Genomics and Transcriptomics Analysis of the Asian Malaria Mosquito *Anopheles stephensi*

Xiaofang Jiang

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Genetics, Bioinformatics, and Computational Biology

Zhijian Tu (Committee Chair)

Igor Sharakhov

David Bevan

Lenwood S. Heath

Liqing Zhang

April 14th 2016

Blacksburg, VA

Keywords: genomic, comparative transcriptomes, dosage compensation, sex-specific expression Iso-Seq, *trans*-splicing

Genomics and Transcriptomics Analysis of the Asian Malaria Mosquito *Anopheles stephensi*

Xiaofang Jiang

Abstract

Anopheles stephensi is a potent vector of malaria throughout the Indian subcontinent and Middle East. *An. stephensi* is emerging as a model for molecular and genetic studies of mosquito-parasite interactions. Here we conducted a series of genomic and transcriptomic studies to improve the understanding of the biology of *Anopheles stephensi* and mosquito in general.

First we reported the genome sequence and annotation of the Indian strain of the “type” form of *An. stephensi*. The 221 Mb genome assembly was produced using a combination of 454, Illumina, and PacBio sequencing. This hybrid assembly method was significantly better than assemblies generated from a single data source. A total of 11,789 protein-encoding genes were annotated using a combination of homology and *de novo* prediction.

Secondly, we demonstrated the presence of complete dosage compensation in *An. stephensi* by determining that autosomal and X-linked genes have very similar levels of expression in both males and females. The uniformity of average expression levels of autosomal and X-linked genes remained when *An. stephensi* gene expression was normalized by that of their *Ae. aegypti* orthologs, strengthening the conclusion of complete dosage compensation in *Anopheles*.

Lastly, we investigated *trans*-splicing events in *Anopheles stephensi*. We identified six *trans*-splicing events and all the *trans*-splicing sites are conserved and present in *Ae. aegypti*. The proteins encoded by the *trans*-spliced mRNAs are also highly conserved and their orthologs are co-linearly transcribed in out-groups of family *Culicidae*. This finding indicates the need to preserve the intact mRNA and protein function of the broken-up genes by *trans*-splicing during evolution.

In summary, we presented the first genome assembly of *Anopheles stephensi* and studied two interesting evolution events – dosage compensation and *trans*-splicing - via transcriptomic analysis.

Genomics and Transcriptomics Analysis of the Asian Malaria Mosquito *Anopheles stephensi*

Xiaofang Jiang

Public Abstract

Malaria is one of the deadliest diseases known to man and is caused by a parasite called *Plasmodium* that is transferred from person to person by mosquitoes. In this dissertation, we studied the Indian malaria mosquito, which as its name suggests is an important malaria vector throughout India and the Middle East. The Indian malaria mosquito is a close relative of the even deadlier African malaria mosquito, so knowledge gleamed from the Indian malaria mosquito can be applied to the African malaria mosquito. In addition, the Indian malaria mosquito is emerging as a model for molecular and genetic studies of mosquito-parasite interactions, so a high-quality genome assembly of this species is invaluable to the mosquito research community. Here, we conducted a series of genomic and transcriptomic studies to improve the basic understanding of the biology of the Indian malaria mosquito.

First, we assembled the genome sequence of the Indian wild-type strain of the Indian malaria mosquito, providing an essential foundation for future research on this species. The Indian malaria mosquito genome is 221 million base pairs and contains a total of 11,789 protein-coding genes. Second, we demonstrated the presence of complete dosage compensation in the Indian malaria mosquito, which means that genes from the X chromosome doubled their expression during the course of evolution. Because dosage compensation functions on a sex-to-sex basis, it may be possible to use the genes responsible for initiating dosage compensation to kill deadly female mosquitoes. As only female mosquitoes bite and transmit disease, killing females is extremely desirable for vector control. Lastly, we identified six trans-splicing events. Trans-splicing is when two or more distinct messenger RNAs are joined into a single coding transcript. All the trans-

splicing sites we identified are conserved in mosquitoes indicating that *trans*-splicing may have evolved to preserve genes broken-up during evolution.

Acknowledgements

It would be difficult for me to accomplish the studies presented in this dissertation without the support of a large number of people.

First and foremost, I would like to thank my committee members, Zhijian Jake Tu, Igor Sharkhov, David Bevan, Lenwood Heath and Liqing Zhang for their input, suggestions and support. I am so grateful to my advisor, Zhijian Tu, for 5 years of mentorship. During this time, he not only provided me guidance and resources when I needed it, but also give me considerable freedom and independence to figure out what I am interested and encourage me to pursue and accomplish it. I would like to thank Dr. Sharakhov for providing me the chance to join the 16 *Anopheles* genome project and give valuable feedback for my papers.

I am very thankful to my lab mates and collaborators that have assisted me with my research: Frank Criscione, James Biedler, Yumin Qi, Wanqi Hu, Randy Saunders, Ashley Peery. I thank all of my friends for the help and suggestion. I have many wonderful memories here and I will miss all the friends I meet here.

The GBCB program provided excellent support throughout my PhD journey. Many thanks to Dr. Bevan and Dennie Munson for their wonderful work.

Finally, and most importantly, I want to thank my parents and sister, who give me unconditional love and invaluable support. I especially want to thank my husband Brantley Hall for help with my research, for invaluable day-to-day support, for tolerating my highs and lows as I write this dissertation.

Attribution

Many colleagues and collaborators were involved in the research presented in chapters 2-3. Here, I have briefly described their respective contributions.

Chapter 2: Genome analysis of a major urban malaria vector mosquito, *Anopheles stephensi*

Chapter 2 was published in Genome Biology.

Jiang, Xiaofang, Ashley Peery, A. Brantley Hall, Atashi Sharma, Xiao-Guang Chen, Robert M. Waterhouse, Aleksey Komissarov et al. "Genome analysis of a major urban malaria vector mosquito, Anopheles stephensi." *Genome biology* 15, no. 9 (2014): 1-18.

Xiaofang Jiang (Program of Genetics, Bioinformatics, and Computational Biology, Virginia Tech; Department of Biochemistry, Virginia Tech) helped to devise and perform the hybrid approach used to assemble the genome, preformed genome and functional annotation, performed additional analysis of gene families, and helped write the manuscript.

Ashley Peery (Department of Entomology, Virginia Tech) was co-first-author on this paper, conducted the physical mapping experiments and the analysis of the distribution of genomic features, helped to generate the figures and write the manuscript.

A. Brantley Hall (Program of Genetics, Bioinformatics, and Computational Biology, Virginia Tech; Department of Biochemistry, Virginia Tech) helped in the assembly effort, generated figures, and helped write the manuscript.

Atashi Sharma (Department of Entomology, Virginia Tech) participated in data generation, analysis and presentation in regards to physical mapping.

Xiao-Guang Chen (Department of Pathogen Biology, Southern Medical University, China) provided resources and tools and critical reviewed manuscript.

Robert M. Waterhouse (Department of Genetic Medicine and Development, University of Geneva Medical School; Swiss Institute of Bioinformatics; Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology; The Broad Institute of MIT and Harvard) participated in data generation, analysis and presentation of orthologous genes.

Aleksey Komissarov (Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, Russia) participated in data generation, analysis and presentation.

Michelle M. Riehl (Department of Microbiology, University of Minnesota) participated in data generation, analysis and presentation.

Yogesh Shouche (National Center for Cell Science, Pune University Campus, India) provided resources and tools and critical reviewed manuscript.

Maria V. Sharakhova (Department of Entomology, Virginia Tech) participated in data generation, analysis and presentation.

Dan Lawson (European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, United Kingdom) participated in data generation, analysis and presentation.

Nazzy Pakpour (Department of Medical Microbiology and Immunology, University of California) participated in data generation, analysis and presentation.

Peter Arensburger (Biological Sciences Department, California State Polytechnic University Pomona) participated in data generation, analysis and presentation.

Victoria L. M. Davidson (Division of Biology, Kansas State University) participated in the data generation, analysis and presentation.

Karin Eiglmeier (Department of Parasitology and Mycology, Unit of Insect Vector Genetics and Genomics, Institut Pasteur) participated in the data generation, analysis and presentation.

Scott Emrich (Department of Computer Science and Engineering, University of Notre Dame) participated in the data generation, analysis and presentation.

Phillip George (Department of Entomology, Virginia Tech) participated in the data generation, analysis and presentation.

Ryan C. Kennedy (Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco) participated in the data generation, analysis and presentation.

Shrinivasrao P. Mane (Virginia Bioinformatics Institute, Virginia Tech) participated in the data generation, analysis and presentation.

Gareth Maslen (European Bioinformatics Institute, Wellcome Trust) participated in the data generation, analysis and presentation.

Chioma Oringanje (Department of Entomology, University of Arizona) participated in the data generation, analysis and presentation.

Yumin Qi (Department of Biochemistry, Virginia Tech) participated in the data generation, analysis and presentation.

Robert Settlage (Virginia Bioinformatics Institute, Virginia Tech) participated in the data generation, analysis and presentation.

Marta Tojo (Department of Physiology, School of medicine – CIMUS, Instituto de Investigaciones Sanitarias, University of Santiago de Compostela) participated in the data generation, analysis and presentation.

Jose M. C. Tubio (Wellcome Trust Sanger Institute) participated in the data generation, analysis and presentation.

Maria F. Unger (Department of Biological Sciences, University of Notre Dame) participated in the data generation, analysis and presentation.

Bo Wang (Department of Medical Microbiology and Immunology, University of California, Davis) participated in the data generation, analysis and presentation.

Kenneth D. Vernick (Department of Parasitology and Mycology, Unit of Insect Vector Genetics and Genomics, Institut Pasteur, Paris) participated in the data generation, analysis and presentation.

Jose M. C. Ribeiro (Section of Vector Biology, Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases) participated in the data generation, analysis and presentation.

Anthony A. James (Departments of Microbiology & Molecular Genetics and Molecular Biology & Biochemistry, University of California, Irvine) participated in the data generation, analysis and presentation.

Kristin Michel (Division of Biology, Kansas State University) participated in the data generation, analysis and presentation.

Michael A. Riehle (Department of Entomology, University of Arizona) participated in the data generation, analysis and presentation.

Shirley Luckhart (Department of Medical Microbiology and Immunology, University of California, Davis) participated in the data generation, analysis and presentation.

Igor V. Sharakhov (Program of Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Department of Entomology, Virginia Tech) conceived and designed the experiments, led the physical mapping effort, helped to write the manuscript.

Zhijian Tu (Program of Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Department of Biochemistry, Virginia Tech) conceived and designed the experiments, led the assembly effort, helped to write the manuscript.

Chapter 3: Complete dosage compensation in *Anopheles stephensi* and the evolution of sex-biased genes in mosquitoes

Chapter 3 was published in Genome Biology Evolution.

Jiang, Xiaofang, James K. Biedler, Yumin Qi, Andrew Brantley Hall, and Zhijian Tu. "Complete dosage compensation in Anopheles stephensi and the evolution of sex-biased genes in mosquitoes." *Genome biology and evolution* 7, no. 7 (2015): 1914-1924.

Xiaofang Jiang (Program of Genetics, Bioinformatics, and Computational Biology, Virginia Tech; Department of Biochemistry, Virginia Tech) helped to conceive and design the experiments performed in this manuscript, performed the bioinformatics analysis and statistical analysis, and wrote the manuscript.

James K. Biedler PhD (Department of Biochemistry, Virginia Tech) helped with experiments and molecular techniques used in this manuscript.

Yumin Qi PhD (Department of Biochemistry, Virginia Tech) helped with experiments and molecular techniques used in this manuscript.

Andrew Brantley Hall (Genetics, Bioinformatics, and Computational Biology program, Virginia Tech) is currently a graduate student and helped with the bioinformatics techniques used in this manuscript.

Zhijian Tu PhD (Department of Biochemistry, Virginia Tech) is a professor of biochemistry and helped to conceive and design the experiments performed in the manuscript.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 *Anopheles stephensi*

Anopheles mosquitoes are the major human malaria vectors. Despite large efforts being made to fight malaria, there are still more than 200 million incidences and around 438,000 mortalities due to malaria in 2015, according to the latest report from WHO (http://www.who.int/malaria/publications/world-malaria-report-2015/en/). The majority of infections and deaths occurred in children under five in sub-Saharan Africa. Consequently, this disease imposes a substantial social and economic burden (White et al. 2011). Therefore, malaria eradication is a goal worth pursuing. Vector control is an essential component to achieve malaria eradication. The knowledge in vector genetics, behavior, and physiology serve as the basis to develop vector control strategies. Understanding basic mosquito biology can be highly accelerated by the study of genome and transcriptome sequence.

Anopheles stephensi belongs to subgenus *Cellia*, the same subgenus as *Anopheles gambiae*, the major malaria vector in Africa. *Anopheles stephensi* is among the 30–40 species that commonly transmit malaria. *Anopheles stephensi* is an established malaria vector predominately in the Indian subcontinent, with a wide distribution across the Middle East and South Asia regions(Sharma 1999; Sinka et al. 2011). A recent study posited that a recent resurgence of human malaria on the African continent could have been caused by the sudden rise of *An. stephensi* in urbanized areas (Faulde, Rueda, and Khaireh 2014). The emergence of *An. stephensi*-transmitted malaria indicates that the key traits determining vectorial capacity of *An. stephensi* such as high susceptibility to malaria infection and quick adaptation to urban habitats make it capable to pose an even greater risk to human health in the future.

There are three ecological variants of *An. stephensi* reported: type, *mysorensis*, and intermediate.

The type form is a competent malaria vector in urban areas due to its anthropophilic nature; the *mysorensis* form is zoophilic, and mainly exists in rural areas; the intermediate form generally stays in rural villages and peri-urban areas, but its vector status is unclear. *An. stephensi* accounts for approximately 12% of all malaria transmission in India (Gakhar, Sharma, and Sharma 2013). Thus, efforts to control *An. stephensi* could result in a significant reduction in malaria transmission, especially in urban environments (Behura et al. 2011).

An. stephensi is amenable to genetic manipulations. Transposon-based germline transformation of *Anopheles stephensi* have been achieved since 2000 (Catteruccia et al. 2000; Nolan et al. 2002). The heritable RNA interference system was first established in *An. stephensi*, which provided an important tool for functional genomic analysis (Brown et al. 2003). In 2011, one study demonstrated that the *piggyBac* transposon is highly active in the germline of *An. stephensi* (O'Brochta et al. 2011), which makes genome-wide mutagenesis technology possible. Site-specific transgene integration systems have also been successfully applied in *An. stephensi* (Isaacs et al. 2012). As to genome-editing, both TALENs and the CRISPR-Cas9 system have been successful established to knockout targeted genes (Smidler et al. 2013; Gantz et al. 2015). Recently, a highly efficient Cas9-mediated gene drive in *An. stephensi* has been constructed and reported (Gantz et al. 2015).

Our understanding of the interactions between *An. stephensi* and the malaria parasites is rapidly improving. Early studies showed that inducible synthesis of nitric oxide in *An. stephensi* limits the development of the malaria parasite and the induction of nitric oxide synthase is mediated by parasite glycosylphosphatidylinositols, which created signaling that is of mimicry but distinctively different from insulin signaling (Luckhart et al. 1998; Lim et al. 2005). The insulin/IGF-1 signaling (IIS) pathway in *Anopheles stephensi* turned out to be critical to determine susceptibility to the malaria parasite (Sanapala et al. 2012). Ingested human insulin suppresses mosquito immune response to parasite by activating mosquito IIS (Pakpour et al. 2012). The malaria parasite has been observed to induce the synthesis of insulin-like peptides to weaken the immune response in their mosquito host (Pietri, Potts, and Pietri 2014). In addition, other factors that contributes to the resistance to malaria parasites such as Caspar, myristoylated Akt and, *Wolbachia* infection have also been extensively studied, even adopted to engineer *An. stephensi* to fight malaria (Garver, Dong, and Dimopoulos 2009; Luckhart et al. 2013; Bian et al. 2013). Therefore, *An. stephensi* is emerging as a model species for genetic and molecular studies.

1.2 Overview of the mosquito genome projects

1.2.1 Brief introduction of sequencing technologies and their applications in mosquitoes

First generation sequencing, or Sanger sequencing, was developed by Frederick Sanger in the mid-1970s. This technology uses the chain-termination method, where DNA fragments of varying length are generated and the last base at the end of each fragment is read (Sanger, Nicklen, and Coulson 1977). Sanger sequencing prevailed for the next two decades, over which period the technique has been greatly advanced. The human genome project started in 1988 and was completed in 2001 (Venter et al. 2001). In this project, 3 billion nucleotide base pairs in the human genome were obtained using Sanger sequencing (Weber and Myers 1997). In shotgun sequencing, genomic DNA fragments are cloned in bacteria and sequenced, and contiguous sequences are assembled from overlapping clones. At the same time, a number of genomes from model organisms such as yeast, fruit flies, and mice were sequenced along with the human genome (Mewes et al. 1997; Adams et al. 2000; Waterston et al. 2002). The first genome of a mosquito, *Anopheles gambiae,* was sequenced with shotgun sequencing and published in 2002 (Holt et al. 2002). In 2007 and 2010, the genomes of two other mosquitoes, *Aedes aegypti* and *Culex* *quinquefasciatus* were sequenced based on Sanger sequencing (Nene et al. 2007; Arensburger et al. 2010).

Although Sanger sequencing can produce reads of high accuracy over long lengths, it is quite slow and expensive. These limitations and problems triggered the development of next-generation sequencing technology. In 2005, the first next-generation sequencing technology, the 454 technology, was released to the market. 454 sequencing is based on the pyrosequencing technique, which reads the sequences by detecting the pyrophosphate release when a new nucleotide is added to a growing chain. To obtain enough light intensity for reliable detection, DNA molecules on each bead are amplified by emulsion PCR. The 454 sequencing technology can analyze a large number of samples in parallel, but the error rate for homopolymer sequences is high. 454 sequencing technology have been used in a variety of applications in many species (Argout et al. 2011; Suen et al. 2011; Velasco et al. 2010). The genome of *Anopheles darlingi* and *Anopheles sinensis* were both sequenced with 454 technology (Osvaldo Marinotti et al. 2013; Zhou et al. 2014). The assembly of our *Anopheles stephensi* genome also relied heavily on 454 sequencing data (Jiang et al. 2014).

Currently, Illumina sequencing is the market leader due to its low cost, its high-throughput, and the flexible nature of the technology. Illumina technology performs sequencing by synthesis using dyed-reversible terminator nucleotides. The sequences are read by detecting fluorescent dye that is cleaved from the last incorporated nucleotide. Along with the dye, the terminal 3' blocker is also chemically removed from the synthesized DNA chain, which allows the next round of synthesis. Illumina sequencing features the biggest output and lowest reagent cost, while the reads produced were generally not long compared to 454 sequencing (the current longest reads for HiSeq2500 is 250bp). In the end of 2014, the genomes analysis of additional 16 *Anopheles* were reported

(Neafsey et al. 2014). All 16 genomes were sequenced using the Illumina platform. The sequencing data are paired end with a range of insert size from 180 bp to 38 kbp. This genome project is another landmark for the field of malaria vector research since the publication of the *An. gambiae* genome (Holt et al. 2002). In 2015, the *Aedes albopictus* genome was sequenced with the Illumina HiSeq2000 platform (X.-G. Chen et al. 2015). The *Aedes albopictus* genome comprises 1,967 Mb, nearly ten-time the size of the genome of *Anopheles*, and is the largest mosquito genome ever sequenced.

Other next generation sequencing technology such as SOLiD sequencing has also been applied to the study of mosquitoes(Scott et al. 2010). However, to our knowledge, no mosquito genome was sequenced using other next generation technology besides 454 and Illumina. The application of next-generation sequencing technologies to mosquito genomics offers exciting opportunities to expand our understanding of mosquito biology in many important vector species and harness the power of comparative genomics.

As to the third generation sequencing technologies, the Single molecule real time (SMRT) sequencing from Pacific Biosciences (PacBio) is the current market leader, as the Heliscope single molecule sequencing company Helicos Biosciences went bankrupt, and the GridION and MinION systems from Oxford Nanopores are still in testing (Munroe and Harris 2010; Han et al. 2015). The SMRT Cell is the unit for PacBio sequencing. Each SMRT cell contains tens of thousands of zero-mode waveguides where a DNA template-polymerase complex is located at the bottom (Eid et al. 2009). While the single stranded DNA is replicated, the sequence will be read. The advantage of PacBio Sequencing is that it is fast, has least GC bias, no amplification bias, and most importantly, the read length is long. Over the past four years, PacBio has greatly improved the read length. With the newest chemistry, reads over 40 kb are possible. However, PacBio sequencing is still costly and the reads produced have a high error rate. The SMRTbell™ template was introduced to fix the error problem by generating high-accuracy consensus sequence from multiple observations of the same single molecules (Travers et al. 2010). *De novo* genome assembly from PacBio reads have been done in multiple species including five model organisms: *Escherichia coli*, *Saccharomyces cerevisiae*, *Neurospora crassa*, *Arabidopsis thaliana*, and *Drosophila melanogaster* (Kim et al. 2014; VanBuren et al. 2015). In mosquitos, PacBio sequencing was performed on *Anopheles gambiae* to obtain Y chromosome sequences (Hall *et al*., in press). In our *An. stephensi* genome project, we used PacBio sequencing data to reduce gaps in the assembly (Jiang et al. 2014).

1.2.2 The *Anopheles gambiae* genome project

Anopheles gambiae is the primary mosquito vector responsible for the transmission of *Plasmodium falciparum*, a malaria parasite of humans prevalent in most of sub-Saharan Africa. The publication of the *Anopheles gambiae* genome ushered in a new era of malaria vector research (Holt et al. 2002).

The availability of the genome accelerated research that has not only enhanced our basic understanding of vector genetics, behavior, and physiology and roles in transmission but also contributed to new strategies for combating malaria. The *An. gambiae* genome was sequenced with shotgun sequencing in a collaboration with several sequencing centers and universities. The PEST strain of *An. gambiae* was sequenced at tenfold coverage and then was assembled into 8,987 scaffolds spanning 278 million base pairs. The largest scaffold is 23 million base pairs. 91% of total sequences can be included with the largest 303 scaffolds. In the most up to update assembly (version 4, https://www.vectorbase.org/organisms/anopheles-gambiae/pest/agamp4), 230 million base pairs have been assigned to chromosomes by in situ mapping. 13,008 protein-encoding and 767 non-coding genes were predicted based on genome assemblies (Holt et al. 2002).

