147.

***De novo* transcriptome assembly:**

Prior to assembly and mapping (described below), we applied filters to remove low quality reads and reads containing suspected poly-Adenine tails from all three paired-end lanes. First, we implemented a 'quality filter' by removing reads where more than 33% of bases were 'N' and reads where more than 34% of the nucleotides had Phred quality scores less than 20, where a Phred score of 20 corresponds to a 1% expected error rate. Next, we removed sequences suspected of containing poly-Adenine tails by discarding any read composed of greater than 33% Adenine.

Sequence assembly was carried out in three steps: 1) iterative *de novo* assembly with Velvet v7.58 [5], 2) 'target-based' clustering using *An. funestus* ESTs to find and unite, where possible, sequences belonging to the same transcription unit but not joined in Velvet, and 3) 'target-based' clustering using *An. gambiae* predicted peptides (Figure 1).

Step 1: Iterative Velvet assembly. For step 1, we used the *de novo* assembler Velvet to assemble all three lanes of paired-end Illumina reads. However, we implemented Velvet in a novel way in order to improve the assembly. First, we conducted 'exploratory' assemblies of the paired reads using multiple hash lengths (k = 21, 25, 31, 35, 41, 49, and 59). We then conducted an additional assembly (k = 57) of all unused reads (un-paired) from 4 of the exploratory assemblies (k = 21, 35, 49, 59). Next, we assembled all contigs obtained from all exploratory assemblies and the unused reads assembly in a series of 'summary' assemblies. First, we assembled all contigs at 3 different kmer values (k = 29, 39, 49) and then assembled the contigs obtained from these assemblies in a final summary assembly (k = 39). This "assembly of assemblies" approach may allow inclusion of some misassemblies, but these should result in low-confidence contigs that will be removed in subsequent steps in the workflow. Contigs from the final summary assembly were included in subsequent clustering steps.

Step 2: EST-based clustering. The highly coverage-sensitive nature of contig selection and node connection within the de Bruijn graph utilized by Velvet often results in partially fragmented assemblies. Therefore, we included the following clustering step to ensure all homologous sequence was joined where possible. For step 2, we downloaded all *An. funestus* ESTs from Genbank (n = 2,846 as of November 2009; referred to as "ESTs" below) [19][22][23]. From this larger set of ESTs, We found 1,496 unique ESTs and used this condensed set as targets in a 'target-based' clustering

process in order to join homologous contigs that were not joined in the Velvet assembly. We first used BLASTN from the stand-alone bundle of BLAST algorithms v2.2.23+ [26] to identify all contigs that showed significant similarity (e-value ≤ 1×10^{-6}) with each *An. funestus* EST downloaded from Genbank. Each matching contig was then individually aligned to its EST match using ClustalW [27], and contig-EST matches were discarded if their ClustalW alignment score was not greater than 50 plus 3 times the length of the shorter of the two sequences. All remaining contigs were then grouped by their matching EST and compared to the match with the highest BLAST score by dividing all BLAST scores by the maximum score in the group. Contigs with normalized BLAST scores less than 0.7 were discarded from further clustering steps. If more than one contig remained in an EST-group after the two previous filtering steps, contigs within each EST-group were aligned in a global alignment using ClustalW. To identify good matches between the EST and individual contigs, individual pairwise alignment scores for each EST-contig alignment within the global alignment were divided by the maximum EST-contig alignment score in that group, and all contigs with a normalized pairwise alignment score less than 0.7 were eliminated from further clustering. The cutoff value of 0.7 used in the previous filtering steps was chosen after visual inspection of a subset of alignments suggested that this criteria readily distinguished credible matches from those more likely to be spurious. Contigs that survived all of these filtering steps were then aligned to their EST match, and any sequence that extended further than the edge of the EST was joined to the EST and the total sequence was used in the final contig set. After this clustering process, the resulting contig set contained some contigs that were comprised partly of contig sequence and partly of EST sequence, some contigs that were comprised of two contigs joined in the middle by EST derived sequence, some contigs that resulted from joining two Velvet contigs and many sequences that were

unaffected by the clustering process. Clustering and joining of contigs was accomplished with custom scripts written in the statistical computing environment R [28].

Step 3: *An. gambiae*-based clustering. For the final assembly step, step 3, we used the *An. gambiae* predicted peptide set (release 3.5) downloaded from Vectorbase.org in a second and analogous 'target-based' clustering step. The clustering step based on *An. funestus* ESTs (step 2 above) was helpful, but was not likely to be exhaustive due to the limited number of *An. funestus* ESTs available in public databases, so we performed the following additional clustering step using the much more complete, albeit evolutionarily diverged, *An. gambiae* transcriptome. All contigs that were not joined in the EST-based clustering in step 2 were evaluated in analogous fashion against *An. gambiae* peptides. These contigs were compared to the entire *An. gambiae* predicted peptide set using BLASTX and submitted to the same filtering step as in the EST-clustering step above, where contigs with a normalized BLAST score less than 0.7 were disregarded. Surviving contigs were then grouped based on their peptide match. If more than two contigs matched a peptide, they were globally aligned using ClustalW. Pairwise ClustalW alignment scores within the global alignment of greater than 80 were considered positive matches and these contigs were joined. If only two contigs matched a peptide, they were aligned and if they overlapped with an alignment score greater than 90, they were joined into a single contig. If the contigs did not overlap in alignment, they were joined together by a string of 'N's using the peptide BLAST high scoring pair coordinates of each contig as a guide for the length of the N string. Like the EST-based clustering step, this clustering process was performed using custom R scripts.

At the end of step 3 of the assembly pipeline, the total contig set was comprised of many contigs that were not affected by the clustering process, some contigs that were

the product of one or two contigs having been concatenated with pre-existing EST sequence, some contigs that were joined during the *An. gambiae* peptide clustering step, and some contigs that had been scaffolded around a run of 'N's.

Bioinformatics and contig validation:

To distinguish between valid transcript sequence and spuriously assembled sequence we compared the post-clustering set of contigs to multiple Dipteran insect transcriptomes, searched for open reading frames and compared translated protein sequences to functional protein domain databases as a means to identify contigs with bioinformatic associations with other species. First, we searched our assembled and clustered *An. funestus* contigs for homology to the translated predicted transcriptomes of other Dipteran insects with sequenced genomes. In addition to the *An. gambiae* peptide set used for clustering above, we downloaded the predicted peptide set from *Aedes aegypti* (release 1.2) and *Culex quinquefasciatus* (release 1.2) as well as the full genome sequence of *An. gambiae* (release 3) from Vectorbase.org. We also downloaded the predicted peptide set of *Drosophila melanogaster* (release 5.26) from Flybase.org. For reference, the genus *Anopheles* (Subfamily Anophelinae) is predicted to have shared a common ancestor with *Aedes* and *Culex* (Subfamily Culicinae), between 145 – 200 million years ago [18] and a common ancestor with *Drosophila* 260 million years ago [29]. We compared our final contig set to each of these four translated transcriptomes using BLASTX, as well as to the *An. gambiae* genome using TBLASTX, and high scoring matches with a minimum e-value of 1×10^{-6} were kept for further analysis. As part of functional annotation (described below), we also compared our contig set to the nr database at NCBI as the first step of Gene Ontology [30] (hereafter

referred to as GO) annotation implemented by Blast2GO [31] using an expect value cutoff of 1×10^{-6}.

In addition, we evaluated the post-clustering contig set based the size of the inferred open reading frame (ORF) relative to contig size. We extracted open reading frames from all contigs using the '-getorf' function in the EMBOSS package [32]. To accommodate the uncertainty of whether our contig captured the full ORF, we extracted both translated regions that were flanked by a Methionine and a STOP codon ('-find 1'; hereafter type A) as well as translated regions that were simply free of STOP codons ('-find 0'; hereafter type B). For each contig, we compared the largest ORF from each of these types and kept the ORF that contained a start codon unless the type B ORF extended upstream of the type A ORF to the beginning of the contig representing cases in which the true start codon is likely truncated from the contig. If no type A ORF was found, the type B ORF was chosen. Then, in order to cleanse the contig set of contigs comprised of spuriously assembled sequence, we discarded any contig if its ORF was shorter than 50 amino acids.

To identify putative conserved protein domains and assign putative functional information to the post-clustering contig set, we compared translated protein sequences extracted from our contigs to multiple functional domain databases using RPS-BLAST and Blast2GO [31]. First, the total peptide set was compared to the SMART [33], KOG [34], Pfam [35] and CDD [36] databases using RPS-BLAST with no expect value threshold cutoff, but only matches with an expect value less than 1×10^{-6} were considered in further analyses. We also mapped our contigs to the GO database using Blast2GO. Annotation through Blast2GO is accomplished by first searching for matches to the nr database at NCBI, then mapping to the BLAST results to the GO database and finally selecting a GO annotation using their Annotation Rule that is based on the degree of

similarity to the GO, GO Evidence Code weights (default values used here) and relative weights given to child versus parent terms [31]. In order to simplify the functional annotations to a set of broad terms, we also mapped the GO annotations to the Generic GO-Slim terms using Blast2GO. All results from BLAST comparisons to functional and conserved protein domain databases as well as the GO annotations are presented in Table S1.

We chose a final contig set by comparing results of all of the BLAST and functional domain database comparisons and keeping only sequences that showed a significant association to at least one proteome or database. This resulted in a conservative set of contigs, although it prevents the discovery of novel genes in the *An. funestus* transcriptome. This is an unfortunate consequence of the inherent difficulty in distinguishing novelties from spuriously assembled sequence. Our contig set as reported is composed entirely of high confidence transcription units. All contigs that did not contain any 'N's inserted during contig clustering (n = 14,980) are available in the Transcriptome Assembly Archive at NCBI under the accession numbers EZ966136 - EZ980985. The full final contig set is available at www.jacobecrawford.com and www.lazzaro.entomology.cornell.edu.

After compiling a conservative set of contigs using the bioinformatic and ORF filtering criteria, we performed a reciprocal best-hit analysis to identify 1:1 orthologues between our *An. funestus* contigs and *An. gambiae* predicted proteins. First, we searched the *An. gambiae* peptide set with BLASTX using *An. funestus* contigs as queries with an e-value threshold of 1×10^{-6}. We then performed the reciprocal search with TBLASTN using *An. gambiae* peptides as the queries and the same e-value threshold. *An. gambiae* peptides shorter than 50 amino acids (n = 26) were omitted from this search because the BLAST algorithm is unable to parse such short sequences.

One-directional 'best-hits' were declared for each query if only a single BLAST result was obtained or the ratio of the BLAST score of the 'second-best-hit' to the BLAST score of the first 'best-hit' was less than 0.7. One-directional 'best-hits' were identified in both directions and 5,434 reciprocal 'best-hits' were obtained by comparing these datasets.

