Analysis of anopheline mosquito behavior and identification of vector control targets in the post-genomic era

Author: Adam Jenkins

Persistent link: http://hdl.handle.net/2345/bc-ir:104489

This work is posted on eScholarship@BC,
Boston College University Libraries.

Boston College Electronic Thesis or Dissertation, 2015

Copyright is held by the author, with all rights reserved, unless otherwise noted.

Boston College

The Graduate School of Arts and Sciences

Department of Biology

ANALYSIS OF ANOPHELINE MOSQUITO BEHAVIOR AND IDENTIFICATION OF VECTOR CONTROL TARGETS IN THE POST-GENOMIC ERA

a dissertation

by

ADAM M. JENKINS

submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

May 2015

© copyright by ADAM M. JENKINS
2015

ABSTRACT

ANALYSIS OF ANOPHELINE MOSQUITO BEHAVIOR AND IDENTIFICATION OF VECTOR CONTROL TARGETS IN THE POST-GENOMIC ERA

Adam M. Jenkins

Thesis Advisor: Marc A.T. Muskavitch

The protozoan *Plasmodium falciparum*, the mosquito-borne pathogen that causes human malaria, remains one of the most difficult infectious parasites to combat and control. Campaigns against malaria eradication have succeeded, in most instances, at the level of vector control, rather than from initiatives that have attempted to decrease malaria burden by targeting parasites. The rapid evolution and spread of insecticide-resistant mosquitoes is threatening our ability to combat vectors and control malaria. Therefore, the development, procurement and distribution of new methods of vector control are paramount. Two aspects of vector biology that can be exploited toward these ends are vector behaviors and vector-specific insecticide targets. In this thesis, I describe three aspects of vector biology with potential for the development of improved means of vector control: photopreference behavior, long non-coding RNA (lncRNA) targets and epigenetic gene ensemble targets.

My studies of photopreference have revealed that specific mosquito species within the genus *Anopheles*, *An. gambiae* and An. *stephensi*, exhibit different photopreference behaviors, and that each gender of mosquito in these species exhibits distinct light-dependent resting behaviors. These inter-specific behavioral differences may be affected by differing numbers of long-wavelength sensing *Opsin* genes in each

species, and my findings regarding species-specific photopreferences suggest that some behavioral interventions may need to be tailored for specific vector mosquito species.

Based on the advancement of next-generation sequencing technologies and the generation by others of assembled genomes of many anopheline mosquito species, I have identified a comprehensive set of approximately 3,000 lncRNAs and find that RNA secondary structures are notably conserved within the *gambiae* species complex. As lncRNAs and epigenetic modifiers cooperate to modulate epigenetic regulation, I have also analyzed the conservation of epigenetic gene ensembles across a number of anopheline species, based on identification of homologous epigenetic ensemble genes in *An. gambiae* compared to *Drosophila melanogaster*. Further analyses of these ensembles illustrate that these epigenetic genes are highly stable among many anopheline species, in that I detect only eight gene family expansion or contraction events among 169 epigenetic ensemble genes within a set of 12 anopheline species.

My hope is that my findings will enable deeper investigations of many behavioral and epigenetic processes in *Anopheles gambiae* and other anopheline vector mosquitoes and thereby enable the development of new, more effective means of vector and malaria control.

DEDICATION

To my mother (Sharon), father (Robert), brother (Shawn) and sisters (Stephanie and Jessica) who have always shown unwavering support and to all the “guys” who gave their lives to make this work possible.

ACKNOWLEDGEMENTS

I would like to thank my fellow lab members, both graduate and undergraduate, of the Muskavitch Lab. Especially Kim, who has travelled, vacationed and studied with me since the beginning of our doctoral careers; I could not have asked for a better person to be within ten feet of me. I would like to thank my advisor, Marc A.T. Muskavitch, for his continued support in every facet of my research and for continually advancing my abilities both within and outside of research and academia. Any future success will be built on the foundation that Marc has provided and for that I am forever grateful.

Various individuals in the mosquito research community have played tremendous roles in my research, none more so than Daniel Neafsey and Robert Waterhouse. Both have been incredible mentors and colleagues and I wish both continued success in their future endeavors. Additionally, members of Alan Kopin's, Flaminia Catteruccia's and Welkin Johnson's labs have played large roles in my research and I thank them.

To all of my friends that I have made at Boston College, both young and old, you are undoubtedly the greatest take away I have from my time at Boston College. I will forever remember our long lunches, thirsty Thursdays and odd Fridays, no shaving contests, game nights and all those times I wish I could forget. I thank you all for being a part of my life.

My family arguably deserves the most appreciation for their support during my studies. They left me be when they knew I was frustrated and supportive when they knew I needed it the most. Although spread out, we are closer than ever and I love you all.

Lastly, I would like to thank my committee, Welkin Johnson, Thomas Chiles, and Marc-Jan Gubbels for their insights and direction during my doctoral studies and Tim van Opijnen for taking time out of his schedule as reader for my dissertation.

TABLE OF CONTENTS

ABSTRACT
DEDICATION **i**
ACKNOWLEDGEMENTS **iii**
TABLE OF CONTENTS **v**
LIST OF FIGURES **viii**
LIST OF TABLES **xi**

CHAPTER I
INTRODUCTION

A. Malaria: Epidemiology, Evolution, and Disease Burden **2**
B. Malaria Vectors: Background and Behavior **7**
C. Vector Control Methods **13**
D. Next Generation Sequencing of Malaria Vectors **18**
E. Identification and Evolution of lncRNA and Epigenetic Gene Classes **21**
F. Figures and Legends **26**

CHAPTER II
CREPUSCULAR BEHAVIORAL VARIATION AND PROFILING OF OPSIN GENES IN *ANOPHELES GAMBIAE* AND *ANOPHELES STEPHENSI*

A. Abstract **31**
B. Introduction **32**
C. Results and Discussion **37**
1. DETERMINATION OF PHOTOPREFERENCES IN *AN. GAMBIAE* AND *AN. STEPEHENSI*
2. DIRUNAL VARIATION OF *OPSIN* GENE EXPRESSION
3. DEVELOPMENTAL EXPRESSION AND EVOLUTION OF OPSIN GENE EXPRESSION
D. Methods **48**
1. COLONY
2. PHOTOPREFERENCE ASSAYS
3. STATISTICAL ANALYSIS
4. COLLECTION OF SAMPLES AND qRT-PCR OF SELECTED PHOTOTRANSDUCTION PATHWAY GENES
5. RNA SEQUENCING AND ANALYSIS
E. Tables/Figures and Legends **52**

CHAPTER III
LONG NON-CODING RNA DISCOVERY ACROSS THE GENUS *ANOPHELES* REVEALS CONSERVED SECONARY STRUCTURES

WITHIN AND BEYOND THE GAMBIAE COMPLEX

A. Abstract **67**
B. Introduction **68**
C. Results **71**
1. ALIGNMENT AND VALIDATION OF RNASEQ DATA SETS
2. DE NOVO IDENTIFICATION OF TRANSCRIPTS
3. EVOLUTIONARY CONSERVATION OF LNCRNA SEQUENCES AND SECONDARY STRUCTURES
D. Discussion **78**
E. Methods **86**
1. COLONY AND SEQUENCING
2. RNASEQ READ ALIGNMENT AND ANALYSIS
3. IDENTIFICATION OF NEWLY ANNOTATED TRANSCRIPTS
4. *ANOPHELES* GENOME ALIGNMENTS AND PHYLOCSF SCANNING FOR PROTEIN-CODING POTENTIAL
5. DIFFERENTIAL GENE EXPRESSION AND CATEGORIZATION
6. DETERMINING CONSERVATION AND SECONDARY STRUCTURE OF NEWLY ANNOTATED GENES ACROSS *ANOPHELES* LINEAGES
F. Tables/Figures and Legends **94**

CHAPTER IV
EVOLUTION OF AN EPIGENETIC GENE ENSEMBLE WITH THE GENUS *ANOPHELES*

A. Abstract **129**
B. Introduction **130**
C. Results **133**
1. DEFINING AN EPIGENETIC GENE ENSEMBLE IN *AN. GAMBIAE*
2. GENE FAMILY EXPANSIONS AND CONTRACTIONS ACROSS THE GENUS *ANOPHELES*
3. FUNCTIONAL AND EVOLUTIONARY COMPARISONS OF EPIGENETIC GENE ENSEMBLES
D. Discussion **143**
E. Methods **154**
1. ORTHOLOGOUS GENE IDENTIFICATION
2. PHYLOGENETIC ASSESSMENT AND dN/dS DETERMINATION
3. EXPRESSION OF EPIGENETIC MODIFIERS IN *AN. GAMBIAE* AND *D. MELANOGASTER*
4. PRINCIPAL COMPONENT ANALYSIS (PCA) OF TISSUE-SPECIFIC GENE EXPRESSION
F. Tables/Figures and Legends **160**

APPENDIX
RECTIFICATION OF G-PROTEIN COUPLED RECEPTOR GENE MODELS IN *ANOPHELES GAMBIAE*

A. Introduction **181**
B. Results/Discussion **185**
C. Methods **188**
D. Tables/Figures and Legends **190**

LIST OF ADDITIONAL FILES **198**

REFERENCES **200**

List of Figures:

Figure 1.1: *Plasmodium falciparum* Life Cycle **27**

Figure 1.2: Distribution of *Anopheles* Mosquitoes Across the World **29**

Figure 2.1. *An. gambiae* and *An. stephensi* Binary Photopreference **57**

Figure 2.2. *An. gambiae* and *An. stephensi* Trinary Photopreference **59**

Figure 2.3. Long Wavelength Opsin Gene Organization on *An. gambiae* Chromosome Arm 2R **61**

Figure 2.4. Opsin Expression Profiles Across Zeitgeber Time **63**

Figure 2.5. Heatmap of *An. gambiae Opsin* Gene Expression **65**

Figure 3.1. Validation of RNA-Seq Library and Analysis Techniques **103**

Figure 3.2. GOSLIM2 Terms of Genes that Exhibit Differential Expression Among Life Stages/Genders **105**

Figure 3.3. Flow Chart of lncRNA and Potential Coding Gene Identification and Expression/Exonic Structure of Defined Gene Classes **107**

Figure 3.4: Examples of Newly Annotated Protein-Coding and lncRNA Genes **109**

Figure 3.5: lncRNAs that Exhibit Differential Expression Among Life Stages/Genders **111**

Figure 3.6: Evolutionary Conservation Across the Genus Anopheles **113**

Figure 3.7: Sequence, structural and expression profiles of identified gene classes **115**

Figure 3.8: RNAz Scores of Secondary Structures in lncRNA and Novel Protein Coding Genes After REAPR Realignment **117**

Figure 3.9: Secondary Structures for a Differentially Expressed lncRNA **119**

Figure 3.10: Conservation of lncRNA predicted secondary structure and genomic regions across the *Anopheles* genus **121**

Figure 3.11: Histogram of Number of Genomes Aligned to For High-Confidence Secondary Structure **123**

Figure 3.12: Clustering of Conserved Secondary Structures in lncRNAs that are Present in All *Anopheles* Species **125**

Figure 3.13: Representative Quality Scores of LRD Samples **127**

Figure 4.1: Epigenetic Gene Set Identification and Analysis in Anopheline Species **165**

Figure 4.2: Phylogenetic Relationship of Set-N Chromatin Proteins **167**

Figure 4.3: Phylogenetic Relationships Among *Heterochromatin Protein-1* Orthologs in *D. melanogaster* and *An. gambiae* **169**

Figure 4.4: Retrotransposition of *CC14* within the genus *Anopheles* **171**

Figure 4.5: Alignment of *E2D* ubiquitin-conjugating enzyme genes and homologous retrogenes in anopheline species **173**

Figure 4.6: Epigenetic Gene Ensemble Expression in Tissues and Development **175**

Figure 4.7: Tissue Expression Difference Between *D. melanogaster* and *An. gambiae* **177**

Figure 4.8: GO Terms of Genes with Temporally Unique Expression Profiles Between Species **179**

Figure A.1: Examples of GPCR Gene Model Rectification **193**

Figure A.2: Peptide Alignment of *An. gambiae GPRFSH* Splice Variants and *D. melanogaster* Orthologs **195**

Figure A.3: *GPRFSH* Gene Model Rectification and Splice Junctions **197**

List of Tables:

Table 2.1. *An. gambiae* and *An. stephensi* Binary Photopreference Data **53**

Table 2.2. *An. gambiae* and *An. stephensi* Trinary Photo Preference Data **55**

Table 3.1: Read Alignment of RNA-Sequencing Data Sets **95**

Table 3.2: Genomes Utilized for Whole Genome Alignments and Associated *Anopheles* Species **97**

Table 3.2: Number of 1:1 Conserved lncRNA Regions in Each Anopheline Genome Assembly **99**

Table 3.4: Number of High-Confidence lncRNA Secondary Structures in Each Anopheline Genome Assembly **101**

Table 4.1: Comparison of epigenetic gene ensemble memberships in *D. melanogaster* and *An. gambiae* **161**

Table 4.2: Expansions/Contractions of Epigenetic Modifier Gene Families Across the Genus *Anopheles* **163**

Table A.1: Rectification of *An. gambiae* GPCR Gene Models **191**

Chapter I:

Introduction

A. Malaria: Epidemiology, Evolution, and Disease Burden

Approximately half of the world's population are at risk of contracting the infectious disease malaria (WHO 2014; Hay et al. 2004). The causative agent is a protozoan in the genus *Plasmodium*, and malaria is the result of infection by any of four distinct species within this genus: *P. vivax*, *P. ovale, P. malariae,* and the most prevalent in terms of number of infections caused, *P. falciparum* (White et al. 2014). Given its ability to infect a range of organisms outside of humans, including primates, non-primate mammals, birds and reptiles, and its ability to infect multiple species, such as macaques and humans, malaria can be classified as a zoonotic disease (Njabo et al. 2012; Langhorne et al. 2011; Lapointe et al. 2012; Cornet et al. 2014; Hayakawa et al. 2008; Escalante et al. 1998). Of the four species of *Plasmodium* that infect humans, the most dangerous form of the parasite, in terms of endemicity and mortality, is *P. falciparum* (Olliaro 2008; Elyazar et al. 2011; Gething et al. 2011; Snow et al. 2005). *P. falciparum* is endemic to sub-Saharan Africa, and many species of *Anopheles* mosquitoes transmit the parasite (White 1974; Sinka et al. 2010). Historically, the natural history of *P. falciparum* becoming a human disease is under debate. Research has indicated that *P. falciparum* is closely related to *P. reichenowi*, which infects the *Pan* genus (chimpanzees, bonobos) and diverged from *P. reichenowi* over 5-8 million years ago (MYA) (Krief et al. 2010; Prugnolle et al. 2011). Other studies have argued that *P. falciparum* evolved from a gorilla-infecting ancestor much more recently, approximately 5,000-50,000 years ago (Liu et al. 2010; Prugnolle et al. 2011). Although the exact means and time of initial transmission from chimpanzee/gorilla to human is still unknown, malaria's ability to

function as a zoonotic disease rests in its life cycle, the plasticity of its genome and its mode of transmission via a mosquito vector.

The life cycle of *Plasmodium* consists of two stages, a sexual stage (present in the mosquito primary host) and an asexual stage (present in a secondary host) (Fig. 1.1) (White et al. 2014; Hall et al. 2005; Miller et al. 2002). The sexual stage begins when an *Anopheles* mosquito takes up male and female gametocytes during a blood meal taken from secondary hosts. In the mosquito midgut, male and female gametocytes fuse to form a zygote. Zygotes then develop into ookinetes. Ookinetes must then invade the basal lamina of the midgut and divide, to form an oocyst. Upon oocyst rupture, sporozoites are released into the hemolymph, then migrate to and enter the mosquito's salivary gland, from which they may be deposited into a secondary host during the next blood meal. Once inside a secondary host, sporozoites infect hepatocytes during the exo-erythrocytic cycle, form a schizont consisting of multiple asexual parasitic merozoites and subsequently rupture after multiple rounds of division, releasing the merozoites into circulation. During the asexual erythrocytic cycle, red blood cells are infected by merozoites that develop further into immature trophozoites (ring stage), which then enter the mature trophozoite phase, and subsequently divide to form schizonts (schizont phase). From the schizont phase, merozoites are again released after rupture of the erythrocyte. In rare cases, during the trophozoite phase, the protozoan instead develops into a gametocyte, ensuring the life cycle can continue when taken up in a later blood meal by a mosquito (Kuehn and Pradel 2010; Bousema and Drakeley 2011; Miller et al. 2002).

The erythrocytic cycle is responsible for many of the main symptomatic manifestations of the malaria disease in the secondary host, including fever, chills, sweats, and fatigue (White et al. 2014). In some cases the parasite can cause a neurological complication called cerebral malaria, which results in neurological and cognitive defects, and can lead to death (Idro et al. 2005; Hunt and Grau 2003; Macpherson et al. 1985). Children under the age of five years old are especially at risk due to their inability to fight off the parasitic load, which increases rapidly during infection (Murray et al. 2012; WHO 2014). In 2013, 198 million cases of malaria were estimated worldwide, with approximately 584,000 deaths attributed to the disease (WHO 2014). Of those deaths, 75 percent were children under the age of five, with 96 percent of those children residing in Africa (WHO 2014). Due to the high incidence of infection, morbidity and mortality in Africa, there is particular interest in curbing *P. falciparum* infection on that continent.

Initial large-scale attempts at curbing malaria focused mainly on combating the *Plasmodium* parasite using small molecule therapeutics. Chloroquine, an inhibitor of lysosome maturation, was utilized as a front-line defense against malaria in the first Global Malaria Eradication campaign in 1955 (Alessandro and Buttie 2001; Nájera et al. 2011). Shortly after mass administration, resistance to the drug was first seen in Southeast Asia and spread to other endemic regions (Alessandro and Buttie 2001; Payne 1987; Wellems and Plowe 2001; Bir et al. 2002). Sulfadoxine/pyrimethamine (SP), an antifolate two-drug system that was developed next, reduced the synthesis of folate, important for DNA synthesis (Olliaro 2001; Sibley et al. 2001; Hyde 2002). SP therapy soon suffered the same fate as chloroquine when resistance evolved and subsequently

spread (Roper et al. 2004). Recently, artemisinin and artemisinin-based combination therapies (ACTs) have been developed, although their mechanism of action is not fully understood (Whitty et al. 2008; Muheki et al. 2010). Some studies have implicated inhibition of the calcium pump *Pf*ATP6 as a likely mechanism of action (Eckstein-Ludwig et al. 2003; Krishna et al. 2014; Adhin et al. 2012; Arnou et al. 2011). As with previous administration of malaria drugs, ACTs are now facing an increased prevalence of drug-resistant parasites with mutations in the *kelch* propeller domain being associated with resistance in both Cambodian and African ACT-resistant parasites (Ariey et al. 2014).

The underlying mechanism that forms the basis of the drug-resistant phenotypes differs for each therapy. For chloroquine resistance, *P. falciparum* is able to expel the compound approximately 40-50 times faster than a wild type parasite due to mutations in the *Pf*CRT (*Plasmodium falciparum* chloroquine resistance transporter) locus (Johnson et al. 2004; Lakshmanan et al. 2005; Fidock et al. 2000; Bir et al. 2002). This increased efflux of compound essentially reduces the concentration of chloroquine within the parasite, thus limiting its effectiveness. SP resistance, on the other hand, is based on drug target mutations rather than drug transport (Happi et al. 2005; Duraisingh et al. 1998; Hyde 2002; Olliaro 2001). SP molecules target both dihydrofolate reductase (DHFR) and dihydropteroate synthase (DHPS), causing a reduction in folate synthesis. Mutations in either DHFR or DHPS can alter SP-binding sites so that SP does not effectively bind its their targets. *P. falciparum*'s ability to quickly and efficiently undermine attempts at elimination during interventions can be partially attributed to the plasticity of the

parasite's genome (Goldberg et al. 2012; Bopp et al. 2013; Ekland and Fidock 2007). This plasticity, or the ability of the genome to change either by mutation or genomic rearrangement, when coupled with the high rates of asexual reproduction in the human host allows *P. falciparum* to quickly gain resistance to drugs.

The *P. falciparum* genome was first sequenced in 2002 (Gardner et al. 2002). Consisting of 14 chromosomes with a total content of 23.3 megabases, with less than 20 percent of its base pairs being G or C (Gardner et al. 2002; Rathore et al. 2001; Bourgon et al. 2004). Of the approximate 5,400 genes in the genome, up to 61 encode *var* group genes (Kraemer and Smith 2006; Flick and Chen 2004). These genes are responsible for adherence of infected red blood cells to the vascular endothelium and for the antigenic evasion of the immune system by the parasite (Falk et al. 2009; Avril et al. 2012; Miller et al. 2002; Su et al. 1995; Smith et al. 1995). Most notably, *PfEMP1* is the best characterized *var* gene, as it has been found to be one of the key genes involved in the pathogenicity of the parasite (Pasternak and Dzikowski 2009; Mayer et al. 2012; Baruch et al. 1995). During the asexual parasitic cycle, only a single *var* gene is expressed (Roch et al. 2003; Chen et al. 1998; Scherf et al. 1998). Over time, as an immune response is mounted against the expressed *var* protein, parasites are selected for based on expression of a second *var* gene. These parasites then give rise to the dominant population during subsequent replication cycles. To further increase *var* gene diversity and evasion of the human immune system, clustering of chromosomal telomeres into "bouquets" during mitotic division promotes recombination among *var* group genes (Kraemer et al. 2007; Kraemer and Smith 2003; Freitas-junior et al. 2000). Evasion of the immune system

allows *P. falciparum* to replicate and increase parasitemia levels. Each subsequent replication cycle increases the chance of a parasite acquiring mutations that may counteract or render a drug treatment ineffective. In fact, it has been shown that many drug resistant strains of the parasite arise in Southeast Asia, although no correlative or causative reasoning has yet been associated with why this region may harbor parasites that evolve resistance at such a high rate (Pickard et al. 2003; Roper et al. 2004). As the rise of drug-resistant parasites continues, it is vital that other avenues of malaria control and local eradication be utilized. One such avenue that has shown particular effectiveness in decreasing malaria transmission is vector control, implemented by various lethal and non-lethal means.

B. Malaria Vectors: Background and Behavior

Mosquitoes of the *Anopheles* genus are the only vectors that transmit the causative pathogen of human malaria (Cohuet et al. 2010). *Anopheles* mosquitoes have four life-stages: embryonic, larval, pupal and adult (Clements 1999; Koutsos et al. 2007). After a female lays an egg raft in a body of water, the larvae hatch from the eggs and undergo three molts, increasing the size of the larvae after each molt. After three molts, the mosquito enters the pupal stage, during which the body plan is completely remodeled to that of an adult mosquito. After 48 hours in pupal form, an adult emerges and becomes sexually mature within 12-24 hours. After mating, females will take up a blood meal from a host and lay her eggs (Clements 1999). After this first blood meal, females become refractory to mating, but are able to lay eggs multiple times, taking a subsequent

blood meal before each brood of eggs is laid (Briegel and Horler 1993; Koella et al. 1998). Between the time of initial uptake of a blood meal and any subsequent blood meals, the malaria parasite reproduces sexually and generates progeny that later infect another human host. Without initial and subsequent bites, the life cycle of *P. falciparum* is broken and transmission will be disrupted.

Additional variables affect the rate at which malaria transmission occurs to secondary hosts, including number of human exposures to infected and non-infected mosquitoes, the population densities of humans and mosquitoes, mosquito survival rate, and the incubation times of the parasite in humans and mosquitos, to name a few (Smith et al. 2008; Depinay et al. 2004; Killeen et al. 2006, 2000). The rate of pathogen transmission through a mosquito vector has been described most adequately by Ronald Ross and George Macdonald, who helped develop a mathematical model that describes the likelihood of transmission of the disease (Ross 1910; Macdonald 1957). These models have shown that decreasing any variable that pertains to mosquito vector competence, longevity, or biting rate can decrease the malaria transmission rates (Smith and McKenzie 2004; Smith et al. 2008). These models were used during the Global Malaria Eradication Programme in the 1950's and 1960's, but only during initial attack phases and not during later maintenance phases (Smith and McKenzie 2004; Nájera et al. 2011). The effectiveness of these models during early malaria eradication efforts provided insight and confidence that vector control is a key factor in malaria eradication initiatives. More importantly, these results illustrated that targeted efforts against vectors will need to be predicated upon an in-depth understanding of vector behaviors.

Basic mosquito behavior is at the core of all mosquito-based interventions. In total, there are approximately 3,500 species, with *An. gambiae* (Sub-Saharan Africa), *An. arabiensis* (Eastern Africa) and *An. stephensi* (Indian peninsula) accounting for the majority of *Plasmodium* transmission around the world (Fig. 1.2) (Sinka et al. 2010, 2012; Hay et al. 2010). *Anopheles* mosquitos are present on six continents; only Antarctica lacks a malaria vector species. Estimates indicate that all *Anopheles* species originated from a most recent common ancestor approximately 100 MYA and diverged from *Drosophila melanogaster* approximately 250 MYA (Neafsey et al. 2014, 2013; Zdobnov et al. 2002). For comparison, modern apes diverged from their last common ancestor with mice approximately 90 MYA and from their last common ancestor with the platypus 200 MYA and contain similar rates of orthologous gene content (Necsulea et al. 2014; Warren et al 2008). The mosquitos that are of particular epidemiological importance, mainly those within the gambiae complex (*An. gambiae*, *An. arabiensis, An. quadriannulatus, An. merus*, and *An. melas*) diverged from each other only ~4 MYA (Fontaine et al. 2014). Due to the extended period of time encompassing the emergence of the genus *Anopheles* and their broad environmental distribution, each species possesses many unique behavioral characteristics.

One of the most important mosquito behaviors, as it pertains to malaria transmission, is host biting and blood feeding by females. There are two main sub-behaviors that are associated with blood feeding; host recognition/preference and time of feeding. In host recognition, mosquitos can be anthropophilic (human biting) or zoophilic (non-

human/animal biting), with some species exhibiting both biting behaviors (Pates 2002; Pates and Curtis 2005b). For example, *An. gambiae* is an anthropophilic mosquito, as it blood feeds on humans, only while *An. quadriannulatus* only feeds on bovids and is therefore considered zoophilic (Prior and Torr 2002; Dekker et al. 1998). Mosquitoes that are both zoophilic and anthropophilic increase the number of zoonotic transmission events, such as the spread of *Plasmodium knowlesi* between macaques and humans (Singh et al. 2004). The factors underlying the decision regarding which host a mosquito feeds upon are yet to be fully elucidated, but research has begun to decipher the components underlying this decision. Interestingly, recent work has shown that a mosquito's preference may be affected based on whom they feed upon first, i.e., feeding on a human first increases the chances a mosquito will feed again upon a human subject compared to non-human subjects (Vantaux et al. 2014). Mosquitoes are able to feed on multiple hosts, but often show preferences toward either human or non-human organisms (Takken and Verhulst 2013; Chaves et al. 2010). Other research implies that a mosquito's ability to recognize a potential blood meal source can be attributed directly to each potential host organism's specific odor (Wang et al. 2010).

Volatiles that emanate from an organism give each organism a specific "scent fingerprint" (Cork and Park 1996; Carey et al. 2010). These fingerprints allow mosquitoes to differentiate between favorable and unfavorable hosts to feed upon (Besansky et al. 2004; Dekker et al. 2002; Braks et al. 1999; Costantini et al. 2001; Foster and Takken 2007). Each mosquito possesses many odorant receptors (ORs) that are tuned to recognize specific cues (Pask et al. 2011; Fox et al. 2001; Pitts et al. 2011). For

example, many *An. gambiae's* ORs are tuned to recognize specific alcohols and heterocyclics, such as those secreted within human sweat and by skin microbiota (Rinker et al. 2012; Verhulst et al. 2011; Carey et al. 2010). Among these, ammonia has been identified as one of the most important components of sweat, for host identification by *An. gambiae* (Meijerink et al. 2001; Smallegange et al. 2005). Similar ORs recognize other cues, such as esters that are found in plants and are beneficial for identifying nectar sources (Lu et al. 2007; Gouagna et al. 2014). When activated, ORs send signals to the brain, which then interprets those signals to decide on a feeding behavior (Benton 2006). Not only does the unique scent signature of a potential feeding target heavily influence feeding behaviors of female *Anopheles* mosquitoes, but the time of day influences when a target will be fed upon, as well.

The biting behavior of mosquitoes is highly modulated by Zeitgeber time, or the time relative to light/dark (day/night) cycles (Jones et al. 1967; Githeko et al. 1996). *Anopheles* mosquitoes have been shown to be the most active during the dusk and dawn hours (Rowland 1989; Manouchehri et al. 1976; Kawada et al. 2005; Ribbands 1946; Paaijmans and Thomas 2011). This pattern also coincides with significantly increased biting rates during the night compared to the day (Manouchehri et al. 1976). Not all *Anopheles* species exhibit identical biting profiles, as illustrated by comparing *An. fluviatilis* and *An. stephensi* (Manouchehri et al. 1976). *An. stephensi* bites mainly during the dusk hours, and biting rates slowly decline until almost no bites are seen during late dawn. *An. fluviatilis* engages in increased biting at dawn followed by a second peak in biting rates during the mid-dawn hours. The reasons why biting and activity rates of

Anopheles mosquitos peak during the diurnal cycle are variable (Githeko et al. 1996). During peak daytime temperatures, especially in areas of high malaria endemicity like sub-Saharan Africa and India, where temperatures can exceed 100 degrees Fahrenheit, the rate of vector desiccation is very high due to the high surface-to-volume ratios of mosquitoes (Fouet et al. 2012; Vargas et al. 2014). Therefore, mosquitoes decrease activity during the daytime hours to keep their temperature low. Nighttime provides an environment that is much more conducive to feeding behaviors as many potential blood meals, such as humans, are inactive during this time (Pates and Curtis 2005a)

One nocturnal cue to mosquitoes is the change of light intensity during the dawn and dusk hours (Yohannes and Boelee 2012). In laboratory experiments, it has been shown that mosquito activity and biting levels can be altered using light-pulses during nighttime hours, suggesting that light levels function as a significant modulator of nighttime activity (Das and Dimopoulos 2008). Furthermore, *An. funestus, An. stephensi* and *Aedes aegypti* all exhibit reduced levels of activity during nights that lack moonlight compared to nights with a full moon (Ribbands 1946). These changes of behavior during Zeitgeber time often correlate with circadian dynamics of molecular mechanisms (Rund et al. 2011). Most importantly, circadian expression profiles of many visual transduction and biogenesis pathways coincide with changes in Zeitgeber time (Rund et al. 2011), implying the adaptive evolution of gene expression patterns to be responsive to the environments mosquitos are active within. Further evidence of the influence of Zeitgeber time on the evolution of mosquito behavior is the ability of mosquitoes to detect and visualize ultraviolet light (Spaethe and Briscoe 2004). While much is now known about

insect vision and the wavelength spectrum mosquitoes can detect, little is known about how these mechanisms correlate with mosquito behavior.

In Chapter II of this thesis, I attempt to elucidate the differences in photopreference between mosquitoes that possess different numbers of visual transduction genes, mainly long-wavelength sensing *opsin* genes. By comparing the behaviors of *Am. gambiae* and *An. stephensi*, vectors responsible for malaria transmission in sub-Saharan Africa and the Middle East, we can begin to decipher how the numbers of wavelength-sensing opsin genes affect how a mosquito behaves when exposed to different levels of illumination. An understanding of the underlying behavioral mechanisms, including those previously described (e.g., host identification and preference, and Zeitgeber time-dependent activity), has been paramount to our most successful methods of malaria control, as most of these methods are based on vector behaviors.

C. Vector Control Methods

Successful control of malaria in many areas of the world can be attributed to control of the *Anopheles* mosquito rather than the targeted control of the *Plasmodium* parasite. Notably, the United States of America eradicated the disease from its borders using environmental management techniques, and one of the first large scale applications of an insecticide directed toward *Anopheles* (Andrews et al. 1945; Williams Jr 1963). Malaria was not endemic to the United States originally, but the damming of many rivers and streams created habitats for *Anopheles quadrimaculatus* and *Anopheles freeborni* that

expanded and began to harbor the imported disease as mosquito populations grew. By draining reservoirs and limiting the environment for vector breeding, disease eradication became possible in the United States (Andrews et al. 1945; Williams Jr 1963). Dichloro-diphenyl-trichloroethane (DDT) was the first broadly applicable insecticide that proved effective in controlling vector populations (License et al. 2011; Turusov et al. 2002). Targeting the voltage-gated sodium channel encoded by the *knockdown resistance* (*kdr*) gene, DDT causes prolonged neuronal excitation in insects, ultimately proving lethal to the organism (Davies et al. 2007; Vijverberg et al. 1982). Because the side effects of broad applications of DDT harmed the eggs of many fowl, DDT was ultimately banned in the United States following many additional environmental studies (Fry 1995). Most simply, the success of the United States malaria eradication campaign was enabled by an understanding of vector behaviors and the creation of interventions that directly targeted those behaviors. In addition, this campaign provided insights into the potential harm that broad application of broad spectrum insecticides can cause and the importance of understanding insecticide mechanisms of action. Built on the success of this and similar campaigns, many of today's most successful malaria control methods combine broad-spectrum insecticides with behavior-specific targeting methods.

Currently, indoor-residual spraying (IRS) and long lasting insecticidal nets (LLIN) are the two most effective techniques for reducing vector populations (License et al. 2011; Enayati and Hemingway 2010; Kim et al. 2012). Advancements in the production and discovery of insecticides have produced two important classes of insecticides, pyrethroids and carbamates, to supplement organophosphates. Organophosphates, like DDT, and

pyrethroids act on voltage-gated ion channels, causing dysregulation of neuronal signaling that leads to the death of the mosquito (Coats 1990). The mechanism of action of carbamates, such as bendiocarb, relies upon binding to acetylcholinesterase, an enzyme responsible for degrading acetylcholine. This binding inhibits the enzyme, causing a lethal build-up of acetylcholine (Fukuto 1990). Unlike DDT in the 1940's, pyrethroids and carbamates have not been broadly applied to the environment, but rather have been strategically administered to specific areas that the mosquito will contact.

Indoor residual spraying consists of spraying an insecticide on the interior walls of a dwelling. After a blood meal, mosquitoes rest for a time to allow digestion and diuresis to occur (Lahondère and Lazzari 2012; Gillett 1983). After taking a blood meal indoors, mosquitoes often rest on the interior walls rather than leave the dwelling, reflecting a preference for a relatively safe environment (Sinka et al. 2010). Therefore, following the application of insecticides to interior walls, blood-fed females will come into contact with the insecticides and die. Decreasing the number of females decreases the total mosquito population (due to a direct loss of females and the subsequent decrease in progeny) and ultimately decreases the incidence of malaria transmission, as previously described by the Ross-MacDonald equation (see above). Similarly, long lasting insecticidal nets take advantage of the propensity of females to blood feed. By surrounding the source of a blood meal with a bed net, mosquitoes will come into contact with such nets while attempting to gain access to their blood meal, thus increasing the probability of the mosquito contacting the insecticide (Nevillts et al. 1996). By coating bed nets with an insecticide, mosquitoes that attempt to take a blood meal from a resting human are first

physically prevented from taking a blood meal and consequently exposed to a lethal dose of a given insecticide. This intervention has been especially effective, decreasing the mortality rate of children under five in sub-Saharan Africa by nearly 60% and premature birthrates by similar magnitudes (WHO 2014). Although these methods are highly effective in reducing the transmission of *Plasmodium* parasites, they have not yet been able to fully eliminate the disease (Eisele et al. 2012; Griffin et al. 2010; Bhattarai et al. 2007). In addition, as *P. falciparum* has acquired resistance to drugs at increasingly high rates, mosquitoes have evolved analogous resistance to insecticides currently in use.

Mutations in the *kdr* locus, the target of organochlorines, pyrethroids and carbamates have decreased the effectiveness of both IRS and LLIN interventions (Ranson et al. 2000; Soderlund and Knipple 2003). The L1014F and L1014S mutations in *kdr* have produced a range of resistance phenotypes toward organochlorines, pyrethroids and carbamates, while a G119S mutation in *ace-1*, the target of bendiocarb, has begun to render carbamates ineffective (Berthomieu et al. 2004; Nwane et al. 2011; Ndiath et al. 2012; Ibrahim et al. 2014). Additionally, further mutations have begun to arise in glutathione-S-transferases, esterases and cytochrome P450s that enable the metabolism of insecticides within mosquitoes into harmless molecules (Djouaka et al. 2008; Ranson et al. 2002; Daborn et al. 2012). Each of these mutations is an example of human-driven evolutionary selection that can result from the incomplete or mono-formulated administration of specific insecticides (Barbosa and Hastings 2012). Low-level or geographically patchy distribution of insecticides within a given region imposes a mild evolutionary selective pressure on the regional mosquito population. This mild

selective pressure leads to selection for evolutionarily advantageous alleles that are allowed to propagate throughout the population, as the population size is not reduced quickly enough with such limited uses of insecticides.

In order to stop such genetic sweeps from occurring, dual- or combination-insecticide treatments have begun to be adopted as the resultant increased evolutionary selective pressure is difficult to overcome and will reduce the vector population before advantageous alleles can arise and be selected for (Okumu and Moore 2011; Kleinschmidt et al. 2009; Vitti et al. 2013; Labbe et al. 2005). Other application methods, such as mosaic insecticide treatments, use multiple single insecticide treatments in a mosaic fashion across a given region, never deploying two insecticides at a single location (Read et al. 2009; Hougard et al. 2007). The development of resistance has also motivated the development of other means of vector control, often targeting specific vector behaviors. New methods of vector control have been developed recently along these lines include larvicides, push-pull methods, transgenic mosquitoes, and symbiotic infection with *Wolbachia* – a few of the newly emerging vector-targeted control methods that have shown potential to curb malaria burden (Takken 2010; Burt 2014; Enayati and Hemingway 2010; Fu et al. 2010).

What is becoming increasingly apparent as we attempt to combat malaria using vector control techniques is that we are in dire need of additional new, targeted control measures against specific species of mosquitos. As a result, vector biologists have begun to conceive of the development of new methods that would be devised based on data from

genomic sequences or expressed gene sequences defined using next generation sequencing of mosquito species of interest, because this rich source of genomic and biological information can offer insights into previously unknown facets of vector biology and evolution.

D. Next Generation Sequencing of Malaria Vectors

As techniques to generate next generation sequencing (NGS) data and the tools to analyze it have developed further, it has become possible to define in greater depth mechanisms underlying mosquito behavior and evolution. The first vector to be sequenced, in 2002, was *An. gambiae* (Holt et al. 2002), and this began a trend of sequencing vectors to better understand infectious disease transmission (Nene et al. 2007; Marinotti et al. 2013; Severson and Behura 2012). Researchers quickly noted that approximately half of protein coding genes within the genomes of *An. gambiae* and *D. melanogaster*, which diverged approximately 250-300 MYA, are orthologous, while the evolutionary rate is increased, compared to vertebrates (Zdobnov et al. 2002). The increased evolutionary rates in invertebrates compared to vertebrates indicate that, although morphologically similar, the underlying genomic structures among the *Anopheles* species may be vastly different in terms of gene content, and those genomic differences may underlie behavioral differences among those species.

Recent studies have aimed at identifying the genomic differences between *Anopheles* species, to understand the evolution of the vectors and better define the molecular

underpinnings of behavioral mechanisms. The *Anopheles* 16 Genomes Project (Neafsey et al. 2013, 2014) was formed in an attempt to leverage sequencing of vector genomes, in addition to *An. gambiae*, to pursue these aims. Overall, this consortium has shown that the assembled members of the *Anopheles* genus possess highly divergent genomes, that exhibit substantial chromosomal differences and very high evolutionary rates among gene families, such as the male accessory gland genes (Neafsey et al. 2014; Fontaine et al. 2014). While comparative evolutionary genomics among *Anopheles* species can help identify differences between species, it is often other sequencing-based techniques, such as RNAseq, genome-wide association studies, and microarray analyses that have allowed researchers to define genetic mechanisms of action within different vector species.

Early studies in *Anopheles gambiae* used microarrays to assess the transcriptional profiles of various life-stages (Koutsos et al. 2007; Harker et al. 2012). Cuticular, detoxification, protease and peptidase classes of genes were shown to be up-regulated during larval stages, while genes involved in immunity, odorant recognition and visual transduction pathways were shown to be up-regulated during adult stages (Harker et al. 2012; Koutsos et al. 2007). These studies became the basis of more specialized studies where, for example, the repertoire of odorant receptors were identified and characterized in *An. gambiae* and subsequently across the 16 *Anopheles* species sequenced (Pitts et al. 2011; Rinker et al. 2013). Additionally, microarrays have also facilitated the discovery of genes involved in insecticide resistance in Africa (Edi et al. 2014; Mitchell et al. 2014). The up-regulation of cytochrome P450s in resistant mosquitoes was identified as one of the mechanisms responsible for multi-drug detoxification (Edi et al. 2014). One major

shortcoming of microarray analysis is its dependence on pre-existing knowledge about the genomic sequence and gene models for the design of probes (Šášik et al. 2004; Hoheisel 2006; Jaluria et al. 2007). In order to circumvent this limitation, RNAseq has become the tool of choice for analyzing gene expression, as previous knowledge of genomic sequence and gene models is not required if RNA sequencing has been conducted with sufficient read depth (Crawford et al. 2010; Trapnell et al. 2010).

RNAseq has only recently been utilized to study vector biology, as the technology has been available for less than a decade (Mortazavi et al. 2008). By sequencing cDNA directly rather than annealing cDNA to known sequence probes, researchers are able to identify full-length transcripts, splice junctions, 5' and 3' untranslated regions (UTRs) and sequence variants (Xia et al. 2014; Lu and Bushel 2013). However, few studies in vector mosquitoes have truly leveraged the advances and advantages offered by RNAseq such as *de novo* transcriptome assembly (Crawford et al. 2010; Neafsey et al. 2014).

The transcriptome of *An. funestus* was first defined using RNAseq rather than traditional pipelines, such as MAKER (Crawford et al. 2010) which is a program used to determine protein coding potential and genomic structures. Using RNAseq to determine a transcriptome circumvents many problems that MAKER pipeline poses (Costa et al. 2010; Wang et al. 2009). For instance, many transcribed regions such as UTRs and non-coding RNAs (ncRNA) are often missed by MAKER (Ilott and Ponting 2013; Lu and Bushel 2013). These shortcomings of MAKER identification methods have recently been shown to cause incorrect numbers of gene family members in nearly 40% of gene

families (Denton et al. 2014). It is paramount to identify gene structures and gene families correctly, in order to maximize identification of potential insecticide targets. In Appendix I of this thesis, I provide a detailed analysis of how I have used RNAseq to annotate correctly previously misannotated G protein-coupled receptor gene family members in *An. gambiae*. Overall, the rapid advancement of NGS in vector insects will facilitate a deeper understanding of mechanisms underlying vector behaviors and the identification of genes and gene classes that were previously unknown to researchers.

E. Identification and Evolution of lncRNA and Epigenetic Gene Classes

The repeated rise of insecticide resistance creates the need to identify new approaches to combat malaria vectors. Utilizing NGS technologies, we can identify new classes of genes and assess the evolutionary potential of each gene class as a source for insecticidable targets. Two genes classes that offer potentially fruitful targets for vector control are long non-coding RNA (lncRNA) and epigenetic modifier genes.

The lncRNAs constitute a set of genes that are classified based upon their length (> 200 basepairs) and coding potential (little to no protein coding potential) (Ponting et al. 2009; Kung et al. 2013; Necsulea and Kaessmann 2014). The functionality of lncRNAs lies within the secondary and tertiary structures of the non-coding RNA (ncRNA) molecule (Novikova et al. 2012). The structures formed interact with proteins, modulating the proteins' functions, most often by forming RNA-protein complexes (Chu et al. 2011; Bellucci et al. 2011). Mechanisms that have been attributed to lncRNAs are wide-

ranging. *Cis*- and *trans*-acting transcriptional regulation, chromatin modification, RNA-splicing and genomic imprinting have all been implicated as being modulated by lncRNAs (Elango et al. 2009; Kiefer 2007; Weng et al. 2012).

The first genome-wide scan for lncRNAs was performed in 2009, while investigating chromatin states in mice (Guttman et al. 2009). Since this initial genome-wide scan, lncRNAs have been discovered in all organisms in which they have been sought, including multiple mammalian species, fruit flies and zebrafish (Necsulea et al. 2014; Nam and Bartel 2012; Ulitsky et al. 2011; Young et al. 2012). Prior to 2009, relatively few lncRNAs had been discovered and the mechanisms of those identified, such as *XIST, AIR*, and *HOTAIR*, operated on a pan-genomic scale (Bhan et al. 2013; Seidl et al. 2006; Penny et al. 1996). *XIST* is implicated in X-chromosome inactivation, while *AIR* and *HOTAIR* function in imprinting and *trans*-acting gene regulation, respectively (Seidl et al. 2006; Bhan et al. 2013; Penny et al. 1996). Additionally, in *Drosophila*, lncRNAs have begun to be linked to mechanisms of sleep and locomotive behaviors (Li et al. 2012; Soshnev et al. 2011). One of the major obstacles to the large-scale identification of lncRNAs across the genome has been a lack of suitable technology.

With the introduction of RNAseq and chromatin immunoprecipitation followed by massive parallel sequencing (ChIPseq), previously unknown transcribed portions of the genome have begun to be identified (Guttman et al. 2009). Within *Drosophila*, at least 1,119 long, intergenic, non-coding RNAs (lincRNA) have been identified, while more than 14,000 lncRNAs have been identified in humans (Young et al. 2012; Derrien et al.

2012; Harrow et al. 2012). These lncRNAs exhibit increased rates of evolution compared to protein coding genes in mammals (Kutter et al. 2012; Marques and Ponting 2014). As a result, orthology of lncRNA gene families decreases rapidly across mammalian lineages (Necsulea et al. 2014). Among vector species, lncRNAs have only been identified in the midgut of *An. gambiae*, in a study that suggested that over 10,000 lncRNA transcripts might exist in this mosquito (Padrón et al. 2014).