The PEST strain of *An. gambiae* was chosen for genome sequencing for several reasons: its chromosome arrangement is homogeneous; it has an X-linked, pink-eye mutation that can be used to screen out cross-colony contaminations; it has been previously used in several studies of human reservoirs of malaria; and bacterial artificial chromosome (BAC) libraries construction had been done in the PEST strain (Holt et al. 2002; Land 2003). However, the high level of polymorphism in the PEST strain made the genomic assembly step difficult. The PEST strain was established from crossing a laboratory strain originating in Nigeria with field-collected ones from western Kenya. As a result, some of the assemblies were alternative assemblies of highly divergent regions of the diploid genome (Mongin et al. 2004). Hence, additional efforts were used to identify these polymorphic scaffolds. In addition, the sequencing sample is a mixture of both male and female mosquitoes. The Y chromosome is too repetitive to be assembled and the assembly quality of the X chromosome is compromised due to the reduced coverage caused by males in the sample (Mongin et al. 2004).

The completed *Anopheles gambiae* genome was no doubt a huge achievement and created numerous opportunities for mosquito and malaria research. At the same time, the lessons acquired from the project also have greatly aided future genome projects.

1.2.3 The *Aedes aegypti* and *Culex quinquefasciatus* genome projects

The genome of the yellow fever mosquito *Ae. aegypti* was published in 2007 (Nene et al. 2007). The sequencing was performed with whole-genome shotgun sequencing. A highly inbred strain, Liverpool-IB12, was chosen for sequencing. The genome coverage was around 7.6 fold and 4,758 scaffolds spanning 1,380 million base pairs were assembled. 98% of the genome can be represented with 1,257 large scaffolds. The N50 scaffold size is approximate 1.5 million base pairs. In 2010, the *Culex quinquefasciatus* genome project was completed (Arensburger et al. 2010). 3,171 scaffolds spanning 579 million base pairs were assembled. This assembly is significantly more fragmented compared to the previous two genome assemblies with an N50 scaffold size of only 486 thousand base pairs, perhaps to reduced sequencing coverage (~6.1 fold).

These two genome projects provided additional resources to perform comparative genomics in mosquito. The genome size is drastically different in the three mosquitoes, largely due to transposable elements. 42% to 47% of the *Ae. aegypti* genome was composed of transposable elements, while the number for *C. quinquefasciatus* is 29% and for *An. gambiae* is 11% to 16%. The number of annotated genes and the size of exons are not so different among the three assemblies. In addition, some gene families with crucial functions display dynamic evolution patterns, as they expand or contract quickly among different mosquito lineages. Those genes include olfactory and gustatory receptors, immunity genes, cuticle-related genes, and gene involved in insecticide resistance (Arensburger et al. 2010; Severson and Behura 2012).

1.2.4 Applications of next generation sequencing in mosquitos and the *Anopheles* 16 genome project

The advent of next generation sequencing technology has promoted a series of genome projects including many additional mosquito species. The genome of *Anopheles darlingi*, the main neotropical malaria vector, was published in 2013 (Osvaldo Marinotti et al. 2013), and the genome of *Anopheles sinensis,* an important human parasitic diseases vector in Southeast Asia, was published in 2014 (Zhou et al. 2014). Both these genomes were assembled from 454 sequencing data. For *Anopheles darlingi,* 20x coverage of raw reads were assembled into 8,233 scaffolds spanning 173.9 million base pairs with a N50 scaffold size of 81 thousand base pairs. For *Anopheles sinensis*, 18.8 fold coverage of reads were assembled into 9,594 scaffolds spanning 220.8 million base pairs with an N50 scaffold size of 220.8 thousand base pairs. To overcome the disadvantages of short read length in 454 sequencing, increased sequencing depth and the use of mate pair reads with long insertion sizes were an effective strategy to generate good genome assemblies. Both genomes assemblies are of high quality and provide valuable genomic information for future research.

Inspired by the 12 *Drosophila* genomes project (Neafsey et al. 2013), the *Anopheles* 16 Genomes project has been carried out to provide a comparative framework to study vectorial capacity of the *Anopheline* mosquito (Neafsey et al. 2013; Neafsey et al. 2014). In this project, the genomes and transcriptomes of an additional 16 *Anopheles* were sequenced using Illumina technology. These 16 species have different degrees of vectoral capacity, covering a range of evolutionary distances (100 million years) and a variety of geographic and ecological niches.

The assemblies of the genomes have been greatly improved by three strategies: the use of inbred mosquitos to reduce genetic diversity, the modification of the assembly algorithm to allow high heterozygosity rates, and the utilization of Fosmid-scale Illumina libraries to generate reads with long insert sizes. Remarkable success has been achieved for the project (Neafsey et al. 2013). Although the quality varies among species, the majority of the assemblies were excellent. For example, the whole genome of *An. albimanus* can be represented with 204 scaffolds and the N50 scaffold size is 18 million base pairs.

Based on this project there are two major findings: introgressive hybridization played an important role in mosquito complex evolution (Fontaine et al. 2014); with quick gene turnover and elevated

X-linked gene shuffles, the *Anopheline* genomes are very dynamic, which may contribute to their vectorial capacity and fast adaptation (Neafsey et al. 2014).

1.3 Genome-wide transcriptomics analysis in mosquitoes

1.3.1 The application of microarrays to transcriptome studies in mosquitoes

Microarrays were the first technique to allow genome-wide surveys of the transcriptome (Schulze and Downward 2001). Before that, only a small number of transcripts could be studied at a time. With the advent of microarrays, thousands of transcripts can be monitored at once. The use of microarrays to perform transcriptome profiling is to simply measure the relative concentrations of DNA or RNA sequences (Schulze and Downward 2001). They have been used in a wide variety of research including many projects in mosquitoes.

Spotted microarrays were the first widely available, where "in-house" printed microarrays were produced with presynthesized oligos or PCR products (Schulze and Downward 2001). Spotted microarrays had been mostly used in comparative studies. Microarrays have been used to study genes expression changes in the female midgut in *Aedes aegypti* to test the effects of blood feeding and pathogen infections (Sanders et al. 2003; H. Chen et al. 2004). Microarrays have also been used in *An. gambiae* to study which genes are responsible for the difference between the insecticide resistant strain and the susceptible strain (Vontas et al. 2005) and which genes are regulated by REL2 (Meister et al. 2005). Spotted microarrays have been used in studies in *An. stephensi* where the female midgut response to *Plasmodium* infection has been examined across development stages of the parasite (Xu et al. 2005).

Later, commercial platforms where oligonucleotide sequences were directly synthesized on a chip surface become available. These commercial platforms were largely enabled by access to the annotated gene sets from genome projects. For example, the probe set of Affymetrix

Plasmodium/Anopheles Genome Array includes approximately 14,900 *Anopheles gambiae* transcripts, the NimbleGen custom *Aedes aegypti* 12plex array is designed for 17,494 ORFs from *Aedes aegypti*, and the NimbleGen custom *Culex quinquefasciatus* array is also based on gene annotation from *Culex quinquefasciatus*.

Numerous research studies have been performed using the custom mosquito microarrays. Taking the Affymetrix Plasmodium/Anopheles Genome Array as an example, many aspects of *Anopheles gambiae* biology have been explored, including: desiccation stress, circadian rhythms, sex- and tissue-specificity expression, embryo development, and response to bloodmeal (Wang et al. 2011; Rund et al. 2011; Baker et al. 2011; Goltsev et al. 2009; O. Marinotti et al. 2006). Not only important findings were uncovered in these studies, the microarray data they deposited online provides great resources to a wide community of researchers. Microarray data is commonly shared through Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). Other sharing methods such as individual databases were also used to provide more convenient exploring experiences, such as angaGEDUCI (*Anopheles gambiae* gene expression database http://www.angaged.bio.uci.edu/) (S. N. Dissanayake et al. 2006), aeGEPUCI (gene expression in the dengue vector mosquiteo, Aedes aegypti http://www.aegep.bio.uci.edu/) (S. Dissanayake et al. 2010), Atlas (Expression atlas of sex- and tissue-specificity in malaria vector http://mozatlas.gen.cam.ac.uk/mozatlas/index.html) (Baker et al. 2011).

1.3.2 The application of RNA-Seq to transcriptome studies in mosquitoes

Despite the large number of research studies performed with microarray transcriptome analysis, currently whole-transcriptome RNA-sequencing (RNA-Seq) analysis is a more preferable approach for whole-transcriptome analysis. There are several aspects where RNA-Seq outcompetes microarrays. First, microarrays require preexisting annotations of genomes and transcripts to design, while RNA-Seq can be *de novo* assembled and used for new species with no reference genome. Second, the microarray signal is linear only over a limited range of concentrations, which is not a problem for RNA-Seq based analysis. Third, the microarray oligos are short (60 base pairs for Agilent, 25 base pairs for Affymetrix) and may potentially bind to homologs genes non-specifically, while RNA-Seq reads can be as long as 250 base pairs and have better specificity. In addition, RNA-Seq can provide more information including SNP and indel mutations, alternative spliced isoforms and untranslated region identification.

RNA-Seq has been used in a wide range of mosquito studies (Akbari et al. 2013; Biedler et al. 2012; Jiang et al. 2015). The transcriptome of *An. funestus* was *de novo* assembled from Illumina RNA-Seq in 2010 (Crawford et al. 2010) and from 454 sequencing data in 2011 (Gregory et al. 2011). Both studies provided valuable resources for research in this medically important malaria vector before the genome was sequenced. Similar research has been done in the Asian tiger mosquito, *Aedes albopictus* (Poelchau et al. 2011), and the South East Asia mosquito *An. sinensis* (B. Chen et al. 2014).

Comparative transcriptome studies also provide insights into mosquito research. Rinker et al. compared the transcriptomes of two sibling *An. gambiae* species and identified olfactory receptor differences that could alter host preference (Rinker, Zhou, et al. 2013). Extensive studies have been done to investigate how blood-feeding changes the behavior of females by comparing transcriptomes before and after blood-meals (Vannini et al. 2014; Rinker, Pitts, et al. 2013; Bonizzoni et al. 2011). Some of this research is on global gene expression changes (Bonizzoni et al. 2011); others focus on a particular tissue or a particular protein (Vannini et al. 2014; Rinker, Pitts, et al. 2013). A recent study discovered that odorant receptor AaegOr4 is strongly associated with human host preference by comparing the transcriptome data between two forms of *Aedes aegypti* (McBride et al. 2014).

Rather than focusing on mRNA, some researchers have driven deeper into the transcriptome to identify microRNAs, piRNA, and long noncoding RNAs in mosquito. microRNAs play an important role in embryonic development, antiviral immunity and other gene regulation. A shifted expression profile has been observed and studied in several *Aedes* studies via the sequenced miRNAome (Liu et al. 2015; Campbell et al. 2014; Etebari et al. 2015). piRNA is also involved in anti-viral defense as piRNAs were induced by virus infection (Morazzani et al. 2012; Hess et al. 2011). Long noncoding RNAs have been identified via deep RNA sequencing and then systematically studied and described in *Anopheles gambiae* and across the genus *Anopheles* (A. M. Jenkins et al. 2014; Adam M Jenkins, Waterhouse, and Muskavitch 2015).

In summary, RNA-seq has revolutionized transcriptomics studies in the mosquito community and biological science in general.

1.4 Reference

Adams, M D, S E Celnicker, R A Holt, C A Evans, and J D Gocayne. 2000. "The Genome Sequence of Drosophila Melanogaster." *Science* 287 (5461). American Association for the Advancement of Science: 2185.

Akbari, Omar S, Igor Antoshechkin, Henry Amrhein, Brian Williams, Race Diloreto, Jeremy Sandler, and Bruce a Hay. 2013. "The Developmental Transcriptome of the Mosquito Aedes Aegypti, an Invasive Species and Major Arbovirus Vector." *G3 (Bethesda, Md.)* 3 (9): 1493–1509. doi:10.1534/g3.113.006742.

Arensburger, Peter, Karine Megy, Robert M Waterhouse, Jenica Abrudan, Paolo Amedeo, Beatriz Antelo, Lyric Bartholomay, et al. 2010. "Sequencing of Culex Quinquefasciatus Establishes a Platform for Mosquito Comparative Genomics." *Science (New York, N.Y.)* 330 (6000): 86–88. doi:10.1126/science.1191864.

Argout, Xavier, Jerome Salse, Jean-Marc Aury, Mark J Guiltinan, Gaetan Droc, Jerome Gouzy, Mathilde Allegre, et al. 2011. “The Genome of Theobroma Cacao.” *Nature Genetics* 43 (2). United States: 101–8. doi:10.1038/ng.736.

Baker, Dean A, Tony Nolan, Bettina Fischer, Alex Pinder, Andrea Crisanti, and Steven Russell. 2011. “A Comprehensive Gene Expression Atlas of Sex- and Tissue-Specificity in the Malaria Vector, Anopheles Gambiae.” *BMC Genomics* 12: 296. doi:10.1186/1471-2164-12-296.

Behura, Susanta K., Consuelo Gomez-Machorro, Brent W. Harker, Becky deBruyn, Diane D. Lovin, Ryan R. Hemme, Akio Mori, Jeanne Romero-Severson, and David W. Severson. 2011. “Global Cross-Talk of Genes of the Mosquito Aedes Aegypti in Response to Dengue Virus Infection.” *PLoS Neglected Tropical Diseases* 5 (11): e1385. doi:10.1371/journal.pntd.0001385.

Bian, Guowu, Deepak Joshi, Yuemei Dong, Peng Lu, Guoli Zhou, Xiaoling Pan, Yao Xu, George Dimopoulos, and Zhiyong Xi. 2013. “Wolbachia Invades Anopheles Stephensi Populations and Induces Refractoriness to Plasmodium Infection.” *Science (New York, N.Y.)* 340: 748–51. doi:10.1126/science.1236192.

Biedler, James K., Wanqi Hu, Hongseok Tae, and Zhijian Tu. 2012. “Identification of Early Zygotic Genes in the Yellow Fever Mosquito Aedes Aegypti and Discovery of a Motif Involved in Early Zygotic Genome Activation.” *PLoS ONE* 7 (3): e33933. doi:10.1371/journal.pone.0033933.

Bonizzoni, Mariangela, W Augustine Dunn, Corey L Campbell, Ken E Olson, Michelle T Dimon, Osvaldo Marinotti, and Anthony a James. 2011. “RNA-Seq Analyses of Blood-Induced Changes in Gene Expression in the Mosquito Vector Species, Aedes Aegypti.” *BMC Genomics* 12 (1): 82. doi:10.1186/1471-2164-12-82.

Brown, A E, L Bugeon, A Crisanti, and F Catteruccia. 2003. “Stable and Heritable Gene Silencing in the Malaria Vector Anopheles Stephensi.” *Nucleic Acids Res* 31: e85. doi:10.1093/nar/gng085.

Campbell, C. L., T. Harrison, A. M. Hess, and G. D. Ebel. 2014. “MicroRNA Levels Are Modulated in Aedes Aegypti after Exposure to Dengue-2.” *Insect Molecular Biology* 23 (1):

132–39. doi:10.1111/imb.12070.

Catteruccia, F, T Nolan, T G Loukeris, C Blass, C Savakis, F C Kafatos, and a Crisanti. 2000. “Stable Germline Transformation of the Malaria Mosquito Anopheles Stephensi.” *Nature* 405 (6789): 959–62. doi:10.1038/35016096.

Chen, Bin, Yu-Juan Zhang, Zhengbo He, Wanshun Li, Fengling Si, Yao Tang, Qiyi He, et al. 2014. “De Novo Transcriptome Sequencing and Sequence Analysis of the Malaria Vector Anopheles Sinensis (Diptera: Culicidae).” *Parasites & Vectors* 7 (1): 314. doi:10.1186/1756-3305-7-314.

Chen, Haifeng, Jianxin Wang, Ping Liang, Monica Karsay-Klein, Anthony a James, Daniel Brazeau, and Guiyun Yan. 2004. “Microarray Analysis for Identification of Plasmodium-Refractoriness Candidate Genes in Mosquitoes.” *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada* 47 (6): 1061–70. doi:10.1139/g04-056.

Chen, Xiao-Guang, Xuanting Jiang, Jinbao Gu, Meng Xu, Yang Wu, Yuhua Deng, Chi Zhang, et al. 2015. “Genome Sequence of the Asian Tiger Mosquito, *Aedes Albopictus* , Reveals Insights into Its Biology, Genetics, and Evolution.” *Proceedings of the National Academy of Sciences*, 201516410. doi:10.1073/pnas.1516410112.

Crawford, Jacob E., Wamdaogo M. Guelbeogo, Antoine Sanou, Alphonse Traoré, Kenneth D. Vernick, N’Fale Sagnon, and Brian P. Lazzaro. 2010. “De Novo Transcriptome Sequencing in Anopheles Funestus Using Illumina RNA-Seq Technology.” *PLoS ONE* 5 (12): e14202. doi:10.1371/journal.pone.0014202.

Dissanayake, S, J Ribeiro, M Wang, W Dunn, G Yan, A James, and O Marinotti. 2010. “aeGEPUCI: A Database of Gene Expression in the Dengue Vector Mosquito, Aedes Aegypti.” *BMC Research Notes* 3: 248. doi:10.1186/1756-0500-3-248.

Dissanayake, Sumudu N, Osvaldo Marinotti, Jose Marcos C Ribeiro, and Anthony a James. 2006. “angaGEDUCI: Anopheles Gambiae Gene Expression Database with Integrated Comparative Algorithms for Identifying Conserved DNA Motifs in Promoter Sequences.” *BMC Genomics* 7: 116. doi:10.1186/1471-2164-7-116.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009.

"Real-Time DNA Sequencing from Single Polymerase Molecules." *Science (New York, N.Y.)* 323 (5910): 133–38. doi:10.1126/science.1162986.

Etebari, Kayvan, Solomon Osei-Amo, Simon Phillip Blomberg, and Sassan Asgari. 2015. "Dengue Virus Infection Alters Post-Transcriptional Modification of microRNAs in the Mosquito Vector Aedes Aegypti." *Scientific Reports* 5 (October): 15968. doi:10.1038/srep15968.

Faulde, Michael K., Leopoldo M. Rueda, and Bouh A. Khaireh. 2014. "First Record of the Asian Malaria Vector Anopheles Stephensi and Its Possible Role in the Resurgence of Malaria in Djibouti, Horn of Africa." *Acta Tropica* 139: 39–43. doi:10.1016/j.actatropica.2014.06.016.

Fontaine, Michael C, James B Pease, Aaron Steele, Robert M Waterhouse, Daniel E Neafsey, Igor V Sharakhov, Xiaofang Jiang, et al. 2014. "Extensive Introgression in a Malaria Vector Species Complex Revealed by Phylogenomics." *Science* 347 (November): science.1258524 – . doi:10.1126/science.1258524.

Gakhar, S K, R Sharma, and A Sharma. 2013. "Population Genetic Structure of Malaria Vector Anopheles Stephensi Liston (Diptera: Culicidae)." *Indian J Exp Biol* 51 (4): 273–79.

Gantz, Valentino M, Nijole Jasinskiene, Olga Tatarenkova, Aniko Fazekas, Vanessa M Macias, Ethan Bier, and Anthony A James. 2015. "Highly Efficient Cas9-Mediated Gene Drive for Population Modification of the Malaria Vector Mosquito \textit{Anopheles Stephensi}." *PNAS* 112 (49): E6736–43. doi:10.1073/pnas.1521077112.

Garver, Lindsey S, Yuemei Dong, and George Dimopoulos. 2009. "Caspar Controls Resistance to Plasmodium Falciparum in Diverse Anopheline Species." *PLoS Pathogens* 5 (3): e1000335. doi:10.1371/journal.ppat.1000335.

Goltsev, Yury, Gustavo L. Rezende, Karen Vranizan, Greg Lanzaro, Denise Valle, and Michael Levine. 2009. "Developmental and Evolutionary Basis for Drought Tolerance of the Anopheles Gambiae Embryo." *Developmental Biology* 330 (2): 462–70. doi:10.1016/j.ydbio.2009.02.038.

Gregory, Richard, Alistair C. Darby, Helen Irving, Mamadou B. Coulibaly, Margaret Hughes, Lizette L. Koekemoer, Maureen Coetzee, et al. 2011. "A De Novo Expression Profiling of Anopheles Funestus, Malaria Vector in Africa, Using 454 Pyrosequencing." *PLoS ONE* 6 (2):

e17418. doi:10.1371/journal.pone.0017418.

Han, Yixing, Shouguo Gao, Kathrin Muegge, Wei Zhang, and Bing Zhou. 2015. “Advanced Applications of RNA Sequencing and Challenges.” *Bioinformatics and Biology Insights* 9: 29–46. doi:10.4137/BBI.S28991.

Hess, Ann M, Abhishek N Prasad, Andrey Ptitsyn, Gregory D Ebel, Ken E Olson, Catalin Barbacioru, Cinna Monighetti, and Corey L Campbell. 2011. “Small RNA Profiling of Dengue Virus-Mosquito Interactions Implicates the PIWI RNA Pathway in Anti-Viral Defense.” *BMC Microbiology* 11 (1): 45. doi:10.1186/1471-2180-11-45.

Holt, Robert A., G Mani Subramanian, Aaron Halpern, Granger G Sutton, Rosane Charlab, Deborah R Nusskern, Patrick Wincker, et al. 2002. “The Genome Sequence of the Malaria Mosquito Anopheles Gambiae.” *Science* 298 (5591): 129–49. doi:10.1126/science.1076181.

Isaacs, A. T., N. Jasinskiene, M. Tretiakov, I. Thiery, A. Zettor, C. Bourgouin, and A. A. James. 2012. “Transgenic Anopheles Stephensi Coexpressing Single-Chain Antibodies Resist Plasmodium Falciparum Development.” *Proceedings of the National Academy of Sciences* 109 (28): E1922–30. doi:10.1073/pnas.1207738109.

Jenkins, A. M., R. M. Waterhouse, A. S. Kopin, and M. A. T. Muskavitch. 2014. “Long Non-Coding RNA Discovery in Anopheles Gambiae Using Deep RNA Sequencing.” *bioRxiv*. doi:10.1101/007484.

Jenkins, Adam M, Robert M Waterhouse, and Marc At Muskavitch. 2015. “Long Non-Coding RNA Discovery across the Genus Anopheles Reveals Conserved Secondary Structures within and beyond the Gambiae Complex.” *BMC Genomics* 16 (1): 337. doi:10.1186/s12864-015-1507-3.