Read mapping, SND calling and expression profiling:

We used the short read alignment algorithm BWA [37] to align all three paired-end lanes of Illumina sequence reads to the final contig set established above. Prior to the assembly steps, all sequence reads were screened for low quality and low complexity as described above. To accommodate the global mapping procedure used in BWA and reduce the number of reads not mapped because of sequencing errors in the terminal end of the read, positions 76-87 of all remaining reads in the two lanes of 87 bp paired-end reads were trimmed using the FastX toolkit [http://hannonlab.cshl.edu/fastx_toolkit/], leaving 75 bp reads for mapping. The trimmed 75bp and 60bp paired-end reads were then aligned to the reference final contig set in BWA, with the maximum number of difference between each read and reference sequence set to 5 ('aln -n 5'). Alignment files from the three paired-end lanes were merged, sorted and parsed by contig identification using pileup in the SAMtools package [38]. The consensus base, putative single nucleotide polymorphisms and short indels were called using the pileup '-c' option. We called single nucleotide polymorphisms and indels (hereafter collectively referred to as short nucleotide discrepancies or SNDs) at sites where 1) the mapping quality was greater than or equal to 20, 2) the alternative base occurred at least twice or the equivalent of 0.025 times the coverage at the site when coverage was greater than 80, 3) only one alternative base occurred at or above this frequency and 4) at least 6 reads covered the site. Thus, in order to be considered a

putative SNP, the alternative nucleotide would have to be observed with high confidence at least twice even at a positions covered by 6 reads. In this way, we aimed to decrease false positive SND calls from sequencing errors, which are generally expected to be unique in the read set, but which should accumulate in abundance linearly with sequence depth.

We were interested in determining colony-level genetic variation, recognizing that because the polymorphisms reported here were obtained from a colony of mosquitoes and not a random population sample, true population genetic parameters describing the natural population can not be appropriately estimated from this data. Estimates of genetic variation from high-throughput sequencing data are complicated by the fact that read depth varies among and within contigs and that highly expressed genes are more likely to be completely sequenced and thus represented by more bases. While raw SND counts are presented for the purpose of SND discovery, we applied several corrections and assigned each contig an adjusted nucleotide diversity (hereafter simply referred to as nucleotide diversity) value. First, we treated any bases that were not covered by at least 6 reads as missing data, so we calculated an initial estimate of nucleotide diversity by dividing all SND counts by the number of bases across the contig that were covered by at least 6 reads to obtain an estimate of SNDs per base. Next, to control for ascertainment bias related to variable read depth, or in this case expression level and mapping success, , we adjusted length-corrected SND counts using the read-depth correction (eq. 7) proposed by Jiang et al. [39] that accounts for the possibility of missing data at low coverage sites and the probability of observing a mutant allele in a given sample. This correction is intended for regions of a genome with identical read depth [39], but since this requirement is not applicable to our case, we used the median read coverage per contig.

We were also interested in testing the utility of the assembled transcriptome for measuring gene expression. To estimate mapping success, we quantified the total number of reads mapped and further distinguished between uniquely mapped reads and repetitively mapped reads. We also quantified gene expression in our dataset extracting the number of reads mapped to each contig during the BWA alignment. However, Gene expression levels can be estimated from RNA-seq data with great accuracy e.g. [40], but, since read mapping is sensitive to the size of the target reference sequence, corrections must be applied to adjust for contig length. Therefore, we adjusted the raw read count by the total number of reads mapped and the length of the contig, calculating Reads Per Kilobase per Million mapped reads (RPKM; [40]).

Protein divergence:

To identify functional categories of proteins that show high levels of divergence or conservation, we determined protein divergence between 1:1 *An. funestus*:*An. gambiae* orthologues. First, we aligned orthologous protein sequences using ClustalW. We then calculated protein distance using the 'identity' mode of the dist.alignment function in the R package seqinr [41]. This function calculates protein distance as the square root of the proportion of the sequence that is different between the two sequences. However, automated sequence alignment is unreliable at high divergence levels, so we excluded orthologous pairs with less than 30% identity (leaving 4,975 contigs) to avoid false mismatches introduced by low confidence alignments. Lastly, we assigned each orthologous pair to one of three categories based on its level of divergence (or proportion amino acids that differ): High (≥ 0.138), Intermediate (< 0.138 and > 0.058) and Low (≤ 0.058) divergence, with bin cutoffs empirically determined so that one-third of transcripts fall into each category.

Comparison among functional categories:

We used X^2 analyses to ask whether any functional categories of contigs as assigned above were significantly enriched or depleted of any of the nucleotide diversity categories. As described above, we first assigned contigs to a protein divergence bin (i.e. Low, Intermediate or High), and then to functional categories based on GO-Slim terms. We then, using a X^2 test, asked whether each GO-Slim category was enriched (or depleted) for any of the bins compared to the expectation of equal proportions expected under the binning method. We used a Bonferroni-adjusted α level of 5.26×10^{-4} to assign significance in tests. As the power of X^2 analyses increases with increasing number of observations, we limited our intra-functional category comparisons to categories populated by at least 15 contigs.

Distribution of Data and Scripts for Analysis:

An Excel spreadsheet, modeled after AnoXcel, containing peptide sequences, BLAST results, functional annotation results and other pertinent information for each contig in the final contig set is available online as Table S1 as well as on the websites www.jacobecrawford.com and www.lazzaro.entomology.cornell.edu. All data and results management, manipulations and analyses were carried out using custom scripts written by J. Crawford in the statistical computing environment R unless specified otherwise. Additionally, a Velvet wrapper script written in python by J. Crawford called AssemblyAssembler.py to automate the iterative Velvet assembly used here is available at the websites given above and is also packaged with Velvet, starting with version 0.7.63.

RESULTS AND DISCUSSION

Sequencing and Assembly:

Due to the low cost and ability to obtain both novel sequence for assembly as well as gene expression data, there is great interest in utilizing Illumina RNA-seq data for *de novo* transcriptome assembly and analysis [3]. We sequenced three paired-end lanes of mRNA extracted from 30 whole, sugar-fed female *An. funestus* using the Illumina Genome Analyzer. Approximately 102 million reads (or 51 million paired-end reads) passed Illumina quality filtering totaling roughly 8.1 GB of sequence. We removed 2% of these reads flagged as either low-quality or low-complexity. A first pass Velvet assembly with default parameters and a hash length of 31 yielded over 440,000 contigs and an N50 of 209 bp (i.e. 50% of the total assembled sequence was contained in contigs of this length or longer). We searched a large range of hash values (k = 21 to k = 59) and obtained a slight improvement by setting the hash value to 57, producing approximately 357,000 contigs with an N50 of 228 bp. Further exploration of various parameter settings and data combinations suggested that an iterative assembly in which contigs output from generic Velvet assemblies using various hash lengths are assembled in a final series of Velvet runs produced the best assembly (Figures 1 and 2). This final contig set of 46,987 contigs with an N50 of 1,140 bp comprised 27.8 MB of sequence and was submitted to further downstream filtering and analysis as detailed in Materials and Methods and briefly described below.

While the present manuscript was in review, several independent efforts to optimize transcriptome assembly using RNA-seq data were made publicly available. We were encouraged by the results of one independent study that obtained high quality transcriptome assemblies of Illumina reads using an iterative, varied kmer approach similar, in principle, to ours [42]. Two other efforts that employ an alternative approach

have also been made available (Oases [Schulz and Zerbino, unpublished] and Cufflinks [43]). To determine how the performance our method compares to an alternative method, we assembled the *An. funestus* transcriptome using Oases with standard parameter settings and obtained a high quality assembly. This approach generated approximately twice as many contigs as our pre-clustering contig set, but the contig sets were very similar with respect to the proportion of sequences showing homology to *An. gambiae* and the N50 value. However, one key difference is that Oases relies heavily node scaffolding using paired-end information (74.8% of contigs contain 'N's in Oases contig set generated here), which is not ideal because these ambiguous bases produce 'edge-effects' in short-read mapping analyses.

In principle, the iterative assembly routine employed here is intended overcome the heterogenous coverage distribution inherent to non-normalized RNA-seq data by taking advantage of the fact that some contigs will be assembled best in certain assembly conditions while others are best assemble in different conditions. We and a colleague found anecdotal evidence using independent datasets that high coverage contigs assemble best in high kmer value assemblies, while low coverage contigs assemble best in low kmer value assemblies. Further exploration is needed to determine whether this can be exploited more directly. Importantly, we subsampled our data and found that this assembly routine produced a very respectable assembly (maximum contig length = 12,688 bp and N50 = 784 bp) with only single paired-end lane of Illumina sequence reads, suggesting that significant progress can be made with very little sequencing cost.

Following the iterative assembly step, we used 'target-based' clustering to improve the *de novo* assembly. By clustering contigs around previously described *An. funestus* ESTs and then predicted *An. gambiae* peptides, we searched the contig set for

potential overlaps and joined contigs where possible and appropriate. This 'target-based' contig clustering process resulted in only a modest condensation of the contig set from 46,987 contigs to 45,644 contigs, a 2.9% reduction (Figure 2). In their Illumina-based assembly of the transcriptome of Chinese hamster ovary cells, Birzele et al. [12] utilized a similar assembly workflow that was a hybrid of *de novo* assembly and read mapping to the phylogenetically closest sequenced model system genome to improve assembly and annotation and achieved similarly modest assembly improvements [12]. Birzele et al. [12] used transcripts from the closely related mouse genome to cluster short reads for assembly leading to a reduction in their contig set of approximately 6% to 92,272 contigs with a mean length of 352 bp. Our own experience and the report of Birzele et al [12] suggests that, in general, clustering may not be an extremely effective means of improving *de novo* transcriptome assemblies. In contrast, our pre-clustering iterative assembly process generated 46,987 contigs with a mean length of 591 bp, underscoring the potential gains to be made through alternative *de novo* assembly approaches even in the absence of any clustering.

To purge our contig set of spuriously assembled sequence, we utilized bioinformatic support to validate our contigs. We eliminated any contig that did not show a significant BLAST match to at least one of four insect transcriptomes or functional databases and did not harbor a convincing ORF (Materials and Methods), leaving 15,527 contigs with an N50 of 1,753 bp (Figure 2). For comparison, the predicted transcript sets of the most thoroughly annotated Dipteran genomes, *An. gambiae* and *D. melanogaster*, are comprised of 14,753 and 21,921 transcripts, respectively, with N50s of 2,258 and 2,475 bp, respectively. This suggests that, not surprisingly, our contigs are frequently incomplete and represent only a subset of the potentially expressed transcriptome. We restricted our final contig set to a limited number of conservative

contigs, reducing the total number of contigs relative to other *de novo* transcriptome studies e.g. [9][12]. Even so, our high confidence contig set of 15,527 transcripts represents a marked expansion of the *An. funestus* genetic sequence space over the previously available ~2,800 ESTs (521 of which we were able to extend through our assembly and clustering process).