The ability to identify lncRNAs within the *Anopheles* lineage may spur the discovery of mechanisms underlying a variety of behaviors and a new set of insecticidable targets. In Chapter III of this thesis, I explore the lncRNA repertoire I have identified in *An. gambiae* using deep RNAseq. Further, I present findings that describe the conservation of lncRNA secondary structures across the *Anopheles* genus and argue that the number of homologous genomic regions and conserved secondary structures decay at the same rate across the genus.

A second class of genes that offers important information in understanding vector species is epigenetic modifier genes. Epigenetic modifier genes are crucial to modulation of genomic regulation during development, inheritance of genetic information, and response to environmental factors (Cantone and Fisher 2013; Kiefer 2007; Portela and Esteller 2010; Goldberg et al. 2007; Meissner 2010; Greer et al. 2011). The effects of epigenetic modifiers on development have been characterized extensively in *D. melanogaster*, including many investigations of Polycomb- and Trithorax-Group proteins (Schwartz and Pirrotta 2007; Schuettengruber et al. 2007; Bracken and Helin 2009). The discovery of

modulation of transcriptional regulation via the addition of acetyl or methyl groups to histone tails and subsequent chromatin-state shifts have shed light on the interplay of environmental factors and the genome (Feil and Fraga 2011). The ability of paralogous genes to acquire differing expression patterns through evolution has been attributed to epigenetic mechanisms and may contribute to the diversification seen within related gene families (Klironomos et al. 2013; Sui et al. 2014; Furrow and Feldman 2014; Keller and Yi 2014; Park and Lehner 2014).

Epigenetics in non-model insects has more recently begun to focus, in part, on behavioral caste systems, and the means, if any, by which epigenetic mechanisms contribute to these strict social hierarchies (Weiner and Toth 2012). Although *Anopheles* species, to the best of our knowledge, possess no such caste system, it seems likely that by understanding the epigenetic make-up of these vectors, we will gain insight into complicated behaviors such as mating, blood feeding and host recognition. With these possibilities in mind, in Chapter IV of this thesis, I explore the orthology between *An. gambiae* epigenetic modifier gene sets and the well-annotated *D. melanogaster* epigenetic modifier gene set. The expression profiles, both temporal and tissue specific, are compared between *D. melanogaster* and *An. gambiae* to determine whether the genes play similar roles in both species. Finally, the conservation of the epigenetic modifier genes is analyzed across the entire *Anopheles* genus to determine whether the gene set is under purifying or diversifying selection.

FIGURES AND LEGENDS

Figure 1.1: *Plasmodium falciparum* Life Cycle

Figure depicts the sexual reproductive life cycle (sporogonic cycle) in the mosquito and the asexual reproductive life cycle (exo-erythrocyctic and erythrocytic cycles) of *Plasmodium falciparum* in humans. Figure was taken from the CDC website (http://www.cdc.gov/malaria/about/biology/)

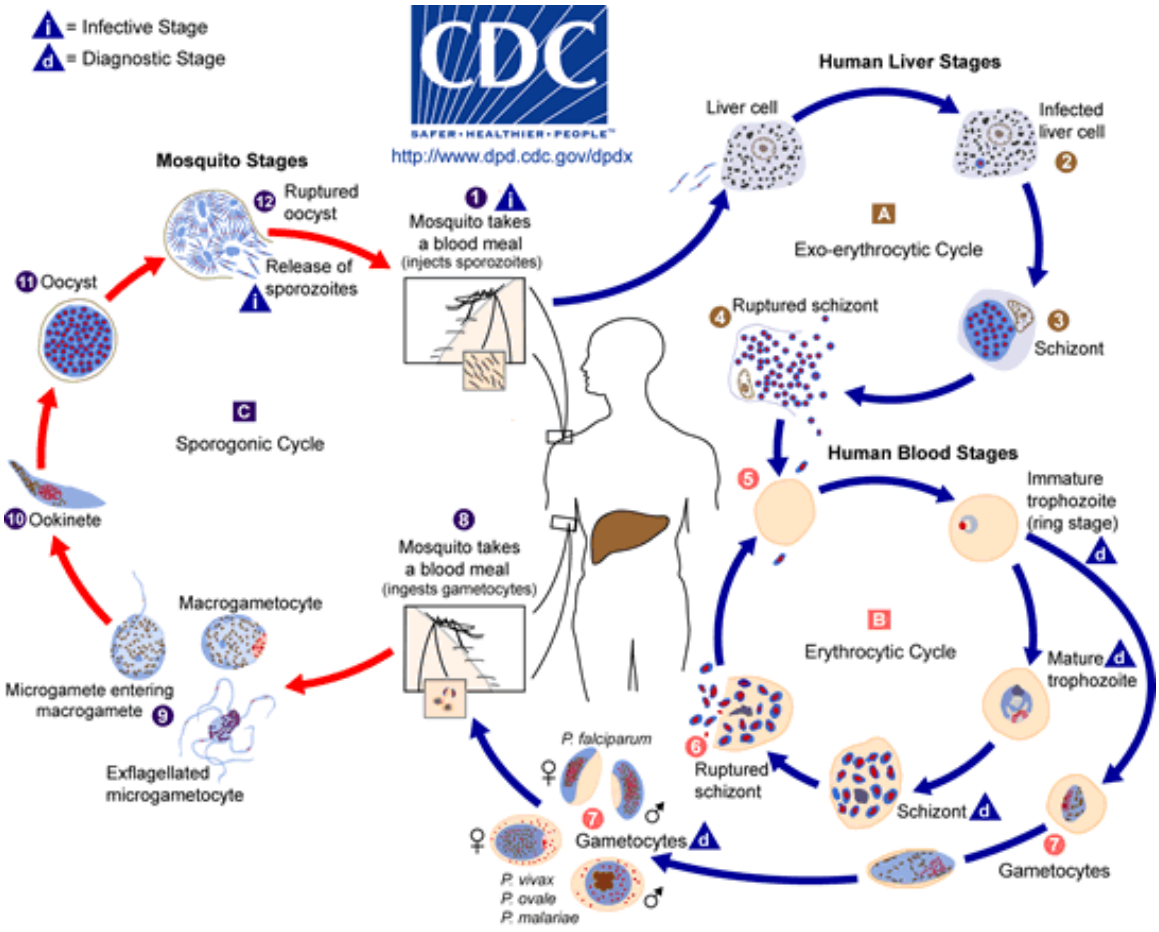


Figure 1.1: *Plasmodium falciparum* Life Cycle

Figure 1.2: Distribution of *Anopheles* Mosquitoes Across the World

Global map showing the distribution of dominant malaria vector species, as depicted in Sinka *et al.* (2012).

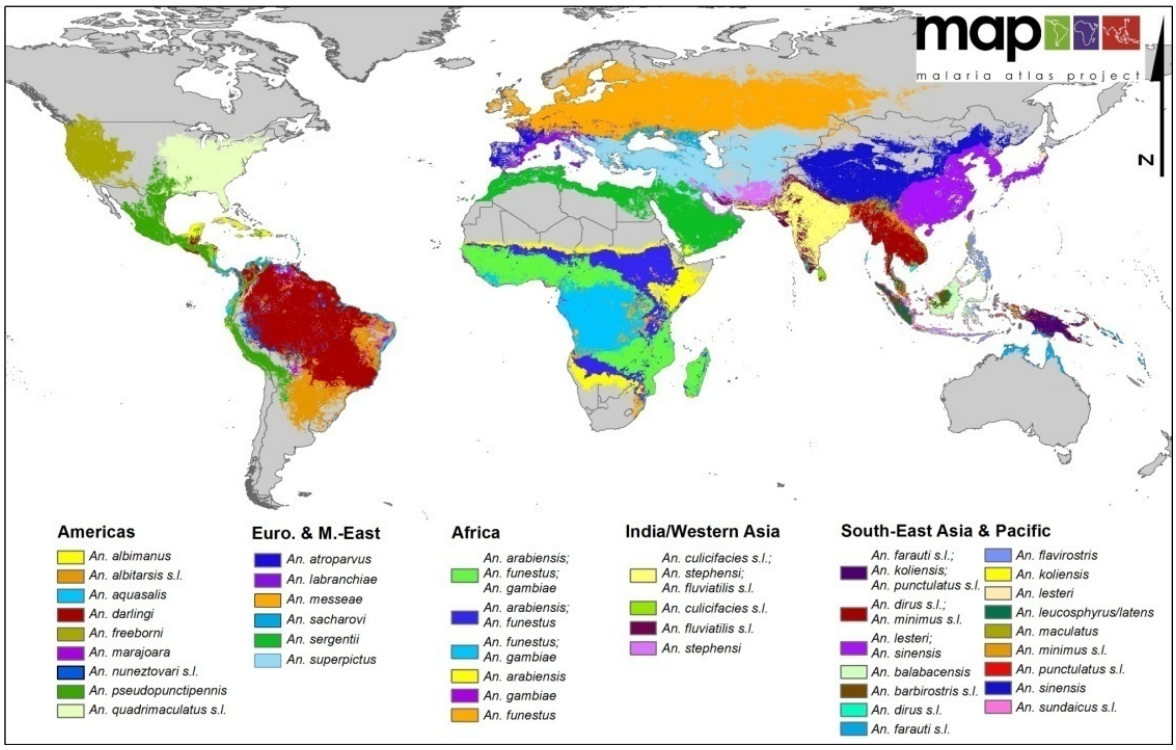


Figure 1.2: Distribution of *Anopheles* Mosquitoes Across the World

Chapter II:

Crepuscular behavioral variation and profiling of opsin genes in *Anopheles gambiae* and *Anopheles stephensi*

ABSTRACT

We understand little about photopreference and the molecular mechanisms governing vision-dependent behavior in vector mosquitoes. Investigations of the influence of photopreference on adult mosquito behaviors such as endophagy/exophagy and endophily/exophily will enhance our ability to develop and deploy vector-targeted interventions and monitoring techniques. Our laboratory-based analyses have revealed that crepuscular period photopreference differs between *An. gambiae* and *An. stephensi*. We employed qRT-PCR to assess crepuscular transcriptional expression patterns of long wavelength-, short wavelength-, and ultraviolet wavelength-sensing opsins (i.e., rhodopsin-class GPCRs) in *An. gambiae* and in *An. stephensi.* Transcript levels do not exhibit consistent differences between species across diurnal cycles, indicating that differences in transcript abundances within this gene set are not correlated with these behavioral differences. Using developmentally staged and gender-specific RNAseq data sets in *An. gambiae*, we show that long wavelength-sensing opsins are expressed in two different patterns (one set expressed during larval stages, and one set expressed during adult stages), while short wavelength- and ultraviolet wavelength-sensing opsins exhibit increased expression during adult stages. Genomic organization of *An. gambiae* opsins suggests paralogous gene expansion of long wavelength-sensing opsins in comparison to *An. stephensi*. We speculate that this difference in gene number may contribute to variation between these species in photopreference behavior (e.g., visual sensitivity).

INTRODUCTION

Among deployable malaria control and prevention techniques, those targeting the primary host of *Plasmodium* – the vector mosquito – continue to constitute our most effective methods of intervention. The use of long-lasting insecticide-treated bed nets (Mittal *et al.* 2012) and indoor residual spraying (Kim *et al.* 2012), along with environmental management (Imbahale *et al.* 2012), have led to significant reductions in malaria-related morbidity and mortality in a number of disease-endemic countries (Fullman *et al.* 2013). However, we must be attentive to impacts on vector-targeted interventions of insecticide resistance (Weill et al. 2000; Reimer et al. 2008). In addition, the inexorable genesis of resistance and extended clearance times of malaria parasites following treatment with drugs such as chloroquine, mefloquine and most recently artemisinin, continue to compromise the utility of anti-malarial drug-based interventions (Dondorp et al. 2009; Bray et al. 1998; Djimde et al. 2001; Alonso and Tanner 2013).

Creation of next-generation vector-targeted interventions that focus on aspects of the mosquito life cycle that are not targeted by present interventions (indoor residual spraying or IRS, and insecticide-treated bednets or ITNs) will depend, in part, on development of a broader understanding of the behaviors of vector mosquitoes. Many mosquito behaviors – including resting, foraging and feeding behaviors, olfactory responses, flight activity and flight patterns – have been studied to identify prospective points of attack for next-generation vector-targeted interventions. Toward that end, we have begun to investigate illumination preferences of Anopheline mosquitoes.

Light traps are often used to monitor vector mosquito population compositions and densities (Overgaard et al. 2012; Tchouassi et al. 2012), and we anticipate that light

sources could be incorporated into push-pull strategies (Takken 2010) for deflecting vector mosquitos from human dwellings. Still, light traps used to monitor biting rates have been known to provide conflicting results that can vary based on study methods, species observed, and geographical location (Mala et al. 2011; Mathenge et al. 2005). By understanding mosquito light preference in greater depth, we will expand our grasp of vector bionomics, and contribute to improvements in the use of light-based tools for monitoring vector populations and for the development of next-generation interventions that will contribute to decreasing the malaria burden in disease-endemic regions.

Anopheles funestus, An. stephensi and *Aedes aegypti* exhibit increased flight activity in dim-light settings compared to a setting of complete darkness, and the illumination intensities that stimulate flight vary among these species (Rowland 1989; Manouchehri *et al.* 1976; Kawada *et al.* 2005; Ribbands 1946). For instance, *An. stephensi* biting rates increase during nighttime hours, and house-entering behavior of *An. funestus* increases on moonlit nights (Ribbands 1946; Manouchehri et al. 1976; Rowland 1989). Mosquito house-entering and resting behaviors have been shown to be dependent on temperature microclimates, inside and outside of dwellings (Paaijmans and Thomas 2011). These resting preferences and illumination-influenced behaviors can impact malaria transmission by vector mosquitos and determine how accurately mosquito-monitoring techniques will reflect species prevalence. Integrative consideration of such bionomic factors has begun to influence the development of multiple interventions, including exposure to surface-applied malathion and fungal biocontrol agents, based on more extensive understanding of mosquito resting and flight behaviors (Perich et al. 2000; Mnyone et al. 2012).

While many innate behaviors have been well-characterized in many vector species, illumination preference is a mosquito behavior that has proven difficult to assay in lab and field settings. We have little molecular insight into possible mechanisms underlying illumination-dependent behavioral differences. For instance, multiple studies have reported conflicting results regarding the attractiveness to mosquitoes of blue/green wavelengths of light. Field studies of *Culex* spp. have reported attraction toward blue light, albeit the least intense of the visible wavelengths with regard to brightness in the study (Ali *et al.* 1989). Other field studies have concluded that a majority of mosquito species (among the genera *Anopheles, Aedes, Coquillettidia, Mansonia, Psorophora* and *Uranotaenia*) prefers green wavelengths, although *Culex nigripalpus* females are reported to prefer blue wavelengths (Bentley *et al.* 2009).

On the other hand, laboratory-based experiments have shown that *Culex nigripalpus* feed for longer periods of time under illumination of 500 and 600 nm, within the green range of the visible spectrum (Burkett *et al.* 2012). Other species such as *Mansonia perturbans* are said to prefer wavelengths of 400-600 nm (blue-green range), while *An. stephensi* is said to be attracted to near-UV and incandescent light rather than to specific wavelengths (Browne and Bennett 1981; Wilton and Fay 1972). At present, we do not understand whether light preference differences among species, or potentially within species, depend on intrinsic genetic and molecular mechanisms, or on features of life history that engender habituation and learned preferences for specific wavelengths.

Within the order Diptera, molecular mechanisms underlying phototransduction and circadian rhythm have been investigated most extensively in *Drosophila melanogaster*, given the genetic and molecular tools available in this model organism

(Montell 2012). We speculate that circadian variation in the expression of mosquito phototransduction genes may underlie diurnally variable mosquito behaviors. In the *Drosophila* head, over 150 genes associated with a variety of biological processes exhibit circadian oscillation in expression (Claridge-Chang *et al.* 2001). Hymenoptera, such as *Apis mellifera*, exhibit circadian fluctuations in expression of a green-sensitive *opsin* gene and an *arrestin* gene, each of which encodes phototransduction components, and their circadian rhythms may be controlled by a mechanism other than that mediated by Cryptochrome-2 (Sasagawa et al. 2003; Yuan et al. 2007).

Given the presence of 11 annotated *opsin* genes in the *An. gambiae* genome, *An. gambiae* has the largest number of *opsin* genes of any of the insects for which genome assemblies exist at present (Holt et al. 2002; Hill et al. 2002). This expanded *opsin* gene set has arisen, in part, due to an early duplication of long wavelength-sensitive *opsin* genes to create a set comprising six long wavelength–sensitive (λmax >500 nm) genes (*GPROP1, GROP3-7*) – in combination with one UV wavelength-sensitive (λmax <400 nm) *opsin* gene (*GPROP8*), one short wavelength-sensitive (λmax 400-500 nm) *opsin* gene (*GPROP9),* one functionally undefined *opsin* gene (*GPROP10*) and two *pteropsin* genes (*GPROP11, GRPOP12*) (Spaethe and Briscoe 2004). To date, none of these *An. gambiae opsin* genes has been shown to exhibit statistically significant circadian variation in expression, although a number do vary in level over the 24-hour circadian cycle (Rund *et al.* 2011). Behavioral analyses of *An. gambiae* have shown that manipulation of light can influence the timing of blood feeding behavior (Das and Dimopoulos 2008). Finally, it has been proposed that variation between *An. gambiae* and *Ae. aegypti* in the localization of *opsin2* and *opsin8* expression within the compound eye

may underlie species-specific behavioral patterns (e.g., photopreference in low light settings) that differ between these two vector mosquito species (Hu *et al.* 2009).

In this study we have developed a simple, laboratory-based assay to assess photopreference of *An. gambiae* and *An. stephensi*. We have employed these photopreference assays to determine that *An. gambiae* and *An. stephensi* exhibit different photopreferences, depending on the time of day and the illumination zone into which they are introduced. Subsequent qRT-PCR analysis fails to reveal significant diurnal differences in *opsin* gene expression, when comparing the two species. RNAseq analysis of *An. gambiae* opsins during four life stages indicates that one-half of the long wavelength-sensing opsins are expressed predominantly during larval stages and the other half during adult life-stages, while ultraviolet wavelength- and short wavelength-sensing opsins are expressed predominantly during adult stages. Further analysis of the organization of the long wavelength-sensitive *opsin* genes in the two species reveals that *An. gambiae* possess two more long wavelength-sensing opsins than *An. stephensi*, and we speculate that this difference in gene number may contribute to the differences in photopreference that we observe in the two species.

RESULTS AND DISCUSSION

Determination of Photopreferences in *An. gambiae* and *An. stephensi*

First, we measured photopreference characteristics of *An. gambiae* and *An. stephensi* to determine whether there are distinctions between the two species. We developed an assay that assesses the photopreference of *An. gambiae* using a binary choice arena (0 Lux vs. 400 Lux, Fig. 2.1, Table 2.1, Additional File 2.1). Introduction of mosquitos into the illuminated end of the apparatus during either dawn or dusk crepuscular periods reveals that females exhibit a significant preference for darkness, while males exhibit no reference between illumination and darkness (Fig. 2.1A,E). Binary choice assays in which *An. gambiae* was introduced into the darkened end of the apparatus reveal that males and females exhibit a significant preference for resting in darkness (Fig. 2.1C,G).

Analogous experiments with *An. stephensi* reveal that females prefer the illuminated portion of the apparatus when added to the illuminated end of the apparatus at dawn, while males prefer darkness (Fig. 2.1B). When introduced into the illuminated end of the apparatus, females exhibit a preference for illumination at dusk, while males no longer display any illumination preference (Fig. 2.1F). When added to the darkened portion of the apparatus at dawn, *An. stephensi* females lack any discernible photopreference, while males display a preference for darkness (Fig. 2.1D). When introduced into darkened end of the apparatus at dusk, *An. stephensi* males exhibit no preference, while females exhibit a preference for the illuminated portion of the apparatus (Fig. 2.1H).

The differences we observed between *An. gambiae* and *An. stephensi* photopreferences are consistent with differences observed in past studies of each species in other physical settings (Rowland 1989; Jones et al. 1967). Female *An. gambiae* generally exhibit a significant preference for a darkened photic zone, which can be attributed to an active avoidance of increased illumination. The active avoidance of illumination by *An. gambiae* females, when they are introduced to the 400 Lux end of the arena (Fig. 2.1A,E), indicates an avoidance of the light rather than a simple, consistent preference toward the end of the apparatus into which the mosquitos are introduced. Given that previous studies of *An. gambiae* indicate that peak flight activity occurs at the dawn and dusk hours, the possibility that *An. gambiae* are not actively moving within our apparatus is unlikely (Jones *et al.* 1967).

Interestingly, *An. stephensi* photopreference differs greatly from that of *An. gambiae.* Female *An. stephensi* prefer the 400 Lux region of the apparatus in all conditions, except when introduced into 0 Lux at dawn, when no significant preference was observed. This suggests a requirement for increased illumination to perform visual-based behaviors, such as identifying a feeding source, an oviposition site, or a mating swarm, or for achieving increased visual acuity. Male *An. stephensi* exhibit a preference for darkness or no preference, for all patterns of introduction, similar to findings for *An. gambiae* males. This suggests that light preference may be less important for *Anopheline* males in the processes of finding mates and food sources. In order to further validate the distinctions in photopreferences we observe between the two species in a binary choice assay, we subsequently conducted trinary choice assays.

Assessment of *An. gambiae* photopreference in a trinary choice assay (0 Lux vs. 100 Lux vs. 400 Lux, Fig. 2.2, Table 2.2), which allows for greater delineation of photopreference, illustrates that females and males prefer 100 Lux illumination during dawn and dusk crepuscular periods, when introduced to the 400 Lux end of the apparatus (Fig. 2.2A,E). When the assay was repeated with the introduction of mosquitos into the 0 Lux end of the apparatus, both sexes of *An. gambiae* prefer to remain in the darkened end of the apparatus during both crepuscular periods (Fig. 2.2C,G). *Anopheles stephensi* display tendencies to rest in 400 and 100 Lux regions of the apparatus, instead of the non-illuminated region, when introduced to the 400 Lux-illuminated region of the apparatus during dawn or dusk (Fig. 2.2B,F). Following introduction into the darkened end of the apparatus during dawn, *An. stephensi* males and females remain in the darkened region (Fig. 2.2D). Females exhibit no preference following introduction into the darkened end during dusk, and males exhibit significant preference toward the 100 Lux-illuminated region when introduced in the same manner (Fig. 2.2H).

With the availability of a photic zone with intermediate illumination in which to rest, both *An. gambiae* and *An. stephensi* photopreferences are altered compared to those measured in the binary photo assay format. Female and male *An. gambiae* exhibit strong preferences for darkness when introduced to the 0 Lux end of the apparatus, as in the binary photo assay. However, both sexes prefer to rest in the intermediate (100 Lux) illumination zone when introduced to the 400 Lux zone (Fig 2. A,E). These results indicate *An. gambiae* males and females still actively avoid the most intensely illuminated region of the apparatus, but do not necessarily prefer complete darkness. Rather, the avoidance of 400 Lux illumination, as seen in the binary assays, can be

achieved by resting in the 100 Lux region rather than the 0 Lux region of the arena. The differing *An. stephensi* trinary preference data indicate a strong preference for an illuminated area when introduced to the 400 Lux end of the arena (Fig. 2.2B,F), consistent with the hypothesis that *An. stephensi* mosquitos require more intense light in order to experience visual perception comparable to that of *An. gambiae.* These data are also consistent with past findings that *An. stephensi* exhibits increased flight activity in a dim-light setting compared to complete darkness (Rowland 1989).

The photopreference differences that we define in binary and trinary assays indicate that our simple photopreference arena – the first of its kind for vector mosquitos – is adequate for assessing differences in photopreferences between species, in a laboratory setting. The simple fabrication, low monetary cost and ease of transportation and setup of the assay arena imply that the assay could be performed with field-captured mosquitoes in a field setting. This strategy would reduce the need to create stable laboratory colonies of field-caught mosquitoes for photopreference behavioral assays and may enable more accurate analysis of a given species' photopreference in the field. Photopreference is of interest as it may inform how insecticides are applied in the field, in addition to expanding our understanding of vector photobiology. Better knowledge of mosquito photopreference may enable the application of insecticides to more specific areas of interest in the home and in the field, in conjunction with control efforts, rather than the use of broad-pattern application that covers many areas without biological relevance to the vector-targeted control. Current insecticide application methods, such as indoor residual spraying, often involve treating the entirety of a dwelling and leaving a residual coating of insecticide for months after treatment. A given vector mosquito

population might experience minimal contact with many of these treated surfaces, depending on its resting patterns within dwellings. By understanding these resting patterns in greater depth, the amount of insecticide needed for spraying may be reduced and better allocated to increase vector contact with insecticides and thereby increase the effectiveness of residual insecticide treatment methods.

Diurnal Variation of *Opsin* Gene Expression

Previous studies have shown that larval swimming behavior in the ascidian *Ciona intestinalis* can be altered by knocking down *Ci-Opsin1*, which results in reduced photoresponsiveness (Inada *et al.* 2003). Given these findings, we chose to determine whether diurnal transcriptional expression patterns of selected *opsin* gene superfamily members in *An. gambiae* and *An. stephensi* are correlated with distinct diurnal photopreferences we observe in these species. The *An. gambiae* haploid genome contains 11 annotated *opsin* genes (Holt et al. 2002; Hill et al. 2002). Eight of the 11 genes have attributable functions, and are defined as long wavelength-sensing, short wavelength-sensing and ultraviolet wavelength-sensing *opsin* genes. Our Reciprocal Best Blast analysis and manual annotation of the *An. stephensi* genome (VectorBase VB-2013-12) using *An. gambiae opsin* genes as query sequences led to the identification of four long wavelength-sensing *opsin* genes, one short wavelength-sensing *opsin* gene, and a single ultraviolet wavelength-sensing *opsin* gene within the *An. stephensi* genome. The organization of a subset of *An. gambiae opsin* genes and homologous genes in *An. stephensi* is depicted in Figure 2.3. On chromosome 2R, *An. gambiae* possesses four long wavelength-sensing *opsin* genes within a gene cluster (*GPROP3, GPROP4,*

GPROP5, GPROP6; Fig. 2.3). *An. stephensi* contains a similar cluster that includes only three long wavelength-sensing genes. The difference between these clusters in the two genomes is an apparent *opsin* gene duplication and inversion of *GPROP4* in *An. gambiae*. In other organisms, mainly primates, increased range of wavelength sensing and trichomatic color vision have been correlated with evolutionary duplications of long wavelength-sensing and medium wavelength-sensing *opsin* genes (Dulai *et al.* 1999). Therefore, the increased number of long wavelength-sensing *opsin* genes in *An. gambiae* as compared to *An. stephensi* may contribute mechanistically to differences in their photopreference behaviors.

We assessed only the long wavelength-sensing *GPROP3* for diurnal expression variation for a number of reasons. First, previous studies by Rund *et al.* 2011 did not suggest diurnal variation in the expression of any opsin (Rund *et al.* 2011). Second, due to sequence conservation among the long wavelength-sensing *opsin* gene set we have defined, *GPROP3* was the only long wavelength-sensing *opsin* gene that could be verified specifically as being expressed using qRT-PCR in *An. gambiae*.

The *GPROP3*, *GPROP8,* and *GPROP9* genes in *An. gambiae*, which are predicted to detect long wavelengths, ultraviolet wavelengths, and short wavelengths, respectively, exhibit no significant diurnal variation in transcription during the 48 hour time period assayed (Fig. 2.4 A,C,E). Among the orthologous genes in *An. stephensi* – annotated as *LW*, *UV*, and *SW* for putative long wavelength-, ultraviolet wavelength-, and short wavelength-responsive *opsin* genes, respectively – the *LW* and *SW* genes fail to exhibit striking diurnal variation in transcription (Fig. 2.4B,F). The *UV* gene transcript levels increase during the dusk crepuscular period compared to levels during other

intervals of Zeitgeber time (Fig. 2.4D). As there are no significant differences in diurnal expression patterns for *opsin* genes we assayed, we can reject the hypothesis that variation in expression of the *opsin* genes assayed is correlated with variations in photopreference that we observe between these two species. Although the transcript levels do not vary throughout diurnal phases, it is possible that protein levels may vary due to translational or post-translational regulation. However, assessment of those possibilities lies beyond the scope of our analysis. Alternatively, as subcellular localization of some opsins in the photoreceptor cells of *Ae. aegypti* and *An. gambiae* has been described, changes in this subcellular localization, again beyond the scope of our analysis, may account for variability in photopreference between species (Hu *et al.* 2009, 2011, 2013).

Developmental Expression and Evolution of *Opsins* in *An. gambiae*

The difference we observe in long wavelength-sensing *opsin* gene number in *An. gambiae* and *An. stephensi* led us to question the potential functional significance the existence of six long wavelength-sensing opsin genes in *An. gambiae* and only four long wavelength-sensing opsin genes in *An. stephensi*. To investigate this question in *An. gambiae*, we utilized RNAseq analysis to assess expression of each of the 11 *opsin* superfamily gene members during first and third larval instars, and in female and male adults (Fig. 2.5, Additional File 2.3). Three annotated long wavelength-sensing *opsin* genes – *GPROP1, GPROP3* and *GPROP4* – are expressed more highly during adult stages, and long wavelength-sensing *opsin* genes *GPROP5-GPROP7* all exhibit increased expression during larval stages, consistent with previous findings from microarray-based

expression analyses (Rund et al. 2011; Marinotti et al. 2006). *GPROP11* and *GPROP12*, pteropsins, are also expressed at low levels during all life stages studied. In contrast, *GPROP10*, an opsin of unknown wavelength sensitivity, is expressed predominantly during adult stages. The remaining *opsin* genes – *GPROP8* and *GPROP9* – which encode one ultraviolet wavelength-sensing opsin and one short wavelength-sensing opsin, respectively, each exhibit higher expression in adults as compared to first and third instar larvae.

The developmental partitioning of *opsin* superfamily gene expression that we observe – most notably the dichotomous expression of long wavelength-sensing *opsin* genes between larval and adult stages – is unexpected and may have functional implications. Past studies of *opsin* gene expression during *An. gambiae* development have utilized the *Plasmodium*/*Anopheles* Genome Array, which groups long wavelength-sensing *GPROP1*, *GPROP3* and *GPROP4* genes into a single probe set (Ag.2R.268.0_CDS_s_at from VectorBase) (Rund et al. 2011; Marinotti et al. 2006). Thus, the respective expression profiles for these three genes have not been defined previously. Each of the other long wavelength-sensing *opsin* genes (*GPROP5*, *GPROP6* and *GPROP7*) is detectable with distinct probes on the array, respectively, allowing for accurate expression profiling of those three *opsin* genes. The use of RNAseq has allowed us to define the expression of each of these *opsin* genes, despite the very limited sequence variation among them, and its use will enable delineation of these paralogs in subsequent analyses.

The fact that half of long wavelength-sensing *opsin* genes are expressed predominantly during larval stages implies that these opsins may mediate functions

specific to larval life stages. In this regard, it is notable that gene structures for the subset of long-wavelength sensing *opsin* genes expressed predominantly during larval stages exhibit structural similarities that distinguish them from those expressed predominantly in adults (Fig 3). Larval-biased *GPROP5, GPROP6* and GRP*OP7* genes each include two exonic CDS regions, and significant 5' UTR and 3' UTR regions are present in *GPROP5* and *GPROP6*. In contrast, adult-biased *GPROP1, GPROP3* and GPR*OP4* each contain a single splice-site within the 5'-UTR of each gene and minimal 3' UTRs and the entireties of their coding capacities reside within a single exon, respectively. These differing structures are consistent with the hypothesis that the two stage-biased *opsin* gene subsets arose from duplication of distinct ancestral genes, with limited subsequent divergence of coding sequences and gene organization within each subset.

However, the life stage-biased functions these long wavelength-sensing opsins mediate remain unclear. Visual acuity may play an important role during larval life stages for the detection of predators within aqueous environments (Klecka and Boukal 2012), while adults may process figures/shapes from the air in search of potential sugar sources, blood meal sources, resting sites and oviposition sites (Allan *et al.* 1987). The predominant expression of some long wavelength-sensing *opsin* genes during larval stages, and the expression of other long wavelength-sensing *opsin* genes, and short wavelength-sensing and ultraviolet wavelength-sensing *opsin* genes only in adults may have arisen because of differing opsin requirements underlying visual acuity in aqueous environments as compared to atmospheric environments.

Subsets of long wavelength-sensing opsins are arranged in homologous loci, which are partially conserved between *An. gambiae* and *An. stephensi* (Fig. 2.3). The

homologous locus in *An. gambiae* that contains two larval-biased genes and one adult-biased gene (i.e., *GPROP4-6*) is highly conserved in *An. stephensi*. If these gene trios are derived from a single gene cluster in the most recent common ancestor (MRCA) of *An. gambiae* and *An. stephensi*, then that MRCA may have possessed similar larval-adult variability in the expression of long wavelength-sensing *opsin* genes. Similarly, *An. stephensi* contains an ortholog of *An. gambiae GPROP7*, and genomic regions surrounding the orthologous gene in each species appear to be syntenic as reflected by the location of *An. gambiae* and *An. stephensi GPROP7* orthologs next to *AGAP002463* and *ASTE008930*, respectively, which are orthologs with homologies to ubiquitin-associated and SH3 domain-containing protein B [**UBASH3B (Megy *et al.* 2012)** , Fig. 2.3]. Taken together, these observations imply that the *GPROP4-6* long wavelength-sensing *opsin* gene cluster and the *GPROP7* orthologs were present in the MRCA of these two species. This invites the hypothesis that the gene family expansion in *An. gambiae* that created *GPROP1* and *GPROP3* occurred after divergence of the two species, and that the differing illumination preferences in the two species also arose following their divergence from a common ancestor, in conjunction with *opsin* gene family expansion. As *GPROP1* and *GPROP3* are expressed predominantly in adults, *An. gambiae* may have been selected during its evolutionary history for greater photosensitivity based on a mechanism mediated by adult *opsin* gene expression. Other organisms, such as butterflies, that exhibit increases in long wavelength-sensing *opsin* gene number also exhibit expanded spectral diversity for visual function (Frentiu et al. 2007; Sison-Mangus et al. 2006). Therefore, the expansion of long wavelength-sensing

opsin gene number may underlie dynamic evolution of visual sensitivity across an expanded spectral range in *An. gambiae*, as compared to *An. stephensi*.

METHODS

Colony

Anopheles gambiae G3 colony (courtesy of Dr. Flaminia Catteruccia, Harvard School of Public Health, Boston, MA, USA) and *An. stephensi* Sind-Kasur strain Nijmegen (courtesy of Dr. Maria Mota, University of Lisbon, Lisbon, Portugal) were used for all experiments. All experiments were performed on mosquitoes 7-10 days post-emergence, that were also aged 3-5 days post-blood feeding and 1-3 days post-egg laying. A Light:Dark (L:D) photoperiod of 11:11 was maintained with 1 hour dawn:dusk transitions between light and dark periods, with a constant temperature of 27° C and 80% relative humidity. Mosquitoes were fed 10% glucose solution *ad libitum* and were kept in the presence of the opposite sex throughout their life cycle.

Photopreference assays

Photopreference assays were performed during the dawn:dusk and dusk:dawn transition periods. Assays were conducted using the arenas illustrated in Additional File 2.1. A 60" long, clear, plexiglass tube with a 2" interior diameter was used for the containment portion of the apparatus. For the trinary assays, photic zones were approximately 20" in length and were illuminated with 0 Lux, 100 Lux or 400 Lux. Illumination levels were based on lux values of a lit room (Yu *et al.* 2007), and lux values obtained from observations outdoors during dawn and dusk hours in Chestnut Hill, MA. Binary assays consisted of a 30" dark zone (0 Lux) and a 30" illuminated zone (400 Lux). There was no temperature change within the tube throughout the course of the experiment, and the dark and illuminated zones of the tube remained at the same

temperature. For each experimental run, approximately 50-75 mosquitoes were aspirated from the colony and introduced to the end of the tube employed for that run. A set of three biological replicates was completed for each pattern of introduction (i.e., illuminated end or dark end introduction). After mosquitoes were allowed to move throughout the tube for 20 minutes, mosquitoes were asphyxiated quickly by rapid exposure to high-concentration CO_2, to avoid alteration of resting patterns, and counts of male and female mosquitoes within each photic zone were then performed. The length of time used for each assay (20 min) was chosen as mosquito activity, i.e., the movement of mosquitoes among regions within the tube, did not change further beyond 20 min following the introduction of mosquitoes (data not shown).

Statistical analysis

Statistical comparisons for the assessment of photopreference were performed using a Chi-Squared test to determine whether observed distributions deviated significantly from a random distribution. Statistical analyses were performed using Prism 5.0 software.

Collection of samples and qRT-PCR of selected phototransduction pathway genes

All gene sequences, nomenclature and identifiers are according to VectorBase VB-2013-12 (https://www.vectorbase.org) (Megy *et al.* 2012). qRT-PCR was performed for genes associated with known functions, including light detection and phototransduction pathways in both *An. gambiae* and *An. stephensi*. Samples were collected over a 48-hour time period in order to encompass two complete diurnal L:D

cycles. Collections were made every 4 hours and consisted of approximately 10-15 female mosquitoes. Mosquito heads were immediately removed, and RNA was extracted using TriReagent (Sigma: St. Louis, MO, USA), for use in subsequent analyses.

RPS7 (AGAP010592) gene expression was used as a reference for both species. Long wavelength-sensing (AGAP012982), short wavelength-sensing (AGAP010089), and ultraviolet wavelength-sensing (AGAP006126) genes were assayed for expression patterns, as compared to control genes, in both species. Sequences and concentrations of primers used for qRT-PCR can be found in Additional File 2.2. *An. stephensi* genes orthologous to those in *An. gambiae* were identified using local BLAST and manual annotation of the of the *An. stephensi* genome (VectorBase VB-2013-12). USB VeriQuest SYBR Green One-Step qRT-PCR Master Mix 2X (Affymetrix: Santa Clara, CA, USA) was used to perform qRT-PCR. Cycling conditions were 50°C for 10 min, 95°C for 10 min, 40 cycles of 95°C for 15 sec and 58°C for 30 sec for *An. gambiae* (61°C for 30 sec for *An. stephensi*). Reactions were run on a 7500 Fast Real-Time PCR System (Applied Biosystems: Grand Island, NY, USA). qRT-PCR reaction products were subsequently sequenced to verify amplification of correct target sequences. All values were normalized to the highest expression value obtained for the given gene, for visualization purposes.

RNA sequencing and analysis

Male and female whole body RNAseq data sets from *An. gambiae* (GASUA strain) mosquitoes were obtained from Dr. Larry Zweibel and Dr. Jason Pitts (Vanderbilt University, Nashville, TN)(Pitts *et al.* 2011). Those mosquitoes, which were reared with

a Light:Dark (L:D) photoperiod of 12:12 in 75% humidity, were collected for sequencing at Zeitgeber time 10-12; therefore, and were therefore exposed to illumination preceding collection of RNA. We collected two biological replicates at the same time points as Pitts *et al.* (2011), i.e., first (L1) and third (L3) instar larvae, as well as single biological replicates of adult males and females (whole body) of *An. gambiae* G3 to compliment the Vanderbilt University data set. We collected only single adult replicates as our goal was to validate expression levels reported by Pitts et al. (2011), rather than define statistically significant differences in transcriptional expression among life stages. RNAseq data sets have been deposited in the European Nucleotide Archive under the SRA accession PRJEB5712 . RNA extraction and sequencing of these collections were performed by Otogenetics Corp. (Norcross, GA, USA) and the Broad Institute (Cambridge, MA, USA). All RNA-seq data were aligned to *An. gambiae* P3 assembly, from VectorBase VB-2013-12, using Tophat2 (Kim *et al.* 2013). FPKM values and comparisons between samples were performed using Cufflinks-Cuffdiff2, and the subsequent heatmap was visualized using CumberBund (Trapnell *et al.* 2013). Genes analyzed included all long wavelength-sensing opsins *GPROP1* (AGAP013149), *GPROP3* (AGAP012982), *GPROP4* (AGAP012985), *GPROP5* (AGAP001162), *GPROP6* (AGAP001161), *GPROP7* (AGAP002462), ultraviolet wavelength-sensing opsin *GPROP8* (AGAP006126), short wavelength-sensing opsin *GPROP9* (AGAP010089), an unknown wavelength-sensing opsin *GPROP10* (AGAP007548) and the two pteropsins *GPROP11* (AGAP002443) and *GPROP12* (AGAP002444).

TABLES/FIGURES AND LEGENDS

Table 2.1. *An. gambiae* and *An. stephensi* Binary Photopreference Data

Tabulation of results presented in Figure 2.1. Zeitgeber time and Introduction site are presented in the left-hand columns, with photic regions represented with 0 Lux and 400 Lux. Values are percent resting in respective region ± SEM.

		An. gambiae				*An. stephensi*			
Zeit. Time	**Int Site**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**
Dawn	**400**	76.1 ±3.3	68.7 ±8.7	25.1 ±2.9	38.0 ±7.7	37.6 ±2.6	70.5 ±3.3	63.0 ±2.8	30.5 ±3.2
Dawn	**0**	69.8 ±4.3	75.4 ±2.6	32.09 ±4.71	25.3 ±2.6	48.2 ±4.2	78.3 ±2.0	53.2 ±4.1	22.3 ±2.1
Dusk	**400**	74.6 ±0.7	57.3 ±5.0	25.5 ±0.7	44.6 ±5.0	29.7 ±4.1	50.7 ±2.0	73.3 ±5.8	49.6 ±2.0
Dusk	**0**	62.0 ±1.3	62.2 ±0.5	38.2 ±1.2	37.8 ±0.5	34.1 ±7.5	44.3 ±7.6	72.3 ±8.0	59.9 ±6.6
		0 Lux		**400 Lux**		**0 Lux**		**100 Lux**	
		PHOTIC PREFERENCE ZONE							

Table 2.1. ***An. gambiae*** **and** ***An. stephensi*** **Binary Photopreference Data**

Table 2.2. *An. gambiae* and *An. stephensi* Trinary Photo Preference Data

Tabulation of results presented in Figure 2.3. Zeitgeber time and Introduction site are presented in the left-hand columns, with photic regions represented with 0 Lux, 100 Lux and 400 Lux. Values are percent resting in respective region ± SEM.