Jiang, Xiaofang, James K Biedler, Yumin Qi, Andrew Brantley Hall, and Zhijian Tu. 2015. “Complete Dosage Compensation in Anopheles Stephensi and the Evolution of Sex-Biased Genes in Mosquitoes.” *Genome Biology and Evolution* 7 (7): 1914–24. doi:10.1093/gbe/evv115.

Jiang, Xiaofang, Ashley Peery, A Brantley Hall, Atashi Sharma, Xiao-Guang Chen, Robert M Waterhouse, Aleksey Komissarov, et al. 2014. “Genome Analysis of a Major Urban Malaria Vector Mosquito, Anopheles Stephensi.” *Genome Biology* 15 (9): 459. doi:10.1186/s13059-

014-0459-2.

Kim, Kristi E, Paul Peluso, Primo Babayan, P. Jane Yeadon, Charles Yu, William W Fisher, Chen-Shan Chin, et al. 2014. "Long-Read, Whole-Genome Shotgun Sequence Data for Five Model Organisms." *Scientific Data* 1: 140045. doi:10.1038/sdata.2014.45.

Land, Kirkwood M. 2003. "The Mosquito Genome: Perspectives and Possibilities." *Trends in Parasitology* 19 (3): 103–5. doi:10.1016/S1471-4922(03)00021-7.

Lim, Junghwa, D. Channe Gowda, Gowdahalli Krishnegowda, and Shirley Luckhart. 2005. "Induction of Nitric Oxide Synthase in Anopheles Stephensi by Plasmodium Falciparum: Mechanism of Signaling and the Role of Parasite Glycosylphosphatidylinositols." *Infection and Immunity* 73 (5): 2778–89. doi:10.1128/IAI.73.5.2778-2789.2005.

Liu, Yanxia, Yanhe Zhou, Jinya Wu, Peiming Zheng, Yiji Li, Xiaoying Zheng, Santhosh Puthiyakunnon, Zhijian Tu, and Xiao-Guang Chen. 2015. "The Expression Profile of Aedes Albopictus miRNAs Is Altered by Dengue Virus Serotype-2 Infection." *Cell & Bioscience* 5 (1): 16. doi:10.1186/s13578-015-0009-y.

Luckhart, S, C Giulivi, A L Drexler, Y Antonova-Koch, D Sakaguchi, E Napoli, S Wong, et al. 2013. "Sustained Activation of Akt Elicits Mitochondrial Dysfunction to Block Plasmodium Falciparum Infection in the Mosquito Host." *PLoS Pathogens* 9 (2): e1003180. doi:10.1371/journal.ppat.1003180.

Luckhart, S, Y Vodovotz, L Cui, and R Rosenberg. 1998. "The Mosquito Anopheles Stephensi Limits Malaria Parasite Development with Inducible Synthesis of Nitric Oxide." *Proceedings of the National Academy of Sciences of the United States of America* 95 (10): 5700–5705. doi:10.1073/pnas.95.10.5700.

Marinotti, O., E. Calvo, Q. K. Nguyen, S. Dissanayake, J. M C Ribeiro, and A. A. James. 2006. "Genome-Wide Analysis of Gene Expression in Adult Anopheles Gambiae." *Insect Molecular Biology* 15 (1): 1–12. doi:10.1111/j.1365-2583.2006.00610.x.

Marinotti, Osvaldo, Gustavo C. Cerqueira, Luiz Gonzaga Paula De Almeida, Maria Inês Tiraboschi Ferro, Elgion Lucio Da Silva Loreto, Arnaldo Zaha, Santuza M R Teixeira, et al. 2013. "The Genome of Anopheles Darlingi, the Main Neotropical Malaria Vector." *Nucleic Acids Research* 41 (15): 7387–7400. doi:10.1093/nar/gkt484.

McBride, Carolyn S., Felix Baier, Aman B. Omondi, Sarabeth A. Spitzer, Joel Lutomiah, Rosemary Sang, Rickard Ignell, and Leslie B. Vosshall. 2014. “Evolution of Mosquito Preference for Humans Linked to an Odorant Receptor.” *Nature* 515 (7526). Nature Publishing Group: 222–27. doi:10.1038/nature13964.

Meister, Stephan, Stefan M Kanzok, Xue-Li Zheng, Coralia Luna, Tong-Ruei Li, Ngo T Hoa, John Randall Clayton, et al. 2005. “Immune Signaling Pathways Regulating Bacterial and Malaria Parasite Infection of the Mosquito Anopheles Gambiae.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (32): 11420–25. doi:10.1073/pnas.0504950102.

Mewes, H W, K Albermann, M Bähr, D Frishman, a Gleissner, J Hani, K Heumann, et al. 1997. “Overview of the Yeast Genome.” *Nature* 387: 7–65. doi:10.1038/42755.

Mongin, Emmanuel, Christos Louis, Robert A. Holt, Ewan Birney, and Frank H. Collins. 2004. “The Anopheles Gambiae Genome: An Update.” *Trends in Parasitology* 20 (2): 49–52. doi:10.1016/j.pt.2003.11.003.

Morazzani, Elaine M, Michael R Wiley, Marta G Murreddu, Zach N Adelman, and Kevin M Myles. 2012. “Production of Virus-Derived Ping-Pong-Dependent piRNA-like Small RNAs in the Mosquito Soma.” *PLoS Pathogens* 8 (1): e1002470. doi:10.1371/journal.ppat.1002470.

Munroe, David J, and Timothy J R Harris. 2010. “Third-Generation Sequencing Fireworks at Marco Island.” *Nat Biotech* 28 (5). Nature Publishing Group: 426–28. http://dx.doi.org/10.1038/nbt0510-426.

Neafsey, D. E., G. K. Christophides, F. H. Collins, S. J. Emrich, M. C. Fontaine, W. Gelbart, M. W. Hahn, et al. 2013. “The Evolution of the Anopheles 16 Genomes Project.” *Genes|Genomes|Genetics* 3 (7): 1191–94. doi:10.1534/g3.113.006247.

Neafsey, D. E., R. M. Waterhouse, M. R. Abai, S. S. Aganezov, M. A. Alekseyev, J. E. Allen, J. Amon, et al. 2014. “Highly Evolvable Malaria Vectors: The Genomes of 16 Anopheles Mosquitoes.” *Science* 347 (6217): 1258522 – . doi:10.1126/science.1258522.

Nene, Vishvanath, Jennifer R Wortman, Daniel Lawson, Brian Haas, Chinnappa Kodira, Zhijian Jake Tu, Brendan Loftus, et al. 2007. “Genome Sequence of Aedes Aegypti, a Major Arbovirus Vector.” *Science (New York, N.Y.)* 316 (5832): 1718–23.

doi:10.1126/science.1138878.

Nolan, T., T. M. Bower, A. E. Brown, A. Crisanti, and F. Catteruccia. 2002. "piggyBac-Mediated Germline Transformation of the Malaria Mosquito Anopheles Stephensi Using the Red Fluorescent Protein dsRED as a Selectable Marker." *Journal of Biological Chemistry* 277 (11): 8759–62. doi:10.1074/jbc.C100766200.

O'Brochta, David A, Robert T Alford, Kristina L Pilitt, Channa U Aluvihare, and Robert A Harrell. 2011. "piggyBac Transposon Remobilization and Enhancer Detection in Anopheles Mosquitoes." *Proceedings of the National Academy of Sciences of the United States of America* 108 (39): 16339–44. doi:10.1073/pnas.1110628108.

Pakpour, Nazzy, Vanessa Corby-Harris, Gabriel P. Green, Hannah M. Smithers, Kong W. Cheung, Michael A. Riehle, and Shirley Luckhart. 2012. "Ingested Human Insulin Inhibits the Mosquito NF-κB-Dependent Immune Response to Plasmodium Falciparum." *Infection and Immunity* 80 (6): 2141–49. doi:10.1128/IAI.00024-12.

Pietri, Jose E, Rashaun Potts, and Eddie Pietri. 2014. "Plasmodium Falciparum Suppresses the Host Immune Response by Inducing Insulin- like Peptide Synthesis in the Mosquito Anopheles Stephensi." *Pending* 53 (1). Elsevier Ltd: 134–44. doi:10.1016/j.dci.2015.06.012.

Poelchau, Monica F, Julie A Reynolds, David L Denlinger, Christine G Elsik, and Peter A Armbruster. 2011. "A de Novo Transcriptome of the Asian Tiger Mosquito, Aedes Albopictus, to Identify Candidate Transcripts for Diapause Preparation." *BMC Genomics* 12 (1): 619. doi:10.1186/1471-2164-12-619.

Rinker, David C, R Jason Pitts, Xiaofan Zhou, Eunho Suh, Antonis Rokas, and Laurence J Zwiebel. 2013. "Blood Meal-Induced Changes to Antennal Transcriptome Profiles Reveal Shifts in Odor Sensitivities in Anopheles Gambiae." *Proceedings of the National Academy of Sciences of the United States of America* 110 (20): 8260–65. doi:10.1073/pnas.1302562110.

Rinker, David C, Xiaofan Zhou, Ronald Jason Pitts, Antonis Rokas, and Laurence J Zwiebel. 2013. "Antennal Transcriptome Profiles of Anopheline Mosquitoes Reveal Human Host Olfactory Specialization in Anopheles Gambiae." *BMC Genomics* 14 (1): 749. doi:10.1186/1471-2164-14-749.

Rund, S S C, T Y Hou, S M Ward, F H Collins, and G E Duffield. 2011. "Genome-Wide Profiling of Diel and Circadian Gene Expression in the Malaria Vector Anopheles Gambiae." *Proceedings of the National Academy of Sciences of the United States of America* 108 (32): E421–30. doi:10.1073/pnas.1100584108.

Sanapala, Shilpa, Jieh Juen Yu, Ashlesh K. Murthy, Weidang Li, M. Neal Guentzel, James P. Chambers, Karl E. Klose, and Bernard P. Arulanandam. 2012. "Perforin- and Granzyme-Mediated Cytotoxic Effector Functions Are Essential for Protection against Francisella Tularensis Following Vaccination by the Defined F. Tularensis Subsp. Novicida ??fopC Vaccine Strain." *Infection and Immunity* 80 (6): 2177–85. doi:10.1128/IAI.00024-12.

Sanders, Heather R., Amy M. Evans, Linda S. Ross, and Sarjeet S. Gill. 2003. "Blood Meal Induces Global Changes in Midgut Gene Expression in the Disease Vector, Aedes Aegypti." *Insect Biochemistry and Molecular Biology* 33 (11): 1105–22. doi:10.1016/S0965-1748(03)00124-3.

Sanger, F, S Nicklen, and A R Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67. doi:10.1073/pnas.74.12.5463.

Schulze, Almut, and Julian Downward. 2001. "Navigating Gene Expression Using Microarrays — a Technology Review." *Nature Cell Biology* 3 (8): E190–95. doi:10.1038/35087138.

Scott, Jaclyn C., Doug E. Brackney, Corey L. Campbell, Virginie Bondu-Hawkins, Brian Hjelle, Greg D. Ebel, Ken E. Olson, and Carol D. Blair. 2010. "Comparison of Dengue Virus Type 2-Specific Small RNAs from RNA Interference-Competent and –Incompetent Mosquito Cells." *PLoS Neglected Tropical Diseases* 4 (10): e848. doi:10.1371/journal.pntd.0000848.

Severson, David W., and Susanta K. Behura. 2012. "Mosquito Genomics: Progress and Challenges." *Annual Review of Entomology* 57 (1): 143–66. doi:10.1146/annurev-ento-120710-100651.

Sharma, V. P. 1999. "Current Scenario of Malaria in India." *Parassitologia* 41 (1-3): 349–53.

Sinka, Marianne E, Michael J Bangs, Sylvie Manguin, Theeraphap Chareonviriyaphap, Anand P Patil, William H Temperley, Peter W Gething, et al. 2011. "The Dominant Anopheles Vectors of Human Malaria in the Asia-Pacific Region: Occurrence Data, Distribution Maps and Bionomic Précis." *Parasites & Vectors* 4 (1). BioMed Central Ltd: 89. doi:10.1186/1756-

3305-4-89.

Smidler, Andrea L., Olivier Terenzi, Julien Soichot, Elena a. Levashina, and Eric Marois. 2013. "Targeted Mutagenesis in the Malaria Mosquito Using TALE Nucleases." *PLoS ONE* 8 (8): 1–9. doi:10.1371/journal.pone.0074511.

Suen, Garret, Clotilde Teiling, Lewyn Li, Carson Holt, Ehab Abouheif, Erich Bornberg-Bauer, Pascal Bouffard, et al. 2011. "The Genome Sequence of the Leaf-Cutter Ant Atta Cephalotes Reveals Insights into Its Obligate Symbiotic Lifestyle." *PLoS Genetics* 7 (2). United States: e1002007. doi:10.1371/journal.pgen.1002007.

Travers, Kevin J, Chen-Shan Chin, David R Rank, John S Eid, and Stephen W Turner. 2010. " A Flexible and Efficient Template Format for Circular Consensus Sequencing and SNP Detection." *Nucleic Acids Research* 38 (15): e159. doi:10.1093/nar/gkq543.

VanBuren, Robert, Doug Bryant, Patrick P Edger, Haibao Tang, Diane Burgess, Dinakar Challabathula, Kristi Spittle, et al. 2015. "Single-Molecule Sequencing of the Desiccation-Tolerant Grass Oropetium Thomaeum." *Nature* 527 (7579). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 508–11. http://dx.doi.org/10.1038/nature15714.

Vannini, Laura, W. Augustine Dunn, Tyler W. Reed, and Judith H. Willis. 2014. "Changes in Transcript Abundance for Cuticular Proteins and Other Genes Three Hours after a Blood Meal in Anopheles Gambiae." *Insect Biochemistry and Molecular Biology* 44 (1): 33–43. doi:10.1016/j.ibmb.2013.11.002.

Velasco, Riccardo, Andrey Zharkikh, Jason Affourtit, Amit Dhingra, Alessandro Cestaro, Ananth Kalyanaraman, Paolo Fontana, et al. 2010. "The Genome of the Domesticated Apple (Malus X Domestica Borkh.)." *Nature Genetics* 42 (10). United States: 833–39. doi:10.1038/ng.654.

Venter, J C, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, et al. 2001. "The Sequence of the Human Genome." *Science (New York, N.Y.)* 291 (5507): 1304–51. doi:10.1126/science.1058040.

Vontas, J., C. Blass, A. C. Koutsos, J. -P. David, F. C. Kafatos, C. Louis, J. Hemingway, G. K. Christophides, and H. Ranson. 2005. "Gene Expression in Insecticide Resistant and Susceptible *Anopheles Gambiae* Strains Constitutively or after Insecticide Exposure." *Insect*

Molecular Biology 14 (5): 509–21. doi:10.1111/j.1365-2583.2005.00582.x.

Wang, Mei-Hui, Osvaldo Marinotti, Anne Vardo-Zalik, Rajni Boparai, and Guiyun Yan. 2011. “Genome-Wide Transcriptional Analysis of Genes Associated with Acute Desiccation Stress in Anopheles Gambiae.” *PLoS ONE* 6 (10): e26011. doi:10.1371/journal.pone.0026011.

Waterston, Robert H, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F Abril, Pankaj Agarwal, Richa Agarwala, et al. 2002. “Initial Sequencing and Comparative Analysis of the Mouse Genome.” *Nature* 420 (6915): 520–62. doi:10.1038/nature01262.

Weber, James L., and Eugene W. Myers. 1997. “Human Whole-Genome Shotgun Sequencing.” *Genome Research* 7: 401–9. doi:10.1101/gr.7.5.401.

White, Michael T, Lesong Conteh, Richard Cibulskis, and Azra C Ghani. 2011. “Costs and Cost-Effectiveness of Malaria Control Interventions - a Systematic Review.” *Malaria Journal* 10 (1): 337. doi:10.1186/1475-2875-10-337.

Xu, X, Y Dong, E G Abraham, A Kocan, P Srinivasan, A K Ghosh, R E Sinden, et al. 2005. “Transcriptome Analysis of Anopheles Stephensi-Plasmodium Berghei Interactions.” *Mol Biochem Parasitol* 142 (1): 76–87. doi:10.1016/j.molbiopara.2005.02.013.

Zhou, Dan, Donghui Zhang, Guohui Ding, Linna Shi, Qing Hou, Yuting Ye, Yang Xu, et al. 2014. “Genome Sequence of Anopheles Sinensis Provides Insight into Genetics Basis of Mosquito Competence for Malaria Parasites.” *BMC Genomics* 15: 42. doi:10.1186/1471-2164-15-42.

Chapter 2: Genome analysis of a major urban malaria vector mosquito, *Anopheles stephensi*

Xiaofang Jiang[1,2]*, Ashley Peery[3]*, A. Brantley Hall[1,2], Atashi Sharma[3], Xiao-Guang Chen[4], Robert M. Waterhouse[5,6,7,8], Aleksey Komissarov[9], Michelle M. Riehl[10], Yogesh Shouche[11], Maria V. Sharakhova[3], Dan Lawson[12], , Nazzy Pakpour[13], Peter Arensburger[14], Victoria L. M. Davidson[15], Karin Eiglmeier[16], Scott Emrich[17], Phillip George[3], Ryan C. Kennedy[18], Shrinivasrao P. Mane[19], Gareth Maslen[12], Chioma Oringanje[20], Yumin Qi[2], Robert Settlage[19], Marta Tojo[21], Jose M. C. Tubio[22], Maria F. Unger[23], Bo Wang[13], Kenneth D. Vernick[16], Jose M. C. Ribeiro[24], Anthony A. James[25], Kristin Michel15, Michael A. Riehle[20], Shirley Luckhart[13], Igor V. Sharakhov[1,3§], Zhijian Tu[1,2§]

[1]Program of Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA, USA

[2]Department of Biochemistry, Virginia Tech, Blacksburg, VA, USA

[3]Department of Entomology, Virginia Tech, Blacksburg, VA, USA

[4]Department of Pathogen Biology, Southern Medical University, Guangzhou, Guangdong, China

[5]Department of Genetic Medicine and Development, University of Geneva Medical School, rue Michel-Servet 1, 1211 Geneva, Switzerland

[6]Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva, Switzerland

[7]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA, USA

[8]The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA, USA

[9]Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, and Institute of Cytology Russian Academy of Sciences, St. Petersburg, Russia

[10]Department of Microbiology, University of Minnesota, Minneapolis, MN, USA

[11]National Center for Cell Science, Pune University Campus, Ganeshkhind, Pune, India

[12]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

[13]Department of Medical Microbiology and Immunology, University of California, Davis, CA, USA

[14]Biological Sciences Department, California State Polytechnic University Pomona, CA, USA

[15]Division of Biology, Kansas State University, Manhattan, KS, USA

[16]Department of Parasitology and Mycology, Unit of Insect Vector Genetics and Genomics, Institut Pasteur, Paris, France and CNRS Unit of Hosts, Vectors and Pathogens (URA3012), Paris, France.

[17]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA

[18]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, California, USA

[19]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA

[20]Department of Entomology, University of Arizona, Tucson, AZ, USA

[21]Department of Physiology, School of medicine – CIMUS, Instituto de Investigaciones Sanitarias, University of Santiago de Compostela, Spain

[22]Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK.

[23]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA

[24]Section of Vector Biology, Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, Rockville, MD, USA

[25]Departments of Microbiology & Molecular Genetics and Molecular Biology & Biochemistry , University of California, Irvine CA, USA

* Equal contribution, listed by alphabetical order

§ Corresponding authors

Email addresses:

ZT: jaketu@vt.edu

IVS: igor@vt.edu

Email addresses of other authors are provided during online submission.

2.1 Author contributions

Conceived and designed experiments: ZT and IVS; Data generation, analysis and presentation: XJ, AP, AS, ABH, MK, MVS, AK, BW, CO, DL, KE, KM, JMCT, JMCR, MAR, MRR, MU, NP, PA, PG, RK, RS, RMW, SL, SM, VLMD, YQ, ZT; Writing of the manuscript: XJ, ABH, AAJ, AP, AS, JMCR, KDV, KM, KP, MK, MAR, MMR, SL, IVS, and ZT; Provided resources and tools and critical reviewed manuscript: XC, YS

2.2 Abstract

2.2.1 Background

Anopheles stephensi is the key vector of malaria throughout the Indian subcontinent and Middle East and an emerging model for molecular and genetic studies of mosquito-parasite interactions. The “type” form of the species is responsible for the majority of urban malaria transmission across its range.

2.2.2 Results

Here we report the genome sequence and annotation of the Indian strain of the “type” form of *An. stephensi*. The 221 Mb genome assembly represents >92% of the entire genome and was produced using a combination of 454, Illumina, and PacBio sequencing. Physical mapping assigned 62% of the genome onto chromosomes, enabling chromosome-based analysis. Comparisons between An. stephensi and An. gambiae revealed that the rate of gene order reshuffling on the X chromosome was three times higher than that on the autosomes. *An. stephensi* has more heterochromatin in pericentric regions but less repetitive DNA in chromosome arms than *An. gambiae*. We also identified a number of Y-chromosome contigs and BACs. Interspersed repeats constitute 7.1% of the assembled genome while LTR retrotransposons alone comprise >49% of the Y contigs. RNA-seq analyses provide new insights into mosquito innate immunity, development, and sexual dimorphism.

2.2.3 Conclusions

We provide a genome resource and platform for fundamental and translational research of a major urban malaria vector. Chromosome-based investigations provide unique perspectives on *Anopheles* chromosome evolution. RNA-seq analysis and studies of immunity genes offer new insights into mosquito biology and mosquito-parasite interactions.