Homology with Dipteran sequences

In order to determine homology with available Dipteran sequences, we compared our contig set to predicted peptide sets extracted from four sequenced Dipteran genomes (*An. gambiae, Ae. aegypti, C. quinquefasciatus* and *D. melanogaster*) using the standalone BLASTX algorithm (e-value ≤ 1×10^{-6}) as well as to the *An. gambiae* genome with the TBLASTX (e-value ≤ 1×10^{-6}). We compared 15,527 *An. funestus* sequences to the closely related *An. gambiae* peptide set, finding 13,137 (84.6%) with significant similarity to an *An. gambiae* sequence, although this percentage may be slightly upwardly biased due to the usage of *An. gambiae* peptides during the clustering process in assembly step 3 (Materials and Methods). And while all contigs showed homology with at least one Dipteran transcriptome consistent with the selection process described above, a core set of 9,929 (63.9%) contigs showed significant matches in all four Dipteran transcriptomes (Figure 3). This high degree of sequence homology is consistent with previous observations of transcriptome conservation among these species [23]. Consistent with the expectation of increased divergence with increased phylogenetic distance, however, the number of contigs showing significant sequence similarity in pairwise comparisons between the *An. funestus* contig set and Dipteran transcriptomes decreased with increasing phylogenetic distance (Figure 3). It should be noted that the annotation process that produced the predicted peptide sets queried here

were not independent since more recent annotations often train their gene model annotation pipeline on gene models from previously annotated genomes e.g. [17]. Highlighting the potentially limiting effect of this dependence, we found 2,360 contigs that showed no matches to the *An. gambiae* predicted peptide set but significant homology to the full *An. gambiae* genome sequence as well as other Dipteran sequences and sequences in functional domain databases. Although this discrepancy could be explained in part by differences between the BLAST algorithms employed in the two comparisons (TBLASTX for the genome versus BLASTX for the peptide set), it implies the presence of unannotated genes or transcribed units in the *An. gambiae* genome. A recent transcriptome profiling study also found many clusters of reads that mapped to unannotated regions of the *Ae. aegypti* genome [44] suggesting empirical validation of transcribed units using next-generation sequencing should be used to complement *in silico* gene prediction pipelines.

To best make direct comparisons between species, we searched for 1:1 orthologous pairs between our *An. funestus* contigs and *An. gambiae* peptides, and putatively assigned 5,434 pairs using the standard reciprocal best-hit criteria. In a comparison between protein sequences of *An. gambiae*, *Ae. aegypti* and *D. melanogaster*, Waterhouse et al. [45] identified 4,951 1:1:1 orthologues and an additional 886 1:1 *Anopheles*:*Aedes* orthologues, suggesting that our contig set harbors about 93.1% of the conserved, single-copy Dipteran orthologues. The median sequence similarity between 1:1 orthologues identified here is 86.9% (σ_{sim} = 0.127), but 57.7% of our *An. funestus* contigs were shorter than their *An. gambiae* orthologue (mean proportion of *An. gambiae* transcript covered = 81.3%, σ_{cov} = 0.237), again suggesting that most of our transcripts are not full-length.

Immune-system genes

When challenged by pathogens such as malaria parasites, mosquitoes mount a strong and effective innate immune response; so immune genes are of particular interest as potential points of exploitation for disruption of disease transmission [16]. To identify putative immune genes within our contig set, we downloaded a list of 414 *An. gambiae* genes annotated as immune genes in the ImmunoDB database [45]. We found significant sequence homology between 345 *An. funestus* contigs and 217 annotated *An. gambiae* immune genes. We also identified contigs with significant homology to 4 *An. gambiae* genes that have been functionally shown to be important in anti-malarial defense but that are not annotated in ImmunoDB: all three of the *APL1* genes (although we are unable to assign strict orthology) [46] and *LRIM1* [47].

Genes in the innate immune system can be split into four broad functional categories: recognition, signaling, regulation and effectors [45][48]. We also included an additional category, 'other', to capture genes involved in other processes such as RNAi or autophagy that have been implicated in immunity. Based on significant BLAST matches to *An. gambiae* genes coding for recognition proteins including those annotated in ImmunoDB (e.g. Thioester-containing Proteins, Gram Negative Binding Proteins; n = 139) as well as the *APL1* paralogues and *LRIM1*, our *An. funestus* transcriptome contains 102 contigs that may function in pathogen recognition. We also recovered 33 contigs (33 in *An. gambiae*) putatively involved in immune signaling (e.g. *Cactus*, *Imd*), 108 putative immune regulatory contigs such as CLIP-domain serine proteases or Serine protease inhibitors (compared to 132 in *An. gambiae*), 33 putative effector genes (compared to 54 in *An. gambiae*; e.g. Cecropins, Lysozymes) and 69 contigs in the 'other' category putatively involved in RNAi (e.g. *Argonaute* and *Dicer*) and autophagy etc. (compared to 45 in *An. gambiae*). All matches between *An. funestus* contigs and

An. gambiae immune genes are listed with their relevant immune annotations in Table S2.

Immune-system genes have been shown to be evolving at a faster rate than other genes in *Drosophila* and mosquitoes [45][48]. We compared protein sequence divergence among 126 1:1 orthologous pairs between *An. funestus* and *An. gambiae* with putative immune function to determine whether this observation holds true for our contig set. We found that orthologous pairs with putative immune function are significantly more diverged than the total set of all 1:1 *An. gambiae*:*An. funestus* orthologous pairs (mean sequence differences for immune gene orthologues = 16.0%, n = 126, mean sequence differences among all orthologues = 10.2%, n = 4,975; $p = 6.95\times10^{-9}$, Mann-Whitney U-test). If we subdivide the analysis based on immune-system function, we find that regulatory proteins are most diverged (mean percent sequence differences = 18.77%), signaling proteins are second-most diverged (mean percent sequence differences = 17.90%), recognition proteins are third-most diverged (mean percent sequence differences = 16.59%) and effector proteins and proteins in the 'other' category are least diverged (mean percent sequence differences = 11.51% for effector, 12.13% for 'other'), although only the regulatory and other categories are significantly different from each other (Figure 4; reg vs. other *p*-value = 0.0032, all other *p*-values range from 0.0566 to 0.8724, pairwise Mann-Whitney U-tests). These results suggest that the immune system genes of *An. funestus* are evolving in a fashion consistent with immune genes in other insects [45][48]. Further studies dissecting the anti-pathogenic role each contig plays will greatly enhance our understanding of the mosquito immune system.

Nucleotide diversity

Single nucleotide polymorphisms and short indels (collectively referred to as single/short nucleotide discrepancies, or SNDs) are very common in natural populations and provide valuable markers for genetic mapping as well as population genetic studies. We identified a set of 366,741 SNDs, suggesting approximately 1.95 SNDs exist in every 100 bp. The mean nucleotide diversity per contig was 0.019 per base before correction for variation in read depth and 0.024, (range of 0 to 0.163 per contig) after correction. This level of variation, particularly high coming from a colony, suggests that the colony may not have suffered the severe loss of genetic variation that could be expected after extended inbreeding. One explanation for this estimate is that the *An. funestus* colony was only recently established and thus too few generations of inbreeding have passed for the effect to be pronounced. Alternatively, our estimate of 0.024 per base may also be upwardly biased, however, by the presence of false-positives in our dataset, resulting from nucleotide mis-incorporation during polymerase chain reaction steps in template library preparation prior to sequencing. A previous Sanger-based re-sequencing study identified 494 SNPs from 20.5 kilobases of sequence (71.4% coding, with 303 SNPs mapping to the coding region), derived from a sample of 21 field and colonized specimens of *An. funestus* [49]. From this survey, they estimated a mean nucleotide diversity level of 0.007 [49], considerably lower than our estimate likely due to their smaller sample size. Estimates of nucleotide diversity in *An. gambiae*, a congeneric species with comparable generation time, geographical distribution and seasonality, are typically similar or perhaps slightly smaller than our estimate for this colony of *An. funestus* (e.g. [50][51]). The colony used in this study does not harbor the chromosomal rearrangements segregating in natural populations, but small and/or unknown inversions may be present and could play a role in preserving genetic variation at some loci. The genetic polymorphisms we have identified as segregating in this colony of *An. funestus*

should be dispersed among all chromosomal arms and suggest that significant natural functional variation can still be found in the colony. Such variation may provide a valuable opportunity for future genetic mapping of phenotypes in the colony.

Functional annotation of the whole transcriptome

To provide a biological foundation on which to begin to globally characterize the transcriptome, we sought to functionally annotate our contig set based on sequence homology to functionally annotated sequences in other species and identification of conserved protein domains. We identified 7,567 contigs that contained regions of significant homology to sequence in at least one database of protein domains (CDD, SMART, and Pfam). Comparisons to the KOG and GO databases provided putative functional information for 10,391 contigs. We were able to assign 3,506 unique GO annotations to 9,026 contigs (36,024 total matches), meaning that 58.1% of our total contigs have affinity to at least one GO term. These GO annotations are quite detailed and provide valuable information for specific contigs, but we were also interested in assigning contigs to broad functional categories that could be used to ask transcriptome-level biological questions. Therefore, we also found 28,781 associations between 119 generic GO-Slim annotations and 9,026 contigs. All annotation information is presented in Table S1, but only the GO-Slim annotations were used in the analyses described below, specifically focusing on 33 functional categories at the Cellular Component level, 39 categories at the Molecular Function level and 49 categories at the Biological Process level.

Transcriptome divergence

An. funestus and *An. gambiae*, estimated to have shared a common ancestor between 30 and 80 MYA [18], exhibit many ecological, behavioral and physiological differences. We examined levels of protein sequence divergence between 1:1 orthologues to determine whether specific functional categories evolve at a rate that is different from the transcriptome as a whole. Of the functional categories tested, 10 Cellular Component categories, 20 Biological Process categories and 12 Molecular Function categories showed significant deviations from expected equal proportions of high, intermediate and low divergence categories at the Bonferroni-adjusted α level. Results for all categories are presented in Table S3, and significant results are presented in Figure 5. In general, significantly deviating categories tended to be enriched with lowly diverged orthologous pairs, indicating a high level of evolutionary conservation within these categories. While no categories were enriched with highly diverged pairs, we found a significant enrichment of intermediately diverged orthologous pairs localizing to the mitochondrion (Figure 5), as might be expected considering the known faster rate of evolution among genes associated with the mitochondrion. We also found significant enrichment of intermediately diverged pairs involved in lipid metabolic processes as well as in contigs with molecular functions involving catalytic activity and binding (Figure 5). A study of protein evolution among single copy orthologues across the phylogeny of the *D. melanogaster* species group identified 12 functional categories putatively under positive selection [52]. The categories identified here as enriched with intermediately diverged orthologous pairs are not among those 12, although immune genes identified outside of the GO-Slim analysis are significantly more diverged than the transcriptome [see above]. We note that our analysis probably has a strong bias toward detecting conserved sequences, since we limited our analysis to high confidence alignments and thus probably excluded highly diverged orthologous pairs. Nonetheless,

our analysis offers the first glimpse into genome level patterns of protein evolution and a step towards a more comprehensive understanding of protein evolution in insect vectors.