Zeit. Time	Ent. Site	*An. gambiae*			*An. stephensi*		
		FEMALES					
Dawn	**400**	27.2 ±6.0	56.1 ±3.7	22.0 ±4.7	59.7 ±6.9	31.5 ± 3.5	12.6 ±2.5
Dawn	**0**	17.1 ±2.5	19.3 ±3.5	66.7 ±5.2	22.0 ±4.8	24.4 ± 3.5	58.2 ±6.7
Dusk	**400**	27.9 ±3.6	52.4 ±4.2	21.5 ±1.5	51.6 ±4.9	46.2 ± 7.5	7.7 ±1.8
Dusk	**0**	14.4 ±2.6	23.4 ±2.2	64.3 ±4.6	46.3 ±9.5	35.7 ± 4.9	25.2 ±3.7
		MALES					
Dawn	**400**	22.9 ±4.0	64.8 ±6.8	16.2 ±2.3	39.8 ±3.2	58.8 ±1.3	6.5 ±0.0
Dawn	**0**	20.4 ±6.6	14.1 ±2.1	76.1 ±4.8	22.2 ±1.3	32.1 ±2.4	54.9 ±10.4
Dusk	**400**	23.4 ±0.7	54.1 ±3.0	23.9 ±3.7	48.6 ±10.2	56.4 ±9.3	8.9 ±0.1
Dusk	**0**	12.3 ±2.2	25.9 ±2.9	63.8 ±1.7	13.1 ±0.2	74.6 ±5.3	17.2 ±4.0
		400 Lux	**100 Lux**	**0 Lux**	**400 Lux**	**100 Lux**	**0 Lux**
		PHOTIC PREFERENCE ZONE					

Table 2.2. ***An. gambiae*** **and** ***An. stephensi*** **Trinary Photo Preference Data**

Figure 2.1. *An. gambiae* and *An. stephensi* Binary Photopreference

Bar graphs depict percent of mosquitos resting in specific photic regions (±SEM, N=3) for each experiment. Left and right columns depict *An. gambiae* and *An. stephensi* resting patterns for each condition, respectively, with males and females being depicted within each column. Dawn and dusk refer to relative crepuscular period. Right hand titles indicate introduction site followed by relative crepuscular period. Black bars represent mosquitos resting in the 0 Lux region of the tube at the end of the experiment, and open bars represent those resting in the 400 Lux region. **A,B**. Introduction into 400 Lux region at dawn **C,D.** Introduction into 0 Lux region at dawn **E,F.** Introduction into 400 Lux region at dusk **G,H.** Introduction into 0 Lux region at dusk ★P<0.05, ★★P<0.01, ★★★P<0.001

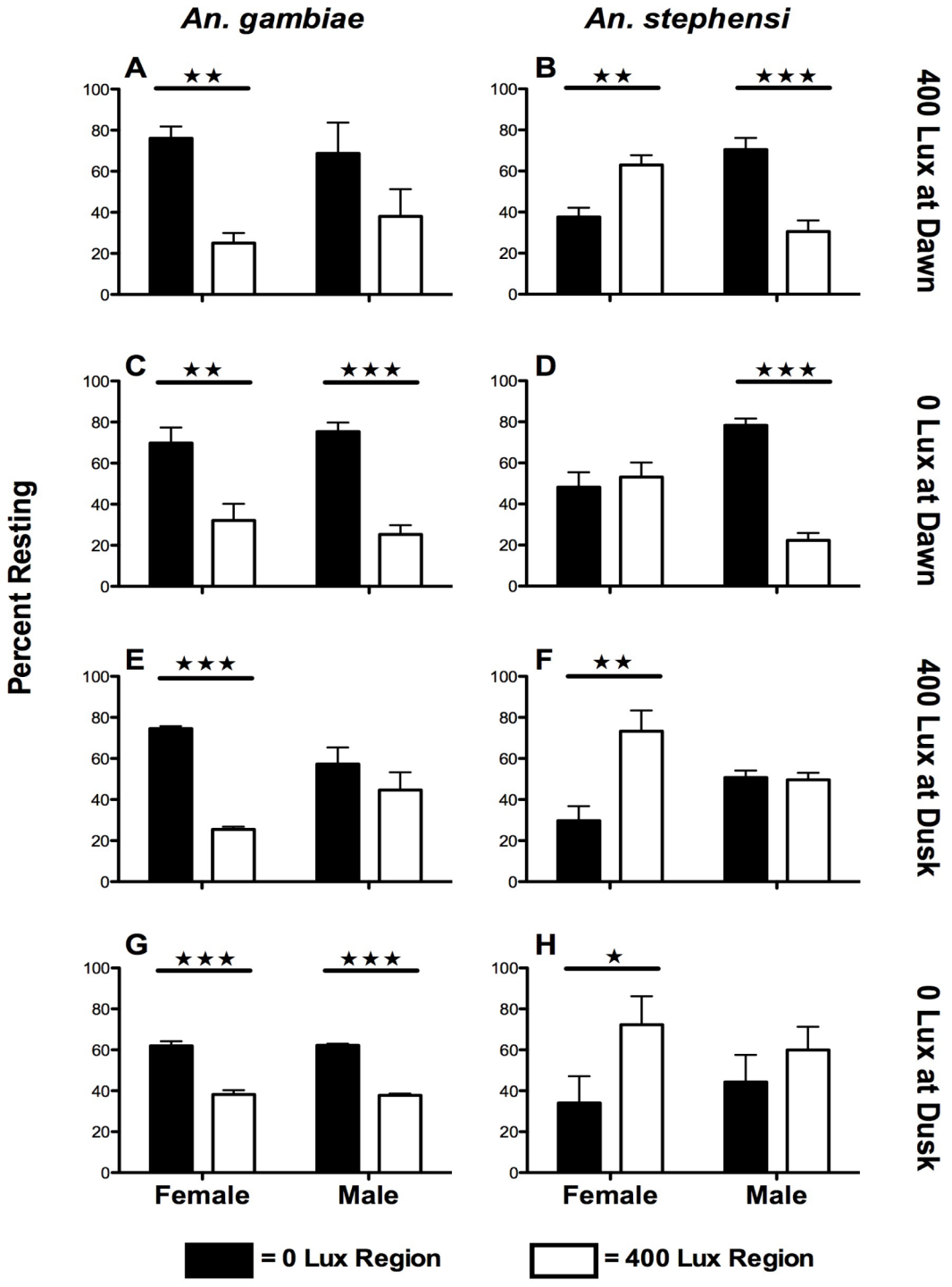


Figure 2.1. ***An. gambiae*** **and** ***An. stephensi*** **Binary Photopreference**

Figure 2.2. *An. gambiae* and *An. stephensi* Trinary Photopreference

Bar graphs depict percent of mosquitos resting in specific photic regions (±SEM, N=3) for each experiment. Left and right columns depict *An. gambiae* and *An. stephensi* resting patterns for each condition, respectively. Dawn and dusk refer to relative crepuscular period. Right hand titles indicate introduction site, followed by relative crepuscular period. Black bars represent mosquitos resting in the 0 Lux region of the tube at the end of the experiment, gray bars represent those resting in the 100 Lux region and open bars represent those resting in the 400 Lux region. **A,B**. Introduction into 400 Lux region at dawn **C,D.** Introduction into 0 Lux at dawn **E,F.** Introduction into 400 Lux at dusk **G,H.** Introduction into 0 Lux at dusk ★P<0.05, ★★P<0.01, ★★★P<0.001

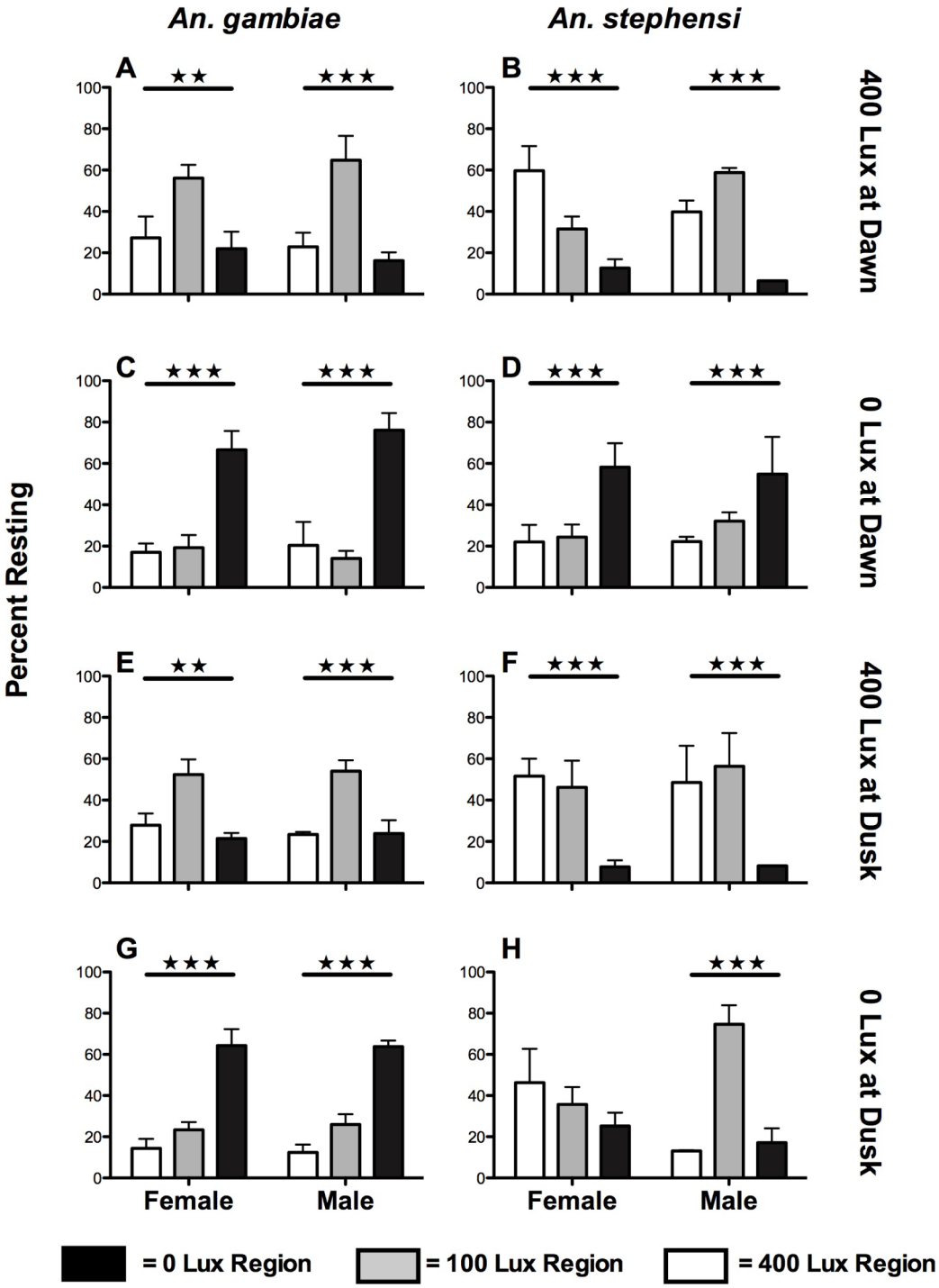


Figure 2.2. *An. gambiae* and *An. stephensi* Trinary Photopreference

Figure 2.3. Long Wavelength Opsin Gene Organization on *An. gambiae* Chromosome Arm 2R

Five of the six long wavelength-sensing *opsin* genes cluster toward the telomeric end of chromosome 2R in *An. gambiae*. This gene number contrasts with the four orthologous long wavelength-sensing *opsin* genes present in *An. stephensi*

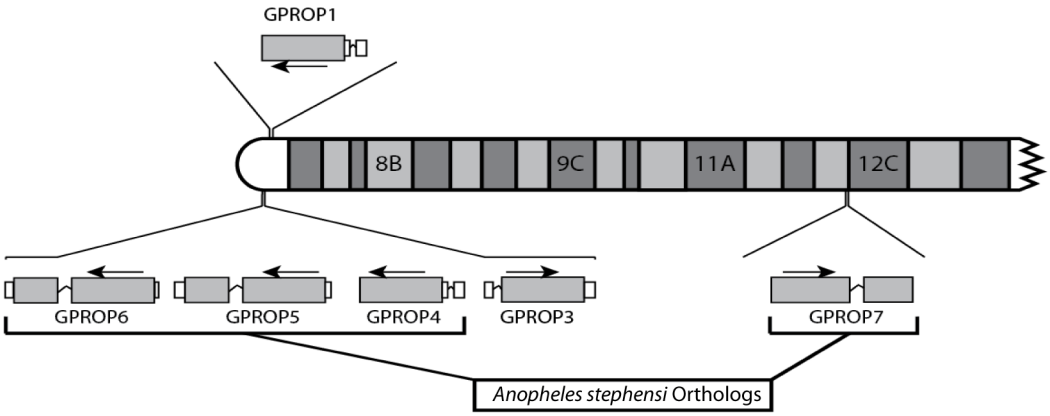


Figure 2.3. Long Wavelength Opsin Gene Organization on *An. gambiae* Chromosome Arm 2R

Figure 2.4. Opsin Expression Profiles Across Zeitgeber Time

Relative quantity ($2^{\Delta Ct}$ ± SEM) of *opsin* gene transcripts normalized to *ribosomal protein subunit-7* transcript, respectively. Time points indicate samples taken every 4 hours, with time point 0 being at the beginning of a 11:11 light:dark cycle with 1 hour dusk:dawn transition periods, spanning two full diurnal cycles. Each time point consists of collections of 10 female mosquitos, with N=3. Values are normalized so the highest level of expression is equal to one for each analysis. Filled bars represent time points sampled during the dark phase of the cycle. Open bars represent time points sampled during the light phase of the cycle.

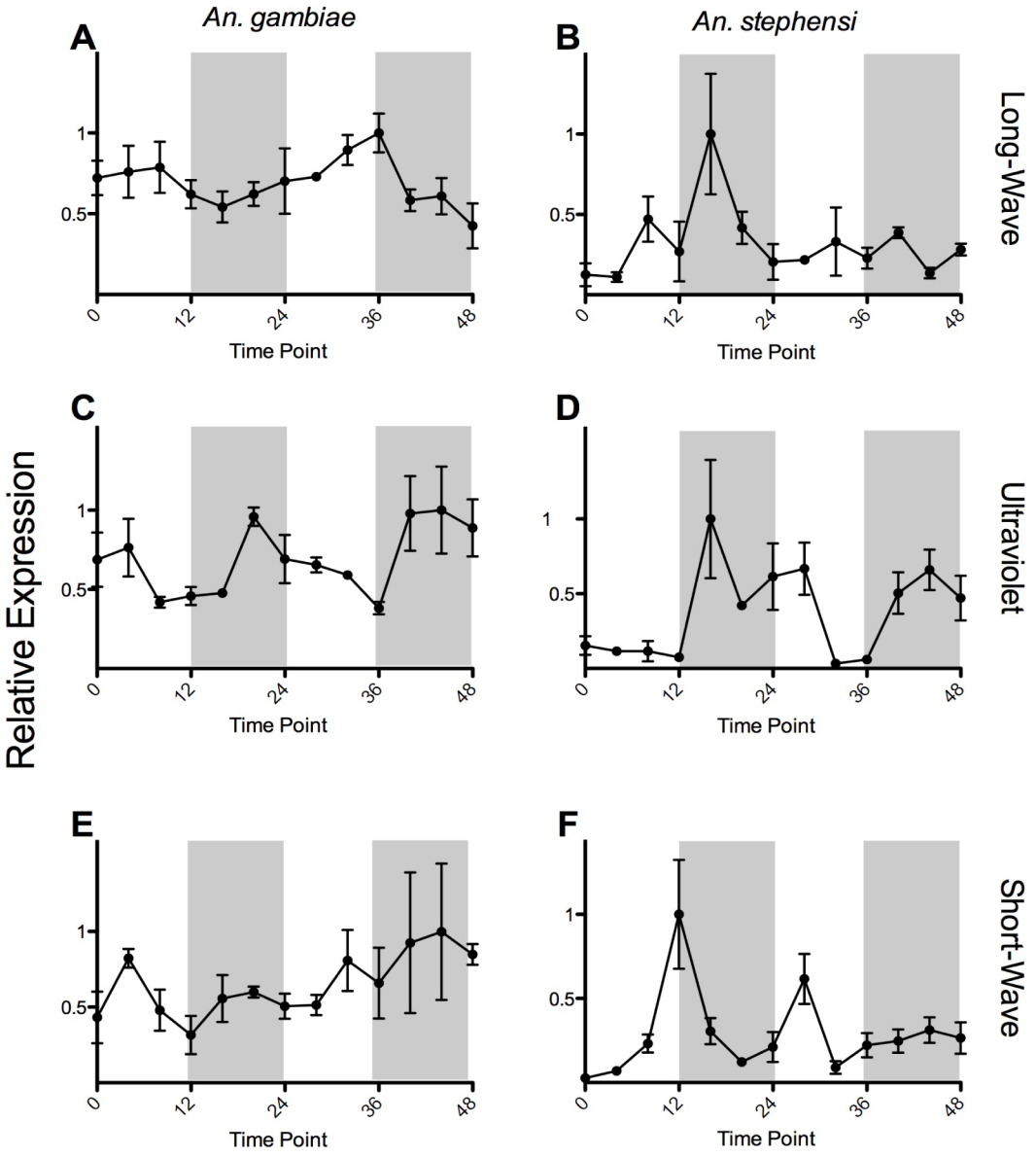


Figure 2.4. Opsin Expression Profiles Across Zeitgeber Time

Figure 2.5. Heatmap of *An. gambiae Opsin* Gene Expression

Expression of *Opsin1, 3-12* in *An. gambiae* in mixed-gender first larval instars (L1), mixed-gender third larval instars (L3), adult females (FB), and adult males (MB). Color intensity scale indicates increasing expression, with yellow reflecting the highest expression, measured as FPKM, and blue reflecting the lowest expression. VectorBase ID identifiers and names are given for each transcript. All *opsin* genes are also grouped based on wavelength detected, PT (pteropsin), UN (unknown), SW (short wavelength), UV (ultraviolet wavelength), LW (long wavelength).

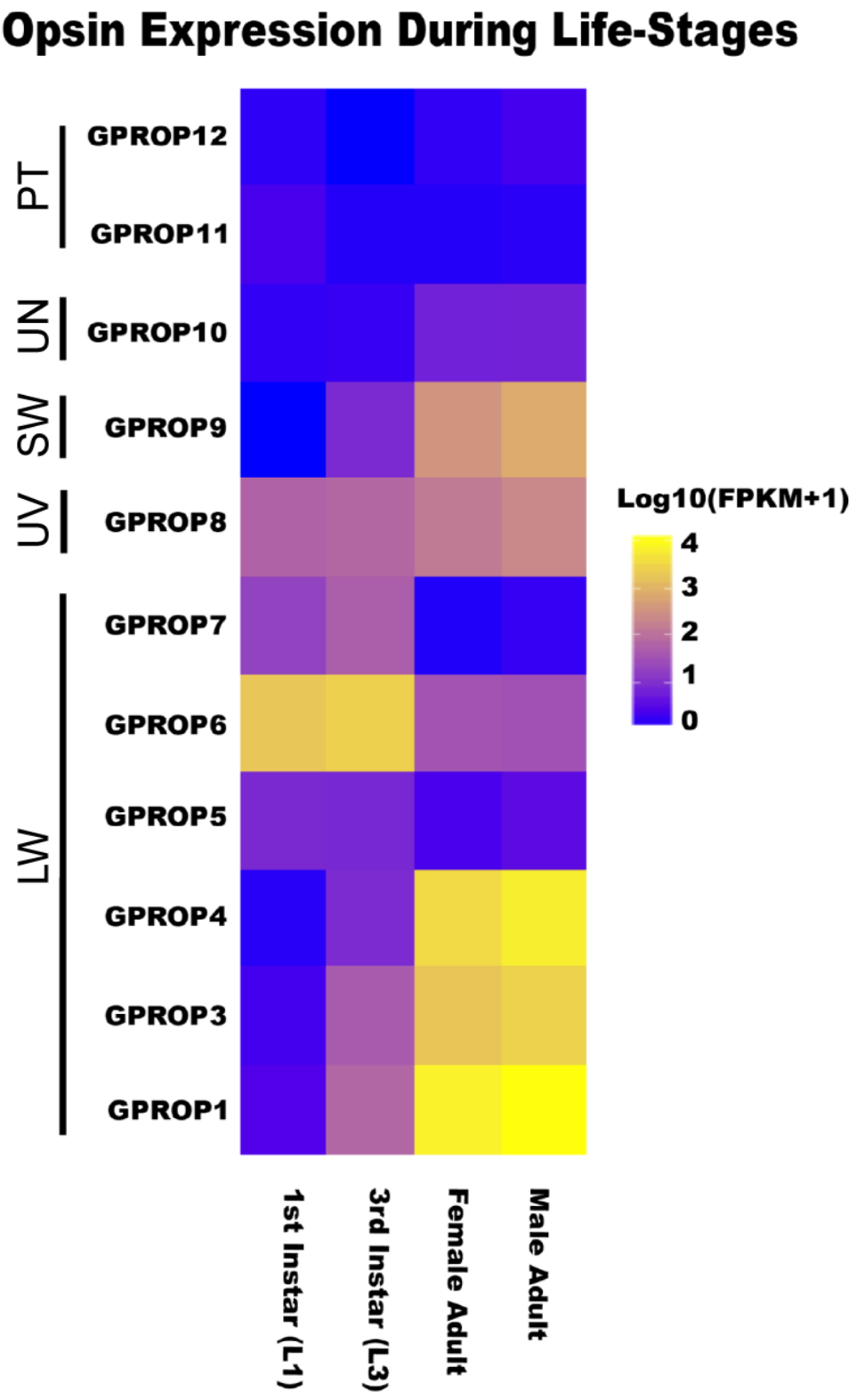


Figure 2.5. Heatmap of *An. gambiae Opsin* Gene Expression

Chapter III:

Long non-coding RNA discovery across the genus *Anopheles* reveals conserved secondary structures within and beyond the Gambiae complex

ABSTRACT

Long non-coding RNAs (lncRNAs) have been defined as mRNA-like transcripts longer than 200 nucleotides that lack significant protein-coding potential, and many of them constitute scaffolds for ribonucleoprotein complexes with critical roles in epigenetic regulation. Various lncRNAs have been implicated in the modulation of chromatin structure, transcriptional and post-transcriptional gene regulation, and regulation of genomic stability in mammals, *Caenorhabditis elegans*, and *Drosophila melanogaster*. The purpose of this study is to identify the lncRNA landscape in the malaria vector *An. gambiae* and assess the evolutionary conservation of lncRNAs and their secondary structures across the *Anopheles* genus. Using deep RNA sequencing of multiple *Anopheles gambiae* life stages, we have identified 2,949 lncRNAs and more than 300 previously unannotated putative protein-coding genes. The lncRNAs exhibit differential expression profiles across life stages and adult genders. We find that across the genus *Anopheles*, lncRNAs display much lower sequence conservation than protein-coding genes. Additionally, we find that lncRNA secondary structure is highly conserved within the Gambiae complex, but diverges rapidly across the rest of the genus *Anopheles*. This study offers one of the first lncRNA secondary structure analyses in vector insects. Our description of lncRNAs in *An. gambiae* offers the most comprehensive genome-wide insights to date into lncRNAs in this vector mosquito, and defines a set of potential targets for the development of vector-based interventions that may further curb the human malaria burden in disease-endemic countries.

INTRODUCTION

Sequencing the genome of the African malaria mosquito, *Anopheles gambiae* (Holt et al. 2002), has fueled many large- and small-scale investigations of the biology of this important vector, in an effort to develop more effective interventions to limit its harmful impacts on human health (Severson and Behura 2012). Functional genomic studies using microarrays have described basic biological processes and stimulus-responsive gene expression by detailing transcriptome profiling during the *An. gambiae* life cycle, in specific tissues, across Zeitgeber time, following blood feeding and infection, and coincident with insecticide resistance (Rund et al. 2011; Koutsos et al. 2007; Harker et al. 2012; Edi et al. 2014; Mitchell et al. 2014; Neira Oviedo et al. 2009; Stamboliyska and Parsch 2011; Phuc et al. 2003; Marinotti et al. 2006). More recent RNA sequencing (RNAseq) studies in *An. gambiae* have described odorant receptor expression in various contexts (Rinker et al. 2013; Pitts et al. 2011) and other RNAseq efforts in vector insects have enabled generation of the first *de novo* transcriptome for *Anopheles funestus* (Crawford et al. 2010). Because they are designed based on existing genome annotations, gene expression microarrays cannot facilitate the discovery of unannotated genes. RNAseq is not constrained in this way, but high read depths are required for significant increases in analytical sensitivity. Most previous RNAseq studies have focused on using reads as a measure of expression of previously annotated genes, rather than discovering new genes, including new classes of genes such as lncRNAs (Nie et al. 2012; Kung et al. 2013; Fatica and Bozzoni 2014). Indeed, recent RNAseq of the *An. gambiae* midgut transcriptome demonstrated that high-depth sequencing can uncover many novel intergenic transcripts, including putative lncRNAs (Padrón et al. 2014).

Large-scale functional genomic projects such as ENCODE and modENCODE, as well as high-throughput genomic screens, have revealed the presence of extensive sets of lncRNAs in humans (approximately 9,300), as well as in model organisms (e.g., approximately 900 in nematodes and 1,100 in fruit flies) (Guttman et al. 2009; Carninci et al. 2005; Young et al. 2012; Ulitsky et al. 2011; Nam and Bartel 2012; Harrow et al. 2012; Bernstein et al. 2012; Hangauer et al. 2013; Pauli et al. 2012). The functions of these lncRNAs, however, remain largely unknown, with a few exceptions that include lncRNAs with defined roles in embryogenesis, development, dosage compensation and sleep behavior (Pauli et al. 2012; Soshnev et al. 2011; Li et al. 2012; Lv et al. 2013; Heard and Disteche 2006; Mercer and Mattick 2013). Part of the difficulty in deciphering the functionality of lncRNAs lies in their rapid evolution and the consequent reduction in levels of primary sequence conservation for lncRNAs among different organisms (Necsulea et al. 2014; Kutter et al. 2012; Necsulea and Kaessmann 2014). While this divergence presents some challenges, the lack of conservation could be exploited in species-specific targeted therapeutics. Indeed, it has been proposed that lncRNAs could be used as targets to regulate gene expression and development, as an alternative to the standard model of using small molecule drugs as antagonists of mRNA-encoded proteins (Wahlestedt 2013). This premise may also be extended to controlling vector-transmitted infectious diseases by identifying and perturbing non-coding RNA (ncRNA) targets in vector insects (Lucas et al. 2013).

Previously successful vector control methods have begun to wane in efficacy with the development of singly and multiply insecticide-resistant mosquitoes in disease-endemic regions (e.g., (Edi et al. 2014; Mitchell et al. 2014)). Future malaria vector control will have to rely on new approaches, some of which may become apparent only as we develop a more complete understanding of the repertoire of mosquito coding and non-coding genes (Lucas et al. 2013; Burt 2014; Padrón et al. 2014). Using RNAseq across multiple mosquito life stages and both genders, our study has developed the most comprehensive deep RNAseq data set for *An. gambiae* to date, encompassing more than 500 million alignable sequence reads. Differential gene expression analysis confirms the roles of different classes of annotated protein-coding genes during key developmental phases, and quantification of protein-coding potential of previously unannotated transcripts identifies 318 new protein-coding genes and 2,949 putative lncRNAs. We find that the lncRNA gene set exhibits much lower sequence conservation across anophelines, when compared with either previously annotated protein-coding genes or protein-coding genes discovered in our study. While these lncRNA genes exhibit low sequence conservation, we provide evidence that the secondary structural features for many lncRNAs have been conserved. These newly identified lncRNAs provide a basis for an expanded understanding of lncRNAs in dipterans, and for future studies of ncRNAs within the genus *Anopheles*.

RESULTS

Alignment and Validation of RNAseq Data Sets

Our transcriptome analysis for each life stage was supported by two RNAseq data sets: one "high read depth (HRD)" set with more than 140 million reads/stage that was used for subsequent lncRNA discovery, and one "low read depth (LRD)" set that contained approximately 30 million reads/stage that constituted biological replicates for the validation of our HRD data sets. In total, over 500 million HRD reads and over 100 million LRD reads were aligned to the *An. gambiae* PEST genome assembly AgamP3 (Table 3.1, see MATERIALS and METHODS). First, Cufflinks' fragments per-kilobase of exonic length per million base pairs mapped (FPKM) expression values were validated against SailFish, an alignment-free quantification method that uses K-mers and defines expression levels based on reads per-kilobase of exonic length per million base pairs mapped (RPKM) (Patro et al. 2014; Trapnell et al. 2010). The average FPKM and RPKM values between the two biological replicates produced by Cufflinks and Sailfish show Pearson correlation coefficients that were all above 0.6 (Fig. 3.1A), indicating a high level of confidence that Cufflinks FPKM values are comparable to other, reference-free quantification methods. Using Cufflinks FPKM values, the number of differentially expressed (DE) genes identified varies greatly depending on the life stages compared, as shown by the clustered FPKM values in Figure 3.1B (Additional File 3.1). Concordant with physiological changes, fewer DE genes were identified between similar life stages, i.e., between larval stages [first larval instar (L1) and third larval instar (L3)] or between adult genders, than between larval and adult stages.

Only three protein-coding genes (*AGAP007089*, *AGAP010068*, *AGAP010708*) exhibit significant decreases in expression in L3 compared to L1, while 61 are significantly up-regulated. In an adult male to adult female comparison, 44 protein-coding genes are down-regulated, while 88 are up-regulated. Adult to larval comparisons range between 133 genes up-regulated between females and L3s, the lowest such difference observed, and up to 388 genes down-regulated between males and L3s, the greatest such difference observed. When these DE genes are grouped based on their GO_Slim2 categories (Hu et al. 2008), a total of 30 major categories are identified, each of which constitutes greater than two percent of the total gene count for a given comparison (Fig. 3.2). Those categories with greater than 2 percent of the gene count are distributed across all life stage and gender comparisons. Any category that is present in less than two percent of the total DE genes for the given comparison is grouped into the "Less Than 2 percent " category; this category is the largest group for many of our comparisons. Due to the expansive nature of these categories, the DE genes were analyzed for functional enrichment using DAVID (**d**atabase for **a**nnotation, **v**isualization and **i**ntegrated **d**iscovery) (Huang et al. 2009) to define biologically relevant groups that are differentially expressed.

Across the adult to larval comparisons, 16 categories possess an enrichment score greater than 1.5 (Fig. 3.1C, Additional File 3.2). Genes associated with cuticle, peptidase activity, chitin/carbohydrate binding and detoxification are enriched during larval stages, when compared to adults. Genes associated with odorant recognition, immunity and visual stimuli are enriched in adults, when compared to larval stages. Overall,

differentially expressed genes and their associated DAVID-enriched terms (Additional File 3.2) are congruent with past studies of *An. gambiae*. (Harker et al. 2012; Koutsos et al. 2007).

De Novo Identification of Transcripts

Cufflinks and Scripture were utilized to produce a reference annotation-based transcript (RABT) assembly – using a merged data set of all HRD RNAseq data sets – in order to identify previously unannotated RNA transcripts (Fig. 3.3A). As the aim of this study was not to identify potential isoforms of previously annotated transcripts, only gene classes of I, U and X (intronic transcript, intergenic transcript, and exonic overlap on opposite strand, respectively) as identified by Cufflinks, were analyzed. A total of 4,690 transcripts possessed assembled transcript support by both Cufflinks and Scripture (Fig. 3.3A). After implementing a length cutoff of 200 nt, a set of 4,477 potential transcribed loci was identified. All genes were given the identifier "Merged" (e.g., Merged.1023), based on the use of merged HRD life stage RNAseq data sets to enable the annotations.

Potential protein-coding mRNAs and lncRNAs were identified based on sequence and amino acid lengths, percent coding sequence and protein-coding potential (using PhyloCSF), as described in MATERIALS and METHODS. This yielded 318 potential protein-coding transcripts (Additional Files 3.3, 3.11, 3.12 and 3.13) and 2,949 potential lncRNAs (Additional Files 3.3, 3.4 and 3.10). Among the 2,949 putative lncRNAs we have identified, most are intergenic transcripts (2059 lncRNAs) (Cufflinks class code "U"), while 108 are in an anti-sense orientation with respect to an exonic region of an

overlapping, protein-coding mRNA (Cufflinks class code "X"), and 782 map within an intron of a protein-coding gene (Cufflinks class code "I") (Additional File 3.5). For transcripts consisting of a single exon, it may be difficult for Cufflinks to predict the correct strandedness of transcript as there is no protein-coding region to validate the strandedness, and the pipeline may generate complementary-strand duplicate gene calls by calling the inferred transcript twice, on each of the complementary strands to which RNAseq reads align. To determine the number of genes that may have been defined as such complementary-strand duplicates we compared all genes identified and found that only 241 genes (i.e., less than 10%) exhibited 50% total overlap (Additional File 3.9). This implies that only a very small proportion of the transcripts identified may constitute complementary-strand duplicates rather than single gene calls. Potential protein-coding genes possess an average of 2.6 exons/gene (Fig. 3.3B), while the lncRNA genes have, on average, 1.2 exons/gene. To further characterize the organization of the newly-annotated genes, respective FPKM expression levels were analyzed (Fig. 3.3C). The FPKM values for the newly annotated protein-coding genes we have identified tend to be lower than those for previously identified protein-coding genes in the reference AgamP3.7 gene set, while newly identified lncRNAs tend to have mean/median FPKM values lower than those for newly annotated protein-coding genes (Fig. 3.3C) (Additional File 3.6). Figure 3.4 illustrates examples of a novel protein-coding gene (Fig. 3.4A), an intronic lncRNA (Fig. 3.4B) and an anti-sense lncRNA (Fig. 3.4C) and an intronic lncRNA (Fig. 3.4C) that were identified in our study. Of the 2,949 lncRNA genes, 39 exhibit significant differences in expression patterns (Fig. 3.5) among life stages (Additional File 3.7). Comparison of our lncRNA gene set to that recently described

based on a gut transcriptome (Padrón et al. 2014) identifies 209 genes that possess at least 50 percent overlap ("Merged" lncRNAs exhibiting overlap can be found in Additional File 3.8).

Evolutionary Conservation of lncRNA Sequences and Secondary Structures

In light of recent studies of the evolutionary conservation, and the lack thereof, among lncRNAs in tetrapods (Necsulea et al. 2014; Necsulea and Kaessmann 2014), we examined the conservation of *An. gambiae* lncRNAs across the *Anopheles* genus. First, we quantified the presence/absence of lncRNA-homologous genomic regions in whole genome multiple sequence alignments across the *Anopheles* phylogeny, based on the presence/absence of an alignable region in our whole genome alignments (WGA) (Fig. 3.6, Tables 3.2 and 3.3). Of the lncRNAs we have identified in *An. gambiae*, almost all exhibit conserved homologous regions within the genomes of the closely-related species within the Gambiae complex, e.g. approximately 97 percent are found in the genome of *Anopheles merus* (Fig. 3.6). At this close evolutionary distance, similarly high percentages of homologous regions are found for the previously annotated protein-coding genes (99 percent) and the newly annotated protein-coding genes (92 percent). In the more distantly-related species, *Anopheles minimus*, of the Myzomia Series, the percentages of protein-coding genes with identifiable homologs drop to 97 percent (previously annotated) and 79 percent (newly annotated), respectively. In the most distantly related species, *Anopheles albimanus*, from the Nysorrhynchus Series, these percentages decline even further to 91 percent and 60 percent, respectively, for previously and newly annotated protein-coding genes (Fig. 3.6). Strikingly, while 77

percent of the *An. gambiae* lncRNAs detect identifiable homologous regions in *An. minimus*, the number of conserved lncRNA-homologous regions drops dramatically, to only 20 percent, in the distant species *An. albimanus*.

To further characterize the conservation of lncRNAs, PhyloP was utilized to determine per-nucleotide conservation p-values across all of the genus members studied (Fig. 3.7A). Previously annotated genes in *An. gambiae* possess higher –log(p-value of conservation) scores compared to both newly identified protein-coding and lncRNA gene classes identified in this study. The previously annotated protein-coding genes exhibit a mean (95 percent CI) value of 122.0 (120.1-123.8), newly identified protein-coding RNAs exhibit a value of 38.34 (31.88-44.80) and lncRNAs exhibit a value of 10.64 (9.958-11.32). All pairwise comparisons of the extent of conservation between all classes were significantly different (Mann-Whitney Test, p-value < 0.001).

Next, we employed REAPR (**rea**lignment for **p**rediction of structural non-coding **R**NA) to examine the conservation of RNA secondary structures in our set of newly identified transcripts. The lncRNA class contains 1,166 conserved secondary structures that possess high-confidence RNA secondary structures according to their RNAz scores (an RNAz score above 0.5 was regarded as a basis for high confidence), distributed among 835 distinct lncRNAs (Fig. 3.7B, 3.8 and 3.9, Table 3.4). By comparison, our set of newly annotated protein-coding genes contains 223 conserved RNA secondary structures among 126 distinct genes. Among the high-confidence secondary structure loci identified among lncRNAs in this study, we next analyzed the conservation of these structures

across the genus *Anopheles* (Fig. 3.10, Fig. 3.11). The genomes of species studied from the Gambiae complex exhibit high numbers of conserved secondary structures, with most genomes retaining similar numbers of conserved structures (Fig. 3.10). Those species outside of the Gambiae complex exhibit much lower numbers of conserved secondary structures compared to *An. gambiae*, especially those species outside of the Pyretophorous Series. The 293 lncRNAs that map to genomic intervals that exhibit primary sequence conservation across all of the anopheline genomes that we analyzed possess 164 distinct secondary structural features. Those features were present in all species within the Gambiae complex, within 129 of the secondary structures we define (Fig. 3.12). Additionally, only two of the secondary structures were present in all 21 genomes analyzed. Overall, the rate of divergence for conserved secondary structures is much greater than for the conserved lncRNA-homologous genomic regions, though the observed difference is not statistically significant (p-value=0.09) (Fig. 3.10B.)

DISCUSSION

Our deep RNA sequencing has facilitated comprehensive transcriptional profiling across four *An. gambiae* life stages, identified multiple previously unannotated protein-coding genes and created the most comprehensive catalog of lncRNAs in any mosquito species, to date. Our quantification of reads mapped to genome assemblies has enabled determination of differential expression among life stages, and our aggregate data set of such genes includes many genes that have been defined as being differentially expressed in previous microarray-based studies of *An. gambiae* gene expression (Harker et al. 2012; Koutsos et al. 2007). First, we compared two quantification methods, Cufflinks and Sailfish, to determine whether an alignment-free quantification method was comparable to Cufflinks and potentially preferable to currently used alignment-based methods due to it's increased speed and accuracy of estimating expression rates (Fig 3.1A). Overall, both Cufflinks FPKM and Sailfish RPKM values are comparable and exhibit correlation values 0.6 or higher (Fig 3.1A). We note that we were unable to produce correlation values between Cufflinks and SailFish that were reported previously when comparing the accuracies of both methods to synthetic and qPCR data sets (Patro et al. 2014). Combined with downstream analyses and visualization packages, we chose to use Cufflinks and its component packages for our lncRNA analysis.

Our differential gene expression profiles (Fig 3.1B, Additional File 3.1) were compared to earlier microarray-based studies to validate our RNAseq data sets. These microarray-based studies identified greater numbers of differentially expressed genes in larval-adult comparisons than in larval-larval or adult-adult comparisons, a trend of differences that is

also clearly observed based on our RNA sequencing approach (Fig. 3.2). Studies by Koutsos *et al.* (2004) and Harker *et al*. (2011) both identified more differentially expressed genes, especially in the L1-L3 comparisons, which can be attributed to the greater number of replicates performed in their microarray studies. Similar to the Koutsos *et al.* (2004) study, we identify more differentially expressed genes between males and larvae than between females and larvae. Functional classes of differentially expressed genes include many cuticular, peptidase and chitin-binding genes that are up-regulated during larval stages, and odorant recognition and immune class genes that are up-regulated in adults (Fig. 3.1C, Additional File 3.2). Similar life stage-related expression patterns have been observed for immunity genes in the pollen beetle, *Meligethes aeneus* (Vogel et al. 2014). Harker *et al.* (2011) described similar larval up-regulation of various gene ensembles in their study of *An. gambiae* using microarrays, including the cuticular gene A*GAP010469* and peptidase-associated genes *AGAP005671*, *AGAP001250*, *AGAP006676* and *AGAP006677*. Koutsos *et al.* (2004) found genes that contain immune-related domains and fall within the pheromone-sensing GO class are up-regulated in adults, and our RNAseq-based analyses have identified similar expression patterns. The consistencies we observe in differential gene expression patterns between life stages, and in functional classes up-regulated during larval and adult life stages, respectively, engender confidence in the quality of our data set.

While approaches for the alignment of RNAseq reads to genomes are relatively mature, the task of grouping such aligned reads into lncRNAs or other gene classes remains challenging and is less well-defined. Previous classifications of lncRNAs have been

based on their lengths, protein-coding potential, and maximum ORF size, and the probability of identifying full-length lncRNA transcripts using RNAseq (Sun et al. 2013; Young et al. 2012; Pauli et al. 2012; Hangauer et al. 2013; Sun et al. 2012). In our study, no FPKM cutoff was utilized, as many lncRNAs have been shown to exhibit very low expression levels (Necsulea and Kaessmann 2014). Implementation of our lncRNA detection pipeline (Fig. 3.3A) identifies 2,949 lncRNAs and 318 protein-coding genes (Additional Files 3.3 and 3.4). The number of lncRNAs we identify in *An. gambiae* is more than double the number identified in *D. melanogaster* and other members of the genus *Drosophila*, for which more than 1,000 long intergenic non-coding RNAs (lincRNAs) have been identified in each species, and many fewer than have been defined in studies of mice and humans, which have identified many thousands of potential lncRNAs (Sun et al. 2012; Derrien et al. 2012). As only long introgenic non-coding RNAs (lincRNAs) have been highly studied in *D. melanogaster,* the total number of lncRNAs may be comparable in *An. gambiae*. Additionally, our putative set of lncRNA genes is smaller than that recently described for the gut transcriptome of *An. gambiae* (Padrón et al. 2014). One of the major reasons for this difference in identified lncRNAs between the two studies is that Padron *et al*. (2014) did not use a peptide length cutoff, and their protein-coding potential analyses did not take advantage of whole genome alignments. By utilizing our peptide length cutoff on their lncRNA data set and only using Cufflinks codes 'I','U', and 'X', the number of lncRNAs identified from their data set is reduced by 62 percent, to 3,740 lncRNA. Among these, only 209 genes exhibit at least 50 percent sequence overlap between the two studies. This limited overlap indicates that tissue-specific RNAseq analysis can yield a vastly different lncRNA population

compared with whole organism RNAseq, which will be an important consideration for the eventual identification of a complete lncRNA gene set in *An. gambiae* and other vector insects.

Members of the lncRNA and putative protein-coding gene classes identified in our study have lower average FPKM levels and lower DNA sequence conservation, in general, than those observed for previously annotated *An. gambiae* protein-coding genes (Fig. 3.3C). This trend of lower observed levels of expression and sequence conservation may explain why genome annotation pipelines have previously missed the putative protein-coding genes that we have defined. In addition, the average number of exons per lncRNA is much lower than the average number of exons per novel protein-coding gene that we have identified in this study (Fig 3.3B). This is similar to the trend in exon number per transcript that has been characterized for human lncRNAs, which have been shown to possess significantly fewer exons per gene compared to protein-coding genes (Derrien et al. 2012).

Previous studies of lncRNA sequence evolution have indicated that primary sequence conservation is very low across tetrapods (Necsulea et al. 2014), while only a few such studies have considered conservation of secondary structure in assessing net evolutionary conservation of lncRNAs (Wood et al. 2013; Engström et al. 2006). Those studies that have considered secondary structure have focused mainly on comparisons between a few species and not on comparisons across complete lineages, such as is now possible within the *Anopheles* genus (Wood et al. 2013; Kutter et al. 2012; Engström et al. 2006). The

ability of RNA to maintain secondary structural features and associated RNA-protein interactions, even in the absence of primary sequence conservation (Necsulea et al. 2014; Kutter et al. 2012), may underlie, in part, the increased rate of divergence for lncRNAs that has been observed in these previous studies.

Our study illustrates that across the sequenced genomes within the genus *Anopheles*, 91 percent of previously annotated protein-coding genes in *An. gambiae* exhibit matching genomic regions in *An. albimanus* (Fig. 3.6). This level of conservation we observe is lower for the set of protein-coding genes we have newly annotated, e.g., 79 percent for *An. minimus* and 60 percent for *An. albimanus*. It is even lower for the lncRNA class, e.g., 77 percent for *An. minimus* and 20 percent for *An. albimanu*s. Furthermore, examining sequence conservation within these genomic regions using PhyloP p-values of conservation scores indicates that lncRNA sequences are much more divergent across the *Anopheles* genus, compared with previously and newly annotated protein-coding classes (Fig. 3.7A). The reduced numbers of identifiable conserved lncRNA-homologous genomic regions is in agreement with previous findings in tetrapods, which illustrated a rapid decrease in 1:1 orthologous lncRNA families across many classes of tetrapods (Necsulea et al. 2014). The proportions of lncRNAs that identify homologous genomic regions in our whole genome alignments are similar to the proportions of conserved protein-coding genes, when considering only the closely-related species within the Gambiae complex (Fig. 3.6). However, beyond the Pyretophorus Series, the proportions of conserved lncRNA-homologous regions decline much more rapidly than those for protein-coding genes. Those putative lncRNA-harboring genomic regions that are

identifiable in other species also show much higher levels of sequence divergence compared with protein-coding genes. Together, these results imply that anopheline lncRNAs diverge at a much higher rate than protein-coding genes. Accordingly, some *An. gambiae* lncRNAs present in the most recent common ancestor of the Pyretophorous Series and the Neocellia and Myzomyia Series, for example, may have diverged beyond recognition within the Neocellia and Myzomyia, while other *An. gambiae* lncRNAs may have arisen relatively recently and are therefore restricted to species within the Gambiae complex.

To extend our analysis beyond primary sequence conservation for lncRNAs, we employed REAPR to identify lncRNA secondary structures and analyze their conservation across the anophelines (Fig. 3.10, Fig. 3.11). Among all putative *An. gambiae* lncRNAs we define, only 28 percent exhibit high-confidence RNA secondary structures. Although it has been proposed that all lncRNAs should possess a functional secondary structure as this structure is what gives a lncRNA its function, this premise has not been validated at the genome-wide level for other sets of related organisms, nor has the conservation of lncRNA secondary structures across multiple related species in other clades been analyzed and described in comparable depth (Novikova et al. 2012; Mercer and Mattick 2013; Will et al. 2013; Smith et al. 2013). The closely related members of the Gambiae species complex, in which homologous genomic regions are found for almost all *An. gambiae* lncRNAs, all exhibit similar proportions of high-confidence RNA secondary structures within these lncRNAs. While these structures are highly conserved within the Gambiae species complex, the numbers of lncRNA secondary structures

conserved relative to *An. gambiae* decline rapidly for species outside of the complex, at an apparent rate even more pronounced than the decline in the numbers of conserved lncRNA-homologous genomic regions (Fig. 3.10A). However, when corrected for the root age of divergence for each species analyzed, we see that primary sequences and secondary structures exhibit similar rates of divergence (Fig. 3.10B). Both of these rates are much higher than those that have been described for lncRNAs in chordates (Necsulea et al. 2014). Increased divergence rates in insects, as compared to chordates, have been noted previously for protein-coding genes (Wyder et al. 2007; Richards et al. 2008). Rapid divergence of lncRNA sequences as compared to protein-coding genes (Fig. 3.6,3.10) has also been reported for rodent species (Kutter et al. 2012).

These differences in the number of conserved lncRNA regions and number of secondary structures across the anophelines, especially evident for those lncRNAs that exhibit conserved genomic regions in all species but secondary structures in only a subset of those species (Fig. 3.12), imply that lncRNA secondary structures tend to evolve after a most recent common ancestor for a given set of species has acquired transcriptional activation of particular genomic loci. This finding is consistent with the long-acknowledged idea of "pervasive transcription" across the genome (Jensen et al. 2013). Pervasive transcription describes the process by which most regions of the genome are transcribed, including those that fail to encode proteins or functional ncRNAs. Through random mutations, these "pervasive" transcripts acquire protein-coding ability or a functional RNA structure, over evolutionary time. Selective pressure causes these altered transcripts to become fixed within a population if they are advantageous for the organism.

Given the evolutionary interval between the onset of transcriptional activation of a particular genomic region and the time at which the transcript becomes functionally beneficial, some lineages/species that have evolved during that time period may express a particular pervasive transcript before it becomes a functionally beneficial transcript within that species or lineage.

Increased evolutionary rates of lncRNA sequences compared to protein-coding genes may contribute to bionomic diversity that has been observed across the genus *Anopheles* by affecting the evolution of species-specific behaviors, such as resting, mating and feeding patterns (Takken and Knols 1999; Paaijmans and Thomas 2011), just as behavioral control has begun to be attributed to variation among *Drosophila* lncRNAs (Soshnev et al. 2011). The notion that lncRNAs modulate the activities of protein-coding genes is well-established (Lee 2012; Fatica and Bozzoni 2014; Ponting et al. 2009). However, we speculate that lncRNA-mediated regulation of gene expression, coupled with the rapid evolution of lineage-specific lncRNA ensembles in mosquitos, may underlie the rapid diversification of vector mosquito behaviors (Pates and Curtis 2005b) for which it has been, thus far, difficult to define differentiating causal mechanisms. By utilizing SNPs in regions outside of protein-coding genes, we may be able to identify these casual variants that were once unknown. Our deep RNA sequencing of *An. gambiae* has provided the most comprehensive catalog of lncRNAs in mosquitoes to date, and presents the prospect of identifying a new generation of targets for approaches to vector control that will enable further reductions in the burden of human malaria.