2.3 Background

Mosquitoes in the genus *Anopheles* are the primary vectors of human malaria parasites and the resulting disease is one of the most deadly and costly in history [1, 2]. Publication and availability of the *Anopheles gambiae* genome sequence accelerated research that has not only enhanced our basic understanding of vector genetics, behavior, and physiology and roles in transmission, but also contributed to new strategies for combating malaria [3]. Recent application of next-generation sequencing technologies to mosquito genomics offers exciting opportunities to expand our understanding of mosquito biology in many important vector species and harness the power of comparative genomics. *Anopheles stephensi* is among the ~60 species considered important in malaria transmission and is the key vector of urban malaria on the Indian subcontinent and the Middle East [4, 5]. The fact that a recent resurgence of human malaria in Africa could have been caused by the sudden appearance of *An. stephensi* indicates that *An. stephensi* may pose an even greater risk to human health in the future [6]. Of the three forms: type, *mysorensis*, and intermediate, the type formis responsible for the majority, if not all, of urban malaria transmission across its range and accounts for approximately 12% of all transmission in India [7]. Thus efforts to control it can be expected to contribute significantly to the malaria eradication agenda [8, 9]. *An. stephensi* is amenable to genetic manipulations such as transposon-based germline transformation [10], genome-wide mutagenesis [11], site-specific integration [12], genome-editing [13] and RNAi-based functional genomics analysis [14]. Our understanding of the interactions between *An. stephensi* and the malaria parasites is rapidly improving [15-20]. Thus *An. stephensi* is emerging as a model species for genetic and molecular studies. We report the genome sequence of the Indian strain of the "type" form of *An. stephensi* as a resource and platform for fundamental and translational research. We also provide unique perspectives on *Anopheles* chromosome evolution and offer new insights into mosquito biology and mosquito-parasite interactions.

2.4 Results and discussion

2.4.1 Draft genome sequence of *An. stephensi*: Assembly and verification

The *An. stephensi* genome was sequenced using 454 GS FLX, Illumina HiSeq, and PacBio RS technologies (Additional file 1: Table S1). The 454 reads comprised 19.4x coverage: 12.2x from single-end reads, 2.2x from 3 kilobase (kb) paired-end reads, 3.4x from 8 kb paired-end reads, and 1.7x from 20 kb paired-end reads. The majority of 454 reads ranged from 194 to 395 base-pairs (bp) in length. A single lane of Illumina sequencing of male genomic DNA resulted in 86.4x coverage of 101 bp paired-end reads with an average insert size of ~200 bp. Ten cells of PacBio RS sequencing of male genomic DNA produced 5.2x coverage with a median length of 1,295 bp. A hybrid assembly combining 454 and Illumina data produced a better overall result than using 454 data alone (Materials and methods). The resulting assembly was further improved by filling gaps with error-corrected PacBio reads and scaffolding with BAC-ends. The current assembly, verified using various methods, contains 23,371 scaffolds spanning 221 Mb. The assembly includes 11.8 Mb (5.3%) of gaps filled with Ns (Table 2.1), which is slightly lower than the size of gaps in the *An. gambiae* assembly (20.7Mb, 7.6%). The N50 scaffold size is 1.59 Mb and the longest scaffold is 5.9 Mb. The number of scaffolds is inflated because we choose to set the minimum scaffold length to 500 bp to include repeat-rich short scaffolds. The assembled size of 221 Mb is consistent with the previous estimate of the *An. stephensi* genome size of ~235 Mb [21].

2.4.2 Physical mapping

Mapping of 227 probes was sufficient to assign 86 scaffolds to unique positions on the *An. stephensi* polytene chromosomes (;Table 2.2; Additional file 2). These 86 scaffolds comprise

137.14 Mb or 62% of the assembled genome. Our physical map includes 28 of the 30 largest scaffolds and we were able to determine the orientation of 32 of the 86 scaffolds. We expect that relatively little of the heterochromatin was captured in our chromosomal assembly based on the morphology of the chromosomes in regions to which the scaffolds mapped. For this reason, subsequent comparisons with *An. gambiae* on molecular features of the genome landscape exclude regions of known heterochromatin from the *An. gambiae* dataset. *An. stephensi* and *An. gambiae* have different chromosome arm associations with 2L of *An. gambiae* homologous to 3L of *An. stephensi* [22]. Therefore, all ensuing discussion of synteny between the two species refers to *An. stephensi* chromosome arms listed in homologous order to those of *An. gambiae*: X, 2R, 3L, 3R, and 2L. While draft genomes also are available for *An. darlingi* and *An. sinensis* [23, 24], we focused our comparative analysis on *An. stephensi* and *An. gambiae,* the only two species that have chromosome-based assembly.

2.4.3 Gene annotation

A total of 11,789 protein-encoding genes were annotated using a combination of homology and *de novo* prediction. These gene models have been submitted to the NCBI (GCA_000300775.2) and are hosted in VectorBase (https://www.vectorbase.org/Anopheles_stephensiI/Info/Index). The average transcript length was 3,666 bp and the average number of exons per transcript was 4.18. Evolutionary relationships among *An. stephensi* and other *dipteran* insects were evaluated by constructing a maximum likelihood molecular species phylogeny using universal single-copy orthologs (Fig. 2A). *An. stephensi* and *An. gambiae* form a well-supported clade representing the subgenus *Cellia* within the genus *Anopheles*. This phylogeny provides the evolutionary context for current and future comparative genomics analysis. A total 10,492 (89.0%) of the 11,789 predicted *An. stephensi* protein-encoding genes had orthologs in *An. gambiae*, *Aedes aegypti* and *Drosophila melanogaster* (Figure 2.2).

2.4.4 Global Transcriptome Analysis

Eleven RNA-seq samples were prepared from 0-1, 2-4, 4-8, and 8-12 hour post-egg deposition embryos, larvae, pupae, adult males, adult females, non-blood-fed ovaries, blood-fed ovaries, and 24 hours post-blood-fed female carcasses without ovaries [25]. The corresponding genes were clustered into 20 distinct groups ranging in size from 8 to 2,106 genes per group on the basis of similar expression patterns (Figure 2.3). Many of the clusters correspond to either a specific developmental stage or sex (Additional files 3 and 4). A search for over-represented gene ontology (GO) terms in the 20 clusters found that many of the co-regulated genes have similar inferred functions or roles. Adult females require a protein-rich blood-meal for oogenesis and thus are the most interesting sex from a health perspective. Genes in clusters 1, 10, and 17 are induced in the female soma after blood-feeding. These clusters are enriched for genes encoding proteins with proteolytic activity, including serine peptidases, and involved in blood-meal digestion. Mosquitoes have undergone lineage-specific amplification of serine peptidases when compared to *Drosophila*, many of which are found in the three clusters described above. Cluster 9 contains 258 genes that showed peak expression in the pupal stage and it is enriched for genes whose products are involved in exoskeleton development. GO analyses of other clusters are described in the supplementary text. We identified 241 and 313 genes with female- or male-biased expression, respectively (Additional file 5). The male-biased genes are enriched for those whose products are involved in spermatogenesis and the auditory perception. Male mosquitoes detect potential mates using their Johnston's organ, which has twice the number of sensory neurons as that of the females [26, 27].

The female-biased genes are enriched for those whose products are involved in proteolysis and other metabolic processes likely relevant to blood digestion.

2.4.5 Immunity genes

Manual annotation was performed on genes involved in innate immunity including those that encode the LRR immune (LRIM) and the *Anopheles Plasmodium*-responsive leucine-rich repeat 1 (APL1) proteins, and the genes of the Toll, immune deficiency (IMD), insulin/insulin-like growth factor signalling (IIS), mitogen-activated protein kinase (MAPK) and TGF-β signalling pathways. A number of studies have demonstrated the importance of these genes or pathways in mosquito defense against parasites or viruses [16-20, 28-30]. Manual analysis showed overall agreement with the automated annotation and improved the gene models in some cases (Additional files 6 and 7). A high level of orthology is generally observed between *An. stephensi* and *An. gambiae* and we highlight here a few potentially interesting exceptions. *An. stephensi* may have only one APL1 gene (ASTEI02571) instead of the three APL1 gene cluster found in *An. gambiae* (Additional file 1: Figure S1). We also observed the apparent lack of TOLL1B and 5B sequences in *An. stephensi*, which in *An. gambiae* are recent duplications of TOLL1A and 5A, respectively.

Expression profiles of all immunity genes were analyzed using the 11 RNA-seq samples to provide insights into their biological functions (Additional file 8). For example, FKBP12, a protein known to regulate both transforming growth factor (TGF)-β and target of rapamycin (TOR) signalling, showed abundant transcript levels across immature stages and adult tissues (Additional file 1: Figure S2). The high expression levels of AsteFKBP12 in all examined stages and tissues were unexpected. Examination of existing publicly-available microarray data confirmed these expression levels and patterns [31]. FKBP12 in mammals forms a complex with rapamycin and FKBP-rapamycin-associated protein (FRAP) to inhibit TOR [32]. Given that TOR signalling is fundamental to many biological functions in mammals [33] and cumulative data support the same for *D. melanogaster* [34], a high level of FBKP12 expression may be critical for tight regulation of TOR activity in *An. stephensi* and perhaps *An. gambiae* [35]. Expression patterns of the *An. gambiae* FKBP12 ortholog, AGAP012184, from microarray datasets (http://funcgen.vectorbase.org/expression-browser/gene/AGAP012184) support the hypothesis that this protein is involved in a broad array of *Anopheline* physiologies including: development, bloodfeeding, molecular form-specific insecticide resistance, circadian rhythms, desiccation resistance, mating status, and possibly also broad regulation of infection based on studies with murine (*Plasmodium berghei*) and human (*Plasmodium falciparum*) malaria parasites. Whether these same physiologies and others are regulated by FKBP12 in *An. stephensi* will require experimental confirmation. Given that signalling pathways regulating embryonic pattern formation in *Drosophila* (e.g., the Toll pathway [36]) have been co-opted in the adult fly for regulation of various physiologies including metabolism and immune defense, the data presented here support the hypothesis that pathways integral to adult biology in adult *Anophelines* also have been similarly co-opted from important developmental roles.

2.4.6 Salivary genes

Saliva of blood feeding arthropods contains a cocktail of pharmacologically active components that disarm vertebrate host’s blood clotting and platelet aggregation, induce vasodilation and affect inflammation and immunity. These salivary proteins are under accelerated evolution due most likely to their host’s immune pressure. A previous salivary gland transcriptome study identified 37 corresponding salivary proteins in *An. stephensi*, most of which are shared with *An. gambiae*, including mosquito and *Anopheles*-specific protein families [37]. A more extensive sialotranscriptome based on ~3,000 EST’s identified the templates for 71 putative secreted proteins for *An. gambiae* [38]. The combined data verify the identity of 71 putative salivary secreted proteins for *An. stephensi,* seven of which have no similarities to *An. gambiae* proteins (Additional file 9). The current assembly of the *An. stephensi* genome shows that, many salivary gland genes are present as tandem repeated genes and represent families that arose by gene duplication events. Tandem repeated gene families often are poorly annotated by automated approaches, therefore, manual annotation was necessary to improve the salivary gland gene models (Additional files 10 and 11). In particular, *An. gambiae* has eight genes of the D7 family, which has modified odorant binding domains (OBD) that strongly bind agonists of platelet aggregation and vasoconstriction (histamine, serotonin, epinephrine and norepinephrine) [39].Three of these genes have two OBD's while the remaining five have only one domain each. As in *An. gambiae*, the short forms are oriented in tandem and in the opposite orientation of the long-form genes. However, *An. stephensi* has apparently collapsed the second long form to create a sixth short form.

2.4.7 Comparative analysis of additional gene families

Functional annotations of a number of gene families in *An. stephensi* were obtained based on their InterPro ID [40] (Additional file 12). We also compared gene numbers in these gene families across several species. *An. stephensi* and *An. gambiae* showed similar gene numbers in most of the gene families [3] and this is consistent with the close phylogenetic relationship between the two species. As observed with manually annotated immunity-related genes (Additional file 1: Figure S3), strong one-to-one relationship was observed between *An. stephensi* and *An. gambiae* genes in odorant binding proteins (OBPs) (Additional file 1: Figure S4A) and other gene families studied. There are a few gene families that showed obvious difference in numbers between *An. stephensi* and *An. gambiae*. We performed phylogenetic analysis of these gene families. The results (Additional file 1: Figure S4B and Figure S4C) indicate gene expansion in the odorant receptors

(OR) and fibrinogen-related proteins in *An. gambiae*. Interestingly, a plurality of expanded genes are physically clustered in *An. gambiae*, supporting the conclusion that the gene expansions may have arisen from local duplications. For example, the *An. stephensi* single-copy OR gene ASTEI08685 has four orthologs in *An. gambiae* (AGAP004354, AGAP004355, AGAP004356 and AGAP004357). The putative orthologs of these "expanded" genes tend to be single- or low-copy in *An. stephensi* and other related species in Vectorbase, supporting the interpretation that the lack of duplicated copies in *An. stephensi* is not due to assembly or annotation error. Further analysis that includes all species in the ongoing 16 *Anopheles* genomes project [41] will facilitate future comparative analysis of gene family expansions and gene losses.

2.4.8 Repeat content

Transposable elements (TEs) and other unclassified interspersed repeats constitute 7.1% of the assembled *An. stephensi* genome (Table 2.3; Additional file 13). TE occupancy of the euchromatic genome in *D. melanogaster* and *An. gambiae* is 2% and 16%, respectively [3]. Thus variations in the size of the genomes correlate with different amounts of repetitive DNA in these three species. More than 200 TEs have been annotated. DNA transposons and miniature inverted-repeat TEs (MITEs) comprise 0.44% of the genome. Non-LTR retrotransposons (or LINEs) comprise 2.36% of the genome. Short intersperse nuclear elements (SINEs), although less than 300 bp in length, are highly repetitive and comprise 1.7% of the genome. There is considerable diversity among the LTR-retrotransposons although they occupy only 0.7% of the genome. Approximately 2% of the genome consists of interspersed repeats that remain to be classified.

2.4.9 Genome landscape: a chromosomal arm perspective

The density of genes, TEs, and short tandem repeats (STRs) for each chromosome were determined based on the physical map (Figure 2.4). The average numbers of genes for each chromosome arm are consistent with those in *An. gambiae*. The X had the lowest number of genes per 100 kb, and the highest densities of genes per 100 kb were seen on 2R and 3L (Figure 2.5; Additional file 1: Tables S2 and S3). Chromosomes 2R and 3L also contain the greatest numbers of polymorphic inversions [42]. Genes functioning as drivers of adaptation could be expected to occur in greater densities on chromosome arms with higher numbers of polymorphic inversions [43].

An. stephensi has a lower density of transposable elements across all chromosome arms than *An. gambiae* (Figure 2.5; Additional file 1: Tables S2 and S3; Additional file 14). The density of transposable elements on the *An. stephensi* X is more than twice that of the autosomes. A comparison of the *An. stephensi* simple repeats with those in *An. gambiae* euchromatin showed that densities in the latter were ~2-2.5x higher (Figure 2.5; Additional file 1: Tables S2 and S3). The greatest densities of simple repeats were found on the X chromosome and this is consistent with a previous study in *An. gambiae* [44]. Although *An. stephensi* shows lower densities of simple repeats across all arms compared to *An. gambiae*, its X appears to harbor an overrepresentation of simple repeats compared to its autosomes. Scaffold/Matrix-associated regions (S/MARs) can potentially affect chromosome mobility in the cell nucleus and rearrangements during evolution [45, 46] and these were found to be enriched in the 2L and 3R arms (Figure 2.5; Additional file 1: Tables S2 and S3).

2.4.10 Molecular organization of pericentric heterochromatin

We observed clear differences in heterochromatin staining patterns when comparing mitotic chromosome squashes prepared from imaginal discs of An. gambiae and An. stephensi. An. stephensi appears to have more pericentric heterochromatin than An. gambiae (Additional file 1: Figure S5). This is particularly evident in the sex chromosomes. Mitotic X chromosomes in An.

stephensi possess much more pericentric heterochromatin compared with X chromosomes from several different strains of An. gambiae. Finally, the Y chromosome in An. stephensi has a large block of heterochromatin. We further investigated whether particular tandem repeats are concentrated in heterochromatin. Aste72A and Aste190A, the two repeats with highest coverage in raw genomic data reads, were selected as probes for FISH analysis (Additional file 15). Aste72A, which comprises approximately 1% of the raw genomic reads, was mapped to the pericentric heterochromatin of X and Y chromosomes (Figure 2.6). Aste190A, which comprises approximately 2% of the raw genomic reads, was mapped to centromere of both autosomes (Additional file 1: Figure S6). The Aste72A tandem repeat has a 26.7% mean GC-content and contributes significantly to the AT-rich peak in the plot of GC distribution of raw genomic reads (Additional file 1: Figure S7).

2.4.11 Y chromosome

Anopheles mosquitoes have heteromorphic sex-chromosomes where males are heterogametic (XY) and females homogametic (XX) [47]. The high repetitive DNA content of Y chromosomes makes them difficult to assemble and they often are ignored in genome projects. An approach called the chromosome quotient [48] was used to identify 57 putative Y sequences spanning 50,375 bp (Additional files 16 and 17). All of these sequences are less than 4,000 bp in length and appear to be highly repetitive. Five BACs that appeared to be Y-linked based on the CQs of their end sequences were analyzed by sequencing and their raw PacBio reads were assembled with the HGAP assembler [49]. Eleven contigs spanning 196,498 bp of predicted Y-linked sequences were obtained (Additional files 18 and 19). The 57 Y-linked sequences and 11 contigs from the Y-linked BACs represent currently the most abundant set of Y sequences in any *Anopheles* species. RepeatMasker analysis using the annotated *An. stephensi* interspersed repeats showed that ~65%

of the *An. stephensi* Y sequences are interspersed repeats. LTR retrotansposons alone occupy ~49% of the annotated Y (Additional files 20 and 21).

2.4.12 Synteny and gene order evolution

We used the chromosomal location and orientation of 6,448 one-to-one orthologs from *An. gambiae* and *An. stephensi* to examine synteny and estimate the number of chromosomal inversions between these two species (Figure 2.7; Additional file 22). Syntenic blocks were defined as those that had at least two genes and all genes within the block had the same order and orientation with respect to one another in both genomes. The X chromosome has markedly more inversions than the autosomes. The number of chromosomal inversions that might have happened since *An. stephensi* and *An. gambiae* last shared a common ancestor were determined with GRIMM [50]. We calculated the density of inversions per chromosome arm ignoring breakpoint reuse and assuming two breakpoints per inversion (Additional file 1: Tables S4 and S5). The length of *An. stephensi* assembly was used as a proxy for the size of the *An. stephensi* chromosomes. The density of inversions per megabase on the X chromosome supports the conclusion that it is much more prone to rearrangement than the autosomes. Genomic segments on the X are ~ 3-fold more likely to change order than those on the autosomes (Figure 2.8 A and Additional file 1: Table S6). The fast rate of X chromosome rearrangements contrasts with the lack of polymorphic inversions in *An. stephensi* and *An. gambiae* (Additional file 1: Table S5). Interestingly, a recent comparative genomic study between *An. gambiae* and *Ae. aegypti* revealed that the homomorphic sex-determining chromosome in *Ae. aegypti* has a higher rate of genome rearrangements than autosomes [51].

2.4.13 Rates of chromosome evolution in Drosophila and Anopheles

Recent studies have established that both *Anopheles* and *Drosophila* species have high rates of chromosomal evolution as compared with mammalian species [44, 52-59]. We compared the number of breaks per megabase for the X chromosome and autosomes chromosomes to understand the differences in the dynamics of chromosome evolution between *Drosophila* and *Anopheles* (Additional file 1: Table S7). These results reveal a higher ratio of the rates of evolution of sex chromosome to all chromosomes in *Anopheles* than *Drosophila*, with means of 2.116 and 1.197, respectively (Figure 2.8 B). We correlated densities of different molecular features including simple repeats, TEs, genes, and S/MARs with the rates of rearrangement calculated for each arm (Additional file 1: Tables S8-S13). The strongest correlations were found among the rates of evolution across all chromosome arms and the densities of microsatellites, minisatellites and satellites in both *An. gambiae* and *An. stephensi*. The highly-positive correlations between rates of inversion across all chromosome arms and satellites of different sizes are due most likely to the co-occurring abundance of satellites and inversions on the X chromosome. Rates of inversions and satellite densities are much lower on the autosomes. S/MARs in autosomes were correlated negatively and genes correlated positively with polymorphic inversions.

2.4.14 Genetic diversity of the genome

The genome sequencing effort reported in the current study is based on an inbred laboratory strain to ensure good assembly. Nonetheless, we performed genome-wide SNP analysis based on the available data. 530,997 SNPs were detected (Additional file 23). 319,751 SNPs were assigned to chromosomes based on mapping information (Additional file 1: Table S14). The SNP calls were assessed for their effect on the primary sequence of transcripts. Variant data and transcript effects are available on the Vectorbase website (www.vectorbase.org). These analyses will help future population genomic studies and facilitate association studies. We found that the X chromosome has a markedly lower frequency of SNPs than the autosomes in agreement with the similar observation in *An. gambiae* [3]. The observed pattern may be explained by a smaller effective population size of the X chromosome due to male hemizygosity and lower sequence coverage of the X chromosome [60].

2.5 Conclusions

The genome assembly of the type-form of the Indian strain of *An. stephensi* was produced using a combination of 454, Illumina, and PacBio sequencing and verified by analysis of BAC clones and ESTs. Physical mapping was in complete agreement with the genome assembly and resulted in a chromosome-based assembly that includes 62% of the genome. Such an assembly enabled analysis of chromosome arm-specific differences that are seldom feasible in next-gen genome projects.

Comparative analyses between *An. stephensi* and *An. gambiae* showed that the *Anopheles* X has a high rate of chromosomal rearrangement when compared with autosomes, despite the lack of polymorphic inversions in the X chromosomes in both species. Additionally, the difference between the rates of X and autosome chromosomal evolution is much more striking in *Anopheles* than in *Drosophila.* The high rate of evolution on the X correlates well with the density of simple repeats. Our data indicate that overall high rates of chromosomal evolution are not restricted to *Drosophila* but may be a feature common to *Diptera*.

The genome landscape of *An. stephensi* is characterized by relatively low repeat content compared to *An. gambiae*. *An. stephensi* appears to have larger amount of repeat-rich heterochromatin in pericentric regions but far less repetitive sequences in chromosomal arms as compared with *An. gambiae*. Using a newly developed chromosome quotient method, we identified a number of Y-chromosome contigs and BACs, which together represent currently the most abundant set of Y sequences in any *Anopheles* species.

The current assembly contains 11,789 predicted protein coding genes, 127 miRNA genes, 434 tRNA genes, and 53 fragments of rRNA genes. *An. stephensi* appears to have fewer gene duplications than *An. gambiae* according to orthology analysis, which may explain the slightly lower number of gene models.