Expression profiling:

Transcriptome sequencing in a non-model system makes it possible to conduct experiments to test hypotheses of differential expression between experimental treatments, for example. RNA-seq provides a powerful means of measuring gene-expression because the depth of sequence coverage of a transcript should be proportional to its expression level [4]. To demonstrate that a transcriptome assembled *de novo* can serve as a reference sequence for short-read mapping, we used the short read alignment program BWA to map three paired-end lanes of Illumina sequence to the final contig set. Of 101 million reads, approximately 51% of the reads were mapped uniquely to the transcriptome, while a fraction of a percent of the reads mapped to more than one location in the transcriptome. Interestingly, the remaining 49% of the reads were not successfully mapped, despite our somewhat liberal mapping criteria. It is possible that this rate of mapping success may reflect problems with this assembly, but a recent study mapping short sequence tags to the *Ae. aegypti* genome reported a comparable rate of mapping success [44] indicating that this rate is not likely to be an artifact of the *de novo* assembly process. Furthermore, we found that the profile of gene expression across contigs, as measured by reads per kilobase per million mapped reads (RPKM), adhered to the expected distribution with 95% of contigs having an RPKM value of 133.83 or less and extreme values that differ by three orders of magnitude (from 2.69 to 19,775.75). Therefore, we failed to find good evidence that transcriptomes assembled with our approach should not be used in short read mapping experiments.

Concluding Remarks

Next generation short-read DNA sequencing has made it possible to explore genome-level questions in non-model organisms, regardless of their phylogenetic proximity to model species [3]. *An. funestus* is a primary vector of human malaria, but, as an experimental system, lags significantly in the availability of research data and scientific resources. To establish a genomic resource that will facilitate future genomic level studies in this species, we used the Illumina GAIIx sequencing platform and a novel assembly workflow to build the adult female *An. funestus* transcriptome. In doing so, we demonstrate the feasibility of Illumina-based transcriptome sequencing low cost ($6,300 in sequencing costs) and with the added value of obtaining quantitative expression and polymorphism data. We assembled a conservative and tractable set of 15,527 expressed *An. funestus* contigs, 5,434 of which could be identified as 1:1 orthologues with the more distantly related species *An. gambiae*. We also identified contigs expressed in *An. funestus* that showed homology with unannotated regions of the *An. gambiae* genome, providing empirical evidence that these may be *bona fide* genes with orthologues that are currently unannotated in the *An. gambiae* genome. We identified almost 367,000 genome-wide polymorphisms segregating in our recently established *An. funestus* colony, and showed that, as expected, most of the *An. funestus* transcriptome is evolutionary constrained and is likely evolving under purifying selection. Our results highlight by example just some of the many questions that can be addressed using next-generation sequencing technology to explore the transcriptome of a non-model organism. We also supply essential tools for future genetic study of *An. funestus* and establish a novel *de novo* transcriptome assembly flow that should be applicable to any eukaryote.

ACKNOWLEDGEMENTS

We thank the entomological team at Centre National de Recherche et de Formation sur le Paludisme for their efforts during this study, our colleagues at the Cornell University Life Sciences Core Facilities, Cornell Computational Biology Service Unit and Cornell Center for Advanced Computing for technical assistance with this work, Daniel Zerbino for guidance in developing our approach and Dieter Best for testing the AssemblyAssembler on an independent dataset. We also thank members of the Lazzaro Lab, Rich Meisel and two anonymous reviewers for helpful comments on earlier versions of this manuscript.

REFERENCES

1. WHO/UNICEF World Malaria Report (2009) Geneva: World Health Organization.
2. Anopheles Genomes Cluster Committee (2008) Genome analysis of vectorial capacity in major Anopheles vectors of malaria parasites. VectorBase.org
3. Hudson HE (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. Molecular Ecology Resources 8: 3-17.
4. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.
5. Zerbino DR and Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821–829
6. Maccallum I, Przybylski D, Gnerre S, Burton J, Shlyakhter I, et al. (2009) ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. Genome Biology 10: R103.

7. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, et al. (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117-23.
8. Toth AL, Varala K, Newman TC, Miguez FE, Hutchison SK et al. (2007) Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. Science 318: 441:444.
9. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, et al. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular Ecology 17: 1636-1647.
10. Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). Molecular Ecology 19(Suppl. 1): 115-131.
11. Collins LJ, Biggs PJ, Voelckel C, Joly S (2008) An approach to transcriptome analysis of non-model organisms using short-read sequences. Genome Informatics 21: 3-14.
12. Birzele F, Schaub J, Werner R, Clemens C, Baum P, et al. (2010) Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. Nucleic Acids Res doi:10.1093/nar/gkq116.
13. Wang X, Luan J, Li J, Bao Y, Zhang C, et al. (2010) *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics 11: 400.
14. Enayati A, Hemingway J (2010) Malaria management: past, present, and future. Annu Rev Entomol 55: 569-591.
15. Cohuet A, Harris C, Robert V, Fontenille D (2010) Evolutionary forces on *Anopheles*: what makes a malaria vector? Trends Parasitol 26: 130-136.

16. Yassine H, Osta MA (2010) *Anopheles gambiae* innate immunity. Cellular Microbiology 12: 1-9.
17. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298: 129-149.
18. Krzywinski J, Grushko OG, Besansky NJ (2006) Analysis of the complete mitochondrial DNA from *Anopheles funestus*: an improved dipteran mitochondrial genome annotation and a temporal dimension of mosquito evolution. Mol Phylogenet Evol 39: 417–423.
19. Sharakhov IV, Serazin AC, Grushko OG, Dana A, Lobo N, et al. (2002) Inversions and gene order shuffling in Anopheles gambiae and A. funestus. Science 298: 182–185.
20. Gillies MT, De Meillon B (1968) The Anophelinae of Africa South of the Sahara. Johannesburg: South African Institute for Medical Research.
21. Coetzee M, Fontenille D (2004) Advances in the study of *Anopheles funestus*, a major vector of malaria in Africa. Insect Biochem Mol Biol 34: 599–605.
22. Calvo E, Dao A, Pham VM, Ribeiro JM (2007) An insight into the sialome of *Anopheles funestus* reveals an emerging pattern in anopheline salivary protein families. Insect Biochem Mol Biol 37: 164–175.
23. Serazin AC, Dana AN, Hillenmeyer ME, Lobo NF, Coulibaly MB, et al. (2009) Comparative analysis of the global transcriptome of *Anopheles funestus* from Mali, West Africa. PLoS One. 4:e7976.
24. Hunt RH, Brooke BD, Pillay C, Koekemoer LL, Coetzee M (2005) Laboratory selection for and characteristics of pyrethroid resistance in the malaria vector *Anopheles funestus*. Med and Vet Entom 19: 271-275.

25. Costantini C, Sagnon NF, Ilboudo-Sanogo E, Coluzzi M, Boccolini D (1999) Chromosomal and bionomic heterogeneities suggest incipient speciation in Anopheles funestus from Burkina Faso. Parassitologia 41:595-611.
26. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
27. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673-4680.
28. R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
29. Gaunt MW, Miles MA (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Mol. Biol. Evol 19: 748–761.
30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
31. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.
32. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite.Trends in Genetics 16: 276-277.

33. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P (2000) SMART: a webbased tool for the study of genetically mobile domains. Nucleic Acids Res 28: 231–234.

34. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.

35. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, et al. (2000) The Pfam protein families database. Nucleic Acids Res 28: 263–266.

36. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, et al. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 30: 281–283.

37. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-60.

38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078-2079.

39. Jiang R, Tavare S, Marjoram P (2009) Population genetic inference from resequencing data. Genetics 181: 187-197.

40. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5: 621–628.

41. Charif D and Lobry JR (2007) Seqin{R} 1.0-2: a contributed package to the R project for statistical computing devoted to biological sciences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural approaches to sequence evolution: Molecules, networks, populations. Springer Verlag, New York. pp. 207-232.

42. Surget-Groba Y, Montoya-Burgos J (2010) Optimization of *de novo* transcriptome

assembly from next-generation sequencing data. Genome Research 20: 1432-1440.

43. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28: 511-515.
44. David J, Coissac E, Melodelima C, Poupardin R, Riaz MA, et al. (2010) Transcriptome response to pollutants and insecticides in the dengue vector *Aedes aegypti* using next-generation sequencing technology. BMC Genomics 11: 216.
45. Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Kanwal S, et al. (2007) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. Science 316: 1738-1743.
46. Riehle MM, Xu J, Lazzaro BP, Rottschaefer SM, Coulibaly B, 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.
47. Fraiture M, Baxter RHG, Steinert S, Chelliah Y, Frolet C, et al. (2009) Two mosquito LRR proteins function as complement control factors in the TEP1-mediated killing in *Plasmodium*. Cell Host & Microbe 5: 273-284.
48. Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, et al. (2007) Dynamic evolution of the innate immune system in *Drosophila*. Nature Genetics 39: 1461-1468.
49. Wondji CS, Hemingway J, Ranson H (2007) Identification and analysis of single nucleotide polymorphisms (SNPs) in the mosquito *Anopheles funestus*, malaria vector. BMC Genomics 8: 5.

50. Obbard DJ, Welch JJ, Little TJ (2009) Inferring selection in the *Anopheles gambiae* species complex: an example from immune-related serine protease inhibitors. Malaria Journal 8: 117.

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

52. Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. Science 450: 203-218.

53. Grimaldi D, Engel MS (2005) Evolution of the insects. Cambridge: Cambridge University Press. 755 p.

FIGURES

Figure 1: *De novo* transcriptome assembly and analysis workflow. Illumina reads were assembled in a series of 'exploratory' Velvet assemblies, the contig output of which was used in a 'summary' assembly. Following iterative assembly with Velvet, contigs were clustered and joined when possible, first using conspecific ESTs, then using the transcriptome of a closely related species. A final contig set was generated by selecting contigs based on bioinformatic support criteria. Illumina reads were then mapped to the final contig set and resulting alignments were used for expression profiling and polymorphism discovery. [a]SND refers to short nucleotide discrepancies including both single nucleotide polymorphisms and indels. [b]RPKM, or reads per kilobase per million mapped reads [40], was calculated for each contig and used to represent expression level.