METHODS:

Colony and Sequencing

Anopheles gambiae G3 colony (courtesy of Dr. Flaminia Catterucia, Harvard School of Public Health, Boston, MA, USA) was reared with an 11:11 Light:Dark (L:D) photoperiod with a one-hour crepuscular period between light and dark stages. Adults were fed 10 percent glucose solution *ad libitum*, and both genders were kept in the same cage. First larval instar (L1) and third larval instar (L3) stages were removed from the colony within 12 hours of emergence from chorion or previous larval cuticle, respectively. Adults were sampled three days post-emergence, and all samples were collected at approximately eight hours into the light cycle of the 11:11 LD photoperiod. All samples were kept in RNA-Later (Ambion, Austin, TX) until RNA extraction and sequencing. The L1 and L3 life stages were chosen because they represent early and late stages during larval development, which can be synchronized clearly, and because previous studies have defined a set of contigs that are differentially expressed between these stages (Koutsos et al. 2007). Future lncRNA discovery studies may include the pupal stage, due to its importance for the completion of morphogenesis that yields the adult mosquito.

High read depth (HRD) paired-end RNA sequencing was performed at the Broad Institute (Cambridge, MA) using a Qiagen RNAeasy Mini Kit for RNA extraction, poly-A tails were selected and the Illumina TruSeq RNA Sample Preparation Kit v2, and libraries were sequenced on the HiSeq 2000 platform. Low read depth (LRD) paired-end RNA sequencing of larval replicates was performed by Otogenetics Corp. (Atlanta, GA), using

the same protocol as the HRD samples. Low read depth adult single-end RNA sequencing data sets were obtained from Pitts *et al.* (2011). All RNA sequencing data produced have been submitted to the European Nucleotide Archive and can be accessed under the SRA Accession number of PRJEB5712.

RNAseq Read Alignment and Analysis

HRD RNAseq reads were soft clipped, and replicate RNAseq reads from Otogenetics Corp. were subsequently hard clipped by 10 bp on both the 5' and 3' ends of each read (Fig. 3.13). First, hard clipping of the LRD replicate samples was performed to reduce the number of potential adapter sequences, even though read quality scores were high overall, as the reads were long enough to support such hard-clipping (~100 bp in length). Second, clipping the reads makes their length more comparable to other replicate reads from Pitts *et al.* (2011) that were trimmed as previously described. Reads were aligned to the *An. gambiae* AgamP3 genome assembly, which was softmasked using RepeatMasker (www.vectorbase.org) (Smit et al.; Megy et al. 2012).. Alignment, transcriptome assembly and analyses were performed using the Tuxedo Suite (Kim et al. 2013; Trapnell et al. 2013, 2010), which comprises Tophat2, Cufflinks, Cuffmerge and Cuffdiff2 programs, Scripture and Sailfish (Guttman et al. 2010; Patro et al. 2014). Splice junction mapping was performed using Tophat2 (version 2.0.10) with a mismatch (-N) appropriation of 3 and a read-edit-dist of 3. Cufflinks (version 2.1.1) was run with default settings using the *An. gambiae* AgamP3.7 annotation –gtf function and a reference annotation-based transcript (RABT) assembly. Scripture (Beta-2 version) was run using default settings. Cuffmerge was used to combine and filter artifacts from the

resulting transcriptome assemblies from Cufflinks, Scripture and the reference *An. gambiae* AgamP3.7 annotation. Cuffdiff2 was used to determine differentially expressed genes of interest with an FDR of 0.05 and the –u (multi-read correct) function, and differentially expressed genes were determined using the Benjamini-Hochberg correction, with two replicates for each life stage (HRD and LRD for each stage). In order to validate the FPKM (fragments per kilobase of exonic length per million reads) values produced by the Tuxedo Suite, Sailfish was used to compare values. Sailfish was run with default parameters and the average RPKM (reads per kilobase exonic length per million reads mapped) was compared to FPKM values determined using Cufflinks.

Identification of Newly Annotated Transcripts

HRD RNAseq data sets for all four stages and genders (L1, L3, Male, Female) were combined and aligned using Tophat2, as previously described (Kim et al. 2013). Cufflinks and Scripture were subsequently used to identify newly annotated transcripts. Cuffcompare was used to compare newly annotated transcripts to the *An. gambiae* AgamP3.7 gene set. To identify probable lncRNAs, class codes "I", "U" and "X" were used in Cufflinks (as this study does not aim to identify potential novel isoforms of known protein-coding genes, the "J" class was not utilized).

***Anopheles* Genome Alignments and PhyloCSF Scanning for Protein-Coding Potential**

A set of 21 available *Anopheles mosquito g*enome assemblies species were retrieved from VectorBase (Megy et al. 2012). These included assemblies of *An. gambiae* PEST (Holt et

al. 2002), *An. gambiae* Pimperena S form and *An. coluzzii* (formerly *An. gambiae* M form) (Lawniczak et al. 2010), the species sequenced as part of the *Anopheles* 16 Genomes Project (Neafsey et al. 2014), *An. darlingi* (Marinotti et al. 2013), and the South Asian species *An. stephensi* (Jiang et al. 2014). Details of assemblies used can be found in Table 3.2. Multiple whole genome alignments of 21 available *Anopheles* assemblies were built using the MULTIZ feature of the Threaded-Blockset Aligner suite of tools (Blanchette et al. 2004), employing a similar approach to that used for other multi-species whole genome alignments such as those for 12 *Drosophila* (Stark et al. 2007) and 29 mammal (Lindblad-Toh et al. 2011) genomes. Before computing the alignments, repetitive regions within each of the input genome assemblies were masked. Assemblies were analysed using RepeatModeler (Smit et al.) to produce repeat libraries that were then combined with known repeats from *An. gambiae* and retrieved from VectorBase, before being used to mask each genome assembly using RepeatMasker (Smit et al.). The 21-species maximum likelihood phylogeny, required to guide the progressive alignment approach of MULTIZ, was estimated using RAxML (Stamatakis 2014) from the concatenated protein sequences of Genewise (Birney et al. 2004) gene predictions using Benchmarking sets of Universal Single-Copy Orthologs (BUSCOs) from OrthoDB (Waterhouse et al. 2013), and rooted with predictions from the genomes of *Aedes aegypti* (Nene et al. 2007) and *Culex quinquefaciatus* (Arensburger et al. 2010). The MULTIZ approach first runs all-against-all pairwise LASTZ alignments (default settings), followed by projections ensuring that the reference species is "single-coverage," with projection steps guided by the species dendrogram to progressively combine the alignments.

Examining patterns of evolutionary conservation across multiple whole genome alignments can help to distinguish protein-coding regions from non-protein-coding regions, e.g., as in the analyses of 12 *Drosophila* (Stark et al. 2007) and 29 mammal (Lindblad-Toh et al. 2011) genomes. Specifically, PhyloCSF (Lin et al. 2011) is a method developed to determine whether a multi-species nucleotide sequence alignment represents a protein-coding region, based on patterns of evolutionary conservation such as codon substitution frequencies (CSF). Thus, PhyloCSF can be used to help distinguish protein-coding and non-coding RNAs represented among new transcript models obtained from high-throughput transcriptome sequencing. Gene transfer format (GTF) files (from Cuffmerge output) defined the required genomic intervals for PhyloCSF analyses per codon, per exon, and per gene. Per-codon analysis scanned each transcript region (plus flanking 50 bp) in the six translational frames to score for protein-coding potential across the entire region. Per-exon analysis identified the best-scoring translational frame for the length of each exon, and per-gene analysis identified the best-scoring, start-codon-to-stop-codon open reading frame of the complete annotated transcript.

Coding transcripts were classified as those new transcripts that possess an open reading frame >100 amino acids in length and a PhyloCSF score greater than ten (i.e., 10 times more likely to be coding than non-coding). Non-coding transcripts were classified as those novel transcripts that possess a maximum open reading frame < 50 amino acids in length, an open-reading frame that is < 35 percent of the total transcript length, a PhyloCSF score less than negative ten, and no recognizable domains as defined by PFAM, TIGRFAM or SUPERFAMILY libraries (Finn et al. 2014; Gough et al. 2001;

Haft et al. 2003), which were searched using HMMER with default settings for e-value cutoffs (website version 1.9) (Finn et al. 2011).

Differential Gene Expression and Categorization

Using the Cuffdiff function as described above, differentially expressed (DE) genes were defined using a false discovery rate of 0.05. Gene Ontology (GO) terms (Consortium 2000) were extracted for those DE genes from VectorBase (Megy et al. 2012). These GO terms were grouped by GO_Slim2 categories with CateGOrizer (Hu et al. 2008). To define the groups or classes of genes that are DE, DAVID (Huang et al. 2009) was utilized to determine enrichment scores. DE genes were compared in order to define genes that were up/down-regulated, regardless of adult gender and regardless of larval life stage.

Determining Conservation and Secondary Structure of Newly Annotated Genes Across *Anopheles* Lineages

In order to quantify the sequence conservation of the lncRNA and newly annotated protein-coding classes of genes, we employed PhyloP. First, PhyloFIT, part of the PHAST package (version 1.3) (Hubisz et al. 2011), was utilized to create a nonconserved substitution model from the multiple genome alignments, using four-fold degenerate sites. Using PhyloP, part of the same PHAST package, the p-value of conservation was then calculated for all genes identified in this study or for genes in the *An. gambiae* AgamP3.7 annotation release, for comparisons. For analysis, only newly annotated genes that had strandedness predicted by Cufflinks were used.

REAPR (**rea**lignment for **p**rediction of structural non-coding **R**NA) was utilized to determine secondary structure scoring of identified lncRNA class members using the RNAz score (Will et al. 2013) . Realignment of the lncRNA genes using REAPR was performed using a delta value of 15 and the --alistat functions. For confident secondary structures, only loci possessing RNAz scores over 0.5 were used, as these correspond to an FDR of ~ 0.04 as described in RNAz 2.1 documentation (Gruber et al. 2010).

Rate of degradation of number of secondary structures and conserved genomic regions was determined using a linear regression and ANCOVA test to determine significance . Analyses were performed using GraphPad Prism 5.0b for Mac, GraphPad Software, San Diego, California USA, www.graphpad.com

Availability of Supporting Data

The data sets supporting the results of this article are available in the European Nucleotide Archive, under accession PRJEB5712 (http://wwwdev.ebi.ac.uk/ena/data/view/PRJEB5712). All files produced by Scripture, PhyloP and REAPR, along with all whole genome alignment and gene alignment files, can be accessed freely at http://bioinformatics.bc.edu/~jenkinad/.

TABLES/FIGURES AND LEGENDS

Table 3.1: Read Alignment of RNA-Sequencing Data Sets

Table of number and percentage of reads mapped for each life stage (1st instar, 3rd instar, male and female) at either a high read depth (HRD) or low read depth (LRD).

Data Set	Raw Read Count	Percentage Mapped	Aligned Read Count
HRD 1st Instar	184,145,330	81.2%	149,517,068
HRD 3rd Instar	143,507,360	76.7%	110,094,659
HRD Female	184,150,422	75.6%	139,217,446
HRD Male	194,179,892	76.8%	149,210,510
LRD 1st Instar	32,425,540	79.8%	25,888,403
LRD 3rd Instar	38,489,668	81.2%	31,269,540
LRD Female	27,877,821	86.7%	24,160,317
LRD Male	31,876,060	82.1%	26,162,196

Table 3.1: Read Alignment of RNA-Sequencing Data Sets

Table 3.2: Genomes Utilized for Whole Genome Alignments and Associated *Anopheles* Species

The assembly names for each genome utilized in this study, along with the species named for the given assembly.

Species	Assembly
Anopheles gambiae PEST	AgamP3
Anopheles gambiae Pimperena S form	AgamS1
Anopheles coluzzii Mali-NIH M form	AgamM1
Anopheles merus	AmerM1
Anopheles arabiensis	AaraD1
Anopheles quadriannulatus A	AquaS1
Anopheles melas	AmelC1
Anopheles chrysti	AchrA1
Anopheles epiroticus	AepiE1
Anopheles minimus A	AminM1
Anopheles culicifacies A	AculA1
Anopheles funestus	AfunF1
Anopheles stephensi	AsteS1
Anopheles stephensi	AsteI2
Anopheles maculatus B	AmacM1
Anopheles farauti	AfarF1
Anopheles dirus A	AdirW1
Anopheles sinensis	AsinS1
Anopheles atroparvus	AatrE1
Anopheles darlingi	AdarC2
Anopheles albimanus	AalbS1

Table 3.2: Genomes Utilized for Whole Genome Alignments and Associated *Anopheles* Species

Table 3.2: Number of 1:1 Conserved lncRNA Regions in Each Anopheline Genome Assembly:

All number of conserved regions are based upon LASTZ identification during WGA alignment.

Species' Genome	Number of 1:1 lncRNA
gambiae (PEST)	2949
gambiae (S-Form)	2729
gambiae (M-Form)	2694
arabiensis	2739
quadriannulatus	2714
merus	2743
melas	2691
christyi	2398
epiroticus	2431
stephensi (I2)	2101
stephensi (S1)	2091
maculatus	1515
culicifacies	2130
minimus	2179
funestus	2176
dirus	1675
farauti	1555
atroparvus	1017
sinensis	877
albimanus	588
darlingi	505

Table 3.2: Number of 1:1 Conserved lncRNA Regions in Each Anopheline Genome Assembly:

Table 3.4: Number of High-Confidence lncRNA Secondary Structures in Each Anopheline Genome Assembly:

Number of high-confidence lncRNA (RNAz Score > 0.50) secondary structures identified in each *Anopheles* genome.

Species' Genome	Number of Secondary Structures
gambiae (PEST)	1129
gambiae (S1)	1091
gambiae (*coluzzi*, M1)	1027
arabiensis	1077
quadriannulatus	1060
merus	1072
melas	1000
christyi	704
epiroticus	664
stephensi (I2)	381
stephensi (S1)	377
maculatus	207
culicifacies	379
minimus	423
funestus	420
dirus	238
farauti	195
atroparvus	87
sinensis	58
albimanus	32
darlingi	21

Table 3.4: Number of High-Confidence lncRNA Secondary Structures in Each Anopheline Genome Assembly:

Figure 3.1: Validation of RNA-Seq Library and Analysis Techniques

A. Life stage comparison of Cufflinks FPKM values to Sailfish RPKM values. Pearson's correlation coefficient is represented for each life stage comparison. Genes used for comparison are those annotated in VectorBase release Agam3.7. B. Clustered FPKM expression (Additional File 3.1) of differentially expressed genes between life stages in *An. gambiae*. Rows and columns were clustered using Pearson correlation method with complete linkage distances. C. DAVID enrichment scores for differentially expressed gene groups between life stage comparisons.

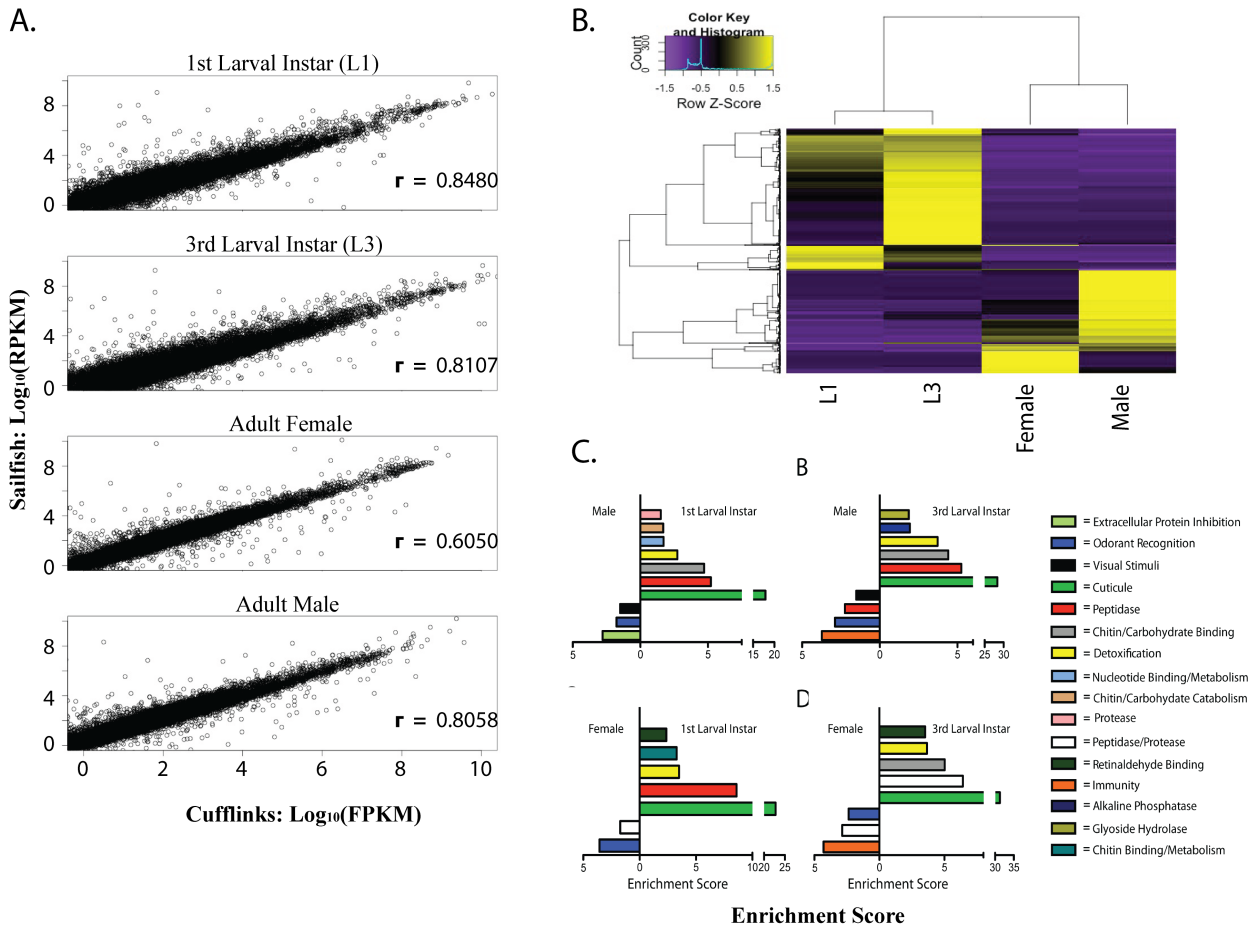


Figure 3.1: Validation of RNA-Seq Library and Analysis Techniques

Figure 3.2: GOSLIM2 Terms of Genes that Exhibit Differential Expression Among Life Stages/Genders

Differentially expressed genes for each pairwise life stage comparison (as indicated on the x-axis) grouped using CateGOrizer into GOSLIM2 terms (Hu et al. 2008). Numbers at top of each group indicate number of differentially expressed genes for the comparison in either the up- or down-regulated direction. Each category is represented as the percentage of total GOSLIM2 terms grouped. The "Less Than 2%" category represents GOSLIM2 categories that represent less than 2% of the total terms grouped for a given comparison. Categories not within this group represent more than 2% of the total genes grouped for a given comparison.

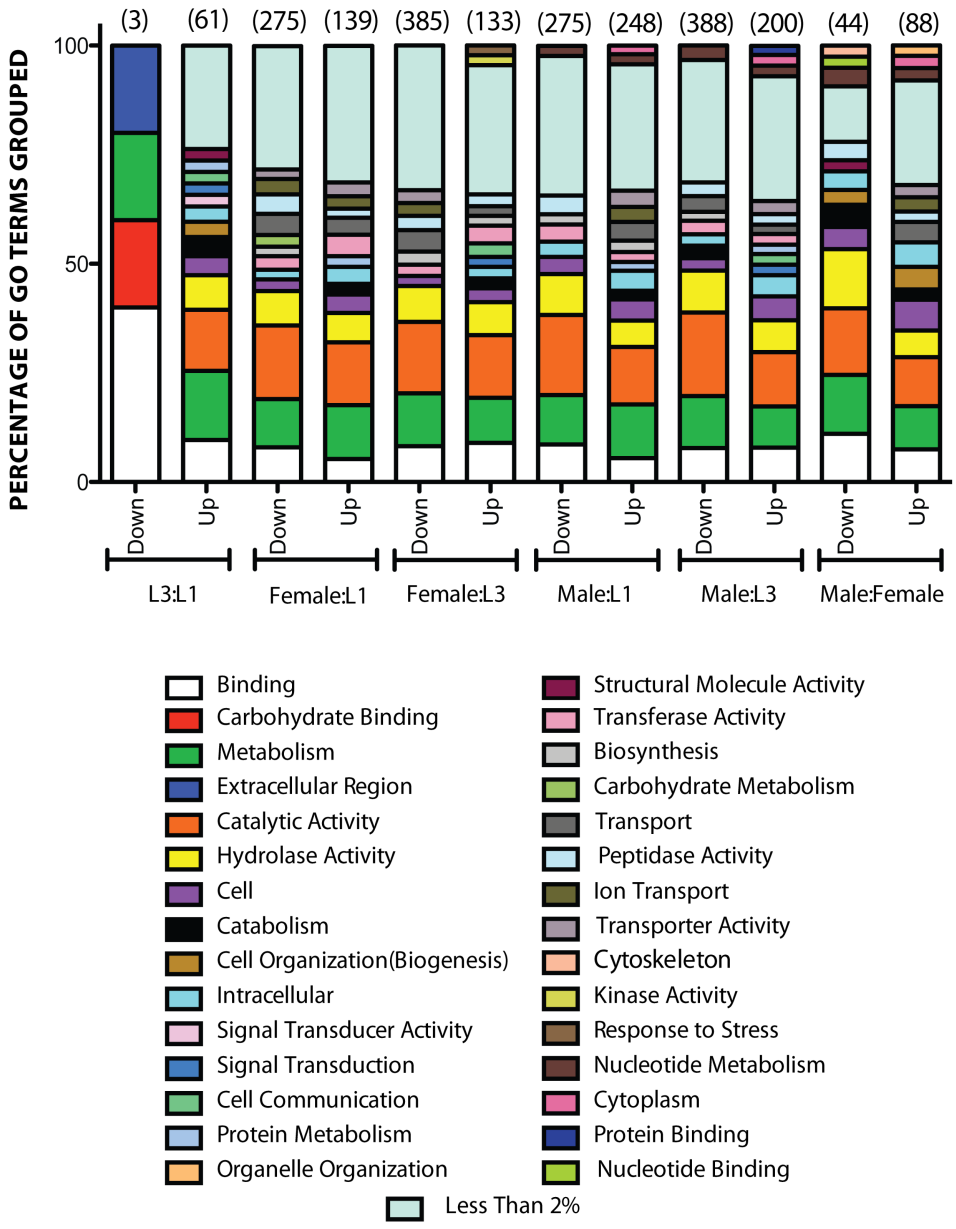


Figure 3.2: GOSLIM2 Terms of Genes that Exhibit Differential Expression Among Life Stages/Genders

Figure 3.3: Flow Chart of lncRNA and Potential Coding Gene Identification and Expression/Exonic Structure of Defined Gene Classes

A. Flow chart of lncRNA and novel protein-coding gene identification. RNAseq data sets were merged and used to produce a transcriptome that was supported by both Cufflinks and Scripture. Length, PhyloCSF score, maximum peptide length, protein domain and total coding-sequence length were used to set inclusion and exclusion criteria for the sets of lncRNAs and putative protein-coding RNAs, among the previously unannotated transcripts. **B.** Density plot of exons per-gene for lncRNAs (blue) and novel protein-coding RNAs (red). **C.** Expression values [Log_{10} (FPKM+1)] calculated by Cufflinks for previously annotated genes in VectorBase (red), lncRNAs (green), and newly identified putative protein-coding RNAs (blue) for all genes that had an FPKM greater than zero for the merged RNAseq data set.

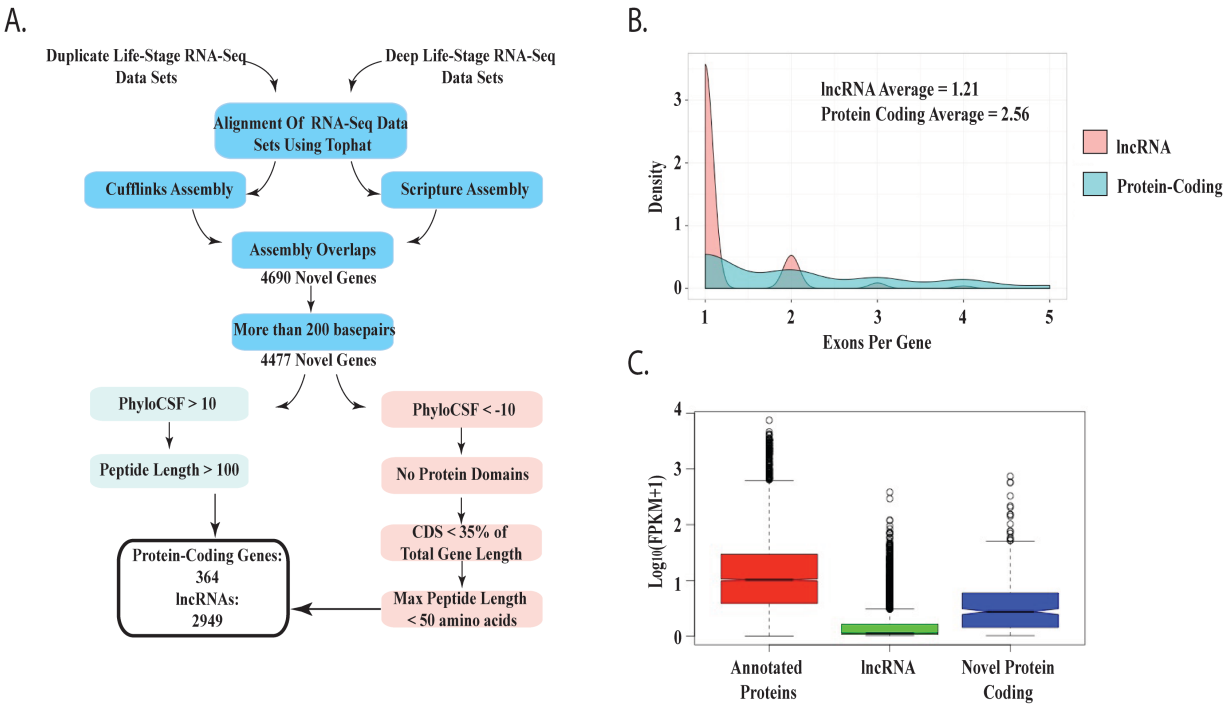


Figure 3.3: Flow Chart of lncRNA and Potential Coding Gene Identification and Expression/Exonic Structure of Defined Gene Classes

Figure 3.4: Examples of Newly Annotated Protein-Coding and lncRNA Genes

Read count profiles of RNAseq alignments to a selected set of newly annotated genes, viewed using IGV (Broad Institute, Cambridge, MA) (Thorvaldsdóttir et al. 2013; Robinson et al. 2011). Chromosomal coordinate scales vary among panels. AGAP designations are given for genes encoding mRNAs (blue boxes for exons) that are complementary to newly annotated antisense lncRNAs (green boxes for exons). Strandedness of lncRNAs is determined by Cufflinks and based on output GTF file (Additional File 3.3). Each panel consists of the top graph indicating read depth (Log scale maximum of 6) with a PhyloCSF track below (scale -70 to 50, red indicating values above 0 and blue indicating values below 0), followed by the gene GTF track. Colored triangles indicate the orientation of the given gene. **A.** Putative protein-coding gene Merged.4500.1 maps antisense to the 3' untranslated region of protein-coding gene AGAP007209. Regions with red boxes of Merged.4500.1 indicate the protein-coding segments of the gene (107 amino acids in length). **B.** lncRNA Merged.6207.1 maps intronically with respect to AGAP002451. **C.** lncRNA Merged.11296.1 is antisense and overlapping to AGAP011074.

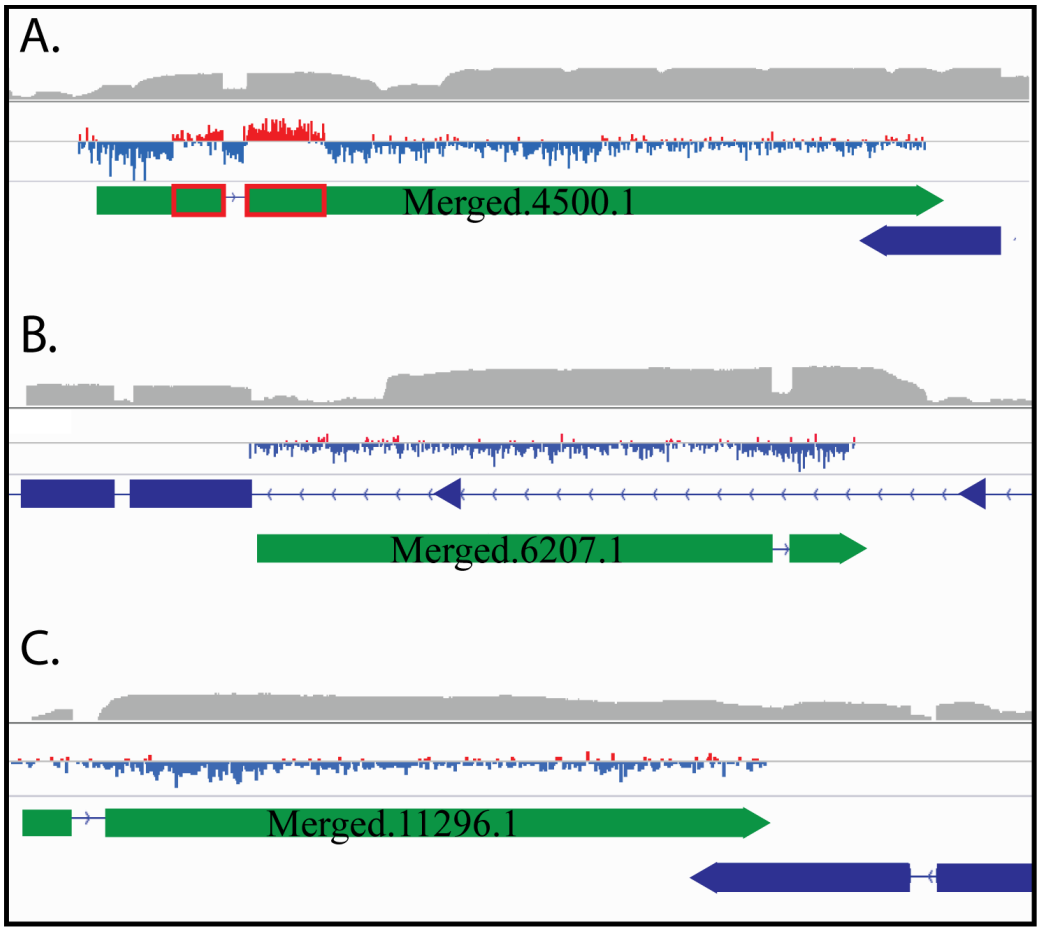


Figure 3.4: Examples of Newly Annotated Protein-Coding and lncRNA Genes

Figure 3.5: lncRNAs that Exhibit Differential Expression Among Life Stages/Genders

Row Z-score expression (FPKM) of differentially expressed lncRNAs, as determined by Cuffdiff2 (Trapnell et al. 2013), between life-stages in *An. gambiae*. Rows were clustered using Pearson correlation method with complete linkage distances (see Materials and Methods)(Supp. File 7)

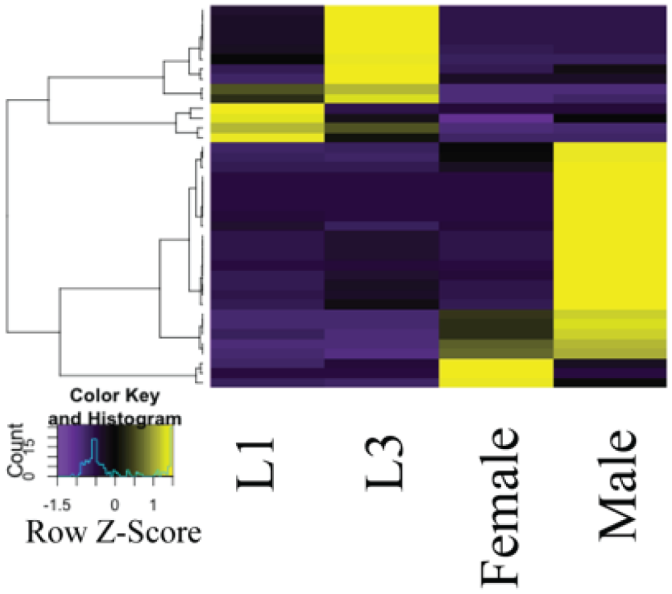


Figure 3.5: lncRNAs that Exhibit Differential Expression Among Life Stages/Genders

Figure 3.6: Evolutionary Conservation Across the Genus Anopheles

Percentage of previously annotated protein-coding genes (left column), newly annotated protein-coding genes (this study, middle column) and newly annotated lncRNAs (this study, right column) that could be aligned among *An. gambiae* and other comparator species using whole genome alignments. Percentages represent percent of total gene class that could be aligned to the genome of each species (heatmap colors are depicted in legend). Number of models for each class of gene, for *An. gambiae*, listed at the top of each column.

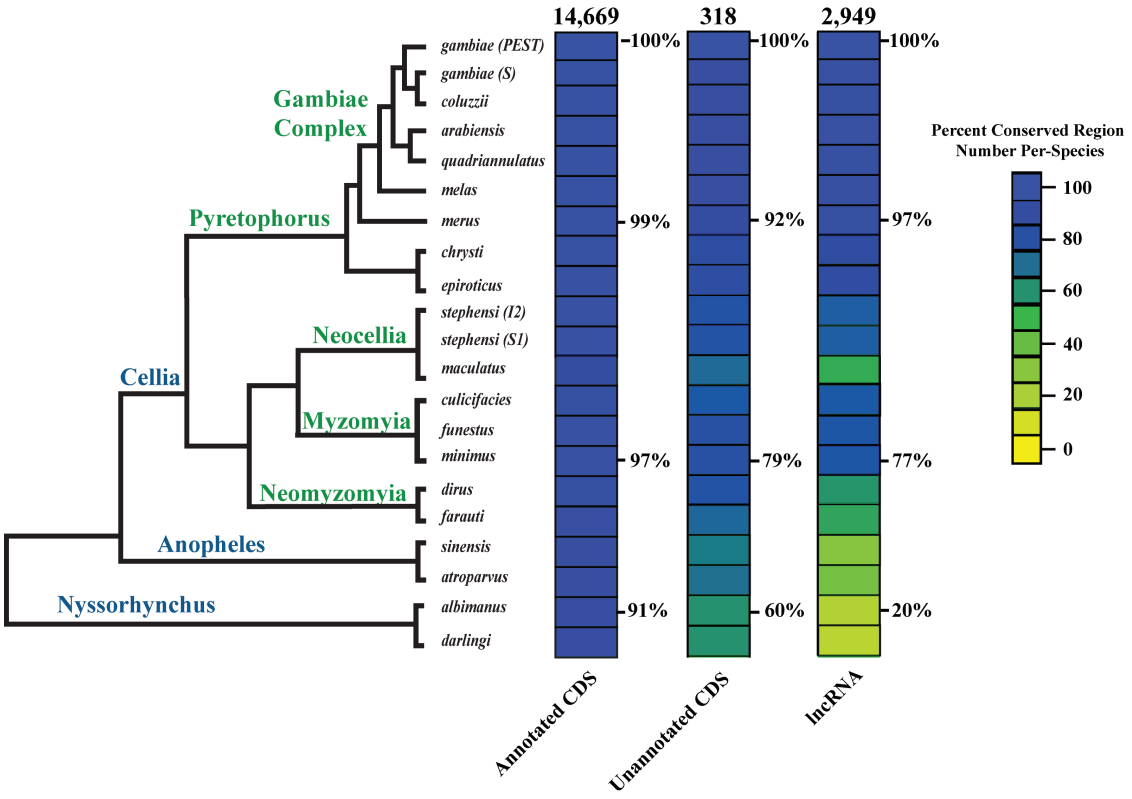


Figure 3.6: Evolutionary Conservation Across the Genus Anopheles

Figure 3.7: Sequence, structural and expression profiles of identified gene classes

A. Characterization of sequence conservation across the genus *Anopheles* performed using PhyloP. The $-\log_{10}$(PhyloP Conservation P-Value) was calculated for each gene within each respective gene class and statistical significance was determined using a Mann-Whitney T-Test. Starred bars denote p-value <0.001. **B.** Stacked histogram of RNAz score output from REAPR analysis (delta value of 10) for lncRNA (red bars) and novel protein-coding genes (blue bars). Insert shows confident RNA secondary structure calls with an RNAz score above 0.5.

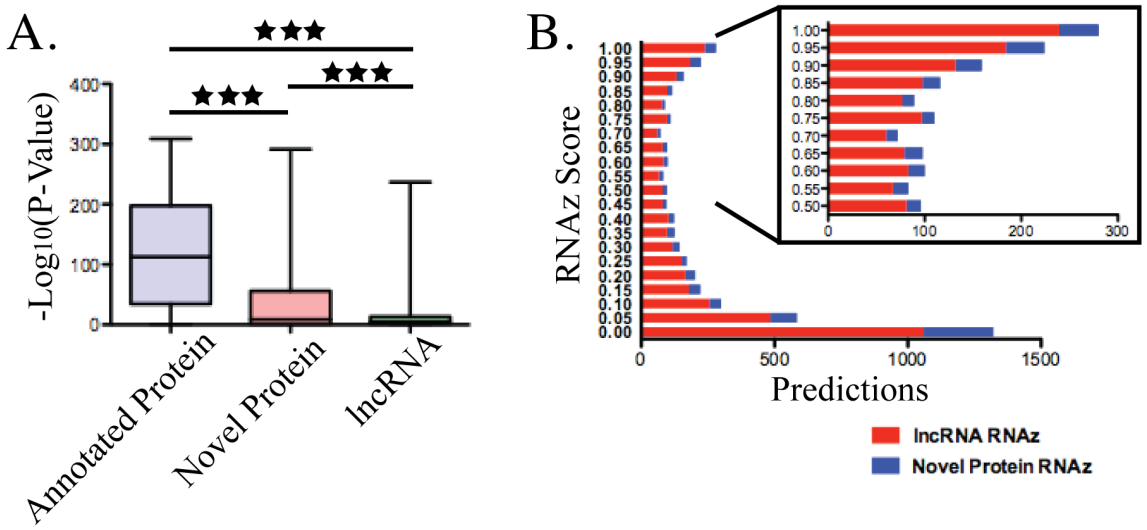


Figure 3.7: Sequence, structural and expression profiles of identified gene classes

Figure 3.8: RNAz Scores of Secondary Structures in lncRNA and Novel Protein Coding Genes After REAPR Realignment

RNAz scores for loci identified during REAPR analysis (Will et al. 2013). RNAz scores were calculated using a delta value of 10 for secondary structure realignment based on original whole genome alignments.

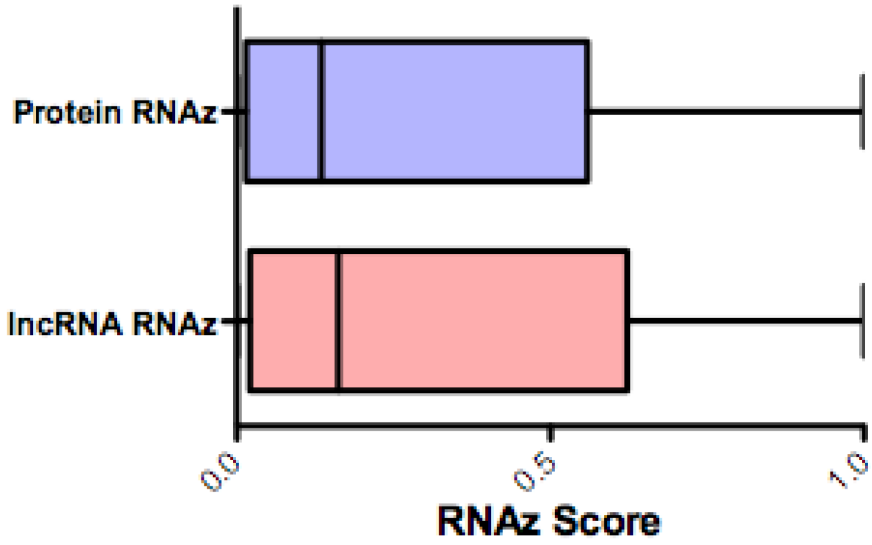


Figure 3.8: RNAz Scores of Secondary Structures in lncRNA and Novel Protein Coding Genes After REAPR Realignment

Figure 3.9: Secondary Structures for a Differentially Expressed lncRNA

Differentially expressed lncRNA Merged.20523.1 is shown with the gene structure and coordinates on the X-chromosome (visualized using IGV) (Thorvaldsdóttir et al. 2013; Robinson et al. 2011). REAPR analyses reveal multiple high confidence secondary structure loci within the gene, four of which are depicted. RNA secondary structures were visualized using VARNA (Blin et al. 2009).

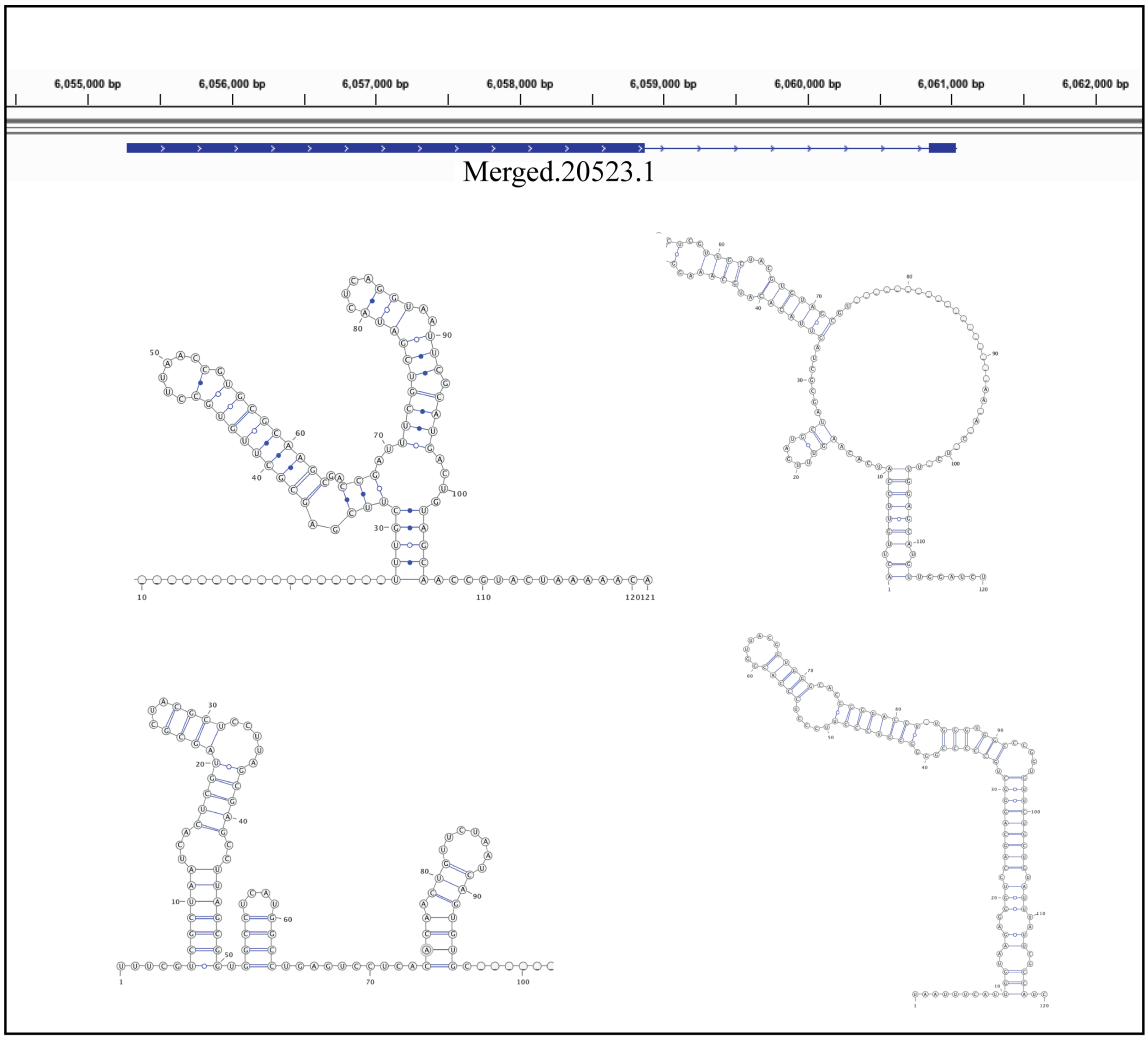


Figure 3.9: Secondary Structures for a Differentially Expressed lncRNA

Figure 3.10: Conservation of lncRNA predicted secondary structure and genomic regions across the *Anopheles* genus

A. The number of lncRNA secondary structures and conserved genomic regions per-species that are present within members of the *Anopheles* genus in relation to *An. gambiae*. Plots represent RNA sequences that possess high confidence (RNAz score > 0.5) secondary structures as identified during REAPR analysis (left, blue line) and conserved genomic regions of the lncRNA gene set (right, red line). For each species in the lineage (phylogenetic tree indicates species), the relative width of the plot corresponds to the number of confident RNA secondary structures or number of conserved genomic regions that were predicted. **B.** Change in the number of conserved genomic regions (red) and secondary lncRNA structures (blue) over time. Root age of divergence times were determined by Neafsey *et al.* 2014 (Neafsey et al. 2014) and lines represent linear regression. Differences in slopes between the linear regression lines are not significant based upon an ANCOVA test (P-value=0.09).