This genome project is accompanied by the first comprehensive RNAseq-based transcriptomic analysis of an *Anopheles* mosquito. Twenty gene clusters were identified according to gene expression profiles, many of which are stage or sex-specific. GO term analysis of these gene clusters provided biological insights and leads for important research. For example, male-biased genes were enriched for genes involved in spermatogenesis and the auditory perception.

Close attention was paid to genes involved in innate immunity including LRIMs, APL1, and proteins in the Toll, IMD, insulin, and TGF-β signalling pathways. A high level of orthology is generally observed between *An. stephensi* and *An. gambiae.* RNAseq analysis, which was corroborated by other expression analysis methods, provided novel insights. For example, a protein known to interact with both TOR and TGF-β signalling pathways showed abundant mRNA expression in a wide range of tissues, providing new leads for insights into both TOR and TGF-β signalling in mosquitoes.

2.6 Methods

2.6.1 Strain selection

The Indian strain of *An. stephensi*, a representative of the type form was sequenced. The lab colony from which we selected mosquitoes for sequencing was originally established from wild mosquitoes collected in India. The lab colony has been maintained continuously for many generations so we did not attempt to inbreed it.

2.6.2 Sample collection

DNA was isolated from more than 50 adult male and female *An. stephensi* using the Qiagen (Hilden, Germany) DNeasy Blood and tissue kit following the suggested protocol. The integrity of the DNA was verified by running an aliquot on a 1% agarose gel to visualize any degradation. Total RNA was isolated using the standard protocol of the mirVana RNA isolation kit (Life Technologies, Carlsbad, CA) and quality was verified using Bioanalyzer(Agilent Technologies, Santa Clara, CA)

2.6.3 Sequencing

The *An. stephensi* genome was sequenced to 19.4x coverage using 454 FLX Titanium sequencing performed by the Virginia Bioinformatics Institute (VBI) core laboratory. Sequencing was performed on four different libraries: a single-end shotgun library, and 3 kb, 8 kb and 20 kb mate-pair libraries. A 200bp insert size library produced from male *An. stephensi* genomic DNA was prepared and subjected to a single lane of Illumina HiSeq. Genomic DNA from male *An.* sequence was subjected to 10 SMRT cells of Pacific Biosciences (PacBio) v1 sequencing. Only male were sequenced with PacBio because we are interested in increasing the probability of finding Y chromosome sequences. Sanger sequencing performed by Amplicon Express was used to sequence 7,263 BAC-ends.

2.6.4 Genome assembly

We used several approaches to combine the Illumina and 454 data to generate a better assembly. Newbler can take raw Illumina data as input, so we tried a Newbler assembly with the 454 and Illumina data. However, this resulted in a worse assembly than 454 alone. We had much more success with the strategy used to assemble the *Solenopsis invicta* genome [61]. We assembled the Illumina data first, and then cut the assembly into pseudo-454 reads. These reads were then used along with the real 454 data as input to Newbler [62].

2.6.5 *De novo* Illumina assembly with Celera

We assembled the paired-end Illumina reads using the Celera assembler [63] with the parameters: ”overlapper = ovl; unitigger = bogart; utgBubblePopping = 1; kickOutNonOvlContigs = 1; cgwDemoteRBP = 0; cgwMergeMissingThreshold = 0.5; merSize = 14”. The Celera assembler output comprise 41,213 contigs spanning 212.8 Mb. The N50 contig size of this assembly was 16.8 kb.

2.6.6 *De novo* 454 and Illumina pseudo-454 reads assembly with Newbler 2.8

The contigs of the aforementioned Illumina assembly were shredded informatically into 400 bp pieces with overlapping 200 bp to approximate 454 reads. To artificially simulate coverage depth, we started the shredding at offsets with the values of 0, 10, and 20. Shredding the Illumina assembly resulted in 2,452,038 pseudo-454 reads simulating 4.17x coverage.

We generated an assembly of the 454 and pseudo-454 reads with Newbler 2.8 using the “-het -scaffold -large -s 500” parameters. The resulting assembly contained 23,595 scaffolds spanned 221 Mb. The scaffold N50 size was 1.34 Mb. Mitochondrial DNA (1 scaffold), and other contamination (87 scaffolds) were identified by blastn and removed from the assembly.

2.6.7 Gap-filling with PacBio reads

PacBio data was used to fill gaps in the scaffolds to further improve the genome assembly. We error-corrected raw PacBio reads using the 454 sequencing data with the Celera pacBioToCa pipeline. pacBioToCa produced 0.88 Gb of error-corrected PacBio reads. Using the error-corrected PacBio data as input, Pbjelly [64] was used to fill gaps with parameters: “-minMatch 30 -

minPctIdentity 98 -bestn 10 -n Candidates 5 -maxScore -500 -nproc 36-noSplitSubreads". Pbjelly filled 1,310 gaps spanning 5.4 Mb.

2.6.8 Further scaffolding with BAC-ends

The scaffolds of the assembly were improved subsequently through the integration of 3,527 BAC-end pairs (120 kb ± 70 kb) using the Bambus scaffolder [65] (Additional file 25). The BAC-end sequences were mapped to the scaffolds using Nucmer [66]. The output files were used to generate the ".contig" format files required for Bambus. In total, 275 links between scaffolds were detected. Of these, 169 were retained as potential valid links, which are links connected by uniquely mapped BAC-ends. Links confirmed by less than two BAC-ends were rejected. A total of 46 links were retained that together connected 22 scaffolds, increasing the N50 scaffold size from 1,378 kb to 1,572 kb.

2.6.9 Assembly validation

CEGMA (Core Eukaryotic Genes): We used CEGMA [67] to search for the number of core eukaryotic genes to test the completeness and correctness of the genome assembly. CEGMA provides additional information as to whether the entire core eukaryotic genes are present (>70%) or only partially present (>20% and <70%). In total, CEGMA found 96.37% of the 248 core eukaryotic genes to be present, and 97.89% of the core eukaryotic genes to be partially present.

BAC-ends: We checked whether BAC-ends align concordantly to the genome to study the structural correctness of the *de novo* assembly. BAC-ends were aligned to the scaffolds using NUCMER. In order to ensure unambiguous mapping, only sequences that aligned to a unique location with >95% coverage and 99% identity were used. In total, 21.6% of the BAC-end sequence pairs could be aligned to a unique position in the *An. stephensi* genome with these stringent criteria. Pairs of BAC-end sequence that aligned discordantly to a single scaffold were considered indicative of potential misassembly. Only four of 717 aligned BAC-end pairs aligned discordantly with the assembly confirming overall structural correctness.

ESTs*: An. stephensi* EST sequences were downloaded from both the NCBI and VectorBase. We screened the EST sequences to remove any residual vector sequence. The screened ESTs were aligned to the assembly with GMAP [68]. In total, 35,367 of 36,064 ESTs aligned to the assembly. Of these, 26,638 aligned over at least 95% of their length with an identity of >98%. The high percentage of aligned ESTs demonstrates the near-completeness of the *An. stephensi* genome assembly.

Fluorescent in situ hybridization (FISH) - Slides were prepared from ovaries of lab reared, half-gravid females of the *An. stephensi* Indian wild-type strain. Slide preparation and hybridization experiments followed the techniques described in Sharakhova *et al* [69]. Fluorescent microscope images were converted to black and white and inverted in Adobe Photoshop. FISH signals were mapped to specific bands or interbands on the physical map for *An. stephensi* presented [70].

2.6.10 Constructing the physical map

For the chromosomal based genome assembly, all probes mapped by *in situ* hybridization by Sharakhova [70] and this study were aligned to the final version of the *An. stephensi* genome using NCBI blast+ blastn. Different blastn parameters were used for probes from different sources to determine if the probe was kept in the final assembly. An e-value of 1e-40 and an identity of >95% was required for probes from *An. stephensi*. An e-value of 1e-5 was required for probes from species other than *An. stephensi*. Probes that mapped to more than one location in the genome were discarded. The work by Sharakhova *et al* [70] hybridized 345 probes however, only ~200 probes from that study were maintained in the final chromosomal assembly. An additional 27 PCR products and BAC clones were hybridized to increase the coverage of our chromosomal assembly.

2.6.11 Annotation

The genome assembly was annotated initially using the MAKER pipeline [71]. This software synthesizes the results from *ab initio* gene prediction with experimental gene evidence to produce final annotations. Within the MAKER framework, RepeatMasker [72] was used to mask low-complexity genomic sequence based on the repeat library from previous prediction. First, ESTs and proteins were aligned to the genome by MAKER using BLASTn and BLASTx, respectively. MAKER uses the program Exonerate to polish BLAST hits. Next, within the MAKER framework, SNAP [73] and AUGUSTUS [74] were run to produce *ab initio* gene predictions based on the initial training data. SNAP and AUGUSTUS were run once again inside of MAKER using the initial training obtained from the ESTs and protein alignments to produce the final annotations.

2.6.12 Orthology and molecular species phylogeny

Orthologs of predicted *An. stephensi* genes were assigned by OrthoDB [75]. Information about orthologous genes for *An. gambiae*, *Ae. aegypti*, and *D. melanogaster* also were downloaded from OrthoDB. Enrichment analysis was performed for categories of orthologs using the methods provided in the ontology section. The molecular phylogeny of the 10 selected species was determined from the concatenated protein sequence alignments using MUSCLE [76] (default parameters) followed by alignment trimming with trimAl [77] (automated1 parameters) of 3,695 relaxed single-copy orthologs (a maximum of three paralogs allowed in no more than two species, longest protein selected) from OrthoDB [75]. The resulting 2,246,060 amino acid columns with 932,504 distinct alignment patterns was analysed with RAxML [78] with the PROTGAMMAJTT model to estimate the maximum likelihood species phylogeny with 100 bootstrap samples.

2.6.13 Transcriptomics

RNA-seq from 11 samples including: 0-1, 2-4, 4-8, and 8-12 hour embryos, larva, pupa, adult males, adult females, non-blood-fed ovaries, blood-fed ovaries, and female carcasses without ovaries as described [25] were used for transcriptome analysis. These RNA-seq samples are available from the NCBI SRA (SRP013839). Tophat [79] was used to align these RNA-seq reads to the *An. stephensi* genome and HTSeq-count [80] was used to generate an occurrence table for each gene in each sample. The numbers of alignments to each gene in each sample then were clustered using MBCluster.Seq [81], an R package designed to cluster genes by expression profile based on Poisson or Negative-Binomial models. MBCluster.Seq generated 20 clusters. To visualize these results we performed regularized log transformation to the original occurrence tables for all 20 clusters using DESeq2 [82]. The results were plotted using ggplot2 [83].

2.6.14 Ontology

Gene ontology (GO) terms were assigned for the 20 clusters of predicted *An. stephensi* genes. GO terms were assigned using Blast2Go [84]. The predicted proteins are blasted against the NCBI non-redundant protein database and scanned with InterProScan [85] against InterPro's signatures. After GO terms were assigned, GO-slim results were generated for the available annotation based on the Generic GO slim mapping. The GO terms assigned by Blast2GO were subject to GO term enrichment. Overrepresented GO terms were identified using a hypergeometric test using the GOstats package in R [86].

2.6.15 Functional annotation of key gene families

We obtained the InterPro ID information for proteins in *An. stephensi* from the ontology analysis. We functional annotated gene families based on the assigned InterPro ID. The gene families, including genes involved in immunity, chemosensation and detoxification were studied. For comparative genome analysis, we retrieved the InterPro ID for other 7 species (*An. gambiae*, *An.darlingi, A. aegypti*, *Culex quinquefasciatus*, *D. melanogaster, Bombyx mori* and *Tribolium castaneum*) using Biomart [87] from vectorbase (https://www.vectorbase.org/) and Ensembl Metazoa (http://metazoa.ensembl.org).We compared gene numbers in gene families of interest. For gene families with obvious differences in numbers between *An. stephensi* and *An. gambiae*, we preformed phylogenetic analysis of these genes. First we aligned these genes from *Anopheles* species using MUSCLE [76].Then, we constructed phylogenetic tree using Neighbor-joining method with 1000 bootstrap replicates by CLC Genomics Workbench 4 (http://www.clcbio.com).

2.6.16 noncoding RNA

We used tRNAScan-SE [88] with the default eukaryotic mode to predict 434 tRNAs in the *An. stephensi* genome (Additional file 1: Table S15; Additional file 26). Other noncoding RNAs were predicted with INFERNAL [89] by searching against Rfam database version 11.0 [90]. A total of 53 fragmental ribosomal RNA, 34 snRNA, 7 snoRNA, 127 miRNA, and 148 sequences with homology to the *An. gambiae* self-cleaving riboswitch were predicted with an e-value cutoff of 1e-5.

2.6.17 Transposable elements and other interspersed repeats

Transposable element discovery and classification were performed on the *An. stephensi* scaffold sequences using previously-described pipelines for LTR-retrotransposons, non-LTR-retrotransposons, SINEs, DNA-transposons, and MITEs, followed by manual inspection [91]. The manually-annotated TE libraries then were compared with the RepeatModeler output to remove redundancy and to correct mis-classification by RepeatModler. A repeat library was produced that contains all manually-annotated TEs and non-redundant sequences from RepeatModeler. The repeat library was used to run RepeatMasker at default settings on the *An. stephensi* assembly to calculate TE copy number and genome occupancy.

2.6.18 Simple repeats

The number of microsatellites, minisatellites, and satellites present in the mapped scaffolds for each chromosome were derived by dividing the scaffolds into strings of 100,000 bp and then concatenating them into a multi-FASTA file to represent an *An. stephensi* pseudo chromosome. Scaffolds were oriented when possible, and all unoriented scaffolds were given the default positive orientation for that chromosome. The multiFASTA file for each pseudo-chromosome was analyzed using a local copy of TandemRepeatsFinder v 4.07b [92]. Parameters for the analysis followed those used by Xia *et al* [44]: microsatellites were those of period size 2-6 with copy number of >8. Minisatellites had period size 7-99 while repeats were considered satellites if they had a period size of >100. Both satellites and minisatellites were considered only if they had a copy number of >2. Simple repeats were recorded only if they had at least 80% identity.

2.6.19 Identification of S/MARs

Scaffold/matrix associated regions were identified using the SMARTest bioinformatic tool provided by Genomatix [93]. Densities of genes and TEs per 100 kb window were calculated using Bedtools coverage based on the genome annotation and TE annotation respectively.

2.6.20 Synteny, gene order evolution, and inversions

One-to-one orthologs from *An. gambiae* and *An. stephensi* were identified using OrthoDB [75] and their locations on the *An. gambiae* and *An. stephensi* scaffolds determined. Comparative positions of the genes on the scaffolds based on ontology relationships were plotted using genoPlotR [94]. Scaffolds that mapped using two or more probes were oriented properly, but those anchored by only one probe were used in their default orientation. The number of synteny blocks for each pair of homologous chromosome arms between *An. stephensi* and *An. gambiae* was determined from the images output from genoPlotR. Two criteria were imposed to determine the number of synteny blocks: the orientation of two or more orthologous genes, and whether the genes remained in the same order on the chromosome of *An. stephensi* as in *An. gambiae*. Thus, a group of two or more genes is assigned to the same synteny block if it has the same orientation and order in both species. Synteny blocks were numbered 1,2,3,4 ...*etc*. along the chromosome by assigning *An. gambiae* as the default gene order. *An. stephensi* was considered rearranged compared to *An. gambiae* when the numbering of synteny blocks was the same in both species but the order was rearranged in *An. stephensi*. After quantifying the number of synteny blocks and the amount of gene rearrangement between the two species, we estimated the number of chromosomal inversions between them using the programs Genome Rearrangements in Mouse and Man (GRIMM [50]).

2.6.21 SNP analysis

We used CLC Genomics Workbench 4 (http://www.clcbio.com) to identify SNPs using a combination of the male and female Illumina data (Accession number: SRP013838). The required coverage was 20 and minimum variant frequency was 35. SNP calls made on the assembly were assessed for their effect on transcripts from the gene build using the Ensembl e-hive, variation database and variation consequence pipeline (available from github https://github.com/Ensembl/ensembl-hive +https://github.com/Ensembl/ensembl-variation/). The Ensembl variation consequence pipeline uses the Ensembl API in the same manner as the Variant Effect Predictor [95] and produces equivalent output. The variation consequence pipeline directly loaded the analysis results into an Ensembl MySQL variation database which was used to generate summary statistics of transcript consequences classified using Sequence Ontologs [96].

2.7 Data access

The *An. stephensi* genome assembly has been deposited in GenBank under the accession number ALPR00000000 and is available at www.VectorBase.org. The raw sequence data used for genome assembly is available in the NCBI SRA: 454 - SRP037783, Illumina - SRP037783 and PacBio - SRP037783. The BAC-ends used for scaffolding are available from the NCBI dbGSS accession numbers: KG772729 - KG777469. RNA-Seq data can be accessed at the NCBI SRA with ID SRP013839.

2.8 Additional files

Additional file 1: this file includes supplemental text, supplemental figures, and supplemental Tables.

Additional files 2-22 are provided as Additional_Files2-22.tar.gz. We also provide a link for all additional files in case the reviewers find it useful (http://tu08.fralin.vt.edu/share/Additional_Files/)

Additional file 2: Physical Map Data.xlsx

Additional file 3: Lists of genes in clusters.xlsx

Additional file 4: Cluster ontology.txt

Additional file 5: Sex-biased genes list and GO terms.xlsx

Additional file 6: Revised annotation for immunity-related genes.gff3

Additional file 7: Sequences of immunity-related genes.fasta

Additional file 8: RNA-seq expression profile of immunity-related genes.xlsx

Additional file 9: Automatic annotated salivary genes.txt

Additional file 10: Revised manual annotation for salivary genes

Additional file 11: Manual annotated salivary genes sequences

Additional file 12: Gene families counts table.xlsx

Additional file 13: Repeat sequences

Additional file 14: Genome Landscape.xlsx

Additional file 15: Tandem repeat sequences.fa

Additional file 16: Chromosome quotients of putative Y-linked scaffolds

Additional file 17: Sequences of putative Y-linked scaffolds

Additional file 18: Chromosome quotients of Y-linked BACs

Additional file 19: Sequences of Y-linked BACs

Additional file 20: Repeat masker output of Y-linked BACs and Y-linked scaffolds

Additional file 21: Repeat masker output of Y-linked BACs

Additional file 22: Synteny Blocks.docx

Additional file 23: SNP analysis raw data.cvs

Additional file 24: Summary of transcript consequences for An. stephensi Indian strain SNP calls.xlsx

Additional file 25: BAC-ends dbGSS accession numbers.txt

Additional file 26: Non-coding RNA annotation.txt

2.9 Acknowledgements

This work is supported by the Fralin Life Science Institute and the Virginia Experimental Station, and by NIH grants AI77680 and AI105575 to ZT, AI094289 and AI099528 to IVS, AI29746 to AAJ, AI095842 to KM, AI073745 to MAR, AI080799 and AI078183 to SL, AI042361 and AI073685 to KDV. RMW is supported by Marie Curie International Outgoing Fellowship PIOF-GA-2011-303312. XC is supported by GDUPS (2009). This work was also supported in part by

NSF grant CNS-0960081 and the HokieSpeed and BlueRidge supercomputers at Virginia Tech. YS thanks the Department of Biotechnology, Government of India for the financial support.

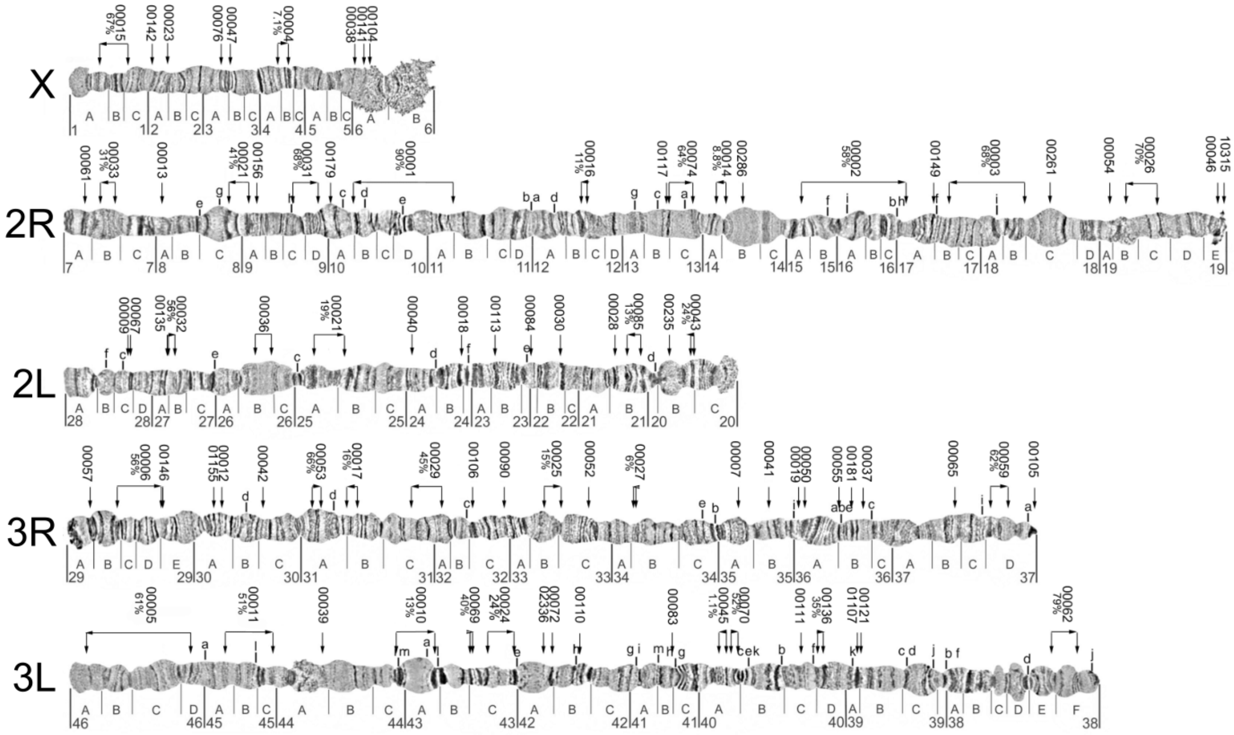

Figure 2.1 Physical Map

A physical map of the *An. stephensi* genome was created from FISH on polytene chromosomes comprising 227 probes and 86 scaffolds. These 86 scaffolds comprise 137.14 Mb or 62% of the *An. stephensi* genome. Orientation was assigned to 32 of the 86 scaffolds. The physical map includes 28 of the 30 largest scaffolds.