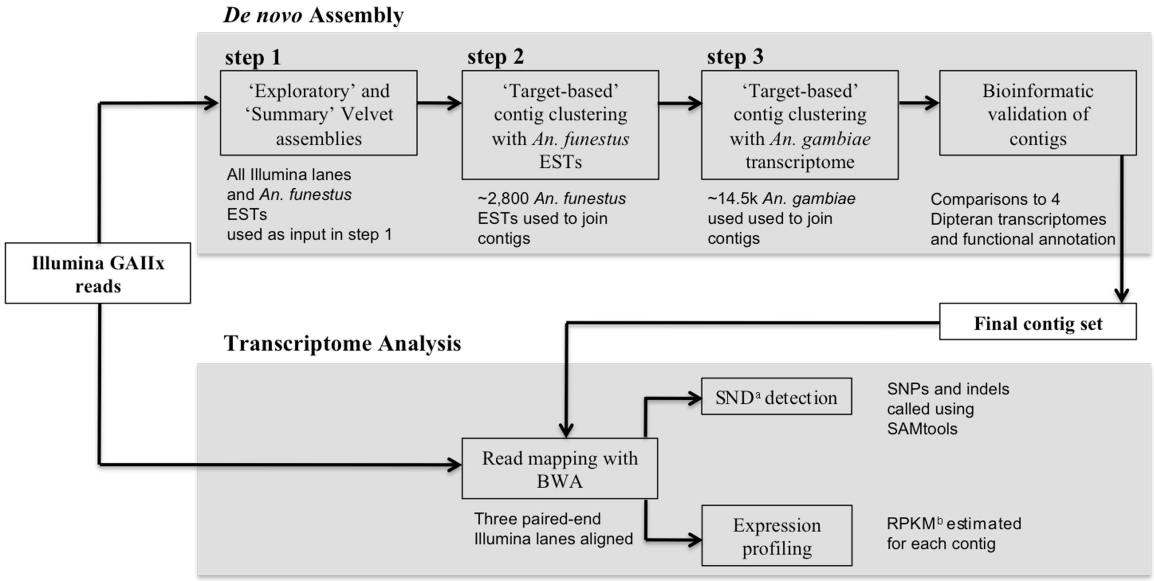


Figure 2: Size distribution of contigs at three points of the assembly. Note that the y-axis is broken between 10,000 and 40,000. White bars indicate the size distribution of contigs generated by the iterative Velvet assembly. Grey bars indicated the size distribution of contigs after 'target-based' clustering to both *An. funestus* ESTs and *An. gambiae* peptides. Black bars indicate the size distribution of the final contig set after quality filtering and bioinformatic analysis. The final contig set contains 15,527 contigs with an N50 of 1,753 bp.

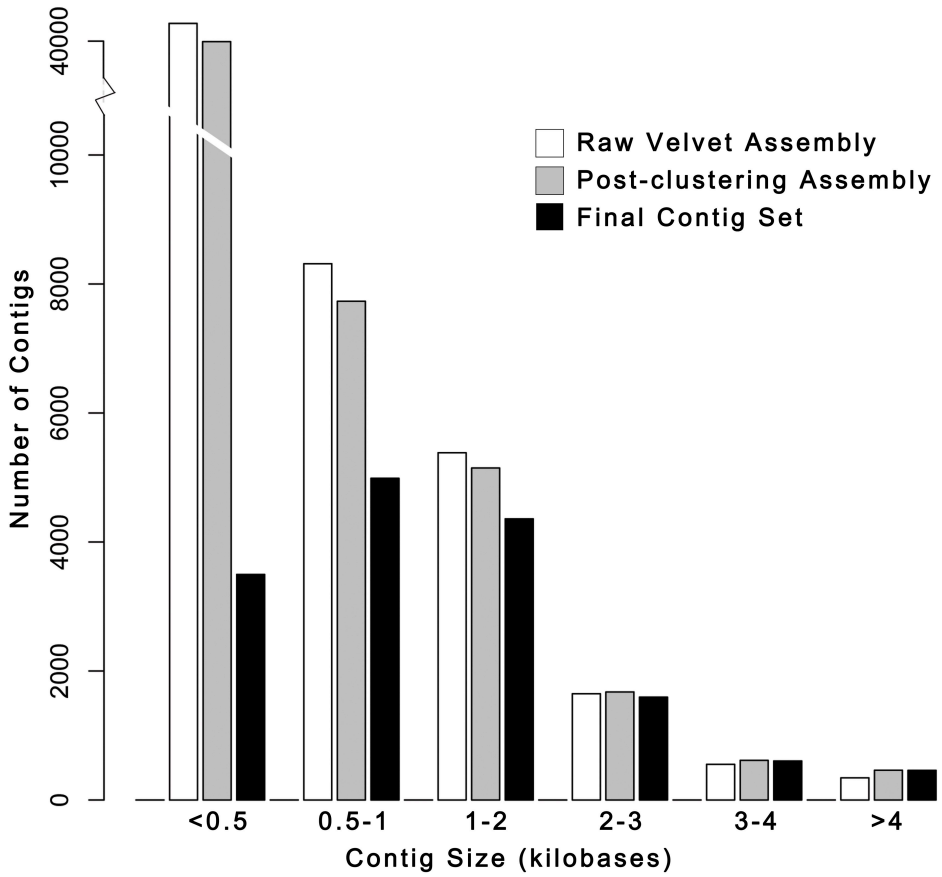


Figure 3: Homology with Dipteran transcriptomes decreases with increasing phylogentic difference. The number of *An. funestus* contigs with significant BLAST hits in pairwise comparisons to *An. gambiae, Ae. aegypti, C. quinquefasciatus* and *D. melanogaster* is plotted. Note that the y-axis only spans 9,000 to 15,000. The solid line indicates the total number of contigs with a significant BLAST hit in each comparison. The dashed line indicates the number of contigs with a significant BLAST hit in all comparisons as phylogenetic distance increases. The phylogenetic tree at the bottom of the panel depicts the evolutionary relationships between the Dipteran insects used in pairwise BLAST comparisons, with estimated divergence times (in millions of years) at each node (adapted from [53]).

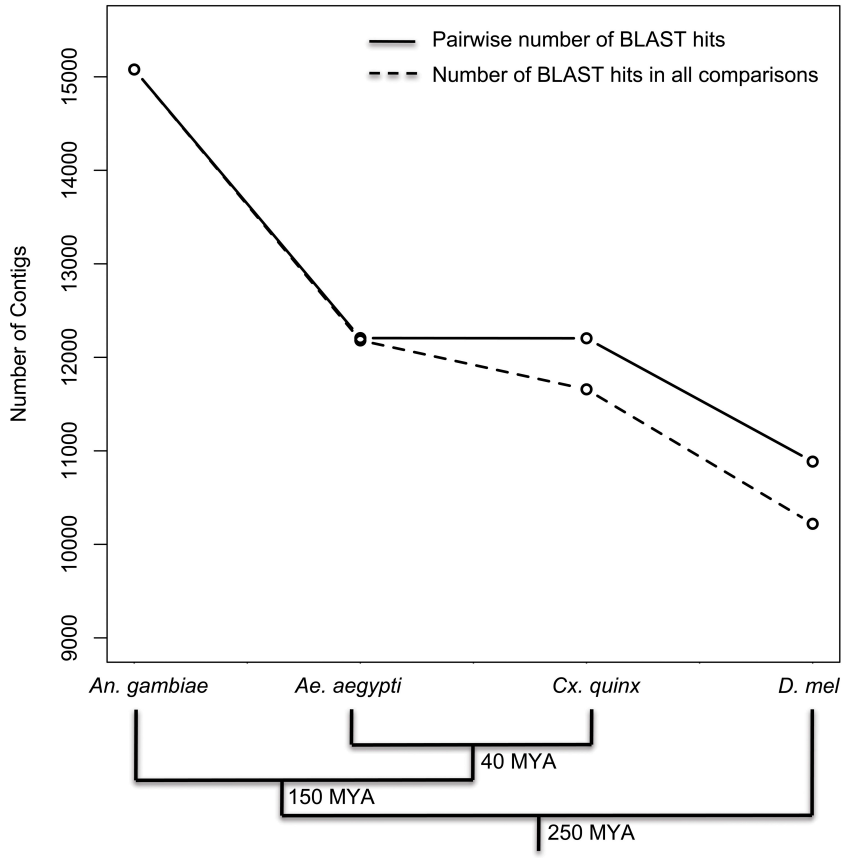


Figure 4: Variation in transcript divergence among immune gene functional classes. Protein sequence divergence was estimated as the proportion of aligned amino acids that differ between 1:1 *An. funestus*:*An. gambiae* orthologues. As a class, immune gene orthologous pairs (dotted line indicates mean divergence between immune gene orthologues) are significantly more diverged than the transcriptome as a whole (solid line indicates mean divergence across the entire transcriptome; p-value = 4.8×10^{-5}, Mann-Whitney U-test). The functional classes within the immune genes are not significantly different from each (p-values > 0.05, pairwise Mann-Whitney U-tests).