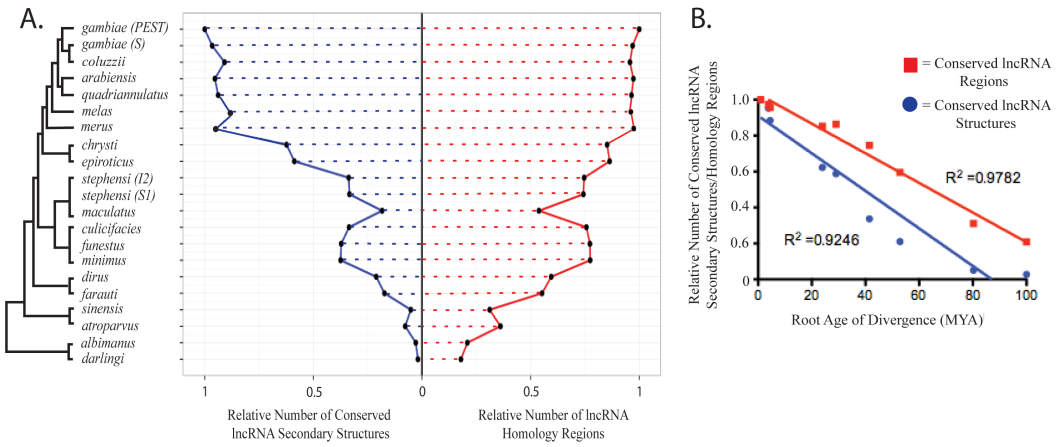


Figure 3.10: Conservation of lncRNA predicted secondary structure and genomic regions across the *Anopheles* genus

Figure 3.11: Histogram of Number of Genomes Aligned to For High-Confidence Secondary Structure

Distribution of the numbers of genomes aligned for each stable RNA secondary structure locus identified during REAPR analysis. REAPR analyses were performed using a delta value of 10, and a high-confidence secondary structure cutoff was placed at a value of 0.5, as described in previous RNAz publications (Gruber et al. 2010).

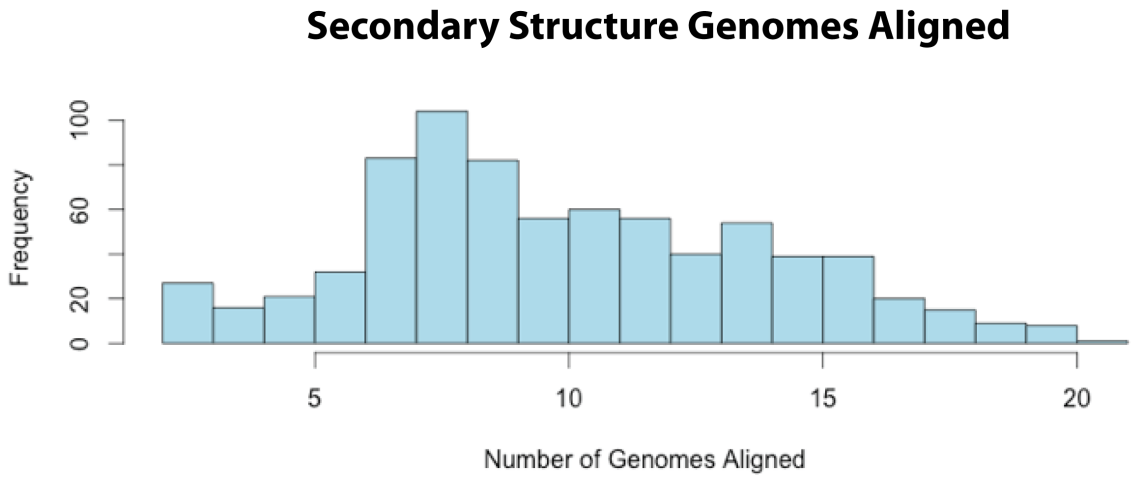


Figure 3.11: Histogram of Number of Genomes Aligned to For High-Confidence Secondary Structure

Figure 3.12: Clustering of Conserved Secondary Structures in lncRNAs that are Present in All *Anopheles* Species:

Of the 293 lncRNAs for which we identify conserved genomic regions in the genome assemblies analyzed, a subset of 90 include a total of 164 distinct secondary structures. These 164 structures were clustered based on presence (yellow) or absence (purple) in each assembly. Each structure was clustered using Pearson correlation method with complete linkage distances. Dendrogram on y-axis indicates the hierarchical clustering relationships. Genome names on x-axis correlate to the species name listed in Table 2.

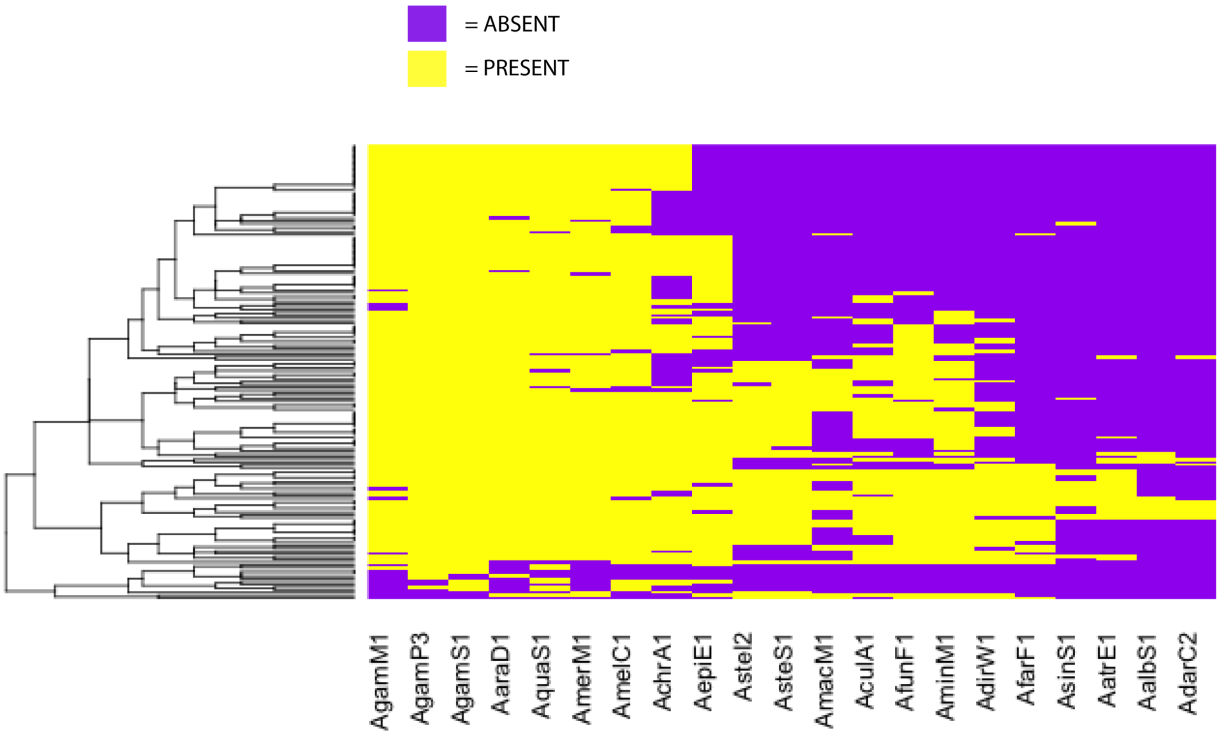
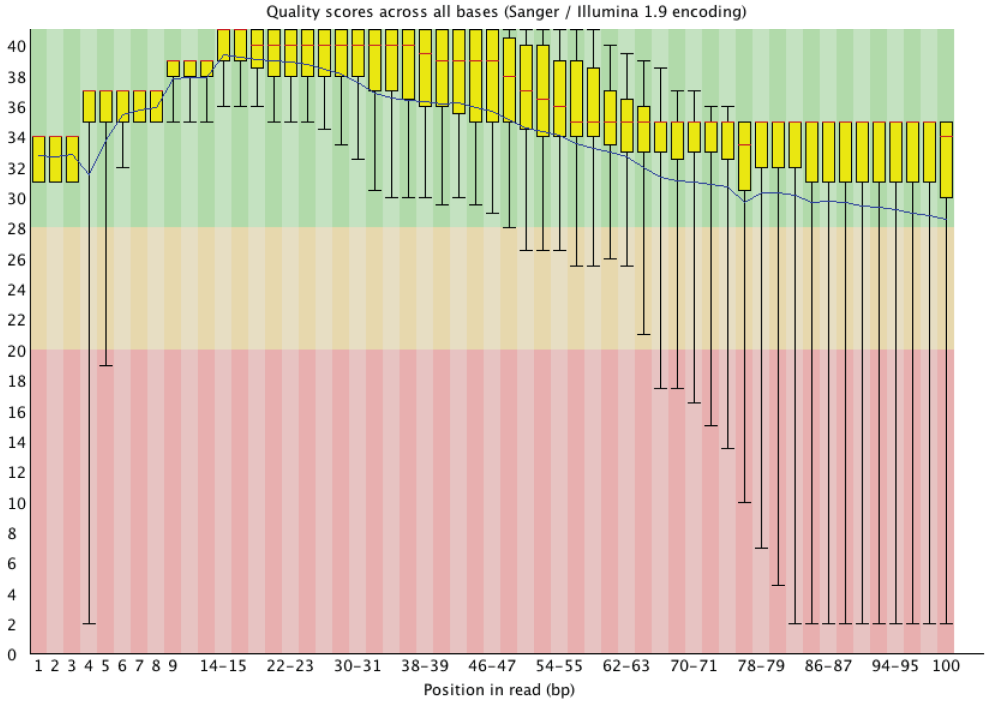


Figure 3.12: Clustering of Conserved Secondary Structures in lncRNAs that are Present in All *Anopheles* Species:

Figure 3.13: Representative Quality Scores of LRD Samples

A. Quality scores of L1 RNAseq reads before trimming. Visualized using FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) **B.** Quality scores of L1 RNAseq reads after trimming 10 nucleotides from each end of the read.

A.



B.

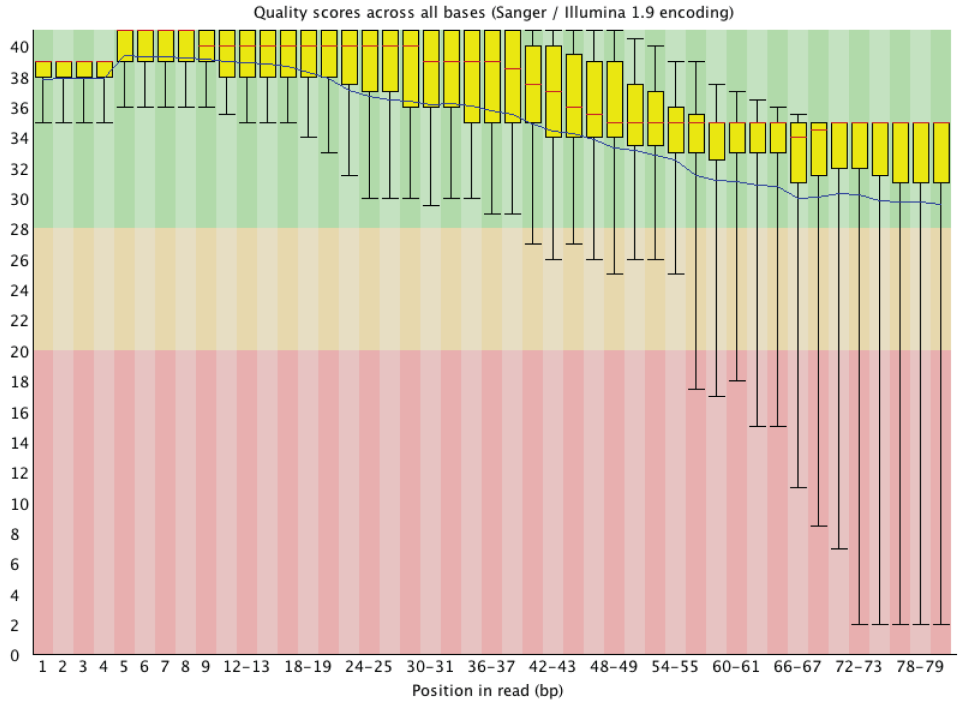


Figure 3.13: Representative Quality Scores of LRD Samples

Chapter IV:

Evolution of an Epigenetic Gene Ensemble within the Genus *Anopheles*

ABSTRACT

Epigenetic control of gene expression has important implications for the regulation of developmental processes, for mediating homeostasis and responses to the external environment, and for transgenerational inheritance of gene expression patterns. Genes that mediate epigenetic control have been well-characterized in *Drosophila melanogaster*, and we have identified and analyzed an orthologous gene ensemble in *Anopheles gambiae* that comprises 169 orthologs related to a 215-member epigenetic gene ensemble in *D. melanogaster*. We find that this ensemble is highly conserved among anopheline mosquitos, as we identify only seven gene family expansion/contraction events within the ensemble among 12 mosquito species we have studied within the genus *Anopheles*. Comparative analyses of the epigenetic gene expression across the genera *Drosophila* and *Anopheles* reveal distinct tissue-associated expression patterns in the two genera, but similar temporal expression patterns. The *An. gambiae* complex and *D. melanogaster* subgroup epigenetic gene ensembles exhibit similar evolutionary rates, as assessed by their respective dN/dS values. These differences in tissue-associated expression patterns, in contrast to similarities in evolutionary rates and temporal expression patterns, may imply that some members of the epigenetic gene ensemble have been redeployed within one or both genera, in comparison to the most recent common ancestor of these two clades. Members of this epigenetic gene ensemble may constitute another set of potential targets for vector control and enable further reductions in the burden of human malaria, by analogy to recent success in development of small molecule antagonists for mammalian epigenetic machinery.

INTRODUCTION

Genome regulation by epigenetic modulation is crucial for many biological processes including development, differentiation, homeostasis, responses to environmental variation and inheritance of gene expression patterns through generations (Kiefer 2007; Cantone and Fisher 2013; Lunyak and Rosenfeld 2008; Meissner 2010; Greer et al. 2011). Epigenetic control of gene expression via histone acetylation and methylation, and DNA methylation, mediates compaction and decompaction of DNA within euchromatic and heterochromatic chromatin (Guil and Esteller 2009; Greer and Shi 2012). The extent of chromatin condensation is often dependent on the extent of specific post-translational modifications to histone tails within nucleosomes (Bártová et al. 2008; Zhou et al. 2011). For instance, regulation of developmentally associated genes is controlled by Polycomb- and Trithorax-Group proteins (Schuettengruber et al. 2007; Bracken and Helin 2009; Schwartz and Pirrotta 2007), which have been well-characterized in *Drosophila melanogaster* (Swaminathan et al. 2012; Schuettengruber et al. 2009; Kennison 1995), and other epigenetic modulators. More recent studies have begun to explore the interplay of epigenetic mechanisms with gene family expansion and evolutionary diversification that enables the acquisition of new functions by paralogous gene family members, through divergence in response to selection (Branciamore et al. 2014; Park and Lehner 2014; Sui et al. 2014; Klironomos et al. 2013; Furrow and Feldman 2014; Keller and Yi 2014).

D. melanogaster has long constituted a model for studies of epigenetic gene regulation because of the extensive genetic tool set available for the species (Lyko et al. 2006) and because the deep genetics of the Bithorax-Complex and other *Drosophila*

developmental genes led to the early discovery of *Polycomb*, *trithorax* and many other genes that have been shown to be central to epigenetic regulation and modulation of chromatin states via histone modification (Gu and Elgin 2013; Kharchenko et al. 2011; van Bemmel et al. 2013; Schulze and Wallrath 2007; Vermaak and Malik 2009; Swaminathan et al. 2012; Zhou et al. 2013; Foglietti et al. 2006; Filion et al. 2010). In contrast, evolution of DNA methylation within the genus *Drosophila* has been investigated based on the presence of a single methytransferase gene, *Dmnt2*, compared to the multiple DNA methyltransferases found in vertebrates (Marhold et al. 2004). Other studies have implicated DNA methylation and histone modification patterns in the differentiation of caste systems in social insects (Weiner and Toth 2012; Hunt et al. 2013; Elango et al. 2009). While these studies have often compared genes of interest to orthologs in model or highly studied organisms (e.g., *Homo sapiens*), few comparisons of epigenetic gene ensembles have been conducted among dipteran species, including species within the malaria vector genus *Anopheles* (Arrowsmith et al. 2012; Talbert et al. 2012; Gregoretti et al. 2004). The pan-genomic homology between *D. melanogaster* and *Anopheles gambiae* gene sets has been well-characterized (Zdobnov et al. 2002) and has been leveraged for the identification and curation of orthologous and paralogous genes in *An. gambiae*, as well as for evaluating rates of gene evolution since the divergence of these two dipteran clades (Dottorini et al. 2007; Gregoretti et al. 2004).

We have defined the membership and rates of evolution for the first comprehensive epigenetic gene ensemble to be described in *An. gambiae, as* compared to *D. melanogaster*. We have identified *An. gambiae* genes orthologous to more than 75 percent of the *D. melanogaster* epigenetic gene ensemble. Our analysis of the *An.*

gambiae epigenetic gene ensemble across the genus *Anopheles* reveals very few gene family expansion and contraction events (i.e., four expansion and three contraction events). Different tissue-associated gene expression profiles we detect for members of *An. gambiae* and *D. melanogaster* ensembles imply that a subset of epigenetic genes may have been redeployed since the divergence of these two dipteran clades to mediate differing mechanisms of developmental and behavioral control, coinciding with the existence of many biological differences between these species (i.e. blood feeding, mating behavior). Our analyses provide strong support for the premise that epigenetic control mechanisms are conserved among Anopheline and Drosophilid species, and invite speculation regarding the existence of potentially insecticidable targets among the epigenetic gene ensembles of *An. gambiae* and other vector insects.

RESULTS

Defining an Epigenetic Gene Ensemble in *An. gambiae*

As the basis for defining an epigenetic gene ensemble in *An. gambiae*, we first identified a comprehensive epigenetic gene set in *D. melanogaster*, as described in MATERIALS AND METHODS (Fig. 4.1). This strategy was motivated by the well-annotated nature of the *Drosophila* genome, the genetic and functional characterizations of many epigenetic modifiers within its genome, and the proximate phylogenetic relationship between these two dipteran species (Lyko et al. 2006; St Pierre et al. 2014; Kharchenko et al. 2011; Zdobnov et al. 2002). We identified 215 total epigenetic ensemble genes in *D. melanogaster*, encompassing genes associated with heterochromatin formation and stability, epigenetic complexes, acetylation and deaceytlation, methylation and demethylation, phosphorylation and dephosphorylation, ubiquitylation and deubiquitylation and other epigenetic functions (Additional File 4.1), based on comparisons with epigenetic genes in humans (Weng et al. 2012; Arrowsmith et al. 2012). Using MRBB, OrthoDB and eggNOG, we identified 169 genes in *An. gambiae* (Table 4.1) that are orthologous to members of the 215-member epigenetic gene ensemble that we had defined in *D. melanogaster* (Additional File 4.1), as described in MATERIALS AND METHODS. We required that at least two of the three ortholog identification methods – MRBB, OrthoDB and/or eggNOG – support the orthologous gene call, in order to define a given gene as being orthologous between the two species. Overall, all three methods positively identified the same ortholog for 146 genes (Additional File 4.1), while 23 orthologs were identified by only two of the three methods. An ortholog was identified by only one method for each of 10 genes, discussed

further below. Finally, all three methods failed to detect an ortholog in *An. gambiae* for 36 genes.

Among the 169 orthologous epigenetic gene ensemble members that we define in *An. gambiae*, many complete or nearly complete functional classes are conserved between fruit flies and mosquitos (Table 4.1). The gene classes within which a plurality of epigenetic modifier genes reside – chromatin acetylation (26 genes in *D. melanogaster*) and chromatin methylation (34 genes in *D. melanogaster*) – are highly conserved, as we identify 22 and 31 orthologous genes for acetylation and methylation classes, respectively, in *An. gambiae*. *An. gambiae* possesses complete sets of orthologs for chromatin deacetylation and demethylation functional classes, including orthologs for all five histone deactylases (Foglietti et al. 2006) and all three arginine-methyltransferases (Boulanger et al. 2004) described in *D. melanogaster*. In total, 68 of the 76 genes that are associated with chromatin methylation/demethylation and chromatin acetylation/deacetylation, including histone demethylases *Kdm4A* and *Kdm4B* and histone methylases *Ash1* and *Ash2*, are conserved between the two species. Among the 28 *D. melanogaster* genes associated with chromatin modifying and remodeling complexes, we identify 25 orthologs in *An. gambiae*. All components of the NuRD and NURF complexes exhibit orthologs in both species, as do nine out of ten other genes involved in the ACF complex and other chromatin-associated complexes. Within the Ino80 complex, seven of nine components exhibit orthologs in both species, as only *CG11970* and *pho* do not exhibit detectable orthologs in *An. gambiae*. All genes in the ubiquitination functional class are conserved, as are five of seven genes within the phosphorylation functional class. Evaluation of heterochromatin-associated genes, i.e.,

centromeric, intercalary and nuclear heterochromatin classes, reveals that *An. gambiae* possesses orthologs for four of six, three of five and three of four *D. melanogaster* genes, respectively, within these three classes.

The multigene Set-N chromatin protein clade in *D. melanogaster* (annotated as CC__ in Additional File 4.1) (van Bemmel et al. 2013) exhibits the greatest absolute and relative reduction in ortholog number within the epigenetic gene ensemble membership in *An. gambiae*. We are unable to identify *An. gambiae* orthologs for 17 of 40 *Set-N* genes that have been defined in *D. melanogaster*, which accounts for 35 percent of the total number of genes for which we cannot identify orthologs between these two species. Other *D. melanogaster* genes for which we cannot identify *An. gambiae* orthologs include those encoding two out of the three *Ada2a*-containing complex components (*Atac1 and Atac2*) and four other histone modification genes (*BEAF-32*, *Incenp*, *Lpt* and *msl-1*). Based on our stringent criteria, we also declined to call *An. gambiae* orthologs of six *D. melanogaster* genes involved in heterochromatin modulation: *e(y)3*, *Lhr*, *Pc*, *Prod*, *Su(var)2 and Su(var)3-7*.

Based on our criteria for ortholog calling (i.e., at least two of the methods among MRBB, eggNOG, OrthoDB must call the same ortholog), there are 10 genes for which only one of these three methods identifies an ortholog in *An. gambiae*: *Borr* (*AGAP0011219*, *AGAP0011220*), *CC34* (*AGAP002753*), *CC35* (*AGAP008006*), *e(y)3* (*AGAP001877*), *HP1b* (*AGAP009444*), *Lpt* (Chromosome 3:18890039-18892840), *Pc* (Chromosome 2:26898592-2757082), *Pcl* (*AGAP003277*), *Su(var)2-HP2* (*AGAP001194*), and *Vig2* (*AGAP013112*). Among these ten genes in *D. melanogaster*, we are able to identify orthologs for seven genes using OrthoDB – *lpt* (7 *Anopheles*

species), *CC34* (4 *Anopheles* species), *Pc* (17 *Anopheles* species), *CC35* (18 *Anopheles* species), *e(y)3* (18 *Anopheles* species), *Vig3* (1 *Anopheles* species) and *Hp1b* (14 *Anopheles* species) (Additional File 4.3) – among members within the genus *Anopheles*. Our ability to identify *lpt*, *Pc, CC35, e(y)3* and *Hp1b* orthologs in many other *Anopheles* species implies that the putative orthologs for these genes that we have identified in *An. gambiae* are valid, despite not satisfying fully our criteria. The remaining five genes may have true orthologs in *An. gambiae* and all other anophelines assembled to date, but we have not called them based on our stringent criteria. For those fruit fly genes for which we fail to detect orthologs in *An. gambiae* with all three methods (N = 36 genes, Additional File 4.1), the apparent absence of an ortholog might reflect assembly errors, as complete *An. gambiae* chromosomes are not yet fully assembled (Holt et al. 2002). However, among the 36 genes that yield no ortholog calls in *An. gambiae* using our methods, only two (*msl-1*, (13 *Anopheles* species) and *CG11970*, (13 *Anopheles* species)) detect putative orthologous genes in other *Anopheles* species in OrthoDB. These findings suggest that the other 34 genes for which we do not detect orthologs in *An. gambiae* may be absent from the *Anopheles* clade.

Determining phylogenetic relationships among all *Set-N* gene family member coding sequences in *D. melanogaster* and orthologous genes in *An. gambiae* by maximum-likelihood using RAxML (Stamatakis 2014) yields inferences regarding differences among species in the evolution of *Set-N* chromatin protein genes (Fig. 4.2). The *D. melanogaster Set-N* chromatin protein gene family includes three related gene clusters for which we do not identify orthologous genes in *An. gambiae*, comprising one group of five *Set-N* genes (*CG15436*, *CG5245*, *CG12744*, *CG17385* and *CG7357*), a

second group of three *Set-N* genes (*CG4936*, *Zif* and *M1BP*) and a third group of two *Set-N* genes (*ssp*, *CG8289*). Overall, there are 17 *Set-N* genes in *D. melanogaster* for which we do not identify orthologs in *An. gambiae* (Fig. 4.2), consistent with expansion of the *Set-N* gene family in the Brachyceran suborder, as compared to the Nematoceran suborder. Of the 17 *Set-N* genes in *D. melanogaster* for which we do not call an ortholog in *An. gambiae*, we do not detect orthologs for 15 genes among any of the *Anopheles* species genomes annotated within OrthoDB. We do call orthologs for both *CC34* and *CC35* in *Anopheles* species outside of *An. gambiae* (see above)

Another gene set that appears to have expanded in the Brachyceran suborder, compared to the Nematoceran suborder, is the *heterochromatin protein-1* (*HP1*) gene family, which has fewer members in *An. gambiae* than in *D. melanogaster*. We identify only two gene family members – *AGAP004723* and *AGAP009444* – in *An. gambiae*, compared to the five *HP1* gene family members – *HP1, HP1b, HP1c, HP1d* (*Rhino*), and *HP1e* – that are present in *D. melanogaster* (Fig. 4.3). In fact, one *HP1b* ortholog (*AGAP009444*) that was identified in *An. gambiae* using MRBB was not supported by either OrthoDB or eggNOG. This reduced *HP-1* gene family membership is also evident among other nematoceran species that span the genus *Anopheles*. Each of the 12 anopheline species we have studied in depth exhibits only two *HP1* gene family members related to the *D. melanogaster HP1* gene family. Comparisons of the expression of orthologous *HP1* family genes in *An. gambiae* and *D. melanogaster* reveal a significant difference in expression patterns of the *D. melanogaster* gene *HP1e* and the *An. gambiae* orthologs *AGAP004723* and *AGAP009444* (Additional File 4.2). *HP1e* exhibits little or no expression across all life stages, while both *AGAP009444* and *AGAP004723* exhibit

significant expression levels among all four life stages/genders assessed, reflective of increased expression of this gene in mosquitos compared to fruit flies.

Gene Family Expansions and Contractions Across the Genus *Anopheles*

Among the set of 12 Anopheline species (listed in MATERIALS AND METHODS) for which high-quality, RNAseq-supported assemblies have been defined (Neafsey et al. 2014), we identify orthologs for all 169 members of the epigenetic gene ensemble we have defined for *An. gambiae* (Additional File 4.1). This implies that the dynamic, widespread evolution of the epigenetic gene ensemble that has occurred since the divergence of the suborders *Nematocera* and *Brachycera* appears not to have continued during species divergence within the genus *Anopheles*. In total, seven gene families exhibit expansions or contractions in one or more Anopheline species (Table 4.2). Gene families that include potential paralogs in *An. gambiae*, but for which one of the putative paralogs maps to the *An. gambiae* UNKN chromosome, were neither studied nor shown on Table 4.2, as the UNKN chromosome in the *An. gambiae* genome represents those contigs that were not mapped during initial assembly, and putative gene duplications that map to this "chromosome" may instead constitute assembly artifacts.

The *D. melanogaster* genes that exhibit duplications in *An. gambiae*, for which one of the *An. gambiae* orthologous family members maps on the UNKN chromosome, are *Chrac-14*, *Mt2*, and *Wds*. Three anopheline gene families exhibit single species expansions in gene number – *Cap-G* (expanded in *An. dirus*), *CG18004* (expanded in *An. atroparvus*) and *Orc2* (expanded in *An. atroparvus*) (Table 4.2). The *EFF* gene has undergone duplication by retrotransposition in multiple anopheline species. We find

CC14 duplications that have arisen via retrotransposition in *An. gambiae*, *An. epiroticus, An. arabiensis*, *An. quadriannulatus*, and *An. merus,* all members of the Pyretophorus Series of Anopheline mosquitoes (Fig. 4.5). Two gene families – *Parg* and *GRO*,– exhibit contractions in gene number among the other anopheline species we have studied, relative to *An. gambiae* as *Parg* is contracted in *An. albimanus* and *GRO* is contracted in *An. epiroticus* and *An. merus* (Table 4.2). All other epigenetic gene ensemble members assessed across the genus *Anopheles* exhibit 1:1 orthologous conservation among all 12 anopheline species analyzed.

Among the epigenetic regulatory genes we have analyzed, the ubiquitin-conjugating enzyme *E2D* (orthologous to *effete* in *D. melanogaster*) has undergone duplication via retrotransposition (Fig. 4.8). Orthologs of this retrogene are found in a subset of anopheline species (Fig. 4.8). The presence of retrogenes in multiple subgenera among the Anophelinae may be consistent with the hypothesis that the initial *E2D* retrotransposition occurred only once after divergence of the subfamilies Anophelinae and Culicinae. If this were the case, the retrogene must have been lost within the series Pyretopherous and Neocellia, and within a subset of the series Myzomyia. Alternatively, the retrotransposition may have occurred independently within two or more subgenera within the subfamily Anophelinae. The presence of two *E2D* retrogenes within *An. dirus* implies that there has been either a second retrotransposition event or a conventional duplication of the *E2D* retrogene within this species. The inference that the retrogene persists as a functional ortholog under selective pressure is supported by the preservation of the full-length *E2D* open reading frame in all eight species in which it is found (Fig. 4.8), with substantial sequence conservation. The identification of this apparently

functional retrogene is consistent with the hypothesis that expansion of gene families through the genesis of functional retrogenes contributes to genetic diversity and phenotypic differences among rapidly divergent anopheline species.

Functional and Evolutionary Comparisons of Epigenetic Gene Ensembles

In order to gain deeper insights into the potential functional similarities and differences between the epigenetic gene ensembles of *An. gambiae* and *D. melanogaster*, we performed a principal component analysis (PCA) on epigenetic gene expression across comparable tissues in both species (Fig. 4.4A). PCA revealed *An. gambiae* and *D. melanogaster* possess two distinct tissue expression profiles. The two principal components identified account for almost 94 percent of the variance between the two species. A subset of tissues comprising carcass, midgut, ovary, head, Malpighian tubules, and salivary gland account for 84.7% of the variance, while the remaining 9.1 percent of variance can be attributed predominantly to expression differences within the testis. To evaluate further possible functional differences between the tissue expression profiles in *D. melanogaster* and *An. gambiae*, we compared relative expression levels between the two species for 144 epigenetic genes in seven tissues (Fig. 4.7). All tissues analyzed exhibited mean increased Log_{10}(fold-change in expression values) in *D. melanogaster* between 0.90 and 1.3, with the exception of the testis, which exhibited an increase of only 0.15. The interspecies differences between the fold-change in expression values in testis and all other tissues analyzed were statistically significant using ANOVA (p-value < 0.0001).

We next compared developmental expression patterns for orthologous genes between these two species to explore functional conservation between *D. melanogaster* and *An. gambiae* of epigenetic gene ensemble members. Similar analyses have been performed on epigenetic modifier gene ensemble expression profiles in human liver and brain tissue to identify clusters of genes with similar expression patterns (Weng et al. 2012). Hierarchical clustering of gene expression in both species reveals two distinct expression classes: those genes that possess high expression (red bar) or low expression (green bar) across developmental life stages (Fig. 4.4B and 4.4C). Among these genes within each species, 119 epigenetic genes reside in the same respective high expression (42 genes) or low expression (77 genes) group in mosquitos and flies, while 50 reside in different expression groups in the two species (Additional File 4.2). Of the 50 genes that exhibit differing expression intensities in these two species, four predominant groups of GO terms are associated with over 75 percent of the 50 genes – acetylation (14 genes), methylation (10 genes), complexes (six genes) and Set-N chromatin protein genes (eight genes, Fig. 4.6, Additional File 4.2). Four other functional classes – heterochromatin (three genes), phosphorylation (one gene), ubiquitination (two genes) and genes that have no attributable GO term descriptors (six genes) – encompass the remaining genes that exhibit differing expression intensities between *An. gambiae* and *D. melanogaster*.

To assess evolutionary conservation of epigenetic gene ensemble members, and gauge any differences in evolutionary rates, we calculated dN/dS for each gene within the *An. gambiae* complex and *D. melanogaster* subgroup (Additional File 4.4). Direct assessment of respective evolutionary rates is tenable because both the *An. gambiae* complex and *D. melanogaster* subgroup are approximately 5 million years old (Obbard et

al. 2012; Neafsey et al. 2014), enabling estimation of relative evolutionary rates across the same time interval. The average dN/dS rate (±SEM) for epigenetic genes in the *An. gambiae* complex was 0.1084 (±0.0089) and while that for the *D. melanogaster* subgroup was 0.1028 (±0.0068), reflecting the absence of a statistically significant difference in evolutionary rates (p-value = 0.61, T-test) (Additional File 4.4).

DISCUSSION

We began this study by assigning 215 genes to the epigenetic gene ensemble of *D. melanogaster* (Fig. 4.1, Additional File 4.1). This ensemble represents approximately 1.5% of the protein coding genes annotated in the *D. melanogaster* genome (among a total of 13,955 genes; St. Pierre *et al.* 2014). We have defined an even smaller epigenetic gene ensemble in *An. gambiae*. The fact that these limited sets of epigenetic genes are sufficient to control many varied and complex pan-genomic processes encourages the premise that these genes have evolved under strong selective pressure. This premise is supported by low dN/dS rates we observe for the epigenetic ensemble genes in *D. melanogaster* and *An. gambiae*, as well as the limited gene family expansion and contraction across the genus *Anopheles* that we observe for members of this ensemble. It has been noted that long noncoding RNAs (lncRNA) and microRNAs (miRNAs) have roles in epigenetic regulation and therefore supplement the epigenetic gene ensemble that mediates chromatin modification (Lee 2012; Kim and Nam 2006; Kim 2005; He and Hannon 2004; Nie et al. 2012). The limited epigenetic gene ensemble we define for *An. gambiae* certainly mediates only a portion of the epigenetic control required to ensure a fully functional genome, while lncRNAs and miRNAs provide other facets of epigenetic control that we and others are only beginning to elucidate (Mercer and Mattick 2013; Lee 2012; Lv et al. 2013; Ponting et al. 2009).

Some proportion of the selective pressure that appears to constrain evolution of the epigenetic gene ensemble may arise from the oft-noted requirement for epigenetic modifiers to operate within the contexts of multicomponent complexes (Conaway and Conaway 2009; Schuettengruber et al. 2007).The structural requirements that must be

satisfied simultaneously for individual members of such complexes to maintain multiple interactions would constitute one such constraint, which could cause epigenetic genes to be less tolerant of increased mutation rates. The sensitivity of epigenetic machinery to mutation is reflected, in part, by the many alterations in body plan patterning in *Drosophila* that result from alterations in dosages of genes the mediate epigenetic regulation of homeotic gene function [*e.g.,* Polycomb, Trithorax; (Schuettengruber et al. 2009, 2007) (Kennison & Tamkun, 1988; Kennison, 2004; Schotta et al., 2002)], and the implication that sometimes subtle alterations in epigenetic gene function in a variety of human neoplasias may contribute to oncogenesis (Dawson and Kouzarides 2012; Portela and Esteller 2010). In these and many other instances, a subtle change in the level of function of one member of an epigenetic gene ensemble may contribute to large changes in the developmental or homeostatic landscape of an entire tissue or organism. As this reasoning pertains to the epigenetic gene ensemble in *D. melanogaster*, it will apply to related gene ensembles in other organisms, as well.

For *An. gambiae*, a dipteran of substantial interest due to its propensity to transmit human malaria parasites (Cohuet et al. 2010) , we have identified a set of 169 genes that are orthologous to genes within a 215-member epigenetic gene ensemble we have defined in *D. melanogaster* (Fig. 4.1, Additional File 4.1). The conservation rate for epigenetic genes of 79% that we observe between these two species is greater than the 62% interspecies conservation rate observed between the completely annotated genomic-wide protein-coding transcriptomes of *An. gambiae* and *D. melanogaster* (Zdobnov et al. 2002). Determination of genome-wide coding transcriptome conservation based on comparisons between *An. gambiae* and each of the other anopheline species we have

analyzed yields an average of 99.1 percent 1:1 orthologous gene number conservation for the 11 pair-wise *Anopheles* species comparisons we have completed (see Table 4.2), including only seven instances of epigenetic gene family expansion or contractions across the genus (Table 4.2). Two species (*An. arabiensis* and *An. quadriannulatus*) exhibit 100 percent 1:1 gene number conservation of the epigenetic gene ensemble when compared to *An. gambiae*. None of the other eleven species compared to *An. gambiae* possesses less than 97.6 percent 1:1 gene number conservation for the epigenetic gene ensemble. This lowest conservation was observed between *An. gambiae* and *An. atroparvus*, one of the most divergent species pairs among those we have analyzed (Neafsey et al. 2014). The most divergent species pair analyzed – *An. gambiae* and *An. albimanus* – exhibits 1:1 gene number conservation of 98.8 percent. The greater rates for epigenetic gene conservation that we observe, compared to those observed for the genome-wide protein-coding transcriptomes, provide further evidence of the action of selective pressure on epigenetic gene ensembles since the divergence of Brachycera and Nematocera, as well as during divergence among Anopheline species. Furthermore, the limited number of paralogs (four in total; *Cap-G* in *An. dirus*, *CG18004* and *Orc2* in *An. atroparvus*, and *CC14* within the Pyretophorus class) that we detect within the epigenetic gene ensembles (Table 4.1) that we define among the anopheline species analyzed implies that the composition of this gene ensemble among these species is relatively stable, as reflected by a nearly constant gene membership. Comparison of the epigenetic gene ensemble membership on the basis of copy number constitutes one measure of the consistency of evolutionary pressure that bears on this gene ensemble. Another useful measure for gauging evolutionary pressure on a given gene set is evolutionary rate.

The inference that the epigenetic gene ensemble has been relatively stable as anophelines have diverged is supported by our finding that evolutionary rates within this gene ensemble are similar between the *An. gambiae* complex and the *D. melanogaster* subgroup (Additional File 4.4). We observe average epigenetic gene ensemble dN/dS values of 0.1084 (±0.008990) for the *An. gambiae* complex and 0.1028 (±0.006837) for the *D. melanogaster* subgroup. Both values are indicative of high levels of purifying selection acting on the epigenetic gene ensembles in both species subgroups (Mugal et al. 2014; Gharib and Robinson-Rechavi 2013). The similar evolutionary rates we observe for both taxa, and the infrequent gene family expansion and contraction events we detect, imply that the gene ensemble is evolutionary stable, for the most part. In striking contrast, however, substantial evolution of gene families encoding the Set-N (Fig. 4.2) and HP1 (Fig. 4.3) proteins has occurred through paralogous expansion and contraction within these two insectan clades. In two other instances of rapid evolution, retrotransposition has led to expansion of the *effete* (Neafsey et al. 2014) and *CC14* gene families (this work, see below) among anopheline mosquitos.

To explore more deeply the functional conservation within the epigenetic gene ensembles in *An. gambiae* and *D. melanogaster*, we investigated the temporal and tissue-specific gene expression patterns of members of the ensembles in these two species. Tissue-specific expression in *D. melanogaster* and *An. gambiae* were compared using principal component analysis (Fig. 4.4A). The two species exhibit well-populated but distinct epigenetic gene expression clusters, respectively, based on PCA analysis. This finding is consistent with the inference that many of these epigenetic modifiers are expressed at different levels in specific tissues within the respective species (Fig. 4.7).

On average, *D. melanogaster* exhibits increased epigenetic gene expression levels for all tissues compared to *An. gambiae* gene expression levels, except for the testis, consistent with the findings of our PCA analysis. These differences in expression levels between organisms are analogous to differences observed in epigenetic gene expression for different human cell types (e.g., liver and brain, Weng *et al.* 2014), suggesting that substantial differences in epigenetic gene expression may be important for cellular distinctions not only between species, but also within single species.

Temporal developmental expression patterns for epigenetic ensemble genes in *D. melanogaster* and *An. gambiae* exhibit broad similarity (Fig. 4.4B and 4.4C). A set of 119 *An. gambiae* genes and their *D. melanogaster* orthologs are clustered within comparable high (green blocks, Fig. 4.4B and 4.4C) or low (red blocks, Fig. 4.4B and 4.4C) expression groups in both species, while 50 *An. gambiae* and *D. melanogaster* orthologs reside within differing respective expression groups (Fig. 4.6, Additional File 4.2). The GO term classes methylation, acetylation, complex components and Set-N chromatin protein are associated with proteins encoded by 75 percent of the genes that exhibit differing expression profiles. This may reflect developmentally dynamic redeployment within these species of a subset of epigenetic functions that modulate methylation and/or acetylation, since the divergence of *Brachycera* and *Nematocera*. The broad similarities of temporal expression patterns we observe for most members of the epigenetic gene ensembles in these two Dipteran species are comparable to similarities that have been noted in other closely related species for genome-wide, 1:1 orthologs (e.g., between human and mouse, Huminiecki and Wolfe 2004).

We find that 17 *D. melanogaster* Set-N chromatin proteins do not have identifiable orthologs in *An. gambiae*, representing 42.5 percent of the total *Set-N* gene set in *D. melanogaster*. When all *Set-N* epigenetic ensemble genes in *D. melanogaster* and *An. gambiae* are compared by maximum likelihood, we find 10 instances of gene multiplication in *D. melanogaster* that are not present in *An. gambiae* (green highlights, Fig. 4.2), consistent with the inference that the majority of non-orthologous genes in *D. melanogaster* evolved after divergence from the most recent common ancestor with *An. gambiae*. We observe acquisition of new expression profiles for the *Set-N* paralogs *AGAP000725* and *AGAP011684* in *An. gambiae*, which are orthologous to the SET-N chromatin protein gene *CC14* in *D. melanogaster*. In *An. gambiae*, *AGAP000725* exhibits increased expression across all life-stages compared to *AGAP011684*, which exhibits much lower expression levels (Additional File 4.2). These variations in expression may reflect acquisition of qualitatively distinct functions for paralogous genes that have been generated by duplication and divergence within the Nematoceran clade. In fact, a retrotransposition event has contributed to paralogous expansion of the *CC14* gene within the *Set-N* gene family in anophelines (Fig. 4.5A). The distinct amino acid profiles we observe within the retrotransposed and original copies (Fig. 4.5B) indicate that the two genes may now be under different evolutionary selective pressures. To further explore this inference, we determined the dN/dS ratios for *AGAP011684* and *AGAP000725*, respectively, as compared to the *D. melanogaster* ortholog *CC14*. The rate of non-synonymous substitutions (dS) was highly saturated (dS >50) for the retrotransposed *AGAP011684*, while being far below saturation for the spliced *AGAP011684* (dS < 1). These findings imply that the evolutionary pressures acting on *AGAP011684* are much

different than those acting on *AGAP000725*, and they correlate with the high number of amino acid substitutions in the retrotransposed *CC14* ortholog *AGAP011684*, as compared to the lower number of substitutions observed for the spliced *CC14* ortholog *AGAP000725* (Fig. 4.5)

While five *HP1* gene family members have been annotated in *D. melanogaster*, only two are present in the *An. gambiae* genome. Based on our phylogenetic analyses, a set of *HP1* genes that is evolutionary orthologous to the *HP1e* gene in *D. melanogaster* (Fig. 4.3, blue highlight) is present in the genus *Anopheles*. A second related set of *HP1*-like genes that we can define among the anophelines (Fig. 4.3, red highlight) is not closely related to any of the *D. melanogaster HP-1* family genes. The predominant expression of *HP1e* in male germline cells in *D. melanogaster* has been proposed to contribute to protection of the male germline genome (Vermaak et al. 2005; Vermaak and Malik 2009). However, the *An. gambiae HP1e* ortholog *AGAP004723* exhibits significantly increased expression in female ovaries, suggesting a function more similar to that of *HP1d* in *D. melanogaster*, which is thought to contribute to protection of the female germline genome (Vermaak et al. 2005; Marinotti et al. 2006; Baker et al. 2011). As previously explored in human and mouse (Huminiecki and Wolfe 2004; Lespinet et al. 2002), intra-specific paralogs often acquire new expression patterns and thereby contribute to evolutionary diversity. This is consistent with the diverse range of expression patterns that members of the *HP1* gene family exhibit in *D. melanogaster*. *HP1d* and *HP1e* exhibit very little to no expression during all life stages, while *HP1, HP1b* and *HP1c* exhibit increased expression during some life stages and lower expression during other life stages (Additional File 4.2). Both *An. gambiae HP1* gene

family orthologs exhibit consistent levels of expression among all life stages, indicating potential functional differences between the orthologous *HP1* genes in these two species. This inference is further supported by differences in temporal expression profiles that we observe between the orthologs *HP1e* and *AGAP004723* (Fig. 4.3). The very limited expression of *HP1e* in fruit flies compared to the increased expression of *AGAP004723* in mosquitos implies that the mosquito ortholog of fruit fly *HP1e* may have acquired a new function during one or more developmental stages, since divergence from the most recent common ancestor of the suborders *Brachycera* and *Nematocera*.

As the *Set-N* and *HP1* gene families expanded among *Brachycera* and *Nematocera* by duplication and divergence, evolutionary constraints bearing on newly arising members of the gene families may have diminished, allowing paralogous genes to diversify and evolve new functions. This is consistent with the premise that paralogous genes contribute to the genesis of increased genetic diversity by serving as substrates for increased rates of sequence evolution and diversification of gene function (Huminiecki and Wolfe 2004).

Sequence orthology is often invoked as the basis for identification of functionally related genes in *An. gambiae* and *D. melanogaster*. However, such identifications, even when further supported by similar expression profiles, remain inferences until validated by functional genomic analysis. While many essential genes within Homeobox (HOX) Complexes, and the Polycomb and Trithorax Groups have been shown to be functionally conserved across a range of insects, it is difficult to posit functional conservation without functional genomic data (Schuettengruber et al. 2007, 2009; Kennison 2004). Our findings regarding strong selective pressure on the epigenetic ensembles in both *An.*

gambiae and *D. melanogaster*, the relative rarity of gene family expansion/contraction events, and similar temporal gene expression profiles between clades provide strong support for the inference that functionality is also conserved for many of these epigenetic genes. Although, admittedly, we do observe differing tissue specific patterns for some epigenetic gene orthologs in each species. Therefore, conclusive statements regarding functional conservation of orthologs should rest on functional genomic validation, which is available in mosquitos at present based on RNA interference approaches (Keene et al. 2004; Michel et al. 2005) and may prove feasible through gene editing (e.g., CRISPR technology, (Cong et al. 2013)) in the future. These approaches to functional validation are particularly important in those instances in which specific epigenetic genes are chosen as potentially druggable targets for insecticide development and vector control.