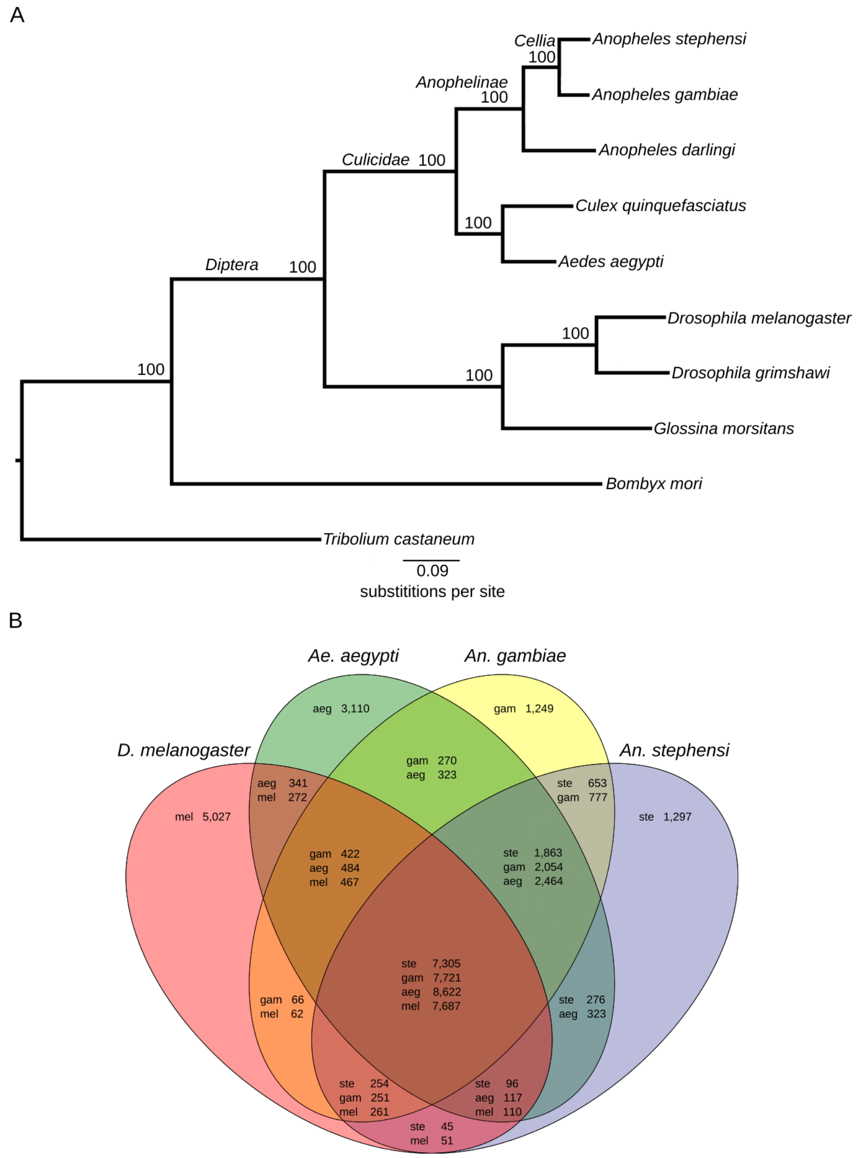

Figure 2.2 Molecular species phylogeny and orthology

(A) The maximum likelihood molecular species phylogeny estimated from universal single-copy orthologs supports the recognised species relationships with *An. stephensi* and *An. gambiae* in subgenus *Cellia* within the genus *Anopheles*. (B) Comparative analysis of orthologs from *An. stephensi*, *An. gambiae*, *Ae. aegypti*, and *D. melanogaster*. Orthologous genes were retrieved from OrthoDB. 7,305 genes were shared among all four species, 1,297 genes were specific to *An. stephensi*, 653 genes were *Anopheles*-specific, and 1,863 genes were mosquito-specific.

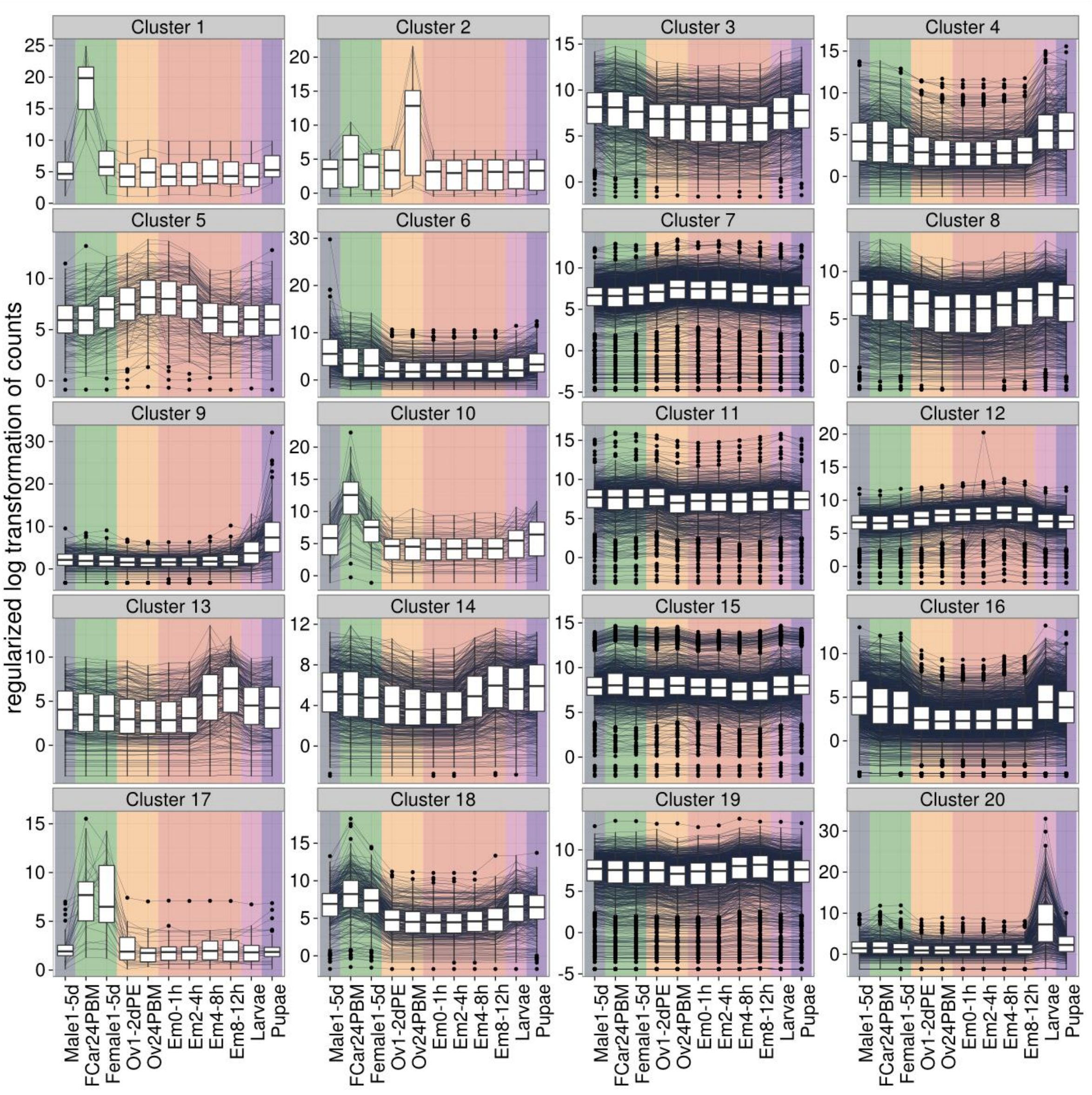

Figure 2.3 Gene clustering according to expression profile.

Twenty groups of genes were clustered by expression profile. The expression profiles used for grouping were generated using 11 RNA-seq samples spanning developmental time points including: 0-1, 2-4, 4-8, and 8-12 hour embryos, larva, pupa, adult males, adult females, non-blood-fed ovaries, blood-fed ovaries, and 24 hours post-blood-fed female carcass without ovaries. Male stage are colored blue, female stages are colored green, ovary samples are colored yellow, embryo samples are colored red, larva samples are colored pink, and pupa samples are colored purple. Many of these clusters correspond to either a specific developmental stage or specific sex.

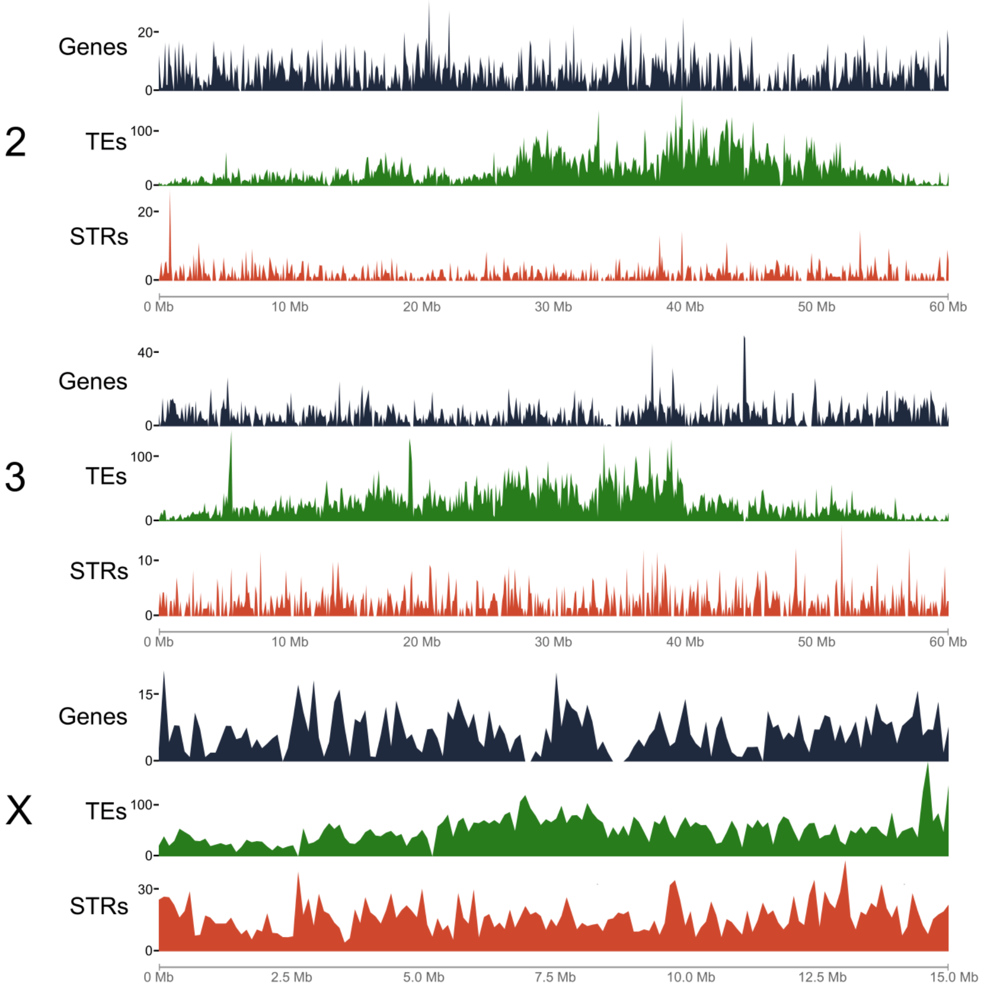

Figure 2.4 Genome Landscape

Density of genes (black vertical lines), transposable elements (TEs; green vertical lines), and short tandem repeats (STRs, red vertical lines) in 100 kb windows of mapped scaffolds. Based on the physical map, scaffolds were ordered and oriented respective to their position in the chromosomes and then 100 kb non-overlapping windows were generated for each scaffold (X-axis). The density of genes and TEs (Y-axis) was determined using coverageBed. Satellite sequences were identified using TandemRepeatFinder. The short tandem repeats track is a combination of the number of microsatellites, minisatellites and satellites per 100 kb window.

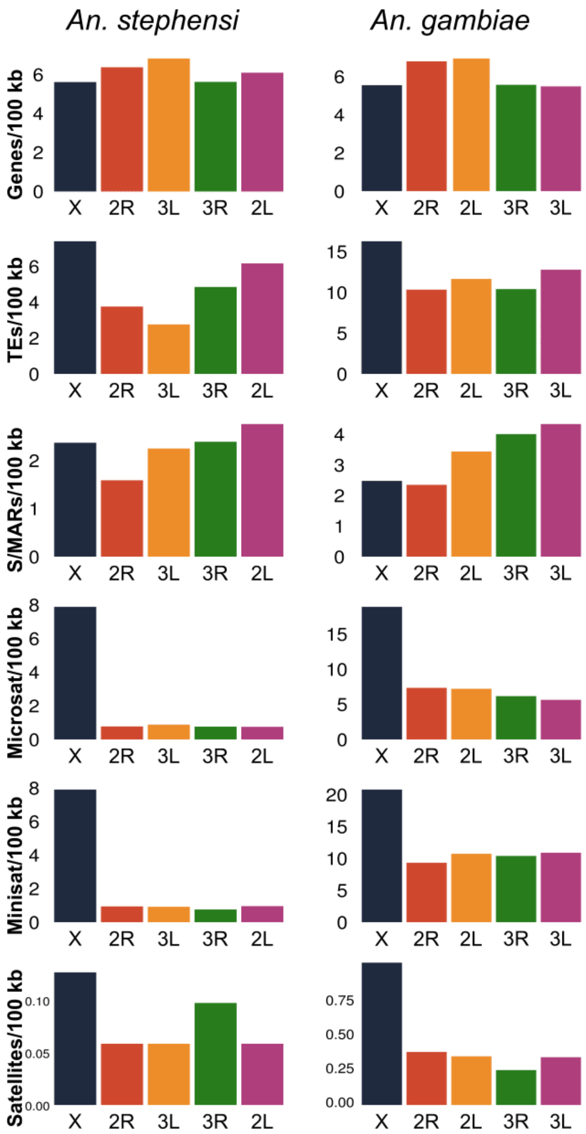

Figure 2.5 Average Density / 100kb / ARM

A comparison of the average density per 100 kb of genes, TEs, S/MARS, microsatellites, minisatellites, and satellites between chromosome arms.

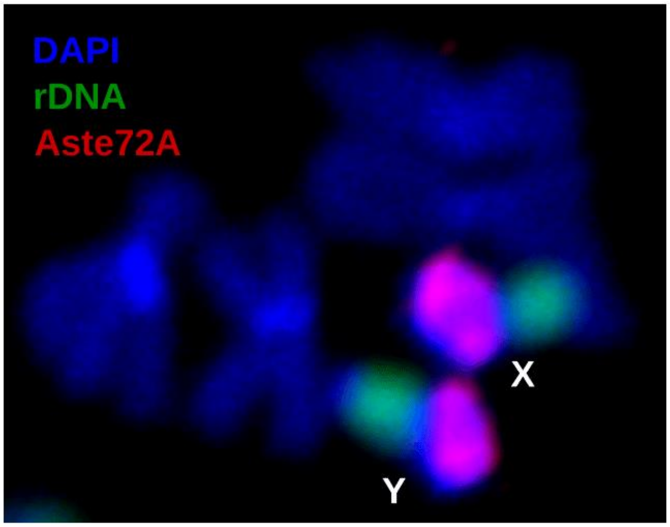

Figure 2.6 FISH with Aste72A, rDNA and DAPI on mitotic chromosomes

The pattern of hybridization for satellite DNA Aste72A on mitotic sex chromosomes of *An. stephensi*. Aste72A hybridizes to pericentric heterochromatin in both X and Y chromosomes while ribosomal DNA locus maps next to the heterochromatin band in sex chromosomes.

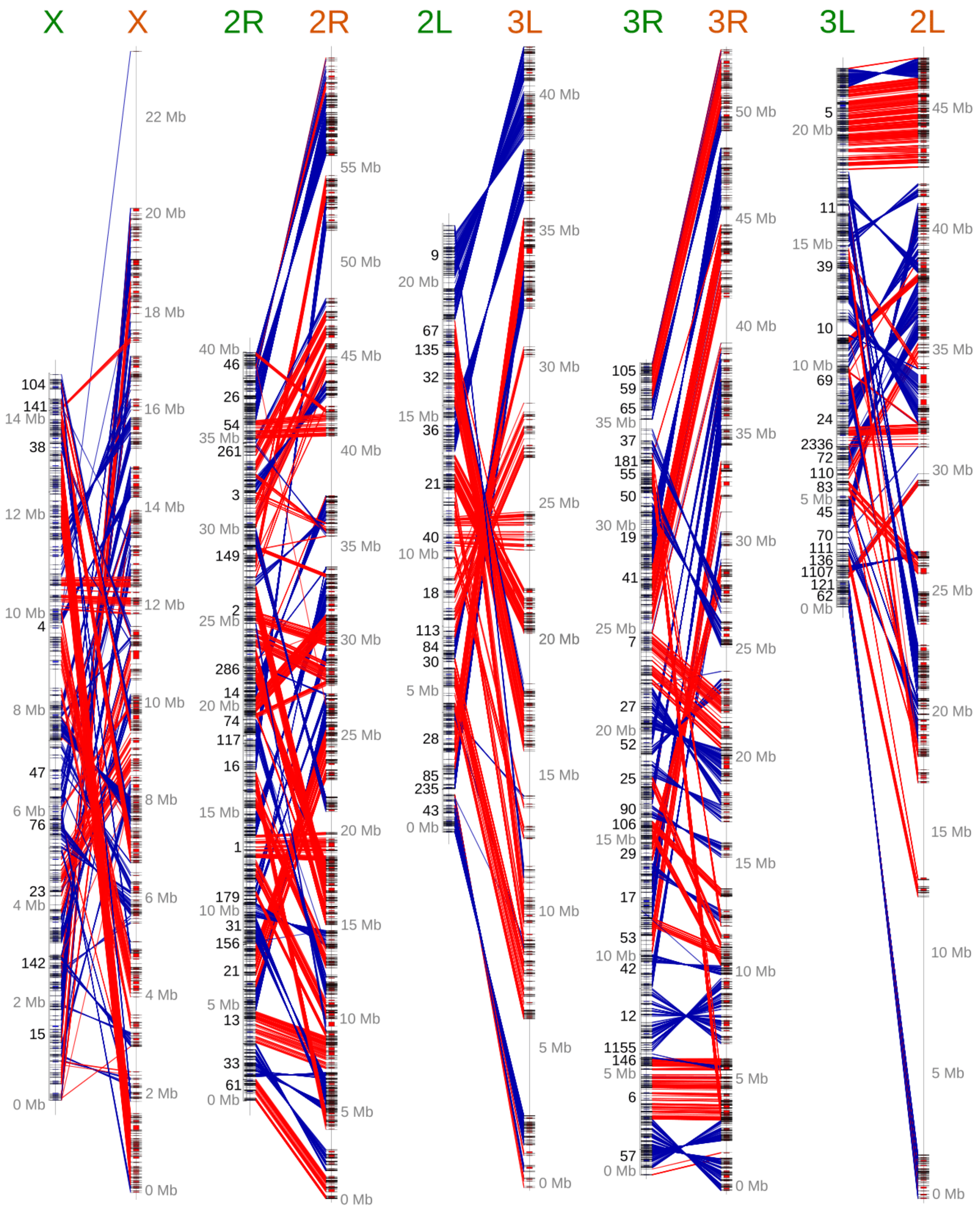

Figure 2.7 Synteny

Synteny between *An. stephensi* and *An. gambiae* based on 6,448 single-copy orthologs. Orthologs with the same orientation in *An. stephensi* and *An. gambiae* are connected with red lines and orthologs with the opposite orientation are connected with blue lines. Orthologous genes from *An. stephensi* and *An. gambiae* were retrieved from OrthoDB. The physical map was used to identify the relative locations of genes on the *An. stephensi* chromosomes. The relationship of the position between the *An. stephensi* and *An. gambiae* orthologs were plotted with GenoPlotR. 66 syntenic blocks were identified on the X chromosome. 104 and 64 syntenic blocks were identified on 2R and 2L (3L in *An. stephensi*). 104 and 42 syntenic blocks were identified on 3R and 3L (2L in *An. stephensi*). Therefore, the X chromosome has undergone the most rearrangements per megabase.

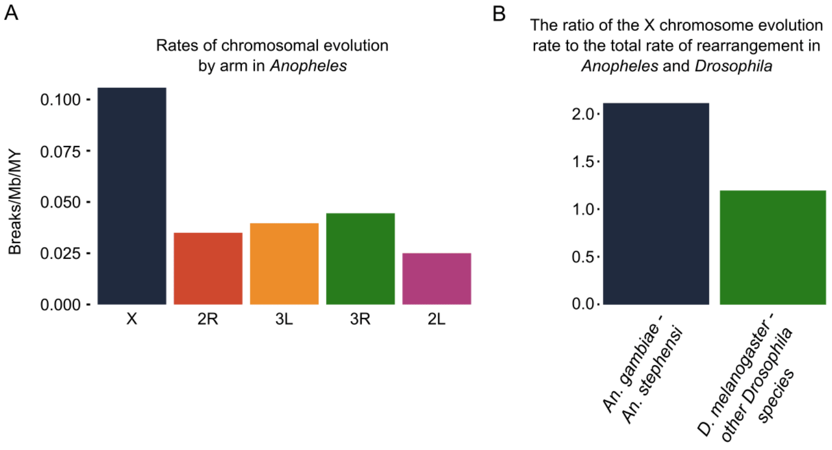

Figure 2.8 Chromosome evolution in Anopheles and Drosophila.

A) Higher rates of rearrangement on the X chromosome compared to autosomes between *An. stephensi* and *An. gambiae*. Arm designations for the figure are according to *An. stephensi*. B) The ratio of the X chromosome evolution rate to the total rate of rearrangement is higher in *Anopheles* than in *Drosophila*.