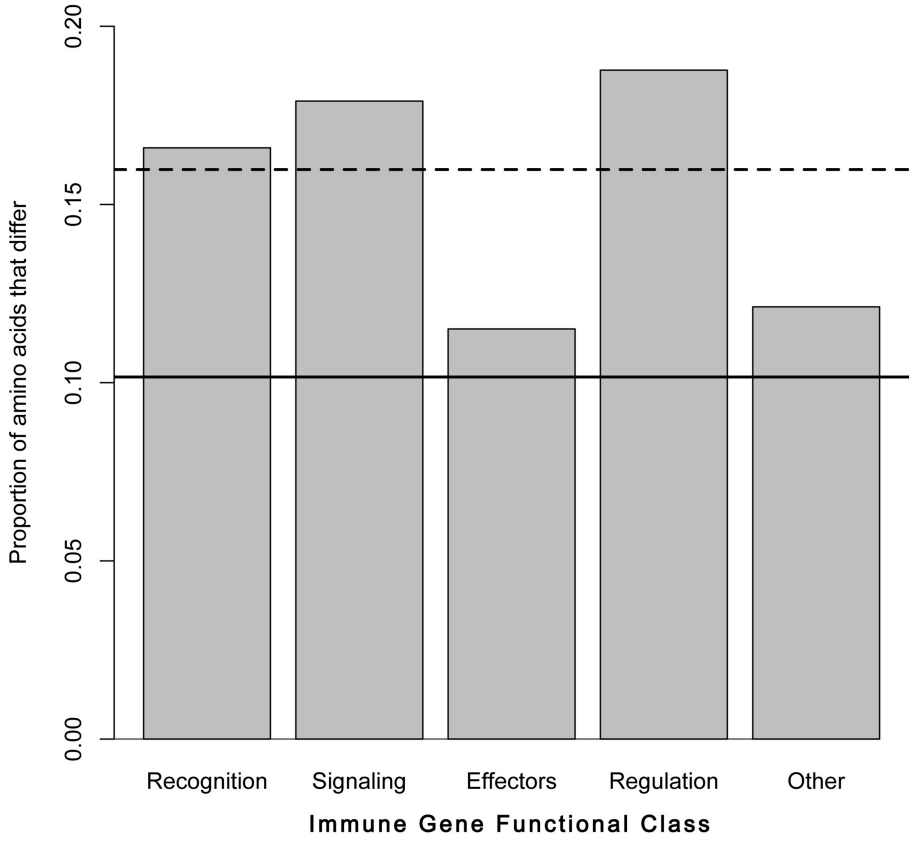
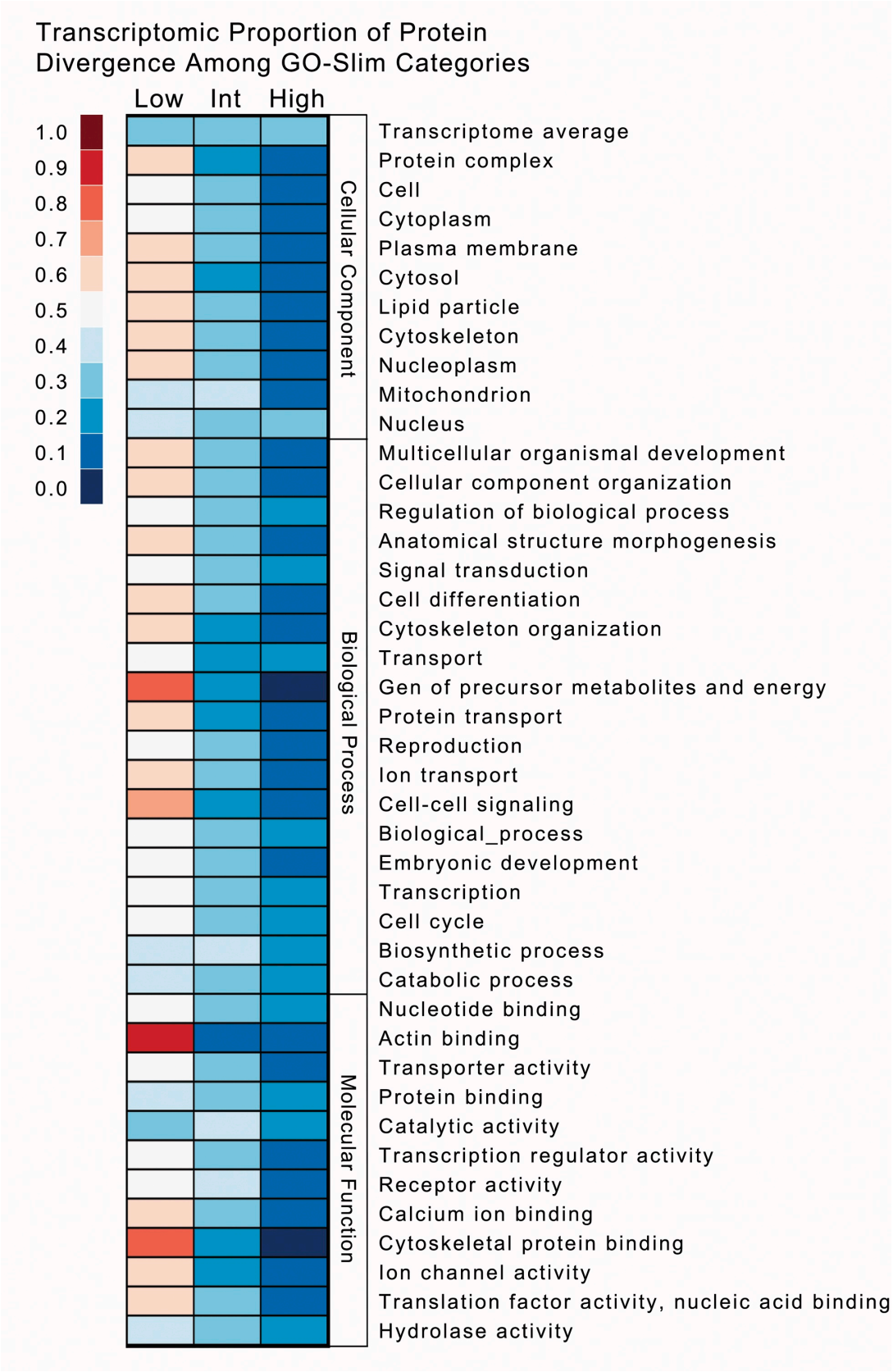


Figure 5: PROTEIN DIVERGENCE is unevenly distributed among GO-Slim categories. The heatplot shows proportion of 1:1 orthologous pairs exhibiting Low, Intermediate and High protein divergence in GO-Slim functional categories. Protein divergence was estimated as the proportion of aligned amino acids that differed between the two orthologues and each orthologous pair was categorized as Low, Intermediate or High (Materials and Methods). Only categories whose proportion of each bin differed from expectations based on all orthologous pairs with a *p* value less than the Bonferroni-adjusted α of 5.26×10^{-4} are presented. The average expected proportions based on all orthologous pairs are presented at the top of the heatplot.


Transcriptomic Proportion of Protein
Divergence Among GO-Slim Categories
Low
Int
High
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Cellular Component
Biological Process
Molecular Function
Transcriptome average
Protein complex
Cell
Cytoplasm
Plasma membrane
Cytosol
Lipid particle
Cytoskeleton
Nucleoplasm
Mitochondrion
Nucleus
Multicellular organismal development
Cellular component organization
Regulation of biological process
Anatomical structure morphogenesis
Signal transduction
Cell differentiation
Cytoskeleton organization
Transport
Gen of precursor metabolites and energy
Protein transport
Reproduction
Ion transport
Cell-cell signaling
Biological_process
Embryonic development
Transcription
Cell cycle
Biosynthetic process
Catabolic process
Nucleotide binding
Actin binding
Transporter activity
Protein binding
Catalytic activity
Transcription regulator activity
Receptor activity
Calcium ion binding
Cytoskeletal protein binding
Ion channel activity
Translation factor activity, nucleic acid binding
Hydrolase activity

SUPPLEMENTARY MATERIAL

Table S1: Contig Information Database. Excel spreadsheet containing information about each contig as well as all results from BLAST and functional domain database comparisons. Each row contains information for a single contig and each column contains a result for a specific analysis. NAs are inserted where there was either no result or the result did not apply to that contig. From left to right, the columns contain descriptive molecular information, BLAST results, functional annotation results and diversity and expression results.

** This table is extremely large (>30MB) and could not be set into this document. It was included in the published version of this work and can be found at the website of PLoS One (doi:10.1371/journal.pone.0014202.s001) or at the Lazzaro Laboratory website (www.lazzaro.entomology.cornell.edu). The most pertinent conclusions from its contents are described in the text making the details useful only to those readers interested in pursuing further analysis with these data.

Table S2: Putative An. funestus immune genes. Excel spreadsheet containing significant BLAST matches between An. funestus contigs and An. gambiae immune genes from ImmunoDB. BLAST e-values as well as relevant immunity annotations are presented. Found at: doi:10.1371/journal.pone.0014202.s002.

Table S3: Protein Divergence GO-Slim Analysis. Excel spreadsheet containing number of contigs in each divergence bin and results of X2 analysis for each GO-Slim functional category. Found at: doi:10.1371/journal.pone.0014202.s003 (0.04 MB XLS)

CHAPTER 7:

RESEARCH SUMMARY

In my thesis, I studied the demographic and selective processes driving the evolution of *Anopheles gambiae*, explored biases and statistical power inherent to next-generation sequencing-based population genomics studies, and sequenced the transcriptome of *Anopheles funestus*. Chapter 2 described the demographic histories of two insipient species of *A. gambiae*, the M and S molecular forms, and explores the implications of these findings for studies of natural selection in this system. Chapters 3 and 4 presented the results of population genetic studies of natural selection at two structural proteins (Chapter 3) and 28 immune-related proteins (Chapter 4), all of which were chosen based on potential involvement in parasite development or susceptibility. Chapter 5 described a simulation study undertaken to assess the biases and relative statistical power inherent to population genomic studies based on next-generation sequencing, in preparation for the shift from candidate gene studies to whole-genome re-sequencing studies. Chapter 6 presented the results of a *de novo* assembly of the transcriptome of *A. funestus* using short-read next-generation sequencing data as well as a full annotation and bioinformatic analysis. By studying the population genetics of *A. gambiae* from both the perspective of demography and natural selection and developing infrastructure, I have brought into focus the population history and selective landscape of this system and developed resources for future studies in *Anopheles* malaria vectors.

Because of their role in vectoring the human malaria parasite *Plasmodium falciparum*, *Anopheles* mosquitoes are the subject of extensive functional study aimed at understanding the physiological and mechanistic interactions between the mosquito and the parasite. Although the results of these studies can sometimes inform our basic

understanding of host-pathogen interactions in general, more often these studies aim towards the development of disease intervention measures. One promising and popular intervention strategy is to identify mosquito molecules that are involved in the anti-*Plasmodium* immune response or in *Plasmodium* development within the mosquito and exploit these mechanisms to enhance the response or block transmission, through the development of a Transmission-Blocking Vaccine or genetically modified mosquitoes (Gwadz 1976; Carter and Chen 1976; Brennan et al. 2000; Arrighi et al. 2005; Dong et al. 2011). In most cases, candidate genes that show a significant phenotypic shift when interrupted or blocked experimentally are identified in laboratory settings, typically using genetically inbred mosquito strains (Vernick et al. 2005; Cirimotich et al. 2010) and potentially missing natural phenotypic variation stemming from genetic variation segregating in the field. Indeed, natural populations harbor substantial genetic variation for parasite susceptibility (Niaré et al. 2002), and natural selection may further complicate the landscape by driving frequency shifts of functionally different genetic variants. Under these conditions, an intervention developed using lab strains may not be effective when applied to genetically heterogeneous natural populations. As such, understanding the evolutionary history and population genetics of candidate genes is essential to maximize the prospects of effective deployments of mosquito-directed malaria intervention technologies. In a population genetic analysis of the salivary gland protein saglin and the basil lamina structural protein laminin, both potential candidates for intervention mechanisms, I found no evidence for adaptive evolution in *A. gambiae*, and therefore no reason that these loci should not be considered for such development. On the contrary, these proteins appear to be under significant purifying selection indicating that they are relatively stable from an evolutionary perspective. This may

imply that functional variation at these genes is limited in natural populations, and that these proteins could be particularly good candidates for development as vaccine targets.