Due to the rapid evolution of insecticide resistance genes in *Anopheles* mosquitos (Mitchell et al. 2014; Edi et al. 2014), the identification of additional proteins that may serve as the bases for new vector-targeted control interventions has assumed paramount importance (Zaim and Guillet 2002). In choosing a candidate target gene that encodes an essential catalytic activity that could be inhibited by small molecule antagonists (i.e., potential insecticides), it is important to consider the evolutionary dynamics of putative target genes. A candidate target gene for which the catalytic domain is highly conserved among a very diverse set of insects may be less tolerant of *de novo* mutations that could confer insecticide resistance. However, an antagonist against a protein that is too broadly conserved may function as an insecticide that kills benign insects as well as vector mosquitos. Therefore, the ideal such proteins will be those that are conserved among members of a vector insect genus, but diverge within benign insect genera (e.g., *Apis*).

This divergence could affect a subset of critical active site residues within an otherwise largely conserved catalytic domain, which would enable identification of vector-selective active site-interacting small molecule antagonists. Alternatively, this divergence could affect regions outside of the catalytic domain, which could be targeted by small molecules that destabilize the target protein or interfere with its interactions with essential protein-protein interaction (PPI) partners. Such proteins could constitute good targets because mutations that arise within a catalytic domain that is highly conserved within the genus and confer insecticide resistance would be difficult to maintain, as they would probably impede wild type protein function. This premise has begun to be investigated for druggable epigenetic targets in cancer and other diseases (Gomez-Diaz et al. 2012; Arrowsmith et al. 2012; Kishore et al. 2013).

Among the epigenetic gene ensemble members we have characterized, the histone methyltransferase *Su(var)3-9* gene encodes a candidate target within the latter group (i.e., divergence outside of the catalytic domain). This protein has similar epigenetic functions across many species, but exhibits a diverse set of structural differences between species, including gene fusions and re-fission with other genes (Krauss et al. 2006). Small molecules that target these divergent non-catalytic domains, and diminish protein stability (Bill et al. 2014) or PPIs with critical interaction partners (Ammosova et al. 2012) in vector species, could be designed to reduce cross-reactivity with closely related proteins in benign non-vector species.

A more conventional approach to insecticide development (e.g., larvicides), based on inhibition of epigenetic functions, would involve identification of small molecules selective for mosquito orthologs within epigenetic gene families essential for

metamorphic development. Many epigenetic modifiers, most notably the Polycomb Group and Trithorax Group genes (Kennison 1995, 2004; Arrowsmith et al. 2012), have been shown to modulate metamorphic development in *D. melanogaster* and other insects. Members of these gene families could be exploited within *An. gambiae* by developing species-selective larvicides and administering them to habitats in which mosquitoes develop.

Another avenue for species-selective mosquito control based on epigenetic genes could involve the incorporation of anopheline epigenetic functions into Anopheles strains analogous to dominant-lethal sterile-insect strains that have been developed for *Aedes aegypti* (Alphey et al. 2010; Phuc et al. 2007). Given the likely functional conservation of epigenetic genes among multiple mosquito species, and potentially among benign insects as well, the use of mass-administered small molecule antagonists to field habitats may produce substantial die-off among multiple off-target insect species. In contrast, the use of sterile-insect strategies that depend on species-restricted genetic transmission of transgenes that mediate directed misexpression of pleiotropic epigenetic genes, which would lead to developmental lethality or adult sterility, would constitute much more selective approaches to mosquito control.

The application of these conceptual and biochemical approaches, coupled with the identification and further characterization of epigenetic gene ensemble members in anopheline species, will continue to deepen our knowledge of vector genetics and biochemistry, and may enable the development of new vector-targeted insecticidal interventions that will reduce the burdens to human health imposed by malaria and other vector-borne diseases.

METHODS

Orthologous Gene Identification

We first defined a comprehensive epigenetic gene ensemble for *D. melanogaster* encompassing genes associated with the Gene Ontology (GO) terms acetyltransferase, ACG/Chrac-complex, beta-heterochromatin, chromatin remodeling, heterochromatin, histone acetylation, histone deacetylation, histone methylation, histone demethylation, histone ubiquitylation, histone deubiquitylation, histone phosphorylation, Ino80 complex, intercalary heterochromatin, Nu4A, nuclear centromeric heterochromatin, nuclear heterochromatin, NuRD complex, RSF complex, Set-N chromatin protein, telomeric heterochromatin and DNA methylation (Consortium, 2000). This set (Table 4.1) was manually augmented to include genes that were described in primary articles and reviews by Filion *et al*. 2010, Greer and Shi 2012, van Bemmel *et al*. 2013, Arrowsmith *et al*. 2012, Schulze *et al*. 2007 and Swaminathan *et al*. 2012. Identification of orthologous genes in *An. gambiae* (Fig. 4.1, Additional File 4.1) was initiated by running TBLASTN using *D. melanogaster* open reading frames as queries against the *An. gambiae* assembly AgamP3.6 from VectorBase (www.vectorbase.org) (Megy et al. 2012), and following this with a modified reciprocal best BLAST (MRBB) analysis. While strict reciprocal best BLAST identifies 1:1 orthologs, we instead used BLAST to identify initial hits with E-values less than 1E-10, for each epigenetic modifier gene. These initial hits were used to BLAST against the reciprocal genome, and aligned genes with the highest E-values were used to define orthologs. This enabled identification of orthologs for genes that have multiple homologs in another species. To further validate putative orthologs, OrthoDB and eggNOG databases were utilized to support MRBB ortholog assignments and to

identify potential missed calls (Waterhouse et al. 2013; Powell et al. 2014). To call conclusively an ortholog between *An. gambiae* and *D. melanogaster*, we required that the putative *An. gambiae* ortholog be identified using at least two of the three assessments we applied, i.e., MRBB analysis, the eggNOG database and/or the OrthoDB database. In instances in which a putative mosquito ortholog did not satisfy this criterion, and in which we did not therefore "call" an ortholog, a true ortholog may exist in *An. gambiae*, but we will not have called it, based on our stringent criteria.

TBLASTN and MRBB analyses were performed among a set of 12 assembled *Anopheles* genomes (*An. gambiae*, *An. epiroticus, An. stephensi, An. funestus, An. arabiensis, An. albimanus, An. dirus, An. minimus, An. quadriannulatus, An. atroparvus, An. merus,* and *An. farauti*) (Megy et al. 2012), based on the *An. gambiae* epigenetic gene ensemble that we defined using TBLASTN, MRBB and eggNOG to identify orthologous genes across the genus Anopheles (Table 4.1). These ortholog calls were then compared to orthologs identified in the OrthoDB database (Waterhouse et al. 2013). Manual curation was performed for all genes that exhibited inconsistencies among TBLASTN, MRBB and OrthoDB calls and for which high-depth RNA sequencing data had been produced by Neafsey *et al.* 2014. We used RNAseq reads for all species (*An. gambiae*, *An. epiroticus, An. stephensi, An. funestus, An. arabiensis, An. albimanus, An. dirus, An. minimus, An. quadriannulatus, An. atroparvus, An. merus,* and *An. farauti)* that are available from SRA accession study PRJNA236161 (Neafsey et al. 2014). Splice junction mapping was performed using TopHat2 (Kim et al. 2013) in relation to the *An. gambiae* P3 genome assembly. A three mismatch maximum was allowed for each read with a maximum -read-edit-dist of three. Gene family expansions that mapped to the *An.*

gambiae UNKN chromosome were not designated true expansions/contractions, as these contigs have not been mapped to any chromosome within the initial assembly, and may reflect assembly artifacts rather than genomic differences (Holt et al. 2002; Megy et al. 2012).

Phylogenetic Assessment and dN/dS Determination

Phylogenetic relationships were analyzed using DNA sequence alignments and based on maximum likelihood, bootstrapped 100 times, performed by RAxML (Stamatakis 2014). The rate of non-synonymous substitutions vs. the rate of synonymous substitution [or dN/dS value (Li et al. 1985; Miyata et al. 1980)] for all 1:1 orthologs was determined for the *An. gambiae* complex (comprising *An. gambiae, An. melas, An. merus, An. arabiensis*, and *An. quadriannulatus*) based on the ratios calculated using data within the OrthoDB database (Waterhouse et al. 2013). The dN/dS values for the *D. melanogaster* subgroup (*D. melanogaster, D. simulans, D. sechellia, D. yakuba* and *D. erecta*) were determined by first extracting open reading frame and protein sequences from all *D. melanogaster* OrthoDB orthologs. A CDS-based alignment was generating using CLUSTAL Omega (Sievers et al. 2011), filtered for at least 60% alignment at any given site using trimAl, and a maximum likelihood tree was generated using RAxML. The alignment and tree were then submitted to PAML for determination of dN/dS values by codeml (Yang 2007). Genes that appeared to have saturated dS values (>1) or no dS value (= 0) were not used. The dN/dS values for single *CC14* paralogs in *An. gambiae* were calculated in comparison to orthologous *D. melanogaster CC14* paralogs using codeml runmode = -2.

Expression of Epigenetic Modifiers in *An. gambiae* and *D. melanogaster*

Gene expression values were obtained for *An. gambiae* by utilizing RNA sequencing reads from SRA accession number PRJEB5712, and from (Pitts et al. 2011). RNA sequencing datasets were aligned using TopHat2 (Kim et al. 2013), as previously described, and FPKM expression values were calculated using CuffDiff (Trapnell et al. 2013; Megy et al. 2012). We utilized the modENCODE expression levels that were given for each gene in FlyBase (www.flybase.org) (St. Pierre *et al.* 2014) to assess *D. melanogaster* gene expression levels. Expression values were grouped among nine distinct life stages, and the average expression level was taken for each life stage. Expression levels were indicated on a scale of 0-6 with the values being 0 = very low/no expression, 1 = low expression, 2 = moderate expression, 3 = moderately high expression, 4 = high expression, 5 = very high expression and 6 = extremely high expression, in accordance with the expression levels described on FlyBase Release 5.48 (St. Pierre *et al.* 2014) . Expression values were then clustered based on the Pearson correlation method using heatmap function in R (R Core Team 2014), for which complete linkage distances and expression classes (high or low expression) were grouped (Fig. 4B and 4C).

Principal Component Analysis (PCA) of Tissue-Specific Gene Expression

Tissue expression values for the epigenetic gene ensembles in *D. melanogaster* and *An. gambiae* were collected from the modENCODE and MozAtlas databases, respectively (Baker et al. 2011; Celniker et al. 2009). Tissues used for PCA analysis in both species

include carcass, midgut, ovary, testis, head, Malpighian tubules, and salivary gland. Expression values for these tissues were normalized to *Act5C* expression, to correct for potential differences in relative magnitudes of expression in each study. We have chosen *Act5C* for the normalization of gene expression values. Although all genes exhibit some variation in expression across different tissues (Vandesompele et al. 2002), *Act5C* tends to exhibit comparable expression levels for specific tissues of interest, respectively, in both *An. gambiae* and *D. melanogaster* (e.g., *D. melanogaster* gut as compared to *An. gambiae* gut), with the exception of the salivary gland (Additional File 4.2), and the *D. melanogaster* ortholog of *Act5C* has been validated as gene for normalization in previous studies (Ponton et al. 2011). Principal component analysis was then performed on the relative expression levels of epigenetic gene ensemble members in the tissues previously specified utilizing the prcomp function in R (R Core Team, 2012).

TABLES/FIGURES AND LEGENDS

Table 4.1: Comparison of epigenetic gene ensemble memberships in *D. melanogaster* and *An. gambiae*.

Gene numbers are based upon orthology between the two species. Functional categorizations are based upon Gene Ontology (GO) terms or known function. The total number of genes in *D. melanogaster* is 215 and in total 169 orthologous genes in *An. gambiae* were identified.

Epigenetic Functional Class Descriptor	Gene number in *D. melanogaster*	Orthologous Gene Number in *An. gambiae*
Acetylation	26	22
Deacetylation	7	7
Methylation	34	31
Demethylation	7	7
DNA Methylation	2	1
Ino80 Complex	9	7
ACF Complex	4	3
NURF Complex	3	3
NuRD Complex	6	6
Other Complexes	6	6
Heterochromatin	13	8
Centromeric Heterochromatin	6	4
Intercalary Heterochromatin	5	3
Nuclear Heterochromatin	4	3
Other Heterochromatin	14	12
Ubiquityation/Phosphorylation	14	12
Set-N Proteins and Misc.	55	34

Table 4.1: Comparison of epigenetic gene ensemble memberships in *D. melanogaster* and *An. gambiae*.

Table 4.2: Expansions/Contractions of Epigenetic Modifier Gene Families Across the Genus Anopheles

Number of orthologous genes that were identified in each of the *Anopheles* species (*Gam*. = *An. gambiae, Epi. = An. epiroticus, Ste. = An. stephensi, Fun.= An. funestus, Ara.= An. arabiensis, Alb. = An. albimanus, Dir. = An. dirus, Min. = An. minimus, Qua. = An. quadriannulatus, Atr. = An. atroparvus, Mer. = An. merus, Far. = An. farauti*) corresponding to the original *An. gambiae* ortholgous gene in *D. melanogaster.*

***D.mel* Gene**	***Gam.***	***Epi.***	***Ste.***	***Fun.***	***Ara***	***Alb.***	***Dir.***	***Min.***	***Qua.***	***Atr.***	***Mer.***	***Far.***
Cap-G	1	1	1	1	1	1	**2**	1	1	1	1	1
Parg	1	1	1	1	1	**0**	1	1	1	1	1	1
CG18004	1	1	1	1	1	1	1	1	1	**2**	1	1
Orc2	1	1	1	1	1	1	1	1	1	**2**	1	1
GRO	2	**1**	2	2	2	2	2	2	2	2	**1**	2
Effete	1	1	**1**	**2**	1	**2**	**2**	**2**	1	**2**	1	**2**
CC14	2	2	**1**	**1**	2	**1**	**1**	**1**	2	**1**	2	**1**

Table 4.2: Expansions/Contractions of Epigenetic Modifier Gene Families Across the Genus Anopheles

Figure 4.1: Epigenetic Gene Set Identification and Analysis in Anopheline Species

Chart illustrating the workflow created to identify and analyze homologous epigenetic gene ensembles in *An. gambiae* and other anopheline species. After compiling an epigenetic gene ensemble for *D. melanogaster*, orthologs were identified in *An. gambiae* using Modified Reciprocal Best BLAST, and eggNOG and OrthoDB databases. Temporal expression patterns of orthologous genes were then compared between the two species. Within the genus *Anopheles,* gene number expansions and contractions were identified, and the dN/dS ratios were calculated and analyzed based on data for multiple members of the *Anopheles* and *Drosophila* clades.

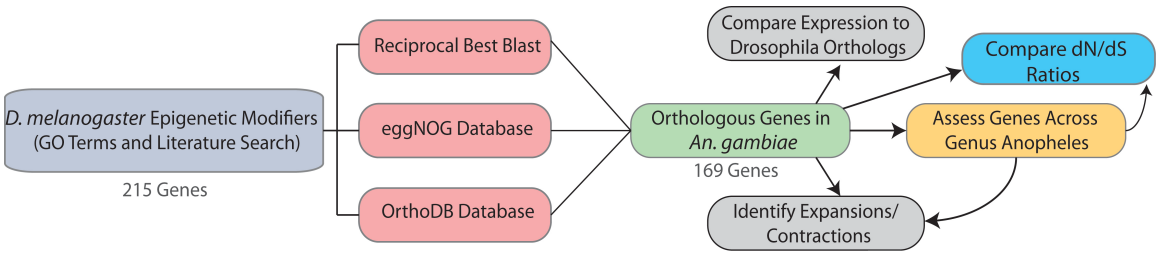


Figure 4.1: Epigenetic Gene Set Identification and Analysis in Anopheline Species

Figure 4.2: Phylogenetic Relationship of Set-N Chromatin Proteins

Relationships among all *D. melanogaster* and *An. gambiae* Set-N chromatin protein coding-sequences determined using maximum-likelihood (Stamatakis 2014). Green boxes indicate *D. melanogaster* genes for which we do not call an ortholog in *An. gambiae*. *An. gambiae* genes are depicted by the identifier AGAP and *D. melanogaster* genes are depicted by CG identifier or gene name, if known. For genes with multiple splice variants, isoform RA is represented.

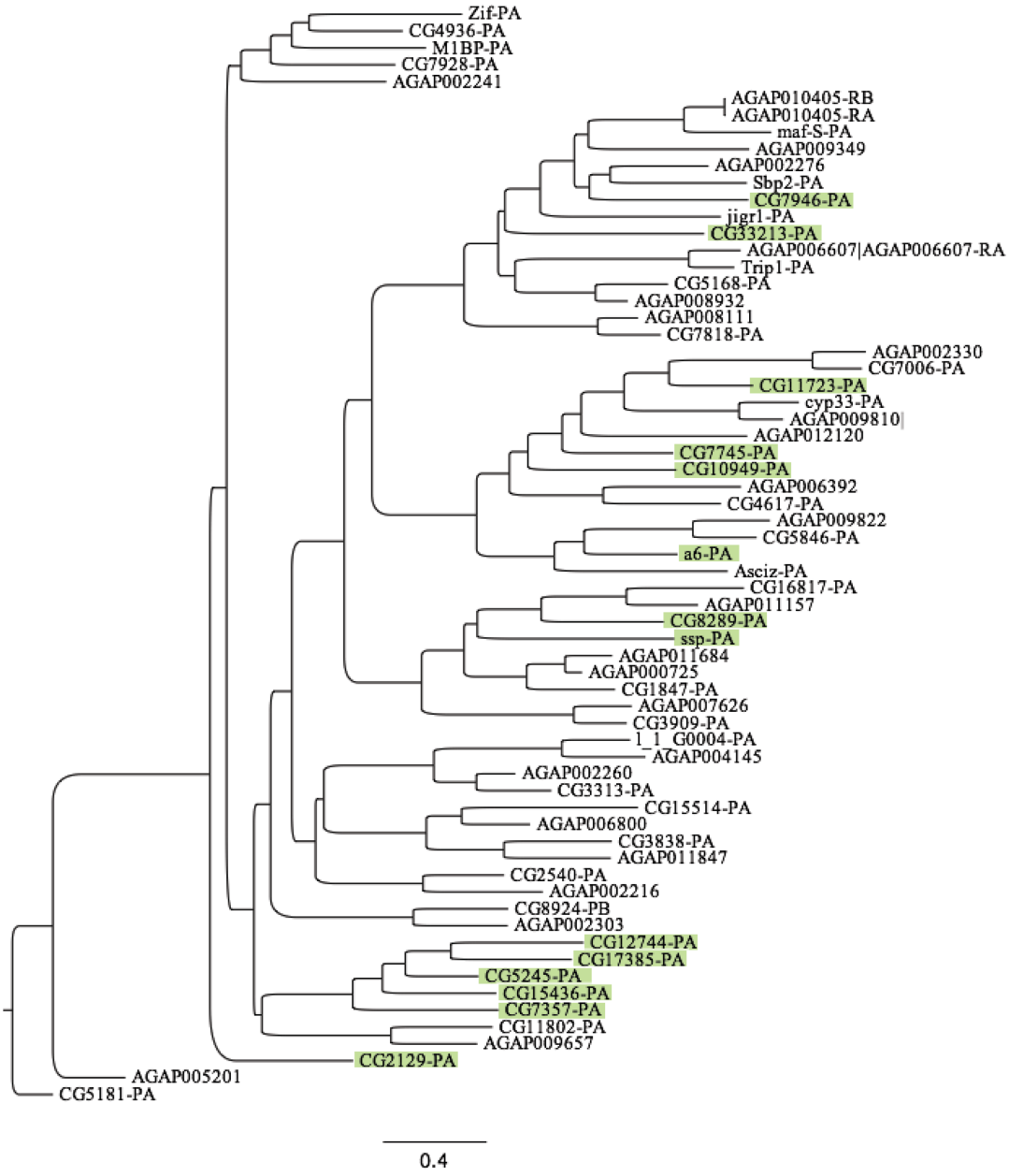


Figure 4.2: Phylogenetic Relationship of Set-N Chromatin Proteins

Figure 4.3: Phylogenetic Relationships Among *Heterochromatin Protein-1* Orthologs in *D. melanogaster* and *An. gambiae*

Phylogenetic tree of the *heterochromatin protein-1* (*HP1*) gene family members in *D. melanogaster* (*HP1, HP1b, HP1c, HP1d, HP1e*), *An. gambiae* (AGAP), *An. arabiensis* (AARA), *An. funestus* (AFUN), *An. dirus* (ADIR), and *An. stephensi* (ASTE) calculated using maximum-likelihood method (Stamatakis 2014). The five *Anopheles* species for which genes are depicted exhibit gene number contractions representative of those we observe in all *Anopheles* species analyzed, for the *HP1* gene family. Blue highlight encompasses genes related to *D. melanogaster HP1e*, and red highlight encompasses all other anopheline *HP1* gene family members.

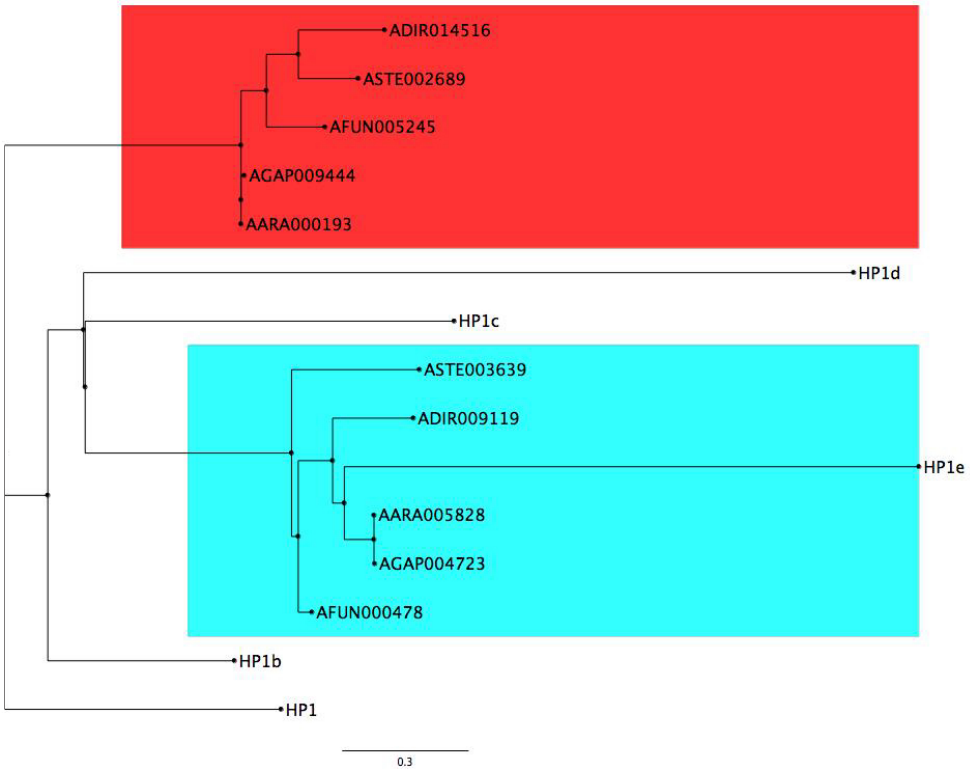


Figure 4.3: Phylogenetic Relationships Among *Heterochromatin Protein-1* Orthologs in *D. melanogaster* and *An. gambiae*

Figure 4.4: Retrotransposition of *CC14* within the genus *Anopheles*

A. Phylogenetic tree depicting retrotransposition event of *CC14* in the Pyretophorus group. Species that possess the retrotransposed gene are annotated with a star and include *An. gambiae, An. arabiensis, An. quadriannulatus, An. merus, An. melas* and *An. epiroticus*. We do not detect a retrotransposed copy of *CC14* in *An. christyi*, but this may be due to a sub-optimal genome assembly for this species (Neafsey et al. 2014). Dendogram is modified from (Neafsey et al. 2013). **B.** Regions of alignment of retrotransposed and original paralogous *CC14* proteins across *Anopheles*. Retrotransposed genes are include "_Retro" at the end of the gene identifier, with red highlight to the left of sequences. Spliced orthologs have a green highlight to left of sequences. Species are given the following identifiers: *An. christyi* (ACHR), *An. gambiae (*AGAP*), An. epiroticus (*AEPI*), An. arabiensis (*AARA*), An. quadriannulatus (*AQUA*), An. merus (*AMEM*), An. stephensi (*ASTE*), An. funestus (*AFUN*), An. albimanus (*ALBI*), An. dirus (*ADIR*), An. atroparvus (*AATE*), An. farauti (*AFAF*), An. melas* (AMEC). Amino acid alignments shown are representations of selected portions of the total open reading frame for each gene, due to the more extensive total lengths of the complete open reading frames. Segments of the open reading frames presented are aa141-180, aa213-252 and aa272-317 in *An. gambiae*.

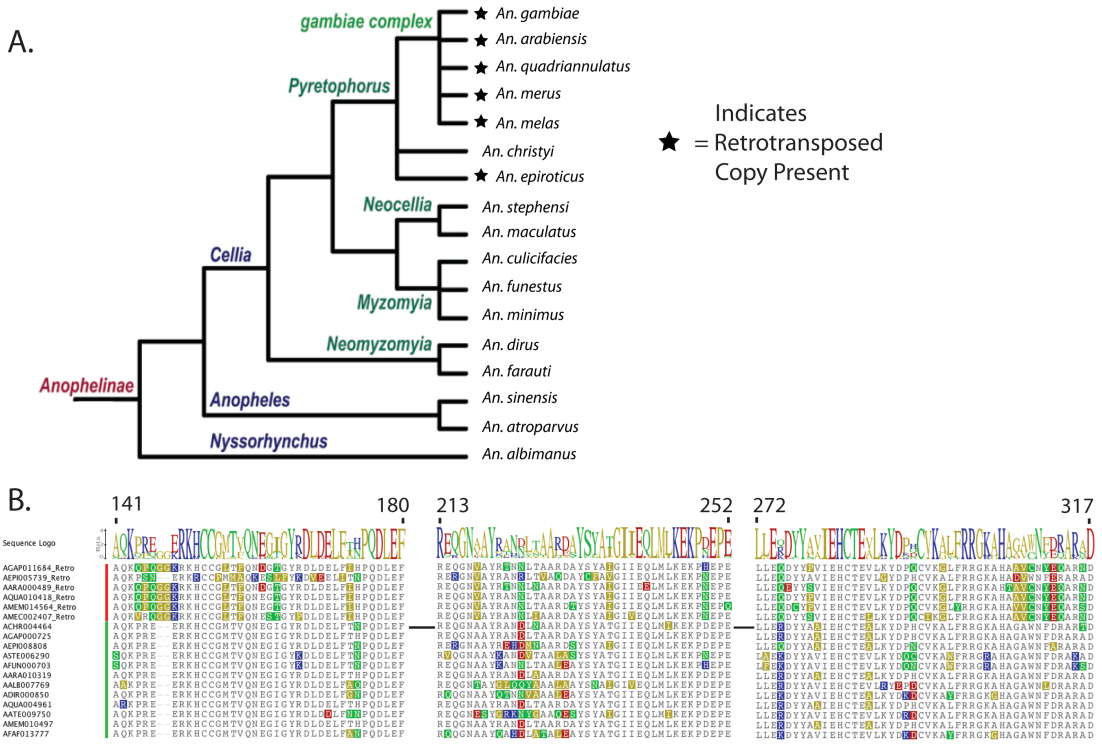


Figure 4.4: Retrotransposition of *CC14* Within the Genus *Anopheles*

Figure 4.5: Alignment of *E2D* ubiquitin-conjugating enzyme genes and homologous retrogenes in anopheline species:

Alignment of orthologous Anopheline proteins, as annotated in VectorBase, to *An. gambiae* ubiquitin-conjugating enzyme E2D (AGAP000145) and to the proteins encoded by retrogenes present in a subset of anopheline species, and by the *D. melanogaster effete* gene. Retrogenes (designated "Species_Retro") were identified, using TBALSTN, in *An. funestus, An. minimus, An. farauti, An. atroparvus, An. darlinigi* and *An. albimanus*. Genes orthologous to the full-length *D. melanogaster effete* gene (FlyBase ID CG7425) and the intron-containing *An. gambiae* ortholog (designated "Species_Effete") have VectorBase IDs ADAC00659, AALB006777, ASIS001446, AATE012345, AFAF019361, ADIR001443, AFUN003878 and AMIN005451. Amino acid substitutions are highlighted based on their polarity (yellow = nonpolar, green = polar, blue = basic, red = acidic). Light blue highlighted boxes indicate regions of increased conservation among genes and retrogenes. Black triangles indicate splice junctions in the *An. gambiae effete* ortholog and other spliced orthologs.

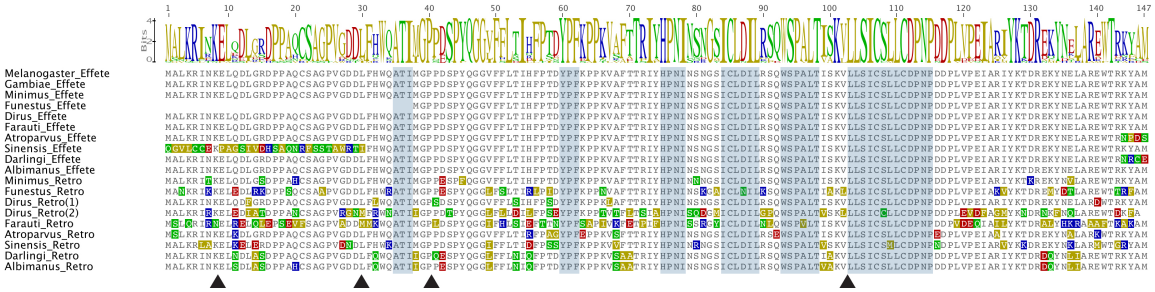


Figure 4.5: Alignment of *E2D* ubiquitin-conjugating enzyme genes and homologous retrogenes in anopheline species:

Figure 4.6: Epigenetic Gene Ensemble Expression in Tissues and Development

A. Principal component analysis (using prcomp function in R (R Core Team 2014)) of Log_{10}(epigenetic modifier gene expression) across tissues in *D. melanogaster* and *An. gambiae*. Expression values were obtained from modENCODE for *D. melanogaster* and MozAtlas for *An. gambiae* (Baker et al. 2011; Celniker et al. 2009). All values were normalized to *Act5C* to control for potential differences relating to magnitude of expression. Arrows indicate tissue-specific components. Topmost vector (30° off-vertical) represents testis expression, next vector clockwise (85° off-vertical) represents ovary expression, while clustered vectors (95° off-vertical) represent carcass, midgut, ovary, head, Malpighian tubules, and salivary gland expression. **B.** Hierarchical clustering of expression of epigenetic gene ensemble members in *An. gambiae* based on RNA sequencing data across four life stages (mixed gender L1, mixed gender L3, adult male, and adult female (Jenkins et al. 2014; Jenkins and Muskavitch 2015). Clustering was performed using Pearson correlation with complete linkage distances. Red bars indicate clustering of the "high expression" gene class (84 genes); green bars indicate the "low expression" gene class (85 genes). **C.** Hierarchical clustering of expression of homologous epigenetic gene ensemble members in *D. melanogaster* based on expression levels identified by modENCODE and listed in FlyBase 5.48 (St Pierre et al. 2014; Celniker et al. 2009). Red bars indicate high expression gene class (50 genes); green bars indicate low expression gene class (119 genes). Comparing heights of same colored bars between panels **B** and **C** reflects the relative number of genes for each class, in each species.

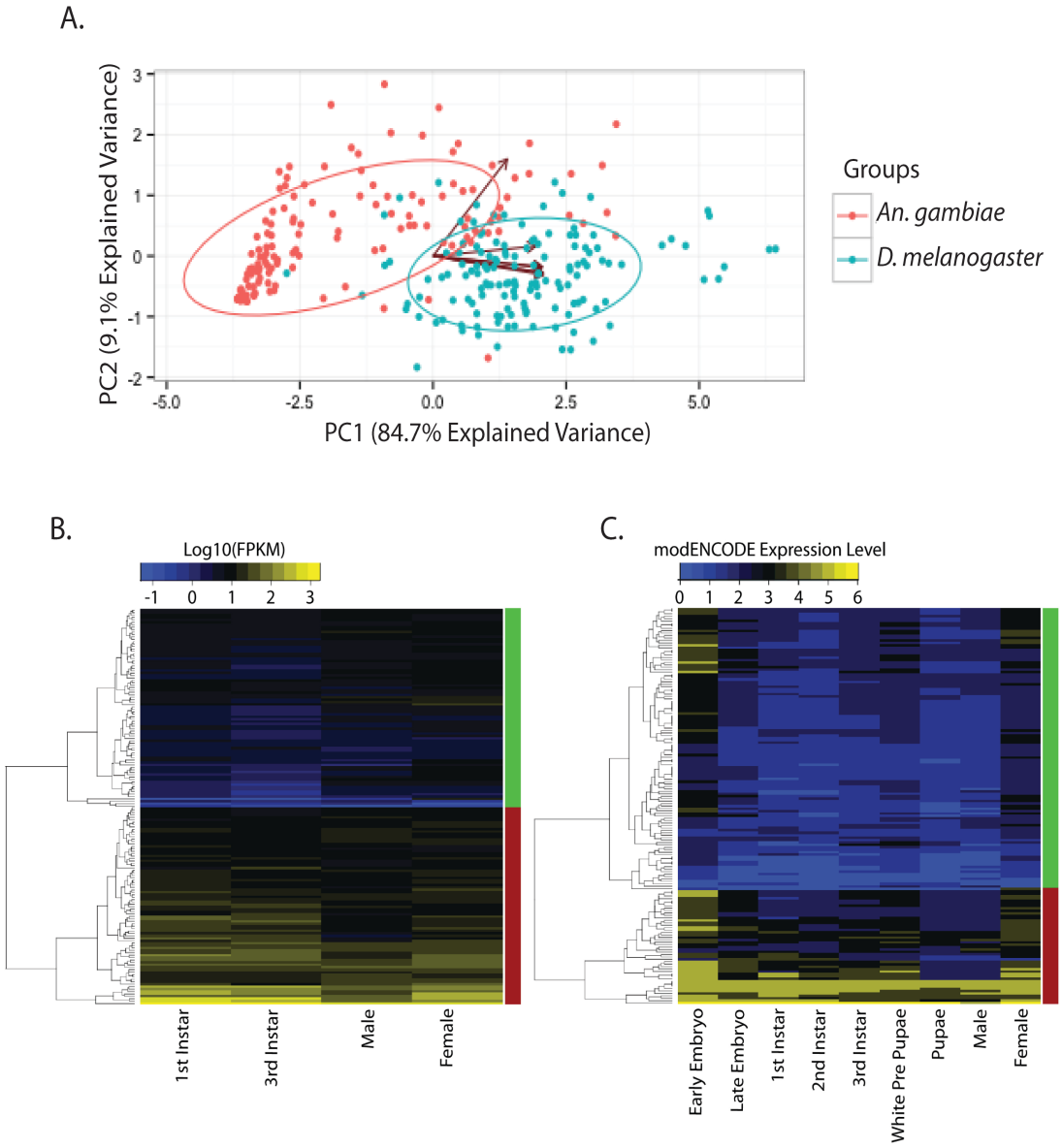


Figure 4.6: Epigenetic Gene Ensemble Expression in Tissues and Development

Figure 4.7: Tissue Expression Difference Between *D. melanogaster* and *An. gambiae*

For each tissue used for principal component analysis (Fig. 4A), the relative expression in *D. melanogaster* was compared to the relative expression in *An. gambiae*. Relative expression was calculated by comparing the gene expression to *ACT5C*. Differences in testis compared to the total group of tissues were statistically significant (p-value <0.0001) using ANOVA.

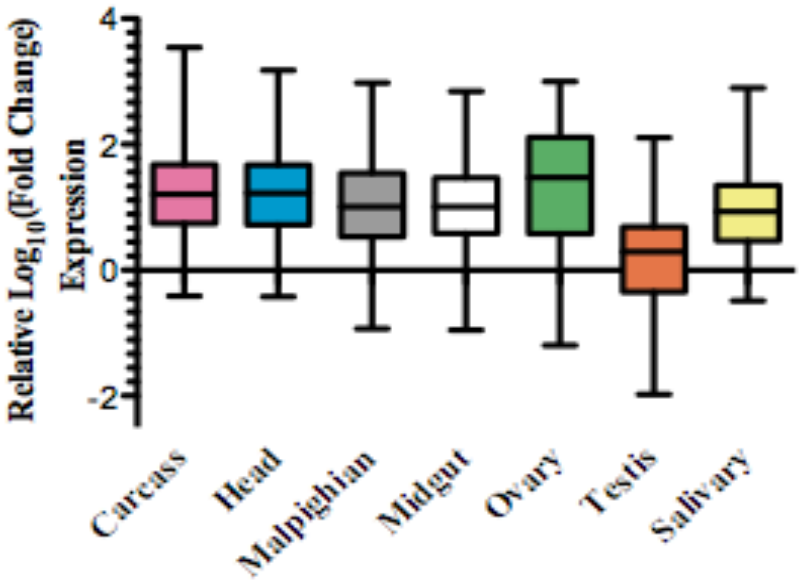


Figure 4.7: Tissue Expression Difference Between *D. melanogaster* and *An. gambiae*

Figure 4.8: GO Terms of Genes with Temporally Unique Expression Profiles Between Species

Epigenetic genes that were not clustered in either high or low expression classes (red or green bars respectively, Fig. 4B,C) in *D. melanogaster* or *An. gambiae* were grouped based upon GO terms. A total of 50 genes had different expression profiles, of which 75 percent possessed GO terms pertaining to acetylation, methylation, SET-N, or complex components.

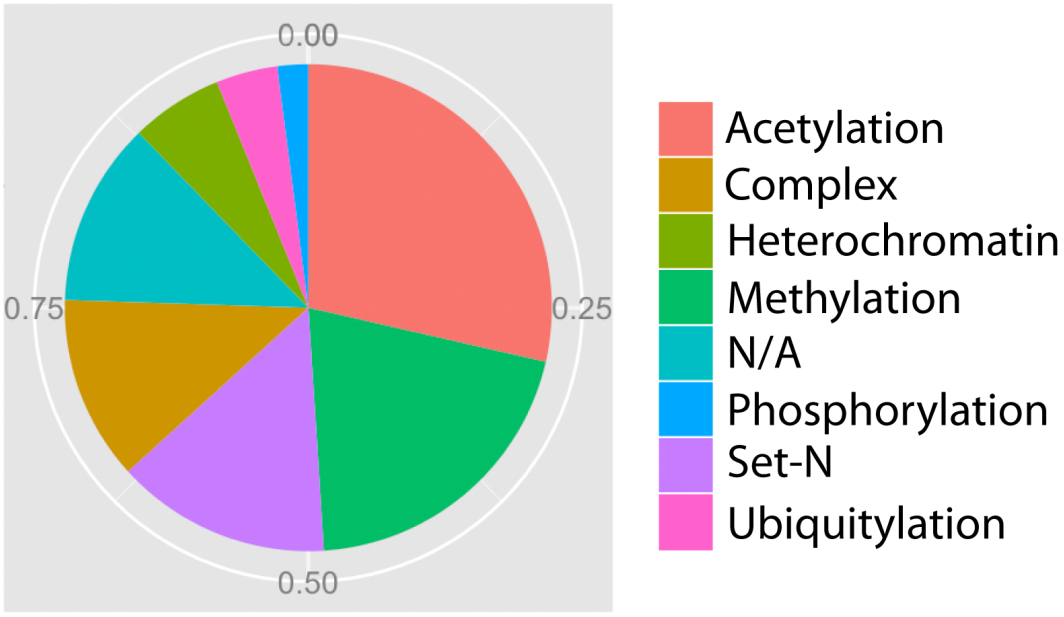


Figure 4.8: GO Terms of Genes with Temporally Unique Expression Profiles Between Species

Appendix:

Rectification of G-Protein Coupled Receptor Gene Models in *Anopheles gambiae*

INTRODUCTION

As disease-carrying vectors of human disease, such as *Aedes aegypti* and *Anopheles gambiae*, begin to harbor insecticide-resistant alleles at greater frequencies, the need to develop novel insecticides has never been greater (Edi *et al*. 2014; Mitchell *et al*. 2014). Currently, many of the drugs used to treat chronic human illnesses target G-protein coupled receptors (GPCRs) (Insel *et al*. 2007; Allen and Roth 2011). GPCRs are proteins that possess seven transmembrane domains, are often stimulated by extracellular ligand recognition, and activate downstream G protein-mediated signal transduction pathways (Katritch *et al*. 2013). Due to the promising pharmacological properties of members of the GPCR superfamily, medical entomologists and vector biologists are now creating insecticides aimed at members of this superfamily expressed in insect vectors (Pates and Curtis 2005b; Meyer et al. 2012; Allen and Roth 2011).

The original annotation of *Anopheles gambiae* GPCR superfamily identified 276 unique genes (Hill *et al*. 2002). Since this initial annotation, multiple additional sub-classes of GPCRs, such as expanded odorant and gustatory receptor families, have been identified *(*Pitts *et al.* 2011*;* Rinker *et al.* 2012*;* Benton 2006*;* Fox *et al*. 2001). In vectors, the odorant and gustatory receptors are perhaps the most extensively studied families of genes due to their importance in vector host-seeking behavior (Lefèvre et al. 2009; Takken and Knols 1999; Carey *et al*. 2010). Multiple studies have identified the specific volatiles lactic acid and ammonia as activating ligands for subsets of odorant receptors and stimuli for host-seeking behavior in both *Aedes* and *Anopheles* mosquitoes (Geier et al. 1999; U. Bernier, D. Kline, S. Allan 2007; Spitzen et al. 2008; Verhulst et al. 2010,

2011). Push-pull vector control methods, in which vectors are pushed away from dwellings with repellents and pulled toward sites away from dwellings with attractants have been developed based on knowledge of gustatory and olfactory preferences (Takken 2010). In addition to gustatory and odorant receptors, GPCRs involved in development, signalling and neuropeptide binding are also of interest as potential insecticide targets.

Annotation of neuropeptides encoded within the *Anopheles gambiae* genome (Riehle *et al*. 2002) occurred contemporaneously with the first annotation of the GPCR superfamily (Hill et al. 2002). Many studies have now characterized individual neuropeptides and their interactions with specific GPCRs. A capa receptor, pyrokin receptors, a FMRFamide receptor, and neuropeptide F receptors – all neuropeptide receptors – have all been cloned from *An. gambiae* and characterized (Duttlinger *et al*. 2003; Olsen *et al*. 2007; Garczynski *et al*. 2005, 2007). The insights gained from these studies have begun to define the roles of specific GPCRs in this mosquito. The dopamine-R2 receptor in *Aedes aegypti* and the octopamine receptor in *An. gambiae* have been characterized, as well, revealing the potential activation of these receptors in neuronal signalling (Conley *et al*. 2015). Recently, high-throughput systems have been used for small-molecule screening against mosquito GPCRs in order to identify molecules that may functions as leads for the development of insecticides (Pridgeon *et al*. 2009; Rinker *et al*. 2012). One highly desirable prerequisite when undertaking the identification of chemical leads directed toward drug targets such as GPCRs via cloning and subsequent downstream signaling assays is knowledge of the correct gene model for the GPCR of interest.

A major problem in genome sequencing is the unambiguous identification of transcribed regions and accurate gene models within the genome (Yandell and Ence 2012; Wilhelm *et al*. 2010). In the past, computational algorithms based on the identification of motifs associated with splice junctions, definition of coding regions etc., coupled with assessment of gene orthologies – like those used in the MAKER program (Holt and Yandell 2011)– were used to predict gene models within the transcriptome. Such pipelines operate at low cost, with relatively high throughput (Holt and Yandell 2011). Yet, these methods are highly error-prone, as up to 40% of gene families defined with such algorithms possess an incorrect number of members and 5' and 3' untranslated regions (UTRs) are often accurately identified due to low sequence homology within UTRs (Denton *et al*. 2014). Advanced transcriptome annotation based on high-coverage RNA sequencing (RNAseq) circumvents the problems associated with algorithm/orthology-based methods, since full-length transcripts, including UTRs and splice junctions, can be directly identified in any organism, tissue or cell type during any developmental stage with sufficient read depth (Dhahbi *et al. 2011*; Lu and Bushe*l 2013*; Kim *et al*. 2013).

In this study, we attempted to improve the accuracies of gene models for GPCRs other than olfactory, gustatory and opsin GPCRs in *An. gambiae*, using deep RNAseq. We find that among the 93 GPCRs reannotated, 83 were represented by inaccurate or incomplete gene models. Among these, 64 genes contained unannotated 5' UTRs and 62 contained unannotated 3' UTRs. In addition, we identified new exons in 55 genes, including 11 new protein-coding exons, and we were able to identify multiple protein-coding splice

variants in 11 genes. These findings illustrate that deep RNAseq is an ideal tool for correcting inaccurate gene annotations within the *An. gambiae* genome for members of the GPCR superfamily, and that deep RNAseq can enable more accurate cloning and characterization of GPCRs in the future.