Table 2.1 Assembly Statistics

Statistic	Value
Scaffolds (n)	23,371
Scaffold N50 size	1,591,355
Maximum Scaffold Length	5,975,090
Minimum Scaffold Length	486
Total Length of Scaffolds	221,309,404
Percent Ns	5.35 %
Contigs (n)	31,761
Contig N50 size	36,511
Maximum Contig Length	475,937
Minimum Contig Length	347
Total Length of Contigs	209,483,518
GC Percent	44.80 %

Table 2.2 Physical Map Information

Arm	Scaffolds per Arm (n)	Length (Mb)	% Mapped Genome	% of Total Genome
X	9	14.95	10.90	6.77
2R	21	39.50	28.80	17.87
2L	15	22.40	16.33	10.14
3R	24	37.83	27.59	17.12
3L	17	22.45	16.37	10.16
Total	86	137.14	100	62.05
Scaffolds mapped to each chromosome, total bp to each chromosome, percent of the predicted genome covered.				

Table 2.3 Transposable elements and other interspersed repeats

Type	Elements (n)	Length Occupied (bp)	Percent of Genome
SINEs	30,514	3,739,253	1.69
LINEs	22,022	5,231,240	2.36
LTR elements	4,359	1,499,282	0.68
DNA elements	4,611	966,667	0.44
Unclassified	30,611	4,322,468	1.95
Total	92,117	15,758,910	7.12

2.10 References

1. Feachem RGA, Phillips AA, Hwang J, Cotter C, Wielgosz B, Greenwood BM, Sabot O, Rodriguez MH, Abeyasinghe RR, Ghebreyesus TA, Snow RW: **Shrinking the malaria map: Progress and prospects.** *The Lancet* 2010, **376:**1566-1578.

2. White MT, Conteh L, Cibulskis R, Ghani AC: **Costs and cost-effectiveness of malaria control interventions--a systematic review.** *Malaria journal* 2011, **10:**337-337.

3. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JMC, Wides R, et al: **The genome sequence of the malaria mosquito Anopheles gambiae.** *Science (New York, NY)* 2002, **298:**129-149.

4. Rafinejad J, Vatandoost H, Nikpoor F, Abai MR, Shaeghi M, Duchen S, Rafi F: **Effect of washing on the bioefficacy of insecticide-treated nets (ITNs) and long-lasting insecticidal nets (LLINs) against main malaria vector Anopheles stephensi by three bioassay methods.** *Journal of vector borne diseases* 2008, **45:**143-150.

5. Sharma VP: **Current scenario of malaria in India.** *Parassitologia* 1999, **41:**349-353.

6. Faulde MK, Rueda LM, Khaireh BA: **First record of the Asian malaria vector Anopheles stephensi and its possible role in the resurgence of malaria in Djibouti, Horn of Africa.** *Acta Trop* 2014, **139C:**39-43.

7. Gakhar SK, Sharma R, Sharma A: **Population genetic structure of malaria vector Anopheles stephensi Liston (Diptera: Culicidae).** *Indian journal of experimental biology* 2013, **51:**273-279.

8. Murray CJL, Rosenfeld LC, Lim SS, Andrews KG, Foreman KJ, Haring D, Fullman N, Naghavi M, Lozano R, Lopez AD: **Global malaria mortality between 1980 and 2010: A systematic analysis.** *The Lancet* 2012, **379:**413-431.

9. Alonso PL, Brown G, Arevalo-Herrera M, Binka F, Chitnis C, Collins F, Doumbo OK, Greenwood B, Hall BF, Levine MM, et al: **A research Agenda to underpin Malaria Eradication.** vol. 8; 2011.

10. Nolan T, Bower TM, Brown AE, Crisanti A, Catteruccia F: **piggyBac-mediated germline transformation of the malaria mosquito Anopheles stephensi using the red fluorescent protein dsRED as a selectable marker.** *The Journal of biological chemistry* 2002, **277:**8759-8762.

11. O'Brochta DA, Alford RT, Pilitt KL, Aluvihare CU, Harrell RA: **piggyBac transposon remobilization and enhancer detection in Anopheles mosquitoes.** In *Proceedings of the National Academy of Sciences*, vol. 108. pp. 16339-16344; 2011:16339-16344.

12. Isaacs AT, Jasinskiene N, Tretiakov M, Thiery I, Zettor A, Bourgouin C, James AA: **PNAS Plus: Transgenic Anopheles stephensi coexpressing single-chain antibodies resist Plasmodium falciparum development.** In *Proceedings of the National Academy of Sciences*, vol. 109. pp. E1922-E1930; 2012:E1922-E1930.

13. Smidler AL, Terenzi O, Soichot J, Levashina EA, Marois E: **Targeted Mutagenesis in the Malaria Mosquito Using TALE Nucleases.** *PLoS ONE* 2013, **8**.

14. Brown AE, Bugeon L, Crisanti A, Catteruccia F: **Stable and heritable gene silencing in the malaria vector Anopheles stephensi.** *Nucleic Acids Res* 2003, **31:**e85.

15. Bian G, Joshi D, Dong Y, Lu P, Zhou G, Pan X, Xu Y, Dimopoulos G, Xi Z: **Wolbachia invades Anopheles stephensi populations and induces refractoriness to Plasmodium infection.** *Science (New York, NY)* 2013, **340:**748-751.

16. Dong Y, Das S, Cirimotich C, Souza-Neto JA, McLean KJ, Dimopoulos G: **Engineered anopheles immunity to plasmodium infection.** *PLoS Pathogens* 2011, **7**.

17. Garver LS, Dong Y, Dimopoulos G: **Caspar controls resistance to plasmodium falciparum in diverse anopheline species.** *PLoS Pathogens* 2009, **5**.

18. Luckhart S, Giulivi C, Drexler AL, Antonova-Koch Y, Sakaguchi D, Napoli E, Wong S, Price MS, Eigenheer R, Phinney BS, et al: **Sustained Activation of Akt Elicits Mitochondrial Dysfunction to Block Plasmodium falciparum Infection in the Mosquito Host.** *PLoS Pathogens* 2013, **9**.

19. Mitri C, Thiery I, Bourgouin C, Paul REL: **Density-dependent impact of the human malaria parasite Plasmodium falciparum gametocyte sex ratio on mosquito infection rates.** *Proceedings Biological sciences / The Royal Society* 2009, **276:**3721-3726.

20. Pakpour N, Corby-Harris V, Green GP, Smithers HM, Cheung KW, Riehle Ma, Luckhart S: **Ingested human insulin inhibits the mosquito NF-κB-dependent immune response to Plasmodium falciparum.** *Infection and immunity* 2012, **80:**2141-2149.

21. Rai KS, Black Iv WC: **Mosquito Genomes: Structure, Organization, and Evolution.** *Advances in Genetics* 1999, **41:**1-33.

22. Sharakhova MV, Xia A, Leman SC, Sharakhov IV: **Arm-specific dynamics of chromosome evolution in malaria mosquitoes.** *BMC Evol Biol* 2011, **11:**91.

23. Marinotti O, Cerqueira GC, de Almeida LG, Ferro MI, Loreto EL, Zaha A, Teixeira SM, Wespiser AR, Almeida ESA, Schlindwein AD, et al: **The genome of Anopheles darlingi, the main neotropical malaria vector.** *Nucleic Acids Res* 2013, **41:**7387-7400.

24. Zhou D, Zhang D, Ding G, Shi L, Hou Q, Ye Y, Xu Y, Zhou H, Xiong C, Li S, et al: **Genome sequence of Anopheles sinensis provides insight into genetics basis of mosquito competence for malaria parasites.** *BMC Genomics* 2014, **15:**42.

25. Criscione F, Qi Y, Saunders R, Hall B, Tu Z: **A unique Y gene in the Asian malaria mosquito Anopheles stephensi encodes a small lysine-rich protein and is transcribed at the onset of embryonic development.** *Insect Mol Biol* 2013, **22:**433-441.

26. Göpfert MC, Robert D: **Active auditory mechanics in mosquitoes.** *Proceedings Biological sciences / The Royal Society* 2001, **268:**333-339.

27. Gibson G, Warren B, Russell IJ: **Humming in tune: sex and species recognition by mosquitoes on the wing.** *Journal of the Association for Research in Otolaryngology : JARO* 2010, **11:**527-540.

28. Xi Z, Ramirez JL, Dimopoulos G: **The Aedes aegypti toll pathway controls dengue virus infection.** *PLoS Pathogens* 2008, **4**.

29. Price I, Ermentrout B, Zamora R, Wang B, Azhar N, Mi Q, Constantine G, Faeder JR, Luckhart S, Vodovotz Y: **In vivo, in vitro, and in silico studies suggest a conserved immune module that regulates malaria parasite transmission from mammals to mosquitoes.** *Journal of Theoretical Biology* 2013, **334:**173-186.

30. Horton AA, Wang B, Camp L, Price MS, Arshi A, Nagy M, Nadler SA, Faeder JR, Luckhart S: **The mitogen-activated protein kinome from Anopheles gambiae: identification, phylogeny and functional characterization of the ERK, JNK and p38 MAP kinases.** vol. 12. pp. 574-574; 2011:574-574.

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

32. Choi J, Chen J, Schreiber SL, Clardy J: **Structure of the FKBP12-rapamycin complex interacting with the binding domain of human FRAP.** *Science (New York, NY)* 1996, **273:**239-242.

33. Laplante M, Sabatini DM: **MTOR signaling in growth control and disease.** In *Cell*, vol. 149. pp. 274-293; 2012:274-293.

34. Grewal SS: **Insulin/TOR signaling in growth and homeostasis: A view from the fly world.** In *International Journal of Biochemistry and Cell Biology*, vol. 41. pp. 1006-1010; 2009:1006-1010.

35. Arsic D, Guerin PM: **Nutrient content of diet affects the signaling activity of the insulin/target of rapamycin/p70 S6 kinase pathway in the African malaria mosquito Anopheles gambiae.** *Journal of Insect Physiology* 2008, **54:**1226-1235.

36. Anderson KV, Bokla L, Nüsslein-Volhard C: **Establishment of dorsal-ventral polarity in the Drosophila embryo: the induction of polarity by the Toll gene product.** *Cell* 1985, **42:**791-798.

37. Valenzuela JG, Francischetti IMB, Pham VM, Garfield MK, Ribeiro JMC: **Exploring the salivary gland transcriptome and proteome of the Anopheles stephensi mosquito.** *Insect Biochemistry and Molecular Biology* 2003, **33:**717-732.

38. Arca B, Lombardo F, Valenzuela JG, Francischetti IM, Marinotti O, Coluzzi M, Ribeiro JM: **An updated catalogue of salivary gland transcripts in the adult female mosquito, Anopheles gambiae.** *J Exp Biol* 2005, **208:**3971-3986.

39. Ribeiro JMC, Mans BJ, Arcà B: **An insight into the sialome of blood-feeding Nematocera.** vol. 40. pp. 767-784; 2010:767-784.

40. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, et al: **InterPro in 2011: new developments in the family and domain prediction database.** *Nucleic Acids Res* 2012, **40:**D306-312.

41. Neafsey DE, Christophides GK, Collins FH, Emrich SJ, Fontaine MC, Gelbart W, Hahn MW, Howell PI, Kafatos FC, Lawson D, et al: **The evolution of the Anopheles 16 genomes project.** *G3 (Bethesda)* 2013, **3:**1191-1194.

42. Mahmood F, Sakai RK: **Inversion polymorphisms in natural populations of Anopheles stephensi.** *Can J Genet Cytol* 1984, **26:**538-546.

43. Hoffmann AA, Sgrò CM, Weeks AR: **Chromosomal inversion polymorphisms and adaptation.** vol. 19. pp. 482-488; 2004:482-488.

44. Xia A, Sharakhova MV, Leman SC, Tu Z, Bailey JA, Smith CD, Sharakhov IV: **Genome landscape and evolutionary plasticity of chromosomes in malaria mosquitoes.** *PLoS ONE* 2010, **5**.

45. Baricheva EA, Berrios M, Bogachev SS, Borisevich IV, Lapik ER, Sharakhov IV, Stuurman N, Fisher PA: **DNA from Drosophila melanogaster β-heterochromatin binds specifically to nuclear lamins in vitro and the nuclear envelope in situ.** *Gene* 1996, **171:**171-176.

46. Dechat T, Pfleghaar K, Sengupta K, Shimi T, Shumaker DK, Solimando L, Goldman RD: **Nuclear lamins: major factors in the structural organization and function of the nucleus and chromatin.** *Genes & development* 2008, **22:**832-853.

47. Baker RH, Sakai RK: **Triploids and male determination in the mosquito, Anopheles culicifacies.** *J Hered* 1979, **70:**345-346.

48. Hall AB, Qi Y, Timoshevskiy V, Sharakhova MV, Sharakhov IV, Tu Z: **Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females.** *BMC Genomics* 2013, **14:**273.

49. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al: **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.** *Nat Methods* 2013, **10:**563-569.

50. Tesler G: **GRIMM: genome rearrangements web server.** *Bioinformatics* 2002, **18:**492-493.

51. Timoshevskiy VA, Kinney NA, deBruyn BS, Mao C, Tu Z, Severson DW, Sharakhov IV, Sharakhova MV: **Genomic composition and evolution of Aedes aegypti chromosomes revealed by the analysis of physically mapped supercontigs.** *BMC Biol* 2014, **12:**27.

52. Sharakhov IV, Serazin AC, Grushko OG, Dana A, Lobo N, Hillenmeyer ME, Westerman R, Romero-Severson J, Costantini C, Sagnon N, et al: **Inversions and gene order shuffling in Anopheles gambiae and A. funestus.** *Science* 2002, **298:**182-185.

53. Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguadé M, Anderson WW, et al: **Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps.** *Genetics* 2008, **179:**1601-1655.

54. Ranz JM, Maurin D, Chan YS, Von Grotthuss M, Hillier LW, Roote J, Ashburner M, Bergman CM: **Principles of genome evolution in the Drosophila melanogaster species group.** *PLoS Biology* 2007, **5:**1366-1381.

55. Ranz JM, Casals F, Ruiz A: **How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila.** *Genome research* 2001, **11:**230-239.

56. Peng Q, Pevzner PA, Tesler G: **The fragile breakage versus random breakage models of chromosome evolution.** *PLoS Comput Biol* 2006, **2:**e14.

57. Chaisson MJ, Raphael BJ, Pevzner PA: **Microinversions in mammalian evolution.** *Proc Natl Acad Sci U S A* 2006, **103:**19824-19829.

58. Bourque G, Pevzner PA: **Genome-scale evolution: reconstructing gene orders in the ancestral species.** *Genome Res* 2002, **12:**26-36.

59. Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, Gelbart WM: **Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes.** *Genetics* 2008, **179:**1657-1680.

60. Lawniczak MK, Emrich SJ, Holloway AK, Regier AP, Olson M, White B, Redmond S, Fulton L, Appelbaum E, Godfrey J, et al: **Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences.** *Science* 2010, **330:**512-514.

61. Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D, et al: **The genome of the fire ant Solenopsis invicta.** *Proc Natl Acad Sci U S A* 2011, **108:**5679-5684.

62. Kumar S, Blaxter ML: **Comparing de novo assemblers for 454 transcriptome data.** *BMC genomics* 2010, **11:**571.

63. Denisov G, Walenz B, Halpern AL, Miller J, Axelrod N, Levy S, Sutton G: **Consensus generation and variant detection by Celera Assembler.** *Bioinformatics* 2008, **24:**1035-1040.

64. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA: **Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology.** *PLoS One* 2012, **7:**e47768.

65. Pop M, Kosack DS, Salzberg SL: **Hierarchical scaffolding with Bambus.** *Genome research* 2004, **14:**149-159.

66. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome biology* 2004, **5:**R12.

67. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics (Oxford, England)* 2007, **23:**1061-1067.

68. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** *Bioinformatics (Oxford, England)* 2005, **21:**1859-1875.

69. Sharakhova MV, George P, Brusentsova IV, Leman SC, Bailey JA, Smith CD, Sharakhov IV: **Genome mapping and characterization of the Anopheles gambiae heterochromatin.** *BMC genomics* 2010, **11:**459.

70. Sharakhova MV, Xia A, Tu Z, Shouche YS, Unger MF, Sharakhov IV: **A physical map for an Asian malaria mosquito, Anopheles stephensi.** *Am J Trop Med Hyg* 2010, **83:**1023-1027.

71. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Res* 2008, **18:**188-196.

72. Tempel S: **Using and understanding RepeatMasker.** *Methods Mol Biol* 2012, **859:**29-51.

73. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5:**59.

74. Stanke M, Steinkamp R, Waack S, Morgenstern B: **AUGUSTUS: a web server for gene finding in eukaryotes.** *Nucleic Acids Res* 2004, **32:**W309-312.

75. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV: **OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs.** *Nucleic acids research* 2013, **41:**D358-365.

76. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792-1797.

77. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T: **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.** *Bioinformatics* 2009, **25:**1972-1973.

78. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30:**1312-1313.

79. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25:**1105-1111.

80. Anders S, Pyl PT, Huber W: **HTSeq A Python framework to work with high-throughput sequencing data.** 2014.

81. Si Y, Liu P, Li P, Brutnell TP: **Model-based clustering for RNA-seq data.** *Bioinformatics (Oxford, England)* 2014, **30:**197-205.

82. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2.** 2014.

83. Wickham H: **ggplot2.** *Wiley Interdisciplinary Reviews: Computational Statistics* 2011, **3:**180-185.

84. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics (Oxford, England)* 2005, **21:**3674-3676.

85. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier.** *Nucleic Acids Res* 2005, **33:**W116-120.

86. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association.** *Bioinformatics* 2007, **23:**257-258.

87. Kasprzyk A: **BioMart: driving a paradigm change in biological data management.** *Database (Oxford)* 2011, **2011:**bar049.

88. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25:**955-964.

89. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25:**1335-1337.

90. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database.** *Nucleic Acids Res* 2003, **31:**439-441.

91. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, et al: **Genome sequence of Aedes aegypti, a major arbovirus vector.** *Science* 2007, **316:**1718-1723.

92. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.

93. Frisch M, Frech K, Klingenhoff A, Cartharius K, Liebich I, Werner T: **In silico prediction of scaffold/matrix attachment regions in large genomic sequences.** *Genome Res* 2002, **12:**349-354.

94. Guy L, Kultima JR, Andersson SG: **genoPlotR: comparative gene and genome visualization in R.** *Bioinformatics* 2010, **26:**2334-2335.

95. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: **Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.** *Bioinformatics* 2010, **26:**2069-2070.

96. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M: **The Sequence Ontology: a tool for the unification of genome annotations.** *Genome Biol* 2005, **6:**R44.

Chapter 3: Complete dosage compensation in *Anopheles stephensi* and the evolution of sex-biased genes in mosquitoes

Authors and affiliations:

Xiaofang Jiang [1,2], James K. Biedler[2], Yumin Qi[2], Andrew Brantley Hall[1,2], Zhijian Jake Tu[1,2,*]

[1]Program of Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA, USA

[2]Department of Biochemistry, Virginia Tech, Blacksburg, VA, USA

*Author for Correspondence: Zhijian Jake Tu, Department of Biochemistry, Virginia Tech, Blacksburg, Virginia, United States of America, jaketu@vt.edu

Data deposition:

Data have been deposited in NCBI SRA under the accession SRP047470 and SRP055921

3.1 Author contributions

Conceived and designed experiments: ZT and XJ; Data analysis and presentation: XJ; Data generation: JB, YQ; Writing of the manuscript: XJ and ZT; Provided resources and tools and critical reviewed manuscript: ABH.

3.2 Abstract

Complete dosage compensation refers to hyper-expression of the entire X or Z chromosome in organisms with heterogametic sex chromosomes (XY male or ZW female) in order to compensate for having only one copy of the X or Z chromosome. Recent analyses suggest that complete dosage compensation, as in *Drosophila melanogaster*, may not be the norm. There has been no systematic study focusing on dosage compensation in mosquitoes. However, analysis of dosage compensation in *Anopheles* mosquitoes provides opportunities for evolutionary insights, as the X chromosome of *Anopheles* and that of its dipteran relative, *D. melanogaster* formed independently from the same ancestral chromosome. Furthermore, *Culicinae* mosquitoes, including the *Aedes* genus, have homomorphic sex-determining chromosomes, negating the need for dosage compensation. Thus, *Culicinae* genes provide a rare phylogenetic context to investigate dosage compensation in *Anopheles* mosquitoes. Here, we performed RNA-seq analysis of male and female samples of the Asian malaria mosquito *An. stephensi* and the yellow fever mosquito *Aedes aegypti*. Autosomal and X-linked genes in *An. stephensi* showed very similar levels of expression in both males and females, indicating complete dosage compensation. The uniformity of average expression levels of autosomal and X-linked genes remained when *An. stephensi* gene expression was normalized by that of their *Ae. aegypti* orthologs, strengthening the finding of complete dosage compensation in *Anopheles*. In addition, we comparatively analyzed the differentially expressed genes between adult males and adult females in both species, investigated sex-biased gene chromosomal distribution patterns in *An. stephensi* and provided three examples where gene duplications may have enabled the acquisition of sex-specific expression during mosquito evolution.

Key words: Comparative transcriptomes, RNA-seq, sex-specific expression, gene duplication

3.3 Introduction

Complete dosage compensation is a mechanism hypothesized to compensate for the loss of one copy of the X/Z chromosome in organisms with heterogametic sex chromosomes by hyper-expressing the entire X/Z chromosome (Ohno 1967; J E Mank 2013). Although complete dosage compensation has been demonstrated in model organisms such as *Drosophila melanogaster* (Straub and Becker 2007; Gelbart and Kuroda 2009), recent transcriptome analyses showed that dosage compensation is highly variable across species (Judith E Mank, Hosken, and Wedell 2011). Although a lack of complete dosage compensation for ZW chromosomes is observed in birds, blood flukes, and snakes (J. E. Mank and Ellegren 2008; Vicoso and Bachtrog 2011; Uebbing et al. 2013; Vicoso et al. 2013), there are also cases where ZW chromosomes displayed dosage compensation (Smith et al. 2014). In addition, there have been new challenges to the earlier conclusion that the eutherian X chromosome exhibits complete dosage compensation (Xiong et al. 2010; Deng et al. 2011; Lin et al. 2012). Analysis based on gene expression in patients with X aneuploidy syndrome showed that dosage sensitive genes are dosage compensated but the remainder of X-linked genes may not be (Pessia et al. 2012; Wright and Mank 2012). It has also been recently demonstrated that complete dosage compensation does not always mean the same level of gene expression in males and females. X-linked genes in male flour beetles are hyper-expressed and on average reached the same expression levels as autosomal genes. X-linked genes in female flour beetles are also hyper-expressed resulting in a higher level of expression of X chromosome genes compared to males (Prince, Kirkland, and Demuth 2010). However, the hyperexpression of the X chromosome in *Tribolium* females has recently been challenged (Mahajan and Bachtrog 2015). All these recent analyses make it clear that the mechanism of dosage compensation employed by one species may not be readily extrapolated to other species because dosage compensation evolves in concert with the formation of heteromorphic sex chromosomes, which could have independent origins and turn over rapidly (Judith E Mank, Hosken, and Wedell 2011).