A slightly more complex picture arose when candidate immune-related genes were studied in *A. gambiae*. It is well established that immune genes are subject to rapid evolution in many organisms (Nielsen et al. 2005; Sackton et al. 2007; Waterhouse et al. 2007; Obbard et al. 2009), and are therefore particularly important to study from the perspective of intervention technology development. An especially promising set of candidate genes came from a series of genetic mapping studies in natural populations from both West and East Africa. These studies pointed to a quantitative trait locus (QTL) on the left arm of chromosome 2 in *A. gambiae* that repeatedly showed significant associations with malaria parasite susceptibility (Riehle et al. 2006; Riehle et al. 2007). Filtering of the gene set under the QTL based on functional and annotation criteria generated a reasonably small set of candidate genes that I studied from a population genetic perspective. I found a stark division of putative selection signals among incipient species of *A. gambiae*. One interpretation of these data is that the demographic strata are exposed to different pathogen repertoires, perhaps at the larval stage, and may be adapting to alternate environments. Specifically, the M molecular form and GOUNDRY showed evidence consistent with adaptive selection at a number of immune genes, while the S molecular form showed no evidence of molecular adaptation at these loci. In concert with the fact that the M and S molecular forms appear to have different demographic histories, with the M form having undergone a more recent population size change (Chapter 2), these results point to a complex ecological and genetic landscape where populations that are largely sympatric at the macro scale (at least in West Africa) are evolving independently and in response to discrete ecological and pathogenic pressures at the micro scale. From the perspective of intervention, this suggests that

immune genes may encode a particularly labile set of proteins that could prove difficult to target reliably with intervention strategies.

The results of these studies are consistent with a developing theme in the *Anopheles gambiae* literature, namely that, while the *Plasmodium* parasite may have adapted to the mosquito host in a species-and-region-specific manor (Billingsley and Sinden 1997; Molina-Cruz et al. 2012), there is less evidence for reciprocal adaptation on the part mosquito host. In fact, substantial cross-talk and overlap exists between the anti-*Plasmodium* immune response and the response to other pathogens (Meister et al. 2005; Dong et al. 2006; Meister et al. 2009; Dong, Manfredini, and Dimopoulos 2009), suggesting that a single gene or set of genes specifically adapted to target the human malaria parasite may not exist. Rather, a class of immune factors that are somewhat broad spectrum is employed to fight both a *Plasmodium* infection as well as infections by other pathogens, but it is currently unclear where the divisions lie within the immune system and within the repertoire of pathogens (Mitri and Vernick 2012). The fact that such stark population-specific selection signals exist among different strata of *Anopheles* suggests that the divisions within the immune system may be evolving differentially between strata in response to population-specific host-pathogen conflicts. Thus far, *A. gambaie* has been the primary focus, but it will be fascinating to explore these divisions and specificities within other vector and non-vector species. For example, experimental infections of *Anopheles quadriannulatus*, a zoophilic member of the *Anopheles gambiae* species complex, revealed a remarkably effective anti-*Plasmodium* response largely based on melanization and involving many of the proteins identified in *Anopheles gambiae*, implying the existence of an ancestral mechanism effective at combating *Plasmodium* that predates the split of these mosquito species (Habtewold et al. 2008). *Anopheles quadriannulatus* is phylogenetically proximal to *A. gambiae*, so it will be even

more insightful to study the immune response of a phylogenetically distant vector species, such as *A. funestus*, where substantial genetic differences have accumulated over a time period much longer than the much shorter and more recent time of exposure to human malaria parasites.

The field of *Anopheles* population genetics, and perhaps non-model systems at large, has been hampered by both a bottleneck in data collection as well as limitations in the development of tools and infrastructure appropriate for this system. For example, a reliable, accurate and sufficiently fine scale genetic map is not available for *A. gambiae* or other vector species, although a coarse scale map has been generated for *A. gambiae* (Zheng et al. 1996). This is in part due to the fact that coarse data suggests that recombination rates are 10-fold higher (~ 1 - 1.5 cM Mb^{-1}) in *A. gambiae* (Zheng et al. 1996; Pombi et al. 2006) than in humans (~ 0.1 cM Mb^{-1}; McVean et al. 2004), so extremely dense marker datasets are required to capture fine-scale changes in recombination rates. Although the sequencing of the *A. gambiae* genome in 2002 (Holt et al. 2002) has undoubtedly accelerated research in many research areas in that species, large-scale genomic technologies such as fine-scale SNP arrays have only recently been developed (Neafsey et al. 2010) precluding the large scale population studies needed to fully discover internal sub-structure within *A. gambiae*, instead resulting in a trickle of data slowly revealing the complex genetic landscape in this system (e.g. Slotman et al. 2007; Caputo et al. 2011). As an example, a coding-sequence-based microarray approach revealed several small regions of extreme genetic differentiation between the M and S molecular forms of *A. gambiae* (Turner, Hahn, and Nuzhdin 2005), leading to the conclusion that these regions represented 'genomic islands of speciation' surrounded by regions of no differentiation. However, subsequent analysis with greater resolution revealed that, while extreme regions exist, genetic

differentiation between these insipient species was widespread across the genome (Lawniczak et al. 2010; Neafsey et al. 2010). Moreover, long held sampling practices based on the assumption of indoor adult resting behavior of *A. gambiae* have lead to population ascertainment biases, as revealed by recent work involving more exhaustive sampling practices that lead to the discovery of the locally abundant and genetically diverged insipient species GOUNDRY in Burkina Faso (Riehle et al. 2011). This finding coupled with the fact that large swaths of the geographic range of many vector species have not been sampled, implies that additional complexity may exist in the field that we have not yet discovered and are not accounting for.

Recent technological advances in DNA sequencing resulting in the ability to obtain massive amounts of data at very low costs are sure to flip the paradigm such that data acquisition is no longer the limiting step, but instead data analysis and mosquito sampling will become limiting. Aside from the dramatic cost and labor reduction that come with next-generation sequencing technologies, they also provide the ability to conduct unbiased genome-wide analyses of allele frequencies and patterns of genetic variation that will allow the circumvention of the candidate gene ascertainment step that is biased by definition. In a recent first step towards truly genomic analysis in this system, Cheng et al. (2012) collected polymorphism data using Illumina next-generation sequencing technology from pools of *A. gambiae* that differ in their karyotype of the large 2La chromosomal inversion known to be adaptive in some populations and were able to characterize differentiation among alternative karyotype chromosomes at single base resolution. However, next-generation sequencing data are complicated by relatively high sequencing and genotype calling error rates are not without their limitations. Their full benefit will only be realized once the statistical and analytical tools capable of accounting for the substantial uncertainty inherent to these data are available.

The complications of next-generation sequencing technologies prohibit straightforward, turn-key application of the technology at present, and care must be taken to avoid introducing biases into downstream analyses that could lead to biased or misleading results. In Chapter 5, I presented the results of a simulation study aimed at quantifying these biases and determining shifts in statistical power when population genetic tests are applied to data with varying depth of read coverage without incorporating data-related uncertainty into the analyses. To explore parameters expected for studies in ecological systems where budgets are likely to be limited, I focused on relatively low read depth (4x – 15x) and a small sample size (n=30). Through comparisons between results based on inferred data and complete data, I found that significant biases are introduced when using low read depths to study demography or measure genetic differentiation, highlighting the necessity of deep sequence read depths for some experimental goals. On the other hand, strong positive selection was easily identified using 4x read depth. As the sequencing technology error rates and analytical tools improve, some of these biases may be mitigated, but my analysis underlines the need for caution when proceeding with whole-genome re-sequencing studies in the short term. The state of the technology will undoubtedly improve, however, and I expect large amounts of data to flood the *Anopheles* system, but I am only cautiously optimistic that this flip in the paradigm will result in dramatic and quick leaps in our understanding of this system.

Another limitation that has mired the study of *Anopheles* malaria vectors besides *A. gambiae* is the unevenness of the phylogeny of this genus, with the approximately seven species within the *Anopheles gambiae* species complex representing very young and closely related species, and most known species outside of the complex being so

genetically diverged that few genetic resources can be shared between the resource-rich *A. gambiae* and species outside the complex (Krzywinski and Besansky 2003). This has been particularly complicating for the study of other primary vectors such as *A. funestus*. *A. funestus* and *A. gambiae* shared a common ancestor between 30 and 80 million years ago (Krzywinski, Grushko, and Besansky 2006) and DNA sequences cannot be aligned at the nucleotide level, precluding many experiments and analyses that require at least some genetic information. Adding to this complication, progress in studying *A. funestus* has been slow to accumulate, probably due to a bias of resource allocation toward the study of *A. gambiae* as well as difficulties in rearing *A. funestus* in the colony. A recent breakthrough has partially alleviated the second limitation in that researchers in Burkina Faso have managed to grow *A. funestus* in the lab, providing the third of three colonies in the world and the opportunity to conduct functional experiments in this species. Towards this goal, I traveled to Burkina Faso to conduct blood-feeding and *Plasmodium*-infection experiments using this new colony in order to identify genes that are transcriptionally regulated following ingestion of an infected blood-meal. The *Plasmodium*-infections failed due to unknown reasons, but I obtained cDNA from this colony, sequenced the messenger RNA using Illumina short-read technology, assembled the short-reads by developing a novel assembly approach, and obtained over 15,000 putative transcripts. From these transcripts, I identified a large number of immune genes and found these to be evolving more rapidly than the transcriptome on average. I also identified over 300,000 putative segregating sites that can be developed into markers for future genetic mapping studies in this system. This effort dramatically increased the genetic resources and bioinformatic data available for this system. The *A. funestus* genome is to eventually be sequenced as part of a genome cluster sequencing project originally funded in 2008 (Besansky and Anopheles Genomes Cluster Commitee

2008), and my transcriptome data provide a substantial genetic foothold into annotation of that genome, provide genomic resources in the *A. funestus* system in the short term, and generally help establish the platform for future studies of this severely understudied primary vector of malaria.

In summary, my thesis addressed the relative roles of demography and natural selection in shaping patterns of genetic variation across the genome of *A. gambiae*, particularly at *Plasmodium*-related genes, and developed infrastructure for future studies of *Anopheles* vectors. It represents the most exhaustive demographic analysis and deepest re-sequencing analysis of immune genes in *A. gambiae* to date, and makes by far the largest contribution of DNA sequencing data and analysis in *A. funestus*. Future studies following up on the putative signals of natural selection at the candidate genes studied here and the transcriptome data in *A. funestus* will further our understanding of demographic and selective forces shaping evolution in this system and will inform efforts to develop mosquito-based intervention technologies.

REFERENCES

Arrighi, Romanico B G, Gareth Lycett, Vassiliki Mahairaki, Inga Siden-Kiamos, and Christos Louis. 2005. “Laminin and the Malaria Parasite’s Journey Through the Mosquito Midgut.” *The Journal of Experimental Biology* 208 (Pt 13) (July): 2497–2502. doi:10.1242/jeb.01664.

Besansky, Nora, and Anopheles Genomes Cluster Commitee. 2008. “Genome Analysis of Vectorial Capacity in Major Anopheles Vectors of Malaria Parasites.”

Billingsley, P.F., and R.E. Sinden. 1997. “Determinants of Malaria-mosquito Specificity.” *Parasitology Today* 13 (8) (August): 297–301. doi:10.1016/S0169-4758(97)01094-6.

Brennan, J D, M Kent, R Dhar, H Fujioka, and N Kumar. 2000. “Anopheles Gambiae Salivary Gland Proteins as Putative Targets for Blocking Transmission of Malaria Parasites.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (25) (December 5): 13859–13864. doi:10.1073/pnas.250472597.