RESULTS/DISCUSSION

In order to determine the annotation accuracy of the currently identified GPCRs within the AgamP3.7 (V3.7) *An. gambiae* transcriptome that are not olfactory receptors, gustatory receptors, or opsins, we aligned our RNAseq reads to the *An. gambiae* PEST genome and compared the AgamP3.7 transcriptome to the transcriptome supported by our RNAseq reads. Among the 93 genes within this subset of the GPCR superfamily, 83 genes were found to possess unannotated regions in comparison to current V3.7 models (Additional File A.1 and Table A.1). Among these 83 genes, a majority of the unannotated regions constituted 5' and 3' UTRs, as 64 genes contained unannotated 5' UTRs and 62 genes contained unannotated 3' UTRs. *GPRMTN* and *GPR5HT1A*, which were found to have unannotated exons in their 5' and 3' UTRs, are examples of such genes (Fig. A.1B and A.1C). Among the 64 genes with previously unannotated 5' UTRs, 51 genes contained previously unannotated exons within 5' UTRs, while only 13 of the 62 genes with previously unannotated 3' UTRs contained previously unannoted exons within 3' UTRs. This finding has many implications, specifically for understanding the transcriptional control of GPCRs. Recent genome-wide studies in mammals have shown the presence of enhancers in genomic regions that were previously annotated as transcriptionally silent (Hallikas et al. 2006). With the introduction of RNAseq, enhancer elements, specifically those that map within the 5' UTRs and 3' UTRs, can be associated with the correct transcript. Our reannotation of 5' and 3' UTRs in *An. gambiae* GPCR gene models will facilitate further studies of the elements controlling GPCR transcriptional expression by enabling the identification of similar motifs, and inference

of the transcription factors that bind within these regions and contribute to gene regulation.

We were able to detect multiple protein-coding differences between the V3.7 reference transcriptome and the revised GPCR gene models supported by our deep RNAseq data, in addition to newly annotated untranslated regions we defined within GPCR gene models. There are 11 GPCR gene models (Additional File A.1 and Table A.1) that include multiple protein-coding splice variants, and we were able to identify 11 new protein-coding exons in total (Additional File A.1 and Table A.1, Fig. A.1A). Among the genes that encode multiple protein isoforms, *GPRoar1* has been cloned and characterized in *An. gambiae* (Kastner *et al*. 2014). The clones reported in this work are consistent with at least one of the protein coding sequences (CDS) that are supported by our RNAseq. As the existence of these splice variants was probably unknown to the group cloning the *GPRoar1* gene, it is possible that the uncloned coding variants possess similar or differing ligand-dependent activation characteristics.

An example of a gene encoding multiple splice variants and previously unannotated 5'and 3' UTRs is *GPRfsh*, which encodes two splice variants that differ in their C-termini (Figure A.2, Fig. A.3A). The RA-form identified by RNAseq is consistent with the previous annotation of *GPRfsh* in V3.7 and VectorBase (Fig. A.2A). The newly identified RB-form encodes a frame shift between exons 9 and 10, leading to differing C-termini for the A and B receptor isoforms. RNAseq reads support expression of the A form during all four life-stages analysed, while the B form is expressed predominantly

during the first larval instar stage (Fig. A.3B and A.3C). The *GPRfsh* ortholog in *D. melanogaster*, *LGR1*, which also encodes two splice variants (Fig. A.2) is required for developmental progression from the late larval stage to the pupal stage (Vandersmissen *et al*. 2014). This suggests that *GPRfsh* could encode an insecticidable target in *An. gambiae*, and that small molecule inhibitors of the receptor could function as insecticides in mosquitoes.

Overall, the ability to combat malaria using vector-targeted interventions will rely on our continuing abilities to understand the mosquito genome and mosquito behavior, and to create novel insecticides targeting specific molecular machinery. This study has shown that previous annotations of GPCRs – a highly promising class of protein targets for insecticides – have often been inaccurate and have missed many transcribed regions that would encompass a complete GPCR transcriptome, particularly 5' and 3' UTRs. By extending efforts such as those we have undertaken and completing an accurate annotation of the GPCR transcriptome of *An. gambiae*, high-throughput assays aimed at identifying agonists and antagonists of *An. gambiae* GPCRs can be pursued with greater confidence in the future, based on more complete and accurate knowledge of correct GPCR gene models in this vector insect.

METHODS

Sequencing of *An. gambiae* RNA and alignment of RNAseq reads was undertaken as described in Chapters 3 and 4 of this thesis. In short, an *An. gambiae* G3 colony (courtesy of Dr. Flaminia Catterucia, Harvard School of Public Health, Boston, MA, USA) was reared with an 11:11 Light:Dark (L:D) photoperiod with a one-hour crepuscular period between light and dark stages, and fed 10 percent glucose solution *ad libitum*. First larval instar (L1) and third larval instar (L3) stages were removed from the colony within 12 hours of emergence from chorion or previous larval cuticle, respectively. Sample preparation and analysis were performed at the Broad Institute (Cambridge, MA), using a Qiagen RNAeasy Mini Kit for RNA extraction and the Illumina TruSeq RNA Sample Preparation Kit v2 for library generation. Then, high read depth (HRD) paired-end RNA sequencing was performed on the HiSeq 2000 platform. HRD reads were soft-clipped and aligned to the *An. gambiae* PEST genome assembly (www.vectorbase.org) (Megy *et al*. 2012). Splice junction mapping/alignment was performed using Tophat2 (version 2.0.10) with a mismatch (-N) appropriation of 3 and a read-edit-dist of 3 (Kim *et al*. 2013). Reads were visualized using Broad Institutes Integrative Genomics Viewer and compared to the *An. gambiae* AgamP3.7 gene annotation.

Unannotated splice junctions were annotated based on a splice junction represented within at least 5 reads with greater than 5 basepairs on either side of the split read. We set the following criteria: each splice junction must possess a minimum of five reads supporting the splice junction, with a minimum of five base pairs on either side of the

splice junction. Untranslated regions were determined to end where less than three reads could be aligned to the genome. Exon read counts were identified using DEXSeq package in R (Anders *et al*. 2012). To validate the discovery of exons, we performed PCR across three new exon junctions in *GPR5HT2A* (new protein coding exon, 335 bp predicted PCR length), *GPRMTN* (new 5' UTR exon, 246 bp predicted PCR length), and *GPR5HT1A* (new 3' UTR exon, 419 bp predicted PCR length) (Fig. A.1D). Amplification was performed using AccuPrime PFX (Life Technologies) using a 58 degree Celsius annealing temperature and 35 cycles of amplification. Primers used were:
5HT2A_F:AACAAAGCGGTCGAGATGAG,
5HT2A_R:GGTACGCTGTTGAGGTGTATC,
MTN_F:TTCACAACCCACCAACCAA,
MTN_R:CCACAATTCCCGTGACCATAA,
5HT1A_F:CTACTTCAACTCCACGCTCAA, and
5HT1A_R:ACGACGACATCCTTACATCATC.

TABLES/FIGURES AND LEGENDS

Table A.1: Rectification of *An. gambiae* GPCR Gene Models

Number of AgamP3.7 GPCR gene models that were found to be missing the described feature

New Exons Identified in 5' UTR	51
Unannotated 5' UTR Without New Exons	13
New Exons Identified in 3' UTR	13
Unannotated 3' UTR Without New Exons	49
Genes Expressing Multiple Isoforms	11
New Protein Coding Exons	11
Total Number of Genes with Unannotated Exons	55

Table A.1: Rectification of *An. gambiae* GPCR Gene Models

Figure A.1: Examples of GPCR Gene Model Rectifications

Sashimi plots and previous model annotations for *GPR5HT2A* (A), *GPRMTN* (B) and *GPR5HT1A* (C). A. *GPR5H2A* includes a previously unannotated protein coding exon. B. *GPTMTN* includes a previously unannotated 5' UTR exon. C. *GPR5HT1A* incl;udes a previously unannotated 3' UTR exon. D. PCR validation across splice junctions between previously unannotated exons and known exons. (Columns: 1. 100 bp ladder, 2. *GPT5HT1A*, 3. *GPR5HT2A*, 4. *GPRMTN*)

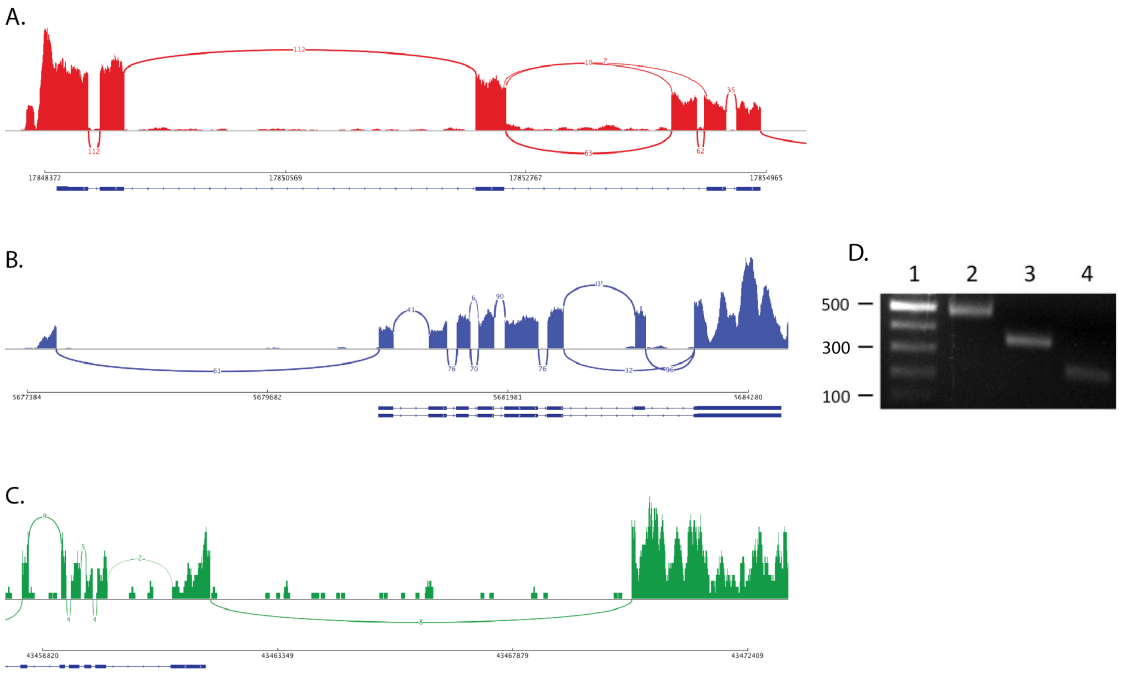


Figure A.1: Examples of GPCR Gene Model Rectifications

Figure A.2: Peptide Alignment of *An. gambiae GPRFSH* Splice Variants and *D. melanogaster* Orthologs

Alignment of two *An. gambiae* splice variants (*GPRFSH* RA form and RB form) with their orthologous genes in *D. melanogaster* (*LGR1* PA and PB forms). Red highlighted boxes indicate putative transmembrane domains.

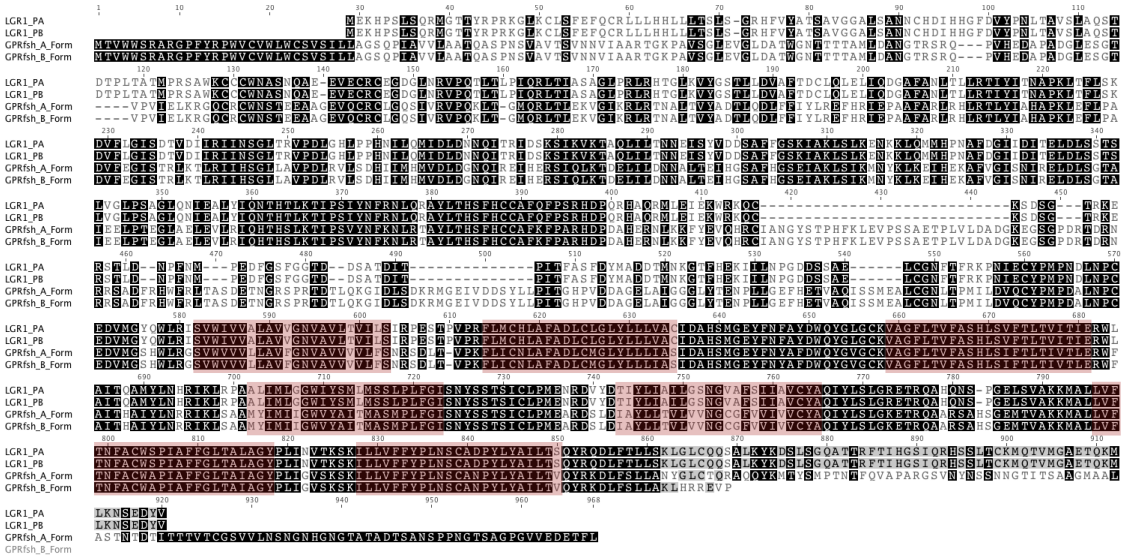
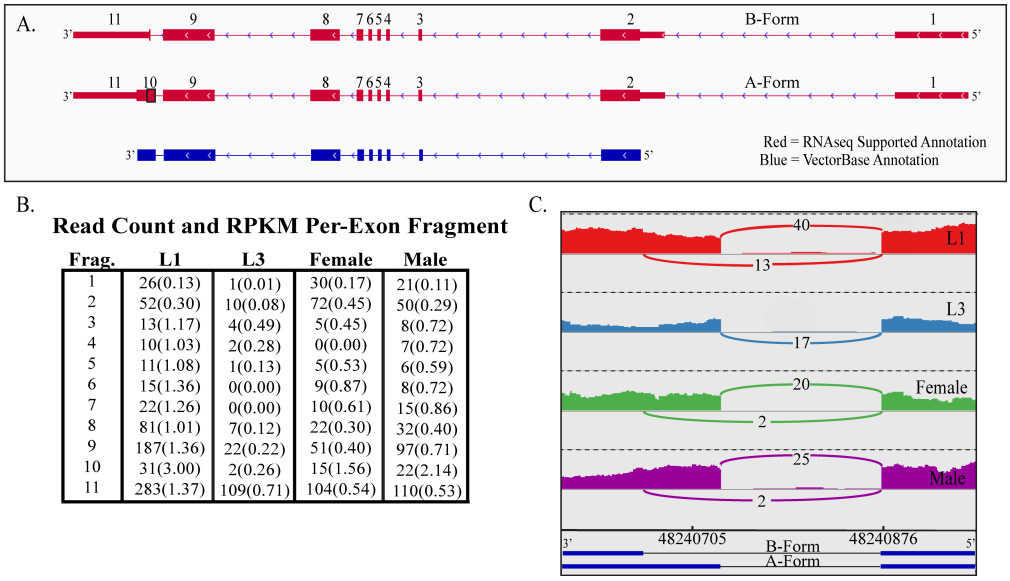


Figure A.2: Peptide Alignment of *An. gambiae GPRFSH* Splice Variants and *D. melanogaster* Orthologs

Figure A.3: *GPRFSH* Gene Model Rectification and Splice Junctions

A. *GPRFSH* models that are RNAseq-supported (red) and the original Vectorbase AgamP3.7 gene model (blue). Exon/fragments each possess a unique number, with section 10 being unique to the RA-Form. **B.** Raw RNAseq counts that align to each fragment, with the RPKM values for each fragment in parentheses. RPKM is the read count normalized to length of the fragment, based on the total number of reads aligned during that developmental stage. **C.** Sashimi plot of the gene model that contains splice junctions for each life stage [L1 (red), L3 (blue), Female (green), and Male (purple)] with portions of the gene models at the bottom.



B.

Read Count and RPKM Per-Exon Fragment

Frag.	L1	L3	Female	Male
1	26(0.13)	1(0.01)	30(0.17)	21(0.11)
2	52(0.30)	10(0.08)	72(0.45)	50(0.29)
3	13(1.17)	4(0.49)	5(0.45)	8(0.72)
4	10(1.03)	2(0.28)	0(0.00)	7(0.72)
5	11(1.08)	1(0.13)	5(0.53)	6(0.59)
6	15(1.36)	0(0.00)	9(0.87)	8(0.72)
7	22(1.26)	0(0.00)	10(0.61)	15(0.86)
8	81(1.01)	7(0.12)	22(0.30)	32(0.40)
9	187(1.36)	22(0.22)	51(0.40)	97(0.71)
10	31(3.00)	2(0.26)	15(1.56)	22(2.14)
11	283(1.37)	109(0.71)	104(0.54)	110(0.53)

C.

S

Figure A.3: *GPRFSH* Gene Model Rectification and Splice Junctions

LIST OF ADDITIONAL FILES:

Additional_File_2.1.png: Photopreference assay schematic

Additional_File_2.2.docx: Primers Used in Diurnal Time Course Analyses

Additional_File_2.3.xlsx: *Opsin* gene expression data

Additional_File_3.1.txt: Expression of *Anopheles gambiae* genes across life-stages.

Additional_File_3.2.xls: DAVID enrichment classes for each life-stage comparison.

Additional_File_3.3.gtf: GTF file produced by Cufflinks of newly annotated genes.

Additional_File_3.4.txt: List of identifiers for lncRNA and putative-protein coding genes identified.

Additional_File_3.5.txt: Cufflinks class codes for all newly identified genes.

Additional_File_3.6.txt: Expression of all *Anopheles gambiae* genes, including newly identified genes across life-stages.

Additional File_3.7.txt: Differential expression analysis of all *Anopheles gambiae* genes, including newly identified genes across all life-stages.

Additional_File_3.8.gtf: GTF file of all lncRNA that exhibit 50% overlap to previously identified lncRNAs in *Anopheles gambiae* midgut.

Additional_File_3.9.xls: List of lncRNA that have 50% overlap to a gene on the complementary strand, as defined by Cufflinks.

Additional_File_3.10.fasta: FASTA file containing sequences of all lncRNA.

Additional_File_3.11.fasta: FASTA file containing coding sequence of all putative protein-coding genes identified in this study.

Additional_File_3.12.fasta: FASTA file containing full length mRNA sequence of all putative protein-coding genes identified in this study.

Additional_File_3.13.xls: File containing putative peptides identified in this study with alignment scores and PhyloCSF scores.

Additional_File_3.14.gff3: GFF3 file of all genes identified in this study.

Additional_File_4.1.docx: Epigenetic modifier orthology table

Additional_File_4.2.xlsx: Epigenetic modifier expression values

Additional_File_4.3.docx: Selected epigenetic gene orthology in *Anopheles* species

Additional_File_4.4.xlsx: dN/dS values of epigenetic modifiers in *Anopheles gambiae* and *Drosophila melanogaster*

Additional_File_A.1.docx: GPCR gene model re-annotation table

REFERENCES:

Adhin MR, Labadie-Bracho M, Vreden SG. 2012. Status of potential PfATP6 molecular markers for artemisinin resistance in Suriname. *Malar J* **11**: 322.

Alessandro UD, Buttie H. 2001. History and importance of antimalarial drug resistance. *Trop Med Int Heal* **6**: 845–848.

Ali A, Nayar JK, Knight JW, Stanley BH. 1989. Attraction of Florida Mosquitoes (Diptera:Culicidae) To Artificial Light in the Field. *57th Annu Conf Calif Mosq Vector Control Assoc Inc*.

Allan S a, Day JF, Edman JD. 1987. Visual ecology of biting flies. *Annu Rev Entomol* **32**: 297–316.

Allen J a, Roth BL. 2011. Strategies to discover unexpected targets for drugs active at g protein-coupled receptors. *Annu Rev Pharmacol Toxicol* **51**: 117–44.

Alonso PL, Tanner M. 2013. Public health challenges and prospects for malaria control and elimination. *Nat Med* **19**: 150–5.

Alphey L, Benedict M, Bellini R, Clark GG, Dame DA, Service MW, Dobson SL. 2010. Sterile-Insect Methods for Control of Mosquito-Borne Diseases:An Analysis. *Vector-Borne Zoonotic Dis* **10**: 295–311.

Ammosova T, Platonov M, Yedavalli VRK, Obukhov Y, Gordeuk VR, Jeang K-T, Kovalskyy D, Nekhai S. 2012. Small molecules targeted to a non-catalytic "RVxF" binding site of protein phosphatase-1 inhibit HIV-1. *PLoS One* **7**: e39481.

Anders S, Reyes A, Huber W. 2012. Detecting differential usage of exons from RNA-seq-npre20126837-2.pdf. 2008–2017.

Andrews JM, Quinby GE, Langmuid AD. 1945. Malaria Eradication in the United States *. *Am J Public Health* **40**.

Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, Antelo B, Bartholomay L, Bidwell S, Caler E, Camara F, et al. 2010. Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. *Science* **330**: 86–8.

Ariey F, Witkowski B, Amaratunga C, Beghain J, Langlois A-C, Khim N, Kim S, Duru V, Bouchier C, Ma L, et al. 2014. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. *Nature* **505**: 50–5.

Arnou B, Montigny C, Morth JP, Nissen P, Jaxel C, Møller J V, Maire M Le. 2011. The Plasmodium falciparum Ca(2+)-ATPase PfATP6: insensitive to artemisinin, but a potential drug target. *Biochem Soc Trans* **39**: 823–31.

Arrowsmith CH, Bountra C, Fish P V, Lee K, Schapira M. 2012. Epigenetic protein families: a new frontier for drug discovery. *Nat Rev Drug Discov* **11**: 384–400.

Avril M, Tripathi AK, Brazier AJ, Andisi C, Janes JH, Soma VL, Sullivan DJ, Bull PC, Stins MF, Smith JD. 2012. A restricted subset of var genes mediates adherence of Plasmodium falciparum-infected erythrocytes to brain endothelial cells. *Proc Natl Acad Sci U S A* **109**: E1782–90.

Baker D a, 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.

Barbosa S, Hastings IM. 2012. The importance of modelling the spread of insecticide resistance in a heterogeneous environment : the example of adding synergists to bed nets. *Malar J* **11**: 1–12.

Bártová E, Krejcí J, Harnicarová A, Galiová G, Kozubek S. 2008. Histone modifications and nuclear architecture: a review. *J Histochem Cytochem* **56**: 711–21.

Baruch DI, Pasolske BL, Singh HB, Bi X, Ma XC, Feldman M, Taraschi TF, Howard RJ. 1995. Cloning the P . falciparum Gene Encoding PfEMPl , a Malarial Variant Antigen and Adherence Receptor on the Surface of Parasitized Human Erythrocytes. *Cell* **82**: 77–87.

Bellucci M, Agostini F, Masin M, Tartaglia GG. 2011. Predicting protein associations with long noncoding RNAs. *Nat Methods* **8**: 444–5.

Van Bemmel JG, Filion GJ, Rosado A, Talhout W, de Haas M, van Welsem T, van Leeuwen F, van Steensel B. 2013. A network model of the molecular organization of chromatin in Drosophila. *Mol Cell* **49**: 759–71.

Bentley MT, Kaufman PE, Kline DL, Hogsette J a. 2009. Response of adult mosquitoes to light-emitting diodes placed in resting boxes and in the field. *J Am Mosq Control Assoc* **25**: 285–91.

Benton R. 2006. On the ORigin of smell: odorant receptors in insects. *Cell Mol Life Sci* **63**: 1579–85.

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Berthomieu A, Marquine M, Raymond M. 2004. The unique mutation in ace-1 giving high insecticide. *Insect Mol Biol* **13**: 1–7.

Besansky NJ, Hill C a, Costantini C. 2004. No accounting for taste: host preference in malaria vectors. *Trends Parasitol* **20**: 249–51.

Bhan A, Hussain I, Ansari KI, Kasiri S, Bashyal A, Mandal SS. 2013. Antisense transcript long noncoding RNA (lncRNA) HOTAIR is transcriptionally induced by estradiol. *J Mol Biol* **425**: 3707–22.

Bhattarai A, Ali AS, Kachur SP, Mårtensson A, Abbas AK, Khatib R, Al-Mafazy A-W, Ramsan M, Rotllant G, Gerstenmaier JF, et al. 2007. Impact of artemisinin-based combination therapy and insecticide-treated nets on malaria burden in Zanzibar. *PLoS Med* **4**: e309.

Bill A, Hall ML, Borawski J, Hodgson C, Jenkins J, Piechon P, Popa O, Rothwell C, Tranter P, Tria S, et al. 2014. Small molecule-facilitated degradation of ANO1 protein: a new targeting approach for anticancer therapeutics. *J Biol Chem* **289**: 11029–41.

Bir A, Sidhu S, Verdier-pinard D, Fidock DA. 2002. Chloroquine Resistance in Plasmodium falciparum Malaria Parasites Conferred by pfcrt Mutations. *Science (80-)* **298**: 210–212.

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res* **14**: 988–95.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF a, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**: 708–15.

Blin G, Denise A, Dulucq S, Herrbach C, Touzet H. 2009. Alignments of RNA structures. *IEEE/ACM Trans Comput Biol Bioinform* **7**: 309–22.

Bopp SER, Manary MJ, Bright a T, Johnston GL, Dharia N V, Luna FL, McCormack S, Plouffe D, McNamara CW, Walker JR, et al. 2013. Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families. *PLoS Genet* **9**: e1003293.

Boulanger M-C, Miranda TB, Clarke S, Di Fruscio M, Suter B, Lasko P, Richard S. 2004. Characterization of the Drosophila protein arginine methyltransferases DART1 and DART4. *Biochem J* **379**: 283–9.

Bourgon R, Delorenzi M, Sargeant T, Hodder AN, Crabb BS, Speed TP. 2004. The serine repeat antigen (SERA) gene family phylogeny in Plasmodium: the impact of GC content and reconciliation of gene and species trees. *Mol Biol Evol* **21**: 2161–71.

Bousema T, Drakeley C. 2011. Epidemiology and infectivity of Plasmodium falciparum and Plasmodium vivax gametocytes in relation to malaria control and elimination. *Clin Microbiol Rev* **24**: 377–410.

Bracken AP, Helin K. 2009. Polycomb group proteins: navigators of lineage pathways led astray in cancer. *Nat Rev Cancer* **9**: 773–84.

Braks MA., Anderson RA, Knols BGJ. 1999. Infochemicals in Mosquito Host Selection: Human Skin Microflorda and Plasmodium Parasites. *Parasitol Today* **15**: 409–413.

Branciamore S, Rodin AS, Riggs AD, Rodin SN. 2014. Enhanced evolution by stochastically variable modification of epigenetic marks in the early embryo. *Proc Natl Acad Sci U S A* **111**: 6353–8.

Bray PG, Mungthin M, Ridley RG, Ward S a. 1998. Access to hematin: the basis of chloroquine resistance. *Mol Pharmacol* **54**: 170–9.

Briegel H, Horler E. 1993. Multiple Blood Meals as a Reproductive Strategy in Anopheles (Diptera: Culicidae). *J Med Entomol* **11**: 975–985.

Browne SM, Bennett GF. 1981. Response of mosquitoes (Diptera: Culicidae) to visual stimuli. *J Med Entomol* **18**: 505–521.

Burkett DA, Butler JF, The S, Entomologist F, Dec N. 2012. Laboratory Evaluation of Colored Light as an Attractant for Female Aedes aegypti , Aedes albopictus, Anopheles quadrimaculatus, and Culex nigripalpus. *Florida Entomol* **88**: 383–389.

Burt A. 2014. Heritable strategies for controlling insect vectors of disease. *Philos Trans R Soc Lond B Biol Sci* **369**: 20130432.

Cantone I, Fisher AG. 2013. Epigenetic programming and reprogramming during development. *Nat Struct Mol Biol* **20**: 282–289.

Carey AF, Wang G, Su C-Y, Zwiebel LJ, Carlson JR. 2010. Odorant reception in the malaria mosquito Anopheles gambiae. *Nature* **464**: 66–71.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309**: 1559–63.

Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, Macalpine DM, et al. 2009. Unlocking the secrets of the genome. **459**: 927–930.

Chaves LF, Harrington LC, Keogh CL, Nguyen AM, Kitron UD. 2010. Blood feeding patterns of mosquitoes: random or structured? *Front Zool* **7**: 3.

Chen Q, Fernandez V, Sundstrom A, Schlichtherle M, Datta S, Hagblom P, Wahlgren M. 1998. Developmental selection of var gene expression in Plasmodium falciparum. *Nature* **394**: 392–395.

Chu C, Qu K, Zhong FL, Artandi SE, Chang HY. 2011. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell* **44**: 667–78.

Claridge-Chang A, Wijnen H, Naef F, Boothroyd C, Rajewsky N, Young MW. 2001. Circadian regulation of gene expression systems in the Drosophila head. *Neuron* **32**: 657–71.

Clements AN. 1999. *The Biology of Mosquitoes*. CUNY Animal Behavior Initiative, New York.

Coats JR. 1990. Mechanisms of Toxic Action and Structure- Activity Relationships for Organochiorine and Synthetic Pyrethroid Insecticides. *Environ Health Perspect* **87**: 255–262.

Cohuet A, Harris C, Robert V, Fontenille D. 2010. Evolutionary forces on Anopheles: what makes a malaria vector? *Trends Parasitol* **26**: 130–6.

Conaway RC, Conaway JW. 2009. The INO80 chromatin remodeling complex in transcription, replication and repair. *Trends Biochem Sci* **34**: 71–7.

Cong L, Ann Ran F, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, et al. 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (80-)* **339**: 819–824.

Conley JM, Meyer JM, Nuss AB, Doyle TB, Savinov SN, Hill CA, Watts VJ. 2015. Evaluation of Aa DOP2 Receptor Antagonists Reveals Antidepressants and Antipsychotics as Novel Lead Molecules for Control of the Yellow Fever Mosquito , Aedes aegypti s.

Consortium TGO. 2000. Gene Ontology : tool for the unification of biology. *Nature* **25**: 25–29.

Cork a., Park KC. 1996. Identification of electrophysiologically-active compounds for the malaria mosquito, Anopheles gambiae, in human sweat extracts. *Med Vet Entomol* **10**: 269–276.

Cornet S, Nicot A, Rivero A, Gandon S. 2014. Evolution of plastic transmission strategies in avian malaria. *PLoS Pathog* **10**: e1004308.

Costa V, Angelini C, De Feis I, Ciccodicola A. 2010. Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotechnol* **2010**: 853916.

Costantini C, Birkett MA, Gibson G, Ziesmann J, Sagnon NF, Mohammed HA. 2001. Electroantennogram and behavioural responses of the malaria vector Anopheles gambiae to human-specificc sweat components. *Med Vet Entomol* **15**: 259–266.

Crawford JE, Guelbeogo WM, Sanou A, Traoré A, Vernick KD, Sagnon N, Lazzaro BP. 2010. De novo transcriptome sequencing in Anopheles funestus using Illumina RNA-seq technology. *PLoS One* **5**: e14202.

Daborn PJ, Lumb C, Harrop TWR, Blasetti A, Pasricha S, Morin S, Mitchell SN, Donnelly MJ, Müller P, Batterham P. 2012. Using Drosophila melanogaster to validate metabolism-based insecticide resistance from insect pests. *Insect Biochem Mol Biol* **42**: 918–24.

Das S, Dimopoulos G. 2008. Molecular analysis of photic inhibition of blood-feeding in Anopheles gambiae. *BMC Physiol* **8**: 23.

Davies TGE, Field LM, Usherwood PNR, Williamson MS. 2007. DDT, pyrethrins, pyrethroids and insect sodium channels. *IUBMB Life* **59**: 151–62.

Dawson M a, Kouzarides T. 2012. Cancer epigenetics: from mechanism to therapy. *Cell* **150**: 12–27.

Dekker T, Steib B, Carde RT, Geier M. 2002. L-lactic acid: a human-signifying host cue for the anthropophilic mosquito Anopheles gambiae. *Med Vet Entomol* **16**: 91–98.

Dekker T, Takken W, Knols BGJ, Bouman E, Laak S, Bever A, Huisman PWT. 1998. Selection of biting sites on a human host by Anopheles gambiae s.s., An. arabiensis and An. quadriannulatus. *Entomol Exp Appl* **87**: 295–300.

Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. 2014. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLoS Comput Biol* **10**: e1003998.

Depinay J-MO, Mbogo CM, Killeen G, Knols B, Beier J, Carlson J, Dushoff J, Billingsley P, Mwambi H, Githure J, et al. 2004. A simulation model of African Anopheles ecology and population dynamics for the analysis of malaria transmission. *Malar J* **3**: 29.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**: 1775–89.

Dhahbi JM, Atamna H, Boffelli D, Magis W, Spindler SR, Martin DIK. 2011. Deep sequencing reveals novel microRNAs and regulation of microRNA expression during cell senescence. *PLoS One* **6**: e20509.

Djimde A, Doumbo OK, Cortese JF, Al. E. 2001. A molecular marker for Chloroquine-resistant falciparum Malaria. *N Engl J Med* **344**: 257–263.

Djouaka RF, Bakare A a, Coulibaly ON, Akogbeto MC, Ranson H, Hemingway J, Strode C. 2008. Expression of the cytochrome P450s, CYP6P3 and CYP6M2 are significantly elevated in multiple pyrethroid resistant populations of Anopheles gambiae s.s. from Southern Benin and Nigeria. *BMC Genomics* **9**: 538.

Dondorp AM, Nosten F, Al. E. 2009. Artemisinin Resistance in Plasmodium falciparum Malaria. *N Engl J Med* **361**: 455–467.

Dottorini T, Nicolaides L, Ranson H, Rogers DW, Crisanti A, Catteruccia F. 2007. A genome-wide analysis in Anopheles gambiae mosquitoes reveals 46 male accessory gland genes, possible modulators of female behavior. *Proc Natl Acad Sci U S A* **104**: 16215–20.

Dulai KS, Dornum M Von, Mollon JD, Hunt DM. 1999. The Evolution of Trichromatic Color Vision by Opsin Gene Duplication in New World and Old World Primates The Evolution of Trichromatic Color Vision by Opsin Gene Duplication in New World and Old World Primates. *Genome Res* 629–638.

Duraisingh MT, Curtis J, Warhurst DC. 1998. Plasmodium falciparum : Detection of Polymorphisms in the Dihydrofolate Reductase and Dihydropteroate Synthetase Genes by PCR and Restriction Digestion. *Exp Parasitol* **8**: 1–8.

Duttlinger A, Mispelon M, Nichols R. 2003. The structure of the FMRFamide receptor and activity of the cardioexcitatory neuropeptide are conserved in mosquito. *Neuropeptides* **37**: 120–126.

Eckstein-Ludwig U, Webb RJ, Van Goethem ID a, East JM, Lee a G, Kimura M, O'Neill PM, Bray PG, Ward S a, Krishna S. 2003. Artemisinins target the SERCA of Plasmodium falciparum. *Nature* **424**: 957–61.

Edi C V, Djogbénou L, Jenkins AM, Regna K, Muskavitch M a T, Poupardin R, Jones CM, Essandoh J, Kétoh GK, Paine MJI, et al. 2014. CYP6 P450 Enzymes and ACE-1 Duplication Produce Extreme and Multiple Insecticide Resistance in the Malaria Mosquito Anopheles gambiae. *PLoS Genet* **10**: e1004236.

Eisele TP, Larsen D a, Walker N, Cibulskis RE, Yukich JO, Zikusooka CM, Steketee RW. 2012. Estimates of child deaths prevented from malaria prevention scale-up in Africa 2001-2010. *Malar J* **11**: 93.

Ekland EH, Fidock D a. 2007. Advances in understanding the genetic basis of antimalarial drug resistance. *Curr Opin Microbiol* **10**: 363–70.

Elango N, Hunt BG, Goodisman M a D, Yi S V. 2009. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera. *Proc Natl Acad Sci U S A* **106**: 11206–11.

Elyazar IRF, Gething PW, Patil AP, Rogayah H, Kusriastuti R, Wismarini DM, Tarmizi SN, Baird JK, Hay SI. 2011. Plasmodium falciparum malaria endemicity in Indonesia in 2010. *PLoS One* **6**: e21315.

Enayati a, Hemingway J. 2010. Malaria management: past, present, and future. *Annu Rev Entomol* **55**: 569–91.

Engström PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzi L, Tan SL, Yang L, et al. 2006. Complex Loci in human and mouse genomes. *PLoS Genet* **2**: e47.

Escalante AA, Freeland DE, Collins WE, Lal AA. 1998. The evolution of primate malaria parasites based on the gene encoding cytochrome b from the linear mitochondrial genome. *PNA* **95**: 8124–8129.

Falk N, Kaestli M, Qi W, Ott M, Baea K, Cortés A, Beck H-P. 2009. Analysis of Plasmodium falciparum var genes expressed in children from Papua New Guinea. *J Infect Dis* **200**: 347–56.

Fatica A, Bozzoni I. 2014. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* **15**: 7–21.

Feil R, Fraga MF. 2011. Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet* **13**: 97–109.

Fidock D a., Nomura T, Talley AK, Cooper R a., Dzekunov SM, Ferdig MT, Ursos LMB, bir Singh Sidhu A, Naudé B, Deitsch KW, et al. 2000. Mutations in the P. falciparum Digestive Vacuole Transmembrane Protein PfCRT and Evidence for Their Role in Chloroquine Resistance. *Mol Cell* **6**: 861–871.

Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, et al. 2010. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**: 212–24.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt Y, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2014. Pfam : the protein families database. **42**: 222–230.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**: W29–37.

Flick K, Chen Q. 2004. var genes, PfEMP1 and the human host. *Mol Biochem Parasitol* **134**: 3–9.

Foglietti C, Filocamo G, Cundari E, De Rinaldis E, Lahm A, Cortese R, Steinkühler C. 2006. Dissecting the biological functions of Drosophila histone deacetylases by RNA interference and transcriptional profiling. *J Biol Chem* **281**: 17968–76.

Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov I V, Jiang X, Hall AB, Catteruccia F, Kakani E, et al. 2014. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science (80-)* **347**.

Foster W a., Takken W. 2007. Nectar-related vs. human-related volatiles: behavioural response and choice by female and male Anopheles gambiae (Diptera: Culicidae) between emergence and first feeding. *Bull Entomol Res* **94**: 145–157.

Fouet C, Gray E, Besansky NJ, Costantini C. 2012. Adaptation to aridity in the malaria mosquito Anopheles gambiae: chromosomal inversion polymorphism and body size influence resistance to desiccation. *PLoS One* **7**: e34841.

Fox a N, Pitts RJ, Robertson HM, Carlson JR, Zwiebel LJ. 2001. Candidate odorant receptors from the malaria vector mosquito Anopheles gambiae and evidence of down-regulation in response to blood feeding. *Proc Natl Acad Sci U S A* **98**: 14693–7.

Freitas-junior ÂH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrbass U, Wellems TE, Scherf A. 2000. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P . falciparum. *Nature* **407**: 1018–1022.

Frentiu FD, Bernard GD, Sison-Mangus MP, Brower AVZ, Briscoe AD. 2007. Gene duplication is an evolutionary mechanism for expanding spectral diversity in the long-wavelength photopigments of butterflies. *Mol Biol Evol* **24**: 2016–28.

Fry DM. 1995. Reproductive Effects in Birds Exposed to Pesticides and Industrial Chemicals. 165–171.

Fu G, Lees RS, Nimmo D, Aw D, Jin L, Gray P, Berendonk TU, White-Cooper H, Scaife S, Kim Phuc H, et al. 2010. Female-specific flightless phenotype for mosquito control. *Proc Natl Acad Sci U S A* **107**: 4550–4.

Fukuto TR. 1990. Mechanism of Action of Organophosphorus and Carbamate Insecticides. *Environ Health Perspect* **87**: 245–254.

Fullman N, Burstein R, Lim SS, Medlin C, Gakidou E. 2013. Nets, spray or both? The effectiveness of insecticide-treated nets and indoor residual spraying in reducing malaria morbidity and child mortality in sub-Saharan Africa. *Malar J* **12**: 62.

Furrow RE, Feldman MW. 2014. Genetic Variation and the Evolution of Epigenetic Regulation. *Evolution (N Y)* **68**: 673–683.

Garczynski SF, Crim JW, Brown MR. 2007. Characterization and expression of the short neuropeptide F receptor in the African malaria mosquito, Anopheles gambiae. *Peptides* **28**: 109–18.

Garczynski SF, Crim JW, Brown MR. 2005. Characterization of neuropeptide F and its receptor from the African malaria mosquito, Anopheles gambiae. *Peptides* **26**: 99–107.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**: 498–511.

Geier M, Bosch OJ, Boeckh J. 1999. Ammonia as an attractive component of host odour for the yellow fever mosquito, Aedes aegypti. *Chem Senses* **24**: 647–53.

Gething PW, Patil AP, Smith DL, Guerra C a, Elyazar IRF, Johnston GL, Tatem AJ, Hay SI. 2011. A new world malaria map: Plasmodium falciparum endemicity in 2010. *Malar J* **10**: 378.

Gharib WH, Robinson-Rechavi M. 2013. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Mol Biol Evol* **30**: 1675–86.

Gillett J. 1983. Diuresis in newly emerged, unfed mosquitoes. II. The basic pattern in relation to escape from the water, preparation for mature flight, mating and the first blood meal. *Proc R Soc Lond B* **217**: 237–242.

Githeko AK, Adungo NI, Karanja DM, Hawley WA, Vulule JM, Seroney IK, Ofulla AVO, Atieli FK, Ondijo SO, Genga IO, et al. 1996. Some Observations on the Biting Behavior of Anopheles gambiae s . s ., Anopheles arabiensis , and Anopheles funestus and Their Implications for Malaria Control. **82**: 306–315.

Goldberg AD, Allis CD, Bernstein E. 2007. Epigenetics: a landscape takes shape. *Cell* **128**: 635–8.

Goldberg DE, Siliciano RF, Jacobs Jr WR. 2012. Outwitting Evolution: Fighting Drug Resistance in the Treatment of TB, Malaria and HIV. *Cell* **148**: 1271–1283.

Gomez-Diaz E, Jorda M, Peinado MA, Rivero A. 2012. Epigenetics of Host – Pathogen Interactions : The Road Ahead and the Road Behind. **8**: e1003007. doi:10.1371/journal.ppat.1003007.

Gouagna LC, Kerampran R, Lebon C, Brengues C, Toty C, Wilkinson D a, Boyer S, Fontenille D. 2014. Sugar-source preference, sugar intake and relative nutritional benefits in Anopheles arabiensis males. *Acta Trop* **132 Suppl**: S70–9.

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313**: 903–19.

Greer EL, Maures TJ, Ucar D, Hauswirth AG, Mancini E, Lim JP, Benayoun B a, Shi Y, Brunet A. 2011. Transgenerational epigenetic inheritance of longevity in Caenorhabditis elegans. *Nature* **479**: 365–71.

Greer EL, Shi Y. 2012. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Genet* **13**: 343–57.

Gregoretti I V, Lee Y-M, Goodson H V. 2004. Molecular evolution of the histone deacetylase family: functional implications of phylogenetic analysis. *J Mol Biol* **338**: 17–31.

Griffin JT, Hollingsworth TD, Okell LC, Churcher TS, White M, Hinsley W, Bousema T, Drakeley CJ, Ferguson NM, Basáñez M-G, et al. 2010. Reducing Plasmodium falciparum malaria transmission in Africa: a model-based evaluation of intervention strategies. *PLoS Med* **7**.

Gruber A., Findeiß S, Washietl S, Hofacker I., Stadlet PF. 2010. RNAz 2.0: improved noncoding RNA detection. *Pacific Symp Biocomput* 69–79.

Gu T, Elgin SCR. 2013. Maternal depletion of Piwi, a component of the RNAi system, impacts heterochromatin formation in Drosophila. *PLoS Genet* **9**: e1003780.

Guil S, Esteller M. 2009. DNA methylomes, histone codes and miRNAs: tying it all together. *Int J Biochem Cell Biol* **41**: 87–95.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**: 223–7.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**: 503–10.

Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. **31**: 371–373.

Hall N, Karras M, Raine JD, Carlton JM, Kooij TWA, Berriman M, Florens L, Janssen CS, Pain A, Christophides GK, et al. 2005. A Comprehensive Survey of the Plasmodium Life Cycle by Genomic, Transcriptomic, and Proteomic Analyses. *Science (80-)* **307**: 82–87.

Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J. 2006. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**: 47–59.

Hangauer MJ, Vaughn IW, McManus MT. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet* **9**: e1003569.

Happi CT, Gbotosho GO, Folarin O a, Akinboye DO, Yusuf BO, Ebong OO, Sowunmi a, Kyle DE, Milhous W, Wirth DF, et al. 2005. Polymorphisms in Plasmodium falciparum dhfr and dhps genes and age related in vivo sulfadoxine-pyrimethamine resistance in malaria-infected patients from Nigeria. *Acta Trop* **95**: 183–93.

Harker BW, Hong YS, Sim C, Dana AN, Bruggner RV, Lobo NF, Kern MK, Sharakhova MV, Collins FH. 2012. Transcription Profiling Associated with Life Cycles of Anopheles gambiae. *J Med Entomol* **49**: 316–325.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760–74.

Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW. 2004. Reviews The global distribution and population at risk of malaria : past , present , and future. **4**: 327–336.

Hay SI, Sinka ME, Okara RM, Kabaria CW, Mbithi PM, Tago CC, Benz D, Gething PW, Howes RE, Patil AP, et al. 2010. Developing global maps of the dominant anopheles vectors of human malaria. *PLoS Med* **7**: e1000209.

Hayakawa T, Culleton R, Otani H, Horii T, Tanabe K. 2008. Big bang in the evolution of extant malaria parasites. *Mol Biol Evol* **25**: 2233–9.

He L, Hannon GJ. 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* **5**: 522–31.

Heard E, Disteche CM. 2006. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes Dev* **20**: 1848–67.

Hill CA, Fox AN, Pitts RJ, Kent LB, Tan PL, Chrystal MA, Cravchik A, Collins FH, Robertson HM, Zwiebel LJ. 2002. G Protein – Coupled Receptors in Anopheles gambiae. *Science (80-)* **298**: 176–178.

Hoheisel JD. 2006. Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* **7**: 200–10.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**: 491.

Holt R, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JMC, Wides R, et al. 2002. The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298**: 129–49.

Hougard J-M, Corbel V, N'Guessan R, Darriet F, Chandre F, Akogbéto M, Baldet T, Guillet P, Carnevale P, Traoré-Lamizana M. 2007. Efficacy of mosquito nets treated with insecticide mixtures or mosaics against insecticide resistant Anopheles gambiae and Culex quinquefasciatus (Diptera: Culicidae) in Côte d'Ivoire. *Bull Entomol Res* **93**: 491–498.

Hu X, England JH, Lani AC, Tung JJ, Ward NJ, Adams SM, Barber K a, Whaley M a, O'Tousa JE. 2009. Patterned rhodopsin expression in R7 photoreceptors of mosquito retina: Implications for species-specific behavior. *J Comp Neurol* **516**: 334–42.