Caputo, Beniamino, Federica Santolamazza, José L Vicente, Davis C Nwakanma, Musa Jawara, Katinka Palsson, Thomas Jaenson, et al. 2011. “The ‘Far-west’ of Anopheles Gambiae Molecular Forms.” *PloS One* 6 (2): e16415. doi:10.1371/journal.pone.0016415.

Carter, Richard, and David H. Chen. 1976. “Malaria Transmission Blocked by Immunisation with Gametes of the Malaria Parasite.” *, Published Online: 02 September 1976; | Doi:10.1038/263057a0* 263 (5572) (September 2): 57–60. doi:10.1038/263057a0.

Cheng, Changde, Bradley J White, Colince Kamdem, Keithanne Mockaitis, Carlo Costantini, Matthew W Hahn, and Nora J Besansky. 2012. “Ecological Genomics of Anopheles Gambiae Along a Latitudinal Cline: A Population-Resequencing Approach.” *Genetics* 190 (4) (April 1): 1417–1432. doi:10.1534/genetics.111.137794.

Cirimotich, Chris M, Yuemei Dong, Lindsey S Garver, Shuzhen Sim, and George Dimopoulos. 2010. “Mosquito Immune Defenses Against Plasmodium Infection.” *Developmental and Comparative Immunology* 34 (4) (April): 387–395. doi:10.1016/j.dci.2009.12.005.

Dong, Yuemei, Ruth Aguilar, Zhiyong Xi, Emma Warr, Emmanuel Mongin, and George Dimopoulos. 2006. “Anopheles Gambiae Immune Responses to Human and Rodent Plasmodium Parasite Species.” *PLoS Pathog* 2 (6) (June 9): e52. doi:10.1371/journal.ppat.0020052.

Dong, Yuemei, Suchismita Das, Chris Cirimotich, Jayme A Souza-Neto, Kyle J McLean, and George Dimopoulos. 2011. “Engineered Anopheles Immunity to Plasmodium Infection.” *PLoS Pathogens* 7 (12) (December): e1002458. doi:10.1371/journal.ppat.1002458.

Dong, Yuemei, Fabio Manfredini, and George Dimopoulos. 2009. “Implication of the Mosquito Midgut Microbiota in the Defense Against Malaria Parasites.” *PLoS Pathogens* 5 (5) (May): e1000423. doi:10.1371/journal.ppat.1000423.

Gwadz, R. W. 1976. “Successful Immunization Against the Sexual Stages of Plasmodium Gallinaceum.” *Science* 193 (4258) (September 17): 1150–1151. doi:10.1126/science.959832.

Habtewold, Tibebu, Michael Povelones, Andrew M Blagborough, and George K Christophides. 2008. “Transmission Blocking Immunity in the Malaria Non-vector

Mosquito Anopheles Quadriannulatus Species A." *PLoS Pathogens* 4 (5) (May): e1000070. doi:10.1371/journal.ppat.1000070.

Holt, Robert A, G Mani Subramanian, Aaron Halpern, Granger G Sutton, Rosane Charlab, Deborah R Nusskern, Patrick Wincker, et al. 2002. "The Genome Sequence of the Malaria Mosquito Anopheles Gambiae." *Science (New York, N.Y.)* 298 (5591) (October 4): 129–149. doi:10.1126/science.1076181.

Krzywinski, Jaroslaw, and Nora J Besansky. 2003. "Molecular Systematics of Anopheles: From Subgenera to Subpopulations." *Annual Review of Entomology* 48: 111–139. doi:10.1146/annurev.ento.48.091801.112647.

Krzywinski, Jaroslaw, Olga G Grushko, and Nora 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 (2) (May): 417–423. doi:10.1016/j.ympev.2006.01.006.

Lawniczak, M K N, S J Emrich, A K Holloway, A P Regier, M Olson, B White, S Redmond, et al. 2010. "Widespread Divergence Between Incipient Anopheles Gambiae Species Revealed by Whole Genome Sequences." *Science (New York, N.Y.)* 330 (6003) (October 22): 512–514. doi:10.1126/science.1195755.

McVean, Gilean A T, Simon R Myers, Sarah Hunt, Panos Deloukas, David R Bentley, and Peter Donnelly. 2004. "The Fine-scale Structure of Recombination Rate Variation in the Human Genome." *Science (New York, N.Y.)* 304 (5670) (April 23): 581–584. doi:10.1126/science.1092500.

Meister, Stephan, Bogos Agianian, Fanny Turlure, Angela Relógio, Isabelle Morlais, Fotis C Kafatos, and George K Christophides. 2009. "Anopheles Gambiae PGRPLC-mediated Defense Against Bacteria Modulates Infections with Malaria Parasites." *PLoS Pathogens* 5 (8) (August): e1000542. doi:10.1371/journal.ppat.1000542.

Meister, Stephan, Stefan M Kanzok, Xue-Li Zheng, Coralia Luna, Tong-Ruei Li, Ngo T Hoa, John Randall Clayton, et al. 2005. "Immune Signaling Pathways Regulating Bacterial and Malaria Parasite Infection of the Mosquito Anopheles Gambiae." *Proceedings of the National Academy of Sciences of the United States of America* 102 (32) (August 9): 11420–11425. doi:10.1073/pnas.0504950102.

Mitri, Christian, and Kenneth D Vernick. 2012. "Anopheles Gambiae Pathogen Susceptibility: The Intersection of Genetics, Immunity and Ecology." *Current Opinion in Microbiology* (April 24). doi:10.1016/j.mib.2012.04.001. http://www.ncbi.nlm.nih.gov/pubmed/22538050.

Molina-Cruz, Alvaro, Randall J Dejong, Corrie Ortega, Ashley Haile, Ekua Abban, Janneth Rodrigues, Giovanna Jaramillo-Gutierrez, and Carolina Barillas-Mury. 2012. "Some Strains of Plasmodium Falciparum, a Human Malaria Parasite, Evade the Complement-like System of Anopheles Gambiae Mosquitoes." *Proceedings of the National Academy of Sciences of the United States of America* (May 23). doi:10.1073/pnas.1121183109. http://www.ncbi.nlm.nih.gov/pubmed/22623529.

Neafsey, D E, M K N Lawniczak, D J Park, S N Redmond, M B Coulibaly, S F Traoré, N Sagnon, et al. 2010. "SNP Genotyping Defines Complex Gene-flow Boundaries Among African Malaria Vector Mosquitoes." *Science (New York, N.Y.)* 330 (6003) (October 22): 514–517. doi:10.1126/science.1193036.

Niaré, Oumou, Kyriacos Markianos, Jennifer Volz, Frederick Oduol, Abdoulaye Touré, Magaran Bagayoko, Djibril Sangaré, et al. 2002. "Genetic Loci Affecting

Resistance to Human Malaria Parasites in a West African Mosquito Vector Population." *Science (New York, N.Y.)* 298 (5591) (October 4): 213–216. doi:10.1126/science.1073420.

Nielsen, Rasmus, Carlos Bustamante, Andrew G Clark, Stephen Glanowski, Timothy B Sackton, Melissa J Hubisz, Adi Fledel-Alon, et al. 2005. "A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees." *PLoS Biol* 3 (6) (May 3): e170. doi:10.1371/journal.pbio.0030170.

Obbard, Darren J, John J Welch, Kang-Wook Kim, and Francis M Jiggins. 2009. "Quantifying Adaptive Evolution in the Drosophila Immune System." *PLoS Genetics* 5 (10) (October): e1000698. doi:10.1371/journal.pgen.1000698.

Pombi, Marco, Aram D Stump, Alessandra Della Torre, and Nora J Besansky. 2006. "Variation in Recombination Rate Across the X Chromosome of Anopheles Gambiae." *The American Journal of Tropical Medicine and Hygiene* 75 (5) (November): 901–903.

Riehle, Michelle M, Wamdaogo M Guelbeogo, Awa Gneme, Karin Eiglmeier, Inge Holm, Emmanuel Bischoff, Thierry Garnier, et al. 2011. "A Cryptic Subgroup of Anopheles Gambiae Is Highly Susceptible to Human Malaria Parasites." *Science (New York, N.Y.)* 331 (6017) (February 4): 596–598. doi:10.1126/science.1196759.

Riehle, Michelle M, Kyriacos Markianos, Louis Lambrechts, Ai Xia, Igor Sharakhov, Jacob C Koella, and Kenneth D Vernick. 2007. "A Major Genetic Locus Controlling Natural Plasmodium Falciparum Infection Is Shared by East and West African Anopheles Gambiae." *Malaria Journal* 6: 87. doi:10.1186/1475-2875-6-87.

Riehle, Michelle M, Kyriacos Markianos, Oumou Niaré, Jiannong Xu, Jun Li, Abdoulaye M Touré, Belco Podiougou, et al. 2006. "Natural Malaria Infection in Anopheles Gambiae Is Regulated by a Single Genomic Control Region." *Science (New York, N.Y.)* 312 (5773) (April 28): 577–579. doi:10.1126/science.1124153.

Sackton, Timothy B, Brian P Lazzaro, Todd A Schlenke, Jay D Evans, Dan Hultmark, and Andrew G Clark. 2007. "Dynamic Evolution of the Innate Immune System in Drosophila." *Nature Genetics* 39 (12) (December): 1461–1468. doi:10.1038/ng.2007.60.

Slotman, M A, F Tripet, A J Cornel, C R Meneses, Y Lee, L J Reimer, T C Thiemann, et al. 2007. "Evidence for Subdivision Within the M Molecular Form of Anopheles Gambiae." *Molecular Ecology* 16 (3) (February): 639–649. doi:10.1111/j.1365-294X.2006.03172.x.

Turner, Thomas L, Matthew W Hahn, and Sergey V Nuzhdin. 2005. "Genomic Islands of Speciation in Anopheles Gambiae." *PLoS Biology* 3 (9) (September): e285. doi:10.1371/journal.pbio.0030285.

Vernick, K D, F Oduol, B P Lazzaro, J Glazebrook, J Xu, M Riehle, and J Li. 2005. "Molecular Genetics of Mosquito Resistance to Malaria Parasites." *Current Topics in Microbiology and Immunology* 295: 383–415.

Waterhouse, Robert M, Evgenia V Kriventseva, Stephan Meister, Zhiyong Xi, Kanwal S Alvarez, Lyric C Bartholomay, Carolina Barillas-Mury, et al. 2007. "Evolutionary Dynamics of Immune-related Genes and Pathways in Disease-vector Mosquitoes." *Science (New York, N.Y.)* 316 (5832) (June 22): 1738–1743. doi:10.1126/science.1139862.

Zheng, L, M Q Benedict, A J Cornel, F H Collins, and F C Kafatos. 1996. “An Integrated Genetic Map of the African Human Malaria Vector Mosquito, Anopheles Gambiae.” *Genetics* 143 (2) (June): 941–952.