Hu X, Leming MT, Whaley M a, O'Tousa JE. 2013. Rhodopsin coexpression in UV photoreceptors of Aedes aegypti and Anopheles gambiae mosquitoes. *J Exp Biol*.

Hu X, Whaley M a, Stein MM, Mitchell BE, O'Tousa JE. 2011. Coexpression of spectrally distinct rhodopsins in Aedes aegypti R7 photoreceptors. *PLoS One* **6**: e23121.

Hu Z-L, Bao J, Reecy J. 2008. CateGOrizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories. *OJB* **9**: 108–112.

Huang DW, Sherman BT, Lempicki R a. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44–57.

Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform* **12**: 41–51.

Huminiecki L, Wolfe KH. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res* **14**: 1870–9.

Hunt BG, Glastad KM, Yi S V, Goodisman M a D. 2013. Patterning and regulatory associations of DNA methylation are mirrored by histone modifications in insects. *Genome Biol Evol* **5**: 591–8.

Hunt NH, Grau GE. 2003. Cytokines: accelerators and brakes in the pathogenesis of cerebral malaria. *Trends Immunol* **24**: 491–499.

Hyde JE. 2002. Mechanisms of resistance of Plasmodium falciparum to antimalarial drugs. *Microbes Infect* **4**: 165–174.

Ibrahim SS, Manu Y a, Tukur Z, Irving H, Wondji CS. 2014. High frequency of kdr L1014F is associated with pyrethroid resistance in Anopheles coluzzii in Sudan savannah of northern Nigeria. *BMC Infect Dis* **14**: 441.

Idro R, Jenkins NE, Newton CR. 2005. Pathogenesis, clincal features, and neurological outcome of cerebral malaria. *Lancet Neurol* **4**: 827–840.

Ilott NE, Ponting CP. 2013. Predicting long non-coding RNAs using RNA sequencing. *Methods* **63**: 50–9.

Imbahale SS, Githeko A, Mukabana WR, Takken W. 2012. Integrated mosquito larval source management reduces larval numbers in two highland villages in western Kenya. *BMC Public Health* **12**: 362.

Inada K, Horie T, Kusakabe T, Tsuda M. 2003. Targeted knockdown of an opsin gene inhibits the swimming behaviour photoresponse of ascidian larvae. *Neurosci Lett* **347**: 167–170.

Insel P a., Tang CM, Hahntow I, Michel MC. 2007. Impact of GPCRs in clinical medicine: Monogenic diseases, genetic variants and drug targets. *Biochim Biophys Acta - Biomembr* **1768**: 994–1005.

Jaluria P, Konstantopoulos K, Betenbaugh M, Shiloach J. 2007. A perspective on microarrays: current applications, pitfalls, and potential uses. *Microb Cell Fact* **6**: 4.

Jenkins AM, Muskavitch MAT. 2015. Crepuscular behavioral variation and profiling of opsin genes in Anopheles gambiae and Anopheles stephensi. *J Med Entomol* In Press.

Jenkins AM, Waterhouse RM, Kopin AS, Muskavitch MAT. 2014. Long non-coding RNA discovery in Anopheles gambiae using deep RNA sequencing. *bioarxiv*.

Jensen TH, Jacquier A, Libri D. 2013. Dealing with pervasive transcription. *Mol Cell* **52**: 473–84.

Jiang X, Peery A, Hall a, Sharma A, Chen X-G, Waterhouse RM, Komissarov A, Riehl MM, Shouche Y, Sharakhova M V, et al. 2014. Genome analysis of a major urban malaria vector mosquito, Anopheles stephensi. *Genome Biol* **15**: 459.

Johnson DJ, Fidock D a, Mungthin M, Lakshmanan V, Sidhu ABS, Bray PG, Ward S a. 2004. Evidence for a central role for PfCRT in conferring Plasmodium falciparum resistance to diverse antimalarial agents. *Mol Cell* **15**: 867–77.

Jones MD., Hill M, Hope AM. 1967. The circadian flight activity of the mosquito Anopheles gambiae: phase setting by the light regime. *J Exp Biol* **47**: 503–511.

Kastner KW, Shoue D a, Estiu GL, Wolford J, Fuerst MF, Markley LD, Izaguirre J a, McDowell MA. 2014. Characterization of the Anopheles gambiae octopamine receptor and discovery of potential agonists and antagonists using a combined computational-experimental approach. *Malar J* **13**: 434.

Katritch V, Cherezov V, Stevens RC. 2013. Structure-function of the G protein-coupled receptor superfamily. *Annu Rev Pharmacol Toxicol* **53**: 531–556.

Kawada H, Takemura S, Arikawa K. 2005. Comparative Study on Nocturnal Behavior of Aedes aegypti and Aedes albopictus. *J Med Entomol* **42**: 312–318.

Keene KM, Foy BD, Sanchez-Vargas I, Beaty BJ, Blair CD, Olson KE. 2004. RNA interference acts as a natural antiviral response to O'nyong-nyong virus (Alphavirus; Togaviridae) infection of Anopheles gambiae. *Proc Natl Acad Sci U S A* **101**: 17240–17245.

Keller TE, Yi S V. 2014. DNA methylation and evolution of duplicate genes. *Proc Natl Acad Sci* **111**: 5932–5937.

Kennison J a. 2004. Introduction to Trx-G and Pc-G genes. *Methods Enzymol* **377**: 61–70.

Kennison J a, Tamkun JW. 1988. Dosage-dependent modifiers of polycomb and antennapedia mutations in Drosophila. *Proc Natl Acad Sci U S A* **85**: 8136–40.

Kennison JA. 1995. THE POL YCOMB AND TRITHORAX GROUP PROTEINS OF DROSOPHILA : Trans-Regulators of Homeotic Gene Function. *Annu Rev Genet* 289–303.

Kharchenko P V, Alekseyenko A a, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov A a, Gu T, et al. 2011. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**: 480–5.

Kiefer JC. 2007. Epigenetics in development. *Dev Dyn* **236**: 1144–56.

Killeen GF, McKenzie FE, Foy BD, Schieffelin C, Billingsley PF, Beier JC. 2000. A simplified model for predicting malaria entomologic inoculation rates based on entomologic and parasitologic parameters relevant to control. *Am J Trop Med Hyg* **62**: 535–544.

Killeen GF, Ross A, Smith T. 2006. Infectiousness of Malaria-Endemic Human Populations to Vectors. *Am J Trop Med Hyg* **75**: 38–45.

Kim D, Fedak K, Kramer R. 2012. Reduction of malaria prevalence by indoor residual spraying: a meta-regression analysis. *Am J Trop Med Hyg* **87**: 117–24.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

Kim VN. 2005. MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol* **6**: 376–85.

Kim VN, Nam J-W. 2006. Genomics of microRNA. *Trends Genet* **22**: 165–73.

Kishore SP, Stiller JW, Deitsch KW. 2013. Horizontal gene transfer of epigenetic machinery and evolution of parasitism in the malaria parasite Plasmodium falciparum and other apicomplexans. *BMC Evol Biol* **13**: 37.

Klecka J, Boukal DS. 2012. Who eats whom in a pool? A comparative study of prey selectivity by predatory aquatic insects. *PLoS One* **7**: e37741.

Kleinschmidt I, Schwabe C, Shiva M, Segura JL, Sima V, Jose S, Mabunda A, Coleman M. 2009. Combining Indoor Residual Spraying and Insecticide-Treated Net Interventions. *Am J Trop Med Hyg* **81**: 519–524.

Klironomos FD, Berg J, Collins S. 2013. How epigenetic mutations can affect genetic evolution: model and mechanism. *Bioessays* **35**: 571–8.

Koella JC, Sörensen FL, Anderson RA. 1998. The malaria parasite , Plasmodium falciparum , increases the frequency of multiple feeding of its mosquito vector , Anopheles gambiae. *Proc R Soc Lond B* **265**: 763–768.

Koutsos AC, Blass C, Meister S, Schmidt S, MacCallum RM, Soares MB, Collins FH, Benes V, Zdobnov E, Kafatos FC, et al. 2007. Life cycle transcriptome of the malaria mosquito Anopheles gambiae and comparison with the fruitfly Drosophila melanogaster. *Proc Natl Acad Sci U S A* **104**: 11304–9.

Kraemer SM, Kyes S a, Aggarwal G, Springer AL, Nelson SO, Christodoulou Z, Smith LM, Wang W, Levin E, Newbold CI, et al. 2007. Patterns of gene recombination shape var gene repertoires in Plasmodium falciparum: comparisons of geographically diverse isolates. *BMC Genomics* **8**: 45.

Kraemer SM, Smith JD. 2006. A family affair: var genes, PfEMP1 binding, and malaria disease. *Curr Opin Microbiol* **9**: 374–80.

Kraemer SM, Smith JD. 2003. Evidence for the importance of genetic structuring to the structural and functional specialization of the Plasmodium falciparum var gene family. *Mol Microbiol* **50**: 1527–1538.

Krauss V, Fassl A, Fiebig P, Patties I, Sass H. 2006. The evolution of the histone methyltransferase gene Su(var)3-9 in metazoans includes a fusion with and a re-fission from a functionally unrelated gene. *BMC Evol Biol* **6**: 18.

Krief S, Escalante A a, Pacheco MA, Mugisha L, André C, Halbwax M, Fischer A, Krief J-M, Kasenene JM, Crandfield M, et al. 2010. On the diversity of malaria parasites in African apes and the origin of Plasmodium falciparum from Bonobos. *PLoS Pathog* **6**: e1000765.

Krishna S, Pulcini S, Moore CM, Teo BH-Y, Staines HM. 2014. Pumped up: reflections on PfATP6 as the target for artemisinins. *Trends Pharmacol Sci* **35**: 4–11.

Kuehn A, Pradel G. 2010. The coming-out of malaria gametocytes. *J Biomed Biotechnol* **2010**.

Kung JTY, Colognori D, Lee JT. 2013. Long noncoding RNAs: past, present, and future. *Genetics* **193**: 651–69.

Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet* **8**: e1002841.

Labbe P, Lenormand T, Raymond M. 2005. On the worldwide spread of an insecticide resistance gene : a role for local selection. *J Evol Biol* **18**: 1471–1484.

Lahondère C, Lazzari CR. 2012. Mosquitoes cool down during blood feeding to avoid overheating. *Curr Biol* **22**: 40–5.

Lakshmanan V, Bray PG, Verdier-Pinard D, Johnson DJ, Horrocks P, Muhle R a, Alakpa GE, Hughes RH, Ward S a, Krogstad DJ, et al. 2005. A critical role for PfCRT K76T in Plasmodium falciparum verapamil-reversible chloroquine resistance. *EMBO J* **24**: 2294–305.

Langhorne J, Buffet P, Galinski M, Good M, Harty J, Leroy D, Mota MM, Pasini E, Renia L, Riley E, et al. 2011. The relevance of non-human primate and rodent malaria models for humans. *Malar J* **10**: 23.

Lapointe D a, Atkinson CT, Samuel MD. 2012. Ecology and conservation biology of avian malaria. *Ann N Y Acad Sci* **1249**: 211–26.

Lawniczak MKN, Emrich SJ, Holloway a K, Regier a P, Olson M, White B, Redmond S, Fulton L, Appelbaum E, Godfrey J, et al. 2010. Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences. *Science* **330**: 512–4.

Lee JT. 2012. Epigenetic regulation by long noncoding RNAs. *Science* **338**: 1435–9.

Lefèvre T, Gouagna L-C, Dabire KR, Elguero E, Fontenille D, Costantini C, Thomas F. 2009. Evolutionary lability of odour-mediated host preference by the malaria vector Anopheles gambiae. *Trop Med Int Health* **14**: 228–36.

Lespinet O, Wolf YI, Koonin E V, Aravind L. 2002. The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes. 1048–1059.

Li M, Wen S, Guo X, Bai B, Gong Z, Liu X, Wang Y, Zhou Y, Chen X, Liu L, et al. 2012. The novel long non-coding RNA CRG regulates Drosophila locomotor behavior. *Nucleic Acids Res* **40**: 11714–27.

Li W-H, Wu C-I, Luo C-C. 1985. A New Method for Estimating Synonymous and Nonsynonymous Rates of Nucleotide Substitution Considering the Relative Likelihood of Nucleotide and Codon Changes. *Mol Biol Evol* **2**: 150–174.

License A, Malaria G, Program E. 2011. A Research Agenda for Malaria Eradication: Vector Control. *PLoS Med* **8**: e1000401.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**: i275–82.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**: 476–82.

Liu W, Li Y, Learn GH, Rudicell RS, Robertson JD, Keele BF, Ndjango J-BN, Sanz CM, Morgan DB, Locatelli S, et al. 2010. Origin of the human malaria parasite Plasmodium falciparum in gorillas. *Nature* **467**: 420–5.

Lu J, Bushel PR. 2013. Dynamic expression of 3' UTRs revealed by Poisson hidden Markov modeling of RNA-Seq: implications in gene expression profiling. *Gene* **527**: 616–23.

Lu T, Qiu YT, Wang G, Kwon JY, Rutzler M, Kwon H-W, Pitts RJ, van Loon JJ a, Takken W, Carlson JR, et al. 2007. Odor coding in the maxillary palp of the malaria vector mosquito Anopheles gambiae. *Curr Biol* **17**: 1533–44.

Lucas KJ, Myles KM, Raikhel AS. 2013. Small RNAs: a new frontier in mosquito biology. *Trends Parasitol* **29**: 295–303.

Lunyak V V., Rosenfeld MG. 2008. Epigenetic regulation of stem cell fate. *Hum Mol Genet* **17**: R28–R36.

Lv J, Liu H, Huang Z, Su J, He H, Xiu Y, Zhang Y, Wu Q. 2013. Long non-coding RNA identification over mouse brain development by integrative modeling of chromatin and genomic features. *Nucleic Acids Res* **41**: 10044–61.

Lyko F, Beisel C, Marhold J, Paro R. 2006. Epigenetic regulation in Drosophila. *Curr Top Microbiol Immunol* **310**: 23–44.

Macdonald G. 1957. *The epidemiology and control of malaria*. Oxford University Press, London, New York.

Macpherson GG, Warrell MJ, White NJ, Looareesuwan S, Warrell DA. 1985. Human Cerebral Malaria: A quantitative Ultrastructural Analysis of Parasitized Erthrocyte Sequestration. *Am J Pathol* **119**: 385–401.

Mala AO, Irungu LW, Shililu JI, Muturi EJ, Mbogo CM, Njagi JK, Mukabana WR, Githure JI. 2011. Plasmodium falciparum transmission and aridity: a Kenyan experience from the dry lands of Baringo and its implications for Anopheles arabiensis control. *Malar J* **10**: 121.

Manouchehri AV, Djanbakhsh B, Eshghi N. 1976. The biting cycle of Anopheles dthali. A. fluviatilis and A. stephensi in southern Iran. *Trop Geogr Med* **28**: 224–7.

Marhold J, Rothe N, Pauli a, Mund C, Kuehle K, Brueckner B, Lyko F. 2004. Conservation of DNA methylation in dipteran insects. *Insect Mol Biol* **13**: 117–23.

Marinotti O, Calvo E, Nguyen QK, Dissanayake S, Ribeiro JMC, James AA. 2006. Genome-wide analysis of gene expression in adult Anopheles gambiae. *Insect Mol Biol* **15**: 1–12.

Marinotti O, Cerqueira GC, de Almeida LGP, Ferro MIT, Loreto ELDS, Zaha A, Teixeira SMR, Wespiser AR, Almeida E Silva A, Schlindwein AD, et al. 2013. The genome of Anopheles darlingi, the main neotropical malaria vector. *Nucleic Acids Res* **41**: 7387–400.

Marques AC, Ponting CP. 2014. Intergenic lncRNAs and the evolution of gene expression. *Curr Opin Genet Dev* **27**: 48–53.

Mathenge EM, Misiani GO, Oulo DO, Irungu LW, Ndegwa PN, Smith T a, Killeen GF, Knols BGJ. 2005. Comparative performance of the Mbita trap, CDC light trap and the human landing catch in the sampling of Anopheles arabiensis, An. funestus and culicine species in a rice irrigation in western Kenya. *Malar J* **4**.

Mayer C, Slater L, Erat MC, Konrat R, Vakonakis I. 2012. Structural analysis of the Plasmodium falciparum erythrocyte membrane protein 1 (PfEMP1) intracellular domain reveals a conserved interaction epitope. *J Biol Chem* **287**: 7182–9.

Megy K, Emrich SJ, Lawson D, Campbell D, Dialynas E, Hughes DST, Koscielny G, Louis C, Maccallum RM, Redmond SN, et al. 2012. VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. *Nucleic Acids Res* **40**: D729–34.

Meijerink J, Braks M a. ., Van Loon JJ. 2001. Olfactory receptors on the antennae of the malaria mosquito Anopheles gambiae are sensitive to ammonia and other sweat-borne components. *J Insect Physiol* **47**: 455–464.

Meissner A. 2010. Epigenetic modifications in pluripotent and differentiated cells. **28**: 1079–1088.

Mercer TR, Mattick JS. 2013. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol* **20**: 300–7.

Meyer JM, Ejendal KFK, Avramova L V, Garland-Kuntz EE, Giraldo-Calderón GI, Brust TF, Watts VJ, Hill C a. 2012. A "genome-to-lead" approach for insecticide discovery: pharmacological characterization and screening of Aedes aegypti D(1)-like dopamine receptors. *PLoS Negl Trop Dis* **6**: e1478.

Michel K, Budd A, Pinto S, Gibson TJ, Kafatos FC. 2005. Anopheles gambiae SRPN2 facilitates midgut invasion by the malaria parasite Plasmodium berghei. *EMBO Rep* **6**: 891–897.

Miller LH, Baruch DI, Marsh K, Doumbo OK. 2002. The pathogenic basis of malaria. *Nature* **415**: 673–679.

Mitchell SN, Rigden DJ, Dowd AJ, Lu F, Wilding CS, Weetman D, Dadzie S, Jenkins AM, Regna K, Boko P, et al. 2014. Metabolic and Target-Site Mechanisms Combine to Confer Strong DDT Resistance in Anopheles gambiae. *PLoS One* **9**: e92662.

Mittal PK, Sood RD, Kapoor N, Razdan RK, Dash a P. 2012. Field evaluation of Icon®Life, a long-lasting insecticidal net (LLIN) against Anopheles culicifacies and transmission of malaria in District Gautam Budh Nagar (Uttar Pradesh), India. *J Vector Borne Dis* **49**: 181–7.

Miyata T, Yasunaga T, Nishida T. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc Natl Acad Sci* **77**: 7328–7332.

Mnyone LL, Lyimo IN, Lwetoijera DW, Mpingwa MW, Nchimbi N, Hancock P a, Russell TL, Kirby MJ, Takken W, Koenraadt CJM. 2012. Exploiting the behaviour of wild malaria vectors to achieve high infection with fungal biocontrol agents. *Malar J* **11**.

Montell C. 2012. Drosophila visual transduction. *Trends Neurosci* **35**: 356–63.

Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. **5**: 1–8.

Mugal CF, Wolf JBW, Kaj I. 2014. Why time matters: codon evolution and the temporal dynamics of dN/dS. *Mol Biol Evol* **31**: 212–31.

Muheki C, Mcintyre D, Barnes KI. 2010. Artemisinin-based combination therapy reduces expenditure on malaria treatment in KwaZulu Natal , South Africa. *Trop Med Int Heal* **9**: 959–966.

Murray CJL, Rosenfeld LC, Lim SS, Andrews KG, Foreman KJ, Haring D, Fullman N, Naghavi M, Lozano R, Lopez AD. 2012. Global malaria mortality between 1980 and 2010: a systematic analysis. *Lancet* **379**: 413–31.

Nájera J a, González-Silva M, Alonso PL. 2011. Some lessons for the future from the global malaria eradication programme (1955-1969). *PLoS Med* **8**: e1000412.

Nam J-W, Bartel DP. 2012. Long noncoding RNAs in C. elegans. *Genome Res* **22**: 2529–40.

Ndiath MO, Sougoufara S, Gaye A, Mazenot C, Konate L, Faye O, Faye O, Sokhna C, Trape J-F. 2012. Resistance to DDT and pyrethroids and increased kdr mutation frequency in An. gambiae after the implementation of permethrin-treated nets in Senegal. *PLoS One* **7**: e31943.

Neafsey DE, Christophides GK, Collins FH, Emrich SJ, Fontaine MC, Gelbart W, Hahn MW, Howell PI, Kafatos FC, Lawson D, et al. 2013. The evolution of the Anopheles 16 genomes project. *G3 (Bethesda)* **3**: 1191–4.

Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Max A, Allen JE, Amon J, Arcà B, Arensburger P, Artemov G, et al. 2014. Highly evolvable malaria vectors : The genomes of 16 Anopheles mosquitoes. *Science (80-)* **347**: 1–19.

Necsulea A, Kaessmann H. 2014. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat Rev Genet*.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**: 635–40.

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**: 382–94.

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, et al. 2007. Genome sequence of Aedes aegypti, a major arbovirus vector. *Science* **316**: 1718–23.

Nevillts CG, Hospital JR, Ox O. 1996. Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast. 139–146.

Nie L, Wu H-J, Hsu J-M, Chang S-S, Labaff AM, Li C-W, Wang Y, Hsu JL, Hung M-C. 2012. Long non-coding RNAs: versatile master regulators of gene expression and crucial players in cancer. *Am J Transl Res* **4**: 127–50.

Njabo KY, Cornel AJ, Bonneaud C, Toffelmier E, Valki G, Russell AF, Smith TB. 2012. Parasite Associations in a Central African Rainforest. *Mol Ecol* **20**: 1049–1061.

Novikova I V, Hennelly SP, Sanbonmatsu KY. 2012. Sizing up long non-coding RNAs: do lncRNAs have secondary and tertiary structure? *Bioarchitecture* **2**: 189–99.

Nwane P, Etang J, Chouaïbou M, Toto JC, Mimpfoundi R, Simard F. 2011. Kdr-based insecticide resistance in Anopheles gambiae s.s populations in. *BMC Res Notes* **4**: 463.

Obbard DJ, Maclennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM. 2012. Estimating divergence dates and substitution rates in the Drosophila phylogeny. *Mol Biol Evol* **29**: 3459–73.

Okumu FO, Moore SJ. 2011. Combining indoor residual spraying and insecticide-treated nets for malaria control in Africa: a review of possible outcomes and an outline of suggestions for the future. *Malar J* **10**: 208.

Olliaro P. 2008. Editorial commentary: mortality associated with severe Plasmodium falciparum malaria increases with age. *Clin Infect Dis* **47**: 158–60.

Olliaro P. 2001. Mode of action and mechanisms of resistance for antimalarial drugs. *Pharmacol Ther* **89**: 207–219.

Olsen SS, Cazzamali G, Williamson M, Grimmelikhuijzen CJP, Hauser F. 2007. Identification of one capa and two pyrokinin receptors from the malaria mosquito Anopheles gambiae q. *Biochem Biophys Res Commun* **362**: 245–251.

Overgaard HJ, Saebø S, Reddy MR, Reddy VP, Abaga S, Matias A, Slotman M a. 2012. Light traps fail to estimate reliable malaria mosquito biting rates on Bioko Island, Equatorial Guinea. *Malar J* **11**.

Paaijmans KP, Thomas MB. 2011. The influence of mosquito resting behaviour and associated microclimate for malaria risk. *Malar J* **10**.

Padrón A, Molina-Cruz A, Quinones M, Ribeiro JM, Ramphul U, Rodrigues J, Shen K, Haile A, Ramirez JL, Barillas-Mury C. 2014. In depth annotation of the Anopheles gambiae mosquito midgut transcriptome. *BMC Genomics* **15**: 636.

Park S, Lehner B. 2014. Epigenetic epistatic interactions constrain the evolution of gene expression. *Mol Syst Biol* **9**: 645–645.

Pask GM, Jones PL, Rützler M, Rinker DC, Zwiebel LJ. 2011. Heteromeric Anopheline odorant receptors exhibit distinct channel properties. *PLoS One* **6**: e28774.

Pasternak ND, Dzikowski R. 2009. PfEMP1: an antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite Plasmodium falciparum. *Int J Biochem Cell Biol* **41**: 1463–6.

Pates H, Curtis C. 2005a. Mosquito Behavior and Vector Control. *Annu Rev Entomol* **50**: 53–70.

Pates H, Curtis C. 2005b. Mosquito behavior and vector control. *Annu Rev Entomol* **50**: 53–70.

Pates H V. 2002. Zoophilic and anthropophilic behavior in the Anopheles gambiae complex. London School of Hygiene & Tropical Medicine.

Patro R, Mount SM, Kingsford C. 2014. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol* **32**: 462–4.

Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, et al. 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* **22**: 577–91.

Payne D. 1987. Spread of chloroquine resistance in Plasmodium Falciparum. *Parasitol Today* **8**: 241–246.

Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N. 1996. Requirement for Xist in X chromosome inactivation. *Nature* **379**: 131–137.

Perich MJ, Davila G, Turner a, Garcia a, Nelson M. 2000. Behavior of resting Aedes aegypti (Culicidae: Diptera) and its relation to ultra-low volume adulticide efficacy in Panama City, Panama. *J Med Entomol* **37**: 541–6.

Phuc HK, Andreasen MH, Burton RS, Vass C, Epton MJ, Pape G, Fu G, Condon KC, Scaife S, Donnelly C a, et al. 2007. Late-acting dominant lethal genetic systems and mosquito control. *BMC Biol* **5**: 1–11.

Phuc HK, Ball a J, Son L, Hanh N V, Tu ND, Lien NG, Verardi a, Townson H. 2003. Multiplex PCR assay for malaria vector Anopheles minimus and four related species in the Myzomyia Series from Southeast Asia. *Med Vet Entomol* **17**: 423–8.

Pickard a. L, Wongsrichanalai C, Purfield a., Kamwendo D, Emery K, Zalewski C, Kawamoto F, Miller RS, Meshnick SR. 2003. Resistance to Antimalarials in Southeast Asia and Genetic Polymorphisms in pfmdr1. *Antimicrob Agents Chemother* **47**: 2418–2423.

Pitts RJ, Rinker DC, Jones PL, Rokas A, Zwiebel LJ. 2011. Transcriptome profiling of chemosensory appendages in the malaria vector Anopheles gambiae reveals tissue- and sex-specific signatures of odor coding. *BMC Genomics* **12**: 271.

Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. *Cell* **136**: 629–41.

Ponton F, Chapuis M-P, Pernice M, Sword G a, Simpson SJ. 2011. Evaluation of potential reference genes for reverse transcription-qPCR studies of physiological responses in Drosophila melanogaster. *J Insect Physiol* **57**: 840–50.

Portela A, Esteller M. 2010. Epigenetic modifications and human disease. *Nat Biotechnol* **28**: 1057–68.

Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldón T, Rattei T, Creevey C, Kuhn M, et al. 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* **42**: D231–9.

Pridgeon JW, Becnel JJ, Clark GG, Linthicum KJ. 2009. A High-Throughput Screening Method to Identify Potential Pesticides for Mosquito Control. *J Med Entomol* **46**: 335–341.

Prior a., Torr SJ. 2002. Host selection by Anopheles arabiensis and An. quadriannulatus feeding on cattle in Zimbabwe. *Med Vet Entomol* **16**: 207–213.

Prugnolle F, Durand P, Ollomo B, Duval L, Ariey F, Arnathau C, Gonzalez J-P, Leroy E, Renaud F. 2011. A fresh look at the origin of Plasmodium falciparum, the most malignant malaria agent. *PLoS Pathog* **7**: e1001283.

R Core Team. 2014. R: A language and environment for statistical computing. *R Found Stat Comput Vienna, Austria*.

Ranson H, Claudianos C, Ortelli F, Abgrall C, Hemingway J, Sharakhova M V, Unger MF, Collins FH, Feyereisen R. 2002. Evolution of supergene families associated with insecticide resistance. *Science* **298**: 179–81.

Ranson H, Jensen B, Vulule JM, Wang X, Hemingway J, Collins FH, Dame N, Biology V. 2000. Identification of a point mutation in the voltage-gated sodium channel gene of Kenyan Anopheles gambiae associated with resistance to DDT and pyrethroids. *Insect Mol Biol* **9**: 491–497.

Rathore D, Wahl AM, Sullivan M, McCutchan TF. 2001. A phylogenetic comparison of gene trees constructed from plastid, mitochondrial and genomic DNA of Plasmodium species. *Mol Biochem Parasitol* **114**: 89–94.

Read AF, Lynch P a, Thomas MB. 2009. How to make evolution-proof insecticides for malaria control. *PLoS Biol* **7**: e1000058.

Reimer L, Fondjo E, Patchoke S, Diallo B, Lee Y, Ng A, Ndjemai HM, Atangana J, Traore SF, Lanzaro G, et al. 2008. Relationship between kdr mutatin and resistance to pyrethroid and DDT insecticides in natural populations of Anopheles gambiae. *J Med Entomol* **45**: 260–266.

Ribbands CR. 1946. Moonlight and house-haunting habits of female anophelines in West Africa. *Bull Entomol Res* **36**: 395–417.

Richards S, Gibbs R a, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Bucher G, Friedrich M, Grimmelikhuijzen CJP, et al. 2008. The genome of the model beetle and pest Tribolium castaneum. *Nature* **452**: 949–55.

Riehle MA, Garczynski SF, Crim JW, Hill CA, Brown MR. 2002. Neuropeptides and Peptide Hormones in Anopheles gambiae. *Science (80-)* **298**: 172–176.

Rinker DC, Jones PL, Pitts RJ, Rutzler M, Camp G, Sun L, Xu P, Dorset DC, Weaver D, Zwiebel LJ. 2012. Novel high-throughput screens of Anopheles gambiae odorant receptors reveal candidate behaviour-modifying chemicals for mosquitoes. *Physiol Entomol* **37**: 33–41.

Rinker DC, Pitts RJ, Zhou X, Suh E, Rokas A, Zwiebel LJ. 2013. Blood meal-induced changes to antennal transcriptome pro fi les reveal shifts in odor sensitivities in Anopheles gambiae. *Proc Natl Acad Sci U S A* **110**: 8260–8265.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Roch KG Le, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, Vega P De, Holder AA, Batalov S, Carucci DJ, et al. 2003. Discovery of Gene Function by Expression Profiling of the Malaria Parasite Life Cycle. *Science (80-)* **301**: 1503–1509.

Roper C, Pearce R, Nair S, Sharp B. 2004. Intercontinental Spread of Pyrimethamine-Resistant Malaria. *Science (80-)* **305**: 2004.

Ross R. 1910. *The prevention of malaria*. E.P. Dutton & Company, New York.

Rowland M. 1989. Changes in the circadian flight activity of the mosquito Anopheles stephensi associated with insemination, blood-feeding, oviposition and nocturnal light intensity. *Physiol Entomol* 77–84.

Rowland M, Boersma E. 1988. Changes in the spontaneous flight activity of the mosquito Anopheles stephensi by parasitization with the rodent malaria Plasmodium yoelii. *Parasitology* **97**: 221–227.

Rund SSC, Hou TY, Ward SM, Collins FH, Duffield GE. 2011. Genome-wide profiling of diel and circadian gene expression in the malaria vector Anopheles gambiae. *PNAS* **108**.

Sasagawa H, Narita R, Kitagawa Y, Kadowaki T. 2003. The expression of genes encoding visual components is regulated by a circadian clock, light environment and age in the honeybee (Apis mellifera). *Eur J Neurosci* **17**: 963–970.

Šášik R, Woelk CH, Corbeil J. 2004. Microarray truths and consequences. *J Mol Endocrinol* **33**: 1–9.

Scherf a, Hernandez-Rivas R, Buffet P, Bottius E, Benatar C, Pouvelle B, Gysin J, Lanzer M. 1998. Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum. *EMBO J* **17**: 5418–26.

Schotta G, Ebert A, Krauss V, Fischer A, Hoffmann J, Rea S, Jenuwein T, Dorn R, Reuter G. 2002. Central role of Drosophila SU(VAR)3-9 in histone H3-K9 methylation and heterochromatic gene silencing. *EMBO J* **21**: 1121–31.

Schuettengruber B, Chourrout D, Vervoort M, Leblanc B, Cavalli G. 2007. Genome regulation by polycomb and trithorax proteins. *Cell* **128**: 735–45.

Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, Tolhuis B, van Lohuizen M, Tanay A, Cavalli G. 2009. Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS Biol* **7**: e13.

Schulze SR, Wallrath LL. 2007. Gene regulation by chromatin structure: paradigms established in Drosophila melanogaster. *Annu Rev Entomol* **52**: 171–92.

Schwartz YB, Pirrotta V. 2007. Polycomb silencing mechanisms and the management of genomic programmes. *Nat Rev Genet* **8**: 9–22.

Seidl CIM, Stricker SH, Barlow DP. 2006. The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. *EMBO J* **25**: 3565–75.

Severson DW, Behura SK. 2012. Mosquito genomics: progress and challenges. *Annu Rev Entomol* **57**: 143–66.

Sibley CH, Hyde JE, Sims PF., Plowe C V, Kublin JG, Mberu EK, Cowman AF, Winstanley P a, Watkins WM, Nzila AM. 2001. Pyrimethamine–sulfadoxine resistance in Plasmodium falciparum: what next? *Trends Parasitol* **17**: 570–571.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**: 539.

Singh B, Sung LK, Matusop A, Radhakrishnan A, Shamsul SSG, Cox-singh J, Thomas A. 2004. A large focus of naturally acquired Plasmodium knowlesi infections in human beings. *Lancet* **363**: 1017–1024.

Sinka ME, Bangs MJ, Manguin S, Coetzee M, Mbogo CM, Hemingway J, Patil AP, Temperley WH, Gething PW, Kabaria CW, et al. 2010. The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. *Parasit Vectors* **3**: 117.

Sinka ME, Bangs MJ, Manguin S, Rubio-Palis Y, Chareonviriyaphap T, Coetzee M, Mbogo CM, Hemingway J, Patil AP, Temperley WH, et al. 2012. A global map of dominant malaria vectors. *Parasit Vectors* **5**: 69.

Sison-Mangus MP, Bernard GD, Lampel J, Briscoe AD. 2006. Beauty in the eye of the beholder: the two blue opsins of lycaenid butterflies and the opsin gene-driven evolution of sexually dimorphic eyes. *J Exp Biol* **209**: 3079–90.

Smallegange RC, Qiu YT, van Loon JJ a, Takken W. 2005. Synergism between ammonia, lactic acid and carboxylic acids as kairomones in the host-seeking behaviour of the malaria mosquito Anopheles gambiae sensu stricto (Diptera: Culicidae). *Chem Senses* **30**: 145–52.

Smit A, Hubley R, Green P. RepeatMasker Open-3.0.

Smit A, Hubley R, Green P. RepeatModeler Open-1.0.

Smith DL, McKenzie FE. 2004. Statics and dynamics of malaria infection in Anopheles mosquitoes. *Malar J* **3**: 13.

Smith JD, Craig AG, Roberts DJ, Hudson-taylor DE, Peterson DS, Pinches R, Newbold C, Miller LH. 1995. Switches in Expression of Plasmodium falciparum var Genes Correlate with Changes in Antigenic and Cytoadherent Phenotypes of Infected Erythrocytes. *Cell* **82**: 101–110.

Smith M a, Gesell T, Stadler PF, Mattick JS. 2013. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res* **41**: 8220–36.

Smith T, Maire N, Ross a, Penny M, Chitnis N, Schapira a, Studer a, Genton B, Lengeler C, Tediosi F, et al. 2008. Towards a comprehensive simulation model of malaria epidemiology and control. *Parasitology* **135**: 1507–16.

Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI. 2005. The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature* **434**: 214–217.

Soderlund DM, Knipple DC. 2003. The molecular biology of knockdown resistance to pyrethroid insecticides. *Insect Biochem Mol Biol* **33**: 563–577.

Soshnev A a, Ishimoto H, McAllister BF, Li X, Wehling MD, Kitamoto T, Geyer PK. 2011. A conserved long noncoding RNA affects sleep behavior in Drosophila. *Genetics* **189**: 455–68.

Spaethe J, Briscoe AD. 2004. Early duplication and functional diversification of the opsin gene family in insects. *Mol Biol Evol* **21**: 1583–94.

Spitzen J, Smallegange RC, Takken W. 2008. Effect of human odours and positioning of CO 2 release point on trap catches of the malaria mosquito Anopheles gambiae sensu stricto in an olfactometer. *Physiol Entomol* **33**: 116–122.

St Pierre SE, Ponting L, Stefancsik R, McQuilton P. 2014. FlyBase 102--advanced approaches to interrogating FlyBase. *Nucleic Acids Res* **42**: D780–8.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–3.

Stamboliyska R, Parsch J. 2011. Dissecting gene expression in mosquito. *BMC Genomics* **12**: 297.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby M a, Rasmussen MD, Roy S, Deoras AN, et al. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450**: 219–32.

Su X, Heatwoie VM, Wertheimer SP, Peterson DS, Ravetch JA, Weilems TE. 1995. The Large Diverse Gene Family w Encodes Proteins Involved in Cytoadherence and Antigenic Variation of Plasmodium falciparum-Infected Erythrocytes. *Cell* **82**: 89–100.

Sui Y, Li B, Shi J, Chen M. 2014. Genomic, regulatory and epigenetic mechanisms underlying duplicated gene evolution in the natural allotetraploid Oryza minuta. *BMC Genomics* **15**: 11.

Sun K, Chen X, Jiang P, Song X, Wang H, Sun H. 2013. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics* **14 Suppl 2**: S7.

Sun L, Zhang Z, Bailey TL, Perkins AC, Tallack MR, Xu Z, Liu H. 2012. Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study. *BMC Bioinformatics* **13**: 331.

Swaminathan A, Gajan A, Pile LA. 2012. Epigenetic regulation of transcription in Drosophila. *Front Biosci* 909–937.

Takken W. 2010. Push-pull strategies for vector control. *Malar J* **9**: I16.

Takken W, Knols BG. 1999. Odor-mediated behavior of Afrotropical malaria mosquitoes. *Annu Rev Entomol* **44**: 131–57.

Takken W, Verhulst NO. 2013. Host preferences of blood-feeding mosquitoes. *Annu Rev Entomol* **58**: 433–53.

Talbert PB, Ahmad K, Almouzni G, Ausió J, Berger F, Bhalla PL, Bonner WM, Cande WZ, Chadwick BP, Chan SWL, et al. 2012. A unified phylogeny-based nomenclature for histone variants. *Epigenetics Chromatin* **5**: 7.

Tchouassi DP, Sang R, Sole CL, Bastos ADS, Cohnstaedt LW, Torto B. 2012. Trapping of Rift Valley Fever (RVF) vectors using light emitting diode (LED) CDC traps in two arboviral disease hot spots in Kenya. *Parasit Vectors* **5**: 94.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**: 178–92.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**: 46–53.

Trapnell C, Williams B a, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–5.

Turusov V, Rakitsky V, Tomatis L. 2002. Dichlorodiphenyltrichloroethane (DDT): Ubiquity, Persistence, and Risks. *Environ Health Perspect* **110**: 125–128.

U. Bernier, D. Kline, S. Allan DB. 2007. Laboratory Comparison of Aedes Aegypti Attraction to Human Odors and to Synthetic Human Odor Compounds and Blends. *J Am Mosq Control Assoc* **23**: 288–293.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**: 1537–50.

Vandersmissen HP, Van Hiel MB, Van Loy T, Vleugels R, Vanden Broeck J. 2014. Silencing D. melanogaster lgr1 impairs transition from larval to pupal stage. *Gen Comp Endocrinol* **209**: 135–147.

Vandesompele J, Preter K De, Poppe B, Roy N Van, Paepe A De. 2002. Accurate normalization of real-time quantitative RT -PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 1–12.

Vantaux A, Lefèvre T, Dabiré KR, Cohuet A. 2014. Individual experience affects host choice in malaria vector mosquitoes. *Parasit Vectors* **7**: 249.

Vargas HCM, Farnesi LC, Martins AJ, Valle D, Rezende GL. 2014. Serosal cuticle formation and distinct degrees of desiccation resistance in embryos of the mosquito vectors Aedes aegypti, Anopheles aquasalis and Culex quinquefasciatus. *J Insect Physiol* **62**: 54–60.

Verhulst NO, Qiu YT, Beijleveld H, Maliepaard C, Knights D, Schulz S, Berg-Lyons D, Lauber CL, Verduijn W, Haasnoot GW, et al. 2011. Composition of Human Skin Microbiota Affects Attractiveness to Malaria Mosquitoes ed. B.S. Schneider. *PLoS One* **6**: e28991.

Verhulst NO, Takken W, Dicke M, Schraa G, Smallegange RC. 2010. Chemical ecology of interactions between human skin microbiota and mosquitoes. *FEMS Microbiol Ecol* **74**: 1–9.

Vermaak D, Henikoff S, Malik HS. 2005. Positive selection drives the evolution of rhino, a member of the heterochromatin protein 1 family in Drosophila. *PLoS Genet* **1**: 96–108.

Vermaak D, Malik HS. 2009. Multiple roles for heterochromatin protein 1 genes in Drosophila. *Annu Rev Genet* **43**: 467–92.

Vijverberg HPM, van der Zalm JM, van den Bercken J. 1982. Similar mode of action of pyrethroids and DDT on sodium channel gating in myelinated nerves. *Nature* **295**: 1–4.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annu Rev Genet* **47**: 97–120.

Vogel H, Badapanda C, Knorr E, Vilcinskas a. 2014. RNA-sequencing analysis reveals abundant developmental stage-specific and immunity-related genes in the pollen beetle Meligethes aeneus. *Insect Mol Biol* **23**: 98–112.

Wahlestedt C. 2013. Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nat Rev Drug Discov* **12**: 433–46.

Wang G, Carey AF, Carlson JR, Zwiebel LJ. 2010. Molecular basis of odor coding in the malaria vector mosquito Anopheles gambiae. *Proc Natl Acad Sci U S A* **107**: 4418–23.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57–63.

Warren W., et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**: 175–183.

Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva E V. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res* **41**: D358–65.

Weill M, Chandre F, Brengues C, Manguin S, Akogbeto M, Pasteur N, Guillet P, Raymond M. 2000. The kdr mutation occurs in the Mopti form of Anopheles gambiae s.s. through introgression. *Insect Mol Biol* **9**: 451–5.

Weiner S a, Toth AL. 2012. Epigenetics in social insects: a new direction for understanding the evolution of castes. *Genet Res Int* **2012**: 609810.

Wellems TE, Plowe C V. 2001. Chloroquine-resistant malaria. *J Infect Dis* **184**: 770–6.

Weng MK, Natarajan K, Scholz D, Ivanova VN, Sachinidis A, Hengstler JG, Waldmann T, Leist M. 2014. Lineage-specific regulation of epigenetic modifier genes in human liver and brain. *PLoS One* **9**: e102035.

Weng MK, Zimmer B, Pöltl D, Broeg MP, Ivanova V, Gaspar J a, Sachinidis A, Wüllner U, Waldmann T, Leist M. 2012. Extensive transcriptional regulation of chromatin modifiers during human neurodevelopment. *PLoS One* **7**: e36708.

White GB. 1974. Anopheles gambiae complex and disease transmission in Africa. *Trop Med Hyg* **68**: 278–298.

White NJ, Pukrittayakamee S, Hien TT, Faiz MA, Mokuolu O a, Dondorp AM. 2014. Malaria. *Lancet* **383**: 723–35.

Whitty CJM, Chandler C, Ansah E, Leslie T, Staedke SG. 2008. Deployment of ACT antimalarials for treatment of malaria: challenges and opportunities. *Malar J* **7 Suppl 1**: S7.

WHO. 2014. *WHO Global Malaria Programme: World Malaria Report 2014*.

Wilhelm BT, Marguerat S, Goodhead I, Bähler J. 2010. Defining transcribed regions using RNA-seq. *Nat Protoc* **5**: 255–266.

Will S, Yu M, Berger B. 2013. Structure-based whole-genome realignment reveals many novel noncoding RNAs. *Genome Res* **23**: 1018–27.

Williams Jr LL. 1963. Malaria Eradication in the United States. *Am J Public Heal Nations Heal* **53**: 17–21.

Wilton DP, Fay RW. 1972. Responses of adult Anopheles stephensi to light of various wavelengths. *J Med Entomol* **9**: 301–304.

Wood EJ, Chin-Inmanu K, Jia H, Lipovich L. 2013. Sense-antisense gene pairs: sequence, transcription, and structure are not conserved between human and mouse. *Front Genet* **4**: 183.

Wyder S, Kriventseva E V, Schröder R, Kadowaki T, Zdobnov EM. 2007. Quantification of ortholog losses in insects and vertebrates. *Genome Biol* **8**: R242.

Xia Z, Donehower L a, Cooper T a, Neilson JR, Wheeler D a, Wagner EJ, Li W. 2014. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun* **5**: 5274.

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**: 329–342.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–91.

Yohannes M, Boelee E. 2012. Early biting rhythm in the Afro-tropical vector of malaria, Anopheles arabiensis, and challenges for its control in Ethiopia. *Med Vet Entomol* **26**: 103–5.

Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu J-L, Ponting CP. 2012. Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. *Genome Biol Evol* **4**: 427–42.

Yu M, Kautz M a, Thomas ML, Johnson D, Hotchkiss ER, Russo MB. 2007. Operational implications of varying ambient light levels and time-of-day effects on saccadic velocity and pupillary light reflex. *Ophthalmic Physiol Opt* **27**: 130–41.

Yuan Q, Metterville D, Briscoe AD, Reppert SM. 2007. Insect cryptochromes: gene duplication and loss define diverse ways to construct insect circadian clocks. *Mol Biol Evol* **24**: 948–55.

Zaim M, Guillet P. 2002. insecticides : an urgent need. **18**: 2001–2003.

Zdobnov EM, Subramanian GM, Mueller H, Birney E, Charlab R, Halpern AL. 2002. Comparative Genome and Proteome Analysis of Anopheles gambiae and Drosophila melanogaster. 149–159.

Zhou Q, Ellison CE, Kaiser VB, Alekseyenko A a, Gorchakov A a, Bachtrog D. 2013. The epigenome of evolving Drosophila neo-sex chromosomes: dosage compensation and heterochromatin formation. *PLoS Biol* **11**: e1001711.

Zhou VW, Goren A, Bernstein BE. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* **12**: 7–18.