The commonly accepted method to determine whether dosage compensation exists is to compare the overall expression level for X- or Z-linked genes to the overall expression level of autosomal genes in the heterogametic sex (J E Mank 2013). If the overall expression levels do not differ significantly, it can be assumed that complete dosage compensation exists. This method assumes that the overall proto-X / proto-Z chromosome expression level is and remains the same as the overall expression level of autosomes. However, these assumptions are not always valid. Proto-X/proto-Z chromosomes do not necessarily have the same overall expression level as other autosomes because the gene content and overall expression patterns differ between individual chromosomes. In addition, this method can be influenced by data processing and choice of statistical analysis if the number of lowly expressed genes differs between the X and autosome. A study of dosage compensation in placental mammals has been criticized for not filtering genes with low or no expression (Xiong et al. 2010; Kharchenko, Xi, and Park 2011; Deng et al. 2011). The conclusions of the aforementioned study can differ based on different filtering criteria (Jue et al. 2013).

A more reliable test for dosage compensation is to compare the expression of X/Z-linked genes to that of their autosomal or pseudoautosomal orthologs in related species. This approach has been applied in mammals (Lin et al. 2012) and in *D. pseudoobscura,* which has a neo-X chromosome (Nozawa et al. 2014).

Incomplete dosage compensation likely has caused the overrepresentation of sex-biased genes on the X/Z chromosomes in several species (Harrison, Mank, and Wedell 2012; Uebbing et al. 2013). In some species with complete dosage compensation, sex-biased genes are not randomly distributed among the sex chromosomes and autosomes. In *Drosophila,* male-biased genes are underrepresented on the X chromosome (Vicoso and Charlesworth 2009; Magnusson et al. 2012). Several different hypotheses have been proposed to explain the paucity of X-linked male-biased genes including: sexual antagonism, the effects of dosage compensation, and male meiotic sex chromosome inactivation (Meiklejohn et al. 2011).

Dosage compensation and the chromosomal distribution of sex-biased genes are well-characterized in *D. melanogaster* (Larschan et al. 2011; Magnusson et al. 2012). A recent publication showed that dosage compensation evolved multiple times, consistently through up-regulation of the single X in males during the numerous transitions of sex chromosomes in diverse fly taxa (Vicoso and Bachtrog 2015). However, these aspects of evolution have not been extensively studied in *Anopheles* mosquitoes, which belong to the same Dipteran order as *Drosophila,* but independently acquired the X chromosome (Toups and Hahn 2010; Pease and Hahn 2012). Unlike *Anopheles*, *Culicinae* mosquitoes including the *Aedes* and *Culex* genera, have a homomorphic sex-determining chromosome with pseudoautosomal regions spanning almost the entire length (Kitzmiller 1963; Hunter Jr and Hartberg 1986). Combined with research based on retrogene movement, it has been hypothesized that the ancestor of *Anopheles* and *Culicinae* mosquitoes lacked heteromorphic sex chromosomes (Toups and Hahn 2010). After the *Culicinae-Anophelinae* divergence, approximately 150 million years ago, heteromorphic X and Y chromosomes formed in *Anopheles* mosquitoes as the non-recombining region around the male-determining locus of the proto-Y expanded (Toups and Hahn 2010). The *Anopheles* X chromosome has persisted for ~100 million years and is present in all extant *Anopheles* species (Neafsey et al. 2015). In *Ae. aegypti*, male-determining gene is located on chromosome 1 (Andrew Brantley Hall et al. 2015). The p arm of chromosome 1 is mostly orthologous to X of *An. stephensi*, and the q arm of chromosome 1 is mostly orthologous to part of 2R of *An. stephensi* (Nene et al. 2007) .

Gene duplication is one common way to generate sex-biased genes (Parsch and Ellegren 2013). Genes can be duplicated through tandem duplication and retrotransposition. Following duplication, the expression pattern of the ancestral gene may remain unchanged, whereas the new duplicate can evolve sex-biased expression. Alternatively, the ancestral one may become specialized to one sex, whereas the new duplicate becomes biased to the other sex. Examples where genes obtained sex-biased expression post-duplication to resolve sex conflicts have only been reported in *Drosophila* and *mice* (Gallach and Betrán 2011; Connallon and Clark 2011; Chen et al. 2012).

Here, we sequenced adult male and female whole body transcriptomes of the Asian malaria mosquito *An. stephensi* and the yellow fever mosquito *Aedes aegypti* (Nene et al. 2007). First, we evaluated the X-to-autosome expression ratio in *An. stephensi* with different filtering criteria to access dosage compensation. Then we performed comparisons with the *Aedes* diploid orthologs of X-linked genes in *An. stephensi* to investigate *Anophelinae* dosage compensation with phylogenetic context. In addition, we also performed comparative analysis of the sex-biased genes in the two species. We have also identified several examples where gene duplication may have enabled the acquisition of sex-specific expression during mosquito evolution.

3.4 Material and Methods

3.4.1 RNA isolation and RNA sequencing

Mosquitoes that emerged over a 24 hr period were either directly collected as 0-1 day old adults, or isolated for later collection time points of 1-2 day old and 2-3 day old adults. For each time point, 5 mosquitoes were homogenized in 300 ul RNA lysis buffer (Zymo Research) and stored at -80 C until RNA isolation. For RNA isolation, equal volumes of homogenate from each time point were combined to represent 0-3 day old mosquitoes. These steps were performed in biological triplicates for males and virgin non-blood-feed females of both *An. stephensi* and *Ae. aegypti*. Illumina paired-end libraries for the resulting samples were prepared using the manufacturer's specific protocol. The libraries were then sequenced using Illumina HiSeq. The resulting samples have been submitted to the NCBI SRA under the accession SRP047470 and SRP055921.

3.4.2 Orthology and Chromosome Assignment

Orthology information was obtained from orthoDB (Waterhouse et al. 2013) (http://cegg.unige.ch/orthodbmoz2). Genomic scaffolds and gene annotations for the *An. stephensi* Indian strain *and Ae. aegypti* were downloaded from VectorBase (https://www.vectorbase.org/). *Ae. aegypti* genome version 3.2 and annotation version 3.2, *An. stephensi* genome version 2 and annotation 2.2 were used. Based on the one-to-one ortholog pairs between *An. stephensi* and *An. gambiae*, *An. stephensi* scaffolds were assigned to chromosome arms. Assignment requires that each scaffold has at least three genes, more than 92% of the genes on the scaffold are orthologous to the same chromosome element (Neafsey et al. 2015), and no more than three contiguous genes are orthologous to other chromosome elements. Chromosomal assignments based on these criteria are reliable in *Anopheles* because there is no large-scale inter-chromosomal gene movement during *Anopheles* evolution (Neafsey et al. 2015). In total, 190 scaffolds containing 11,413 of 12,350 total genes were assigned to chromosome arms. Previous publications reported physical mapping of 60% of the *An. stephensi* Indian strain genome (Jiang et al. 2014). All 155 gene-containing scaffolds based on the reported physical mapping were included and were in complete agreement with our orthology-based chromosomal assignments.

3.4.3 Dosage compensation analysis in *An. stephensi*

RNA-seq reads from triplicate male and female samples for both *An. stephensi* and *Ae. aegypti* were trimmed using trimmomatic (Bolger, Lohse, and Usadel 2014) with parameter "LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36" and then aligned to their respective genomes using Tophat2 (Kim et al. 2013). Read counts for each gene based on the six samples were generated using HTSeq (Anders, Pyl, and Huber 2014). We normalized the read count table through the RPKM (Reads per kb of sequence, per million mapped reads) (Mortazavi et al. 2008) approach to estimate expression level. Normalization was performed with the TMM method in the R package edgeR (Robinson et al. 2011). Read counts for genes were used to calculate the Spearman's correlation coefficient between samples. Because the replicates of the same sex were highly correlated (Supplementary Table S1), the triplicate RNA-seq data from each sex was combined as one single fastq file and the average RPKM values from the combined files were used for the following analysis:

Inactive genes (RPKM = 0 in both sample) and genes with low expression levels (0<RPKM<cutoff value) were removed from the analysis. Different cutoff values including 1 to 4 RPKM were used to define genes with low expression levels. The ratios of the median RPKM value of X-linked genes to the median RPKM value of autosomal genes in both males and females were calculated and used to assess whether dosage compensation is present in *An. stephensi*. The analyses were performed on unfiltered data sets as well as filtered data sets to explore how filtering and filtering

with different criteria affected the analysis. Two-sample Wilcoxon rank sum tests were applied to test the overall difference between X-linked and autosomal gene expression level.

One-to-one ortholog pairs of *An. stephensi* and *Ae. aegypti* were generated from orthoDB (Waterhouse et al. 2013). Of the 7,236 ortholog pairs, 7,035 were assigned to chromosome arms in *An. stephensi*. Genes with RPKM values less than 2 were removed in both species, leaving 5,096 ortholog pairs. The Spearman's correlations of the RPKM values between one-to-one ortholog pairs were examined (Supplementary Table S2). The ratio of *An. stephensi* gene RPKM value to their *Ae. aegypti* ortholog RPKM value were calculated for each ortholog pair. The ratios were linearly adjusted by the same factor to make the median expression levels of *An.* stephensi autosomal genes the same as the median expression levels of their orthologs.

3.4.4 Sex-biased gene analysis

Three commonly used tools: CuffDiff (Trapnell et al. 2012), DESeq2 (Love, Huber, and Anders 2014), and edgeR (Robinson et al. 2011), were used for statistical analysis of differential expression between males and females. Read count tables for each triplicate male and female sample were used as input for both edgeR and DESeq2. The Tophat output from our previous analysis was further processed by Cufflinks and then differentially expressed genes between males and females were identified by CuffDiff. Genes that were detected as differentially expressed by at least two of the three metrics (FDR<0.05 for edgeR, pvalue<0.05 for cuffdiff, qvalue<0.05 for DESeq2) were used as the final set of sex-biased genes in our analysis. The analysis was performed using the triplicate RNA-seq data from *An. stephensi* and *Ae. aegypti.*

The expression level bias between two sexes for individual genes was estimated by the magnitude of the difference in expression levels between the sexes. Sex-biased genes were divided into groups for further analysis based on the magnitude of the difference in the expression levels between the sexes. All protein sequences of *An. stephensi* and *Ae. aegypti* were used as input for Blast2GO (Conesa et al. 2005) to retrieve GO (Gene Ontology) term information. Over-represented GO terms for each group were identified using a hypergeometric test using the GOstats package (Beißbarth and Speed 2004) in R.

3.4.5 Phylogenetic inference

Culicidae orthologous protein sequences of the genes of interest were retrieved from orthoDB and fragmented sequences were removed manually. Sequences were aligned using MUSCLE (Edgar 2004) with default parameters. Alignments were trimmed with trimAl (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) with the parameter "-gt 0.8" to exclude alignment columns with gaps presented in more than 20% of the sequence. The trimmed alignments were used as input for Mrbayes (Huelsenbeck 2001), a program for Bayesian estimation of phylogeny. The rate matrix for amino acid data was set as "mixed" for Mrbayes analysis. Mrbayes performed a Markov Chain Monte Carlo (MCMC) analysis for 1,000,000 generations with four chains with the temperature set to 0.2. The resulting consensus tree was visualized with FigTree (Rambaut 2009). Phylogenetic trees were built to infer the relative time of duplication with respect to speciation of mosquitoes. The trees also assisted to distinguish the ancestral gene and derived duplicates.

3.5 Results

3.5.1 Complete dosage compensation in *An. stephensi* as shown by the X-to-autosome expression ratio

Based on the chromosomal location of *An. gambiae* orthologs, 190 *An. stephensi* scaffolds were assigned to chromosome arms. In total, 1,029 genes were assigned to the X chromosome, and 9,933 genes were assigned to autosomes (Supplementary Table S3). Gene expression levels in RPKM were estimated from triplicate male and female RNA-seq samples. Correlations between samples of the same sex were statistically significant for both males and females (Spearman’s correlation >0.95). Thus, an average expression level was used for comparisons between the two sexes.

The most common method to determine whether dosage compensation is present is to compare the overall expression level of genes on the X/Z chromosome to that of autosomal genes in the heterogametic sex. In *An. stephensi*, this is the X-to-autosome (X:AA) ratio of gene expression in males. Genes with low or no expression can have large effects on the X:AA ratio of gene expression. Consequently, we removed genes with RPKM values below various arbitrary cutoffs and then used the median RPKM values for the X-linked and autosomal genes to derive the X:AA ratio. If there is complete dosage compensation, we expect the X:AA ratio to be approximately 1. In theory when there is not dosage compensation, the X:AA ratio should be 0.5. However, due to the buffering effects of gene regulatory networks, even without complete dosage compensation the ratio could be 0.6-0.75 (Judith E. Mank 2009; Harrison, Mank, and Wedell 2012; Wright, Moghadam, and Mank 2012).

In males, no matter the filtering criteria used, the X:AA ratio was always greater than 0.94 indicating that there was complete dosage compensation. We used the Wilcoxon rank sum test to evaluate whether there was a statistical difference between the expression level of X-linked and autosomal genes. We found no statistical difference between the expression level of X-linked and autosomal genes when no filtering or filtering greater than 2 RPKM was applied (Table 3.1 and Supplementary Table S4).

We chose 2 as the RPKM cutoff for the analysis so that there was enough filtering stringency to remove noise and at the same time retain the maximum number of genes (Figure 3.1). We also performed the same analysis on individual replicates of male and female RNA-seq samples and observed similar results (Supplementary Figure S1). Taken together, these results strongly suggest there is complete dosage compensation in *An. stephensi.*

3.5.2 Complete dosage compensation in *An. stephensi* as shown by the expression ratio of *An. stephensi* X-linked genes to their *Ae. aegypti* orthologs

Although many studies have used the X:AA (or Z:AA) ratio to assess the presence of dosage compensation, concerns have been brought forth (J E Mank 2013). X-linked genes are not homologous to autosomal genes so the difference in gene content may result in differences in chromosome-wide gene expression levels independent of dosage compensation. These problems can be mitigated by comparing the expression levels of X/Z-linked genes to those of their autosomal or pseudoautosomal orthologs in related species. Here, we calculated the ratio of the expression level of X-linked genes in *An. stephensi* to their one-to-one orthologs in *Ae. aegypti* (X:$\underline{XX}$ expression ratio) as well as the expression level of autosomal genes in *An. stephensi* to their one-to-one orthologs in *Ae. aegypti* (AA:$\underline{AA}$) to assess whether dosage compensation is present in *An. stephensi*.

Orthologous genes may have different lengths between species, so we used normalized RPKM values. Genes with RPKM values less than 2 were removed. In total, 5,096 orthologous gene pairs were identified including 421 X-linked genes in *An. stephensi*, providing ample data for our analysis. The overall expression levels of one-to-one orthologs are strongly correlated (Supplementary Table S2), with a correlation of 0.52 from female RNA-seq and 0.64 from male RNA-seq. The correlation is lower for X-linked genes compared with autosomal genes in both males (0.59) and females (0.48) but is still correlated. These lower correlation values may be a result of more adaptive selection pressure acting on the X chromosome causing higher X-linked divergence (faster-X effect) as has been observed in embryos and adult *Drosophila* (Kayserili et al. 2012).

We calculated the RPKM ratio for each pair of orthologs in *An. stephensi* and *Ae. aegypti*. We normalized the median RPKM ratio of *An. stephensi* autosomal genes to their orthologs in *Ae. aegypti* to 1 to adjust for differences in the overall expression levels between the species. After normalization, we observed that the median RPKM ratio of *An. stephensi* X-linked genes to their orthologs in *Ae. aegypti* was close to 1 in both sexes (Figure 3.2). These results indicate that there is dosage compensation for the X chromosome in male *An. stephensi*. The female X-linked gene expression level remained the same indicating that the dosage compensation mechanism is either exclusive to males or has been repressed in females.

3.5.3 Sex-biased genes in *An. stephensi* and their chromosomal distribution

We used three commonly used tools: Cuffdiff, DEseq2, and edgeR to identify differentially expressed genes between male and female samples in *An. stephensi*. We then used a Venn diagram to identify genes that were classified as differentially expressed using all three tools. Of the genes identified as differentially expressed by each method, more than 75% were identified as differentially expressed by the other methods (Figure 3.3). The overlap was greatest between DESeq2 and edgeR (2,018 for female-biased genes; 1,825 for male-biased genes), perhaps due to the similarity of the statistical approaches employed by these two methods. Here, we selected genes that were identified as differentially expressed by at least two methods for further analysis. In *An. stephensi*, 2,112 genes were identified as female-biased and 1,933 genes were identified as male-biased. In *Ae. aegypti*, 3,567 genes were identified as female-biased and 3,660 genes were identified as male-biased. Ninety percent of female-biased genes and 82.7 percent of male-biased genes in *An. stephensi* have orthologs in *Ae. aegypti* (Supplementary Table S5).

We further categorized the sex-biased genes based on the magnitude of the difference in expression level between the sexes (Figure 3.4 A,Figure 3.4 B). We then analyzed the gene ontology (GO) terms enriched in the highly sex-biased genes. The most female-biased genes (log2 RPKM female to male ratio >4) were enriched for genes with molecular functions such as serine-type peptidase activity， proteolysis, and odorant binding, which is consistent with specialization of female mosquitoes for blood-feeding and subsequent blood-meal digestion. For the most part, male-biased genes (log2 RPKM male to female ratio >4) were overrepresented with GO terms such as microtubule-based movement， nucleosome assembly, and dynein complex, indicating involvement in spermatogenesis (Additional File 1).

Previous research has indicated that the distribution of sex-biased genes is non-random between sex chromosomes and autosomes (Parisi et al. 2003; Sturgill et al. 2007; Jaquiéry et al. 2013; Albritton et al. 2014). Magnusson et al. indicated demasculinization of the X chromosome in *An. gambiae* based on microarray data (Magnusson et al. 2012). To test whether the same pattern is observed in *An. stephensi* based on RNA-seq analysis, we investigated the chromosomal distribution of sex-biased genes.

When we set the threshold below a 3-fold difference between the sexes to define the term sex-biased, we observed no obvious demasculinization of the *An. stephensi* X chromosome (Figure 3.4 C). However, when we increased the threshold to more than a 3-fold difference, we began to see an obvious demasculinization of the X chromosome similar to what was observed in *An. gambiae*. We observed both feminization and demasculinization of the X chromosome, when the threshold was greater than an 11.5-fold difference between males and females.

3.5.4 Examples of sex-specific subfunctionalization post gene duplication

Sex-biased genes often arise from gene duplications. A gene that codes for the protein actin, *Actin-4,* has been well-characterized in *Ae. aegypti* and *Ae. albopictus* (Muñoz et al. 2004; Fu et al. 2010; Labbé et al. 2012). In *Ae. aegypti*, the *Actin-4* gene (*AAEL001951*) has two isoforms: the female isoform, which is highly expressed, and the male isoform, which is expressed at a lower level than the female isoform and codes for a non-functional protein. Our RNA-seq data show that *Actin-4* is female-biased in adults, and its paralog *Actin-3* (*AAEL009451*) is male-biased (Vyazunova and Lan 2004). In *An. stephensi*, there are two *Actin-4* orthologs: *ASTEI10165,* which is extremely female-biased, and *ASTEI03074*, which is extremely male-biased (Supplementary Table S6). We retrieved all available *Actin-4* orthologs from OrthoDB (Group MZ20123647) and built a phylogenetic tree (Supplementary Figure S2). As the tree shows, the duplication and sub-functionalization of this family of actin genes likely occurred before the divergence of *Aedes* and *Anopheles*.

A recent publication (A B Hall et al. 2014) identified a gene, *myo-sex*, which is tightly linked to the M-locus, is male-specific, and highly expressed in pupae of *Ae. aegypti*. *Myo-sex* and its paralog *AAEL005656* likely originated from duplications of *AAEL005733* (A B Hall et al. 2014). Based on our RNA-seq data, *myo-sex* is male-specific in adults and *AAEL005733* is male-biased with a two-fold greater expression in adult males than in adult females. *AAEL005656* is extremely female-biased (Supplementary Table S6). In *Anopheles*, only one ortholog of *AAEL005733* (group MZ20123647) exists, suggesting there has not been a duplication. In *An. stephensi*, the ortholog of *AAEL005733* is *ASTEI08310*. The expression pattern of *ASTEI08310* in both sexes is similar to that of *AAEL005733*, which may suggest that in *Ae. aegypti* the expression profile of the parental gene *AAEL005733* remains unchanged, whereas the two new copies specialized, each to a different sex.

Another example of duplicated genes becoming specialized in two sexes is the orthology group MZ22302531. Genes in this group are orthologs to venom allergens of wasps and fire ants, and belong to the family of cysteine-rich secretory proteins (Lu et al. 1993). This family is also related to mammalian testis-specific protein (Tpx-1), which is required for sperm capacitation (Kasahara et al. 1989). Based on RNA-seq data, the expression levels of paralogs of this group vary significantly between adult males and females. Of the four paralogs in *An. stephensi, ASTEI10265* is highly female-biased in adults, *ASTEI10266* is male-specific in adults, and the other two paralogs are barely expressed in adults (Supplementary Table S6). In *Ae. aegypti*, in which there are six orthologs, *AAEL000793* is female-specific, *AAEL002693* is male-biased, *AAEL009239* is male-specific, and the rest are not significantly expressed in adults. We also checked the microarray data of *An. gambiae* on VectorBase. There are six paralogs in *An. gambiae*, and the genomic location indicates that all should have arisen as tandem duplications of the same ancestral gene. Also, the expression levels of these six paralogs vary between adult males and females: two are female-biased and two are male-biased. Based on phylogenic analysis (Supplementary Figure S3), the sub-functionalization of venom allergens occurred independently in *Aedes* and *Anopheles.*

3.6 Discussion

3.6.1 Dosage compensation

We used RNA-seq to provide conclusive evidence that *An. stephensi* has complete dosage compensation. Dosage compensation is thought to evolve in concert with the formation of heteromorphic sex chromosomes (Charlesworth 1996; J E Mank 2013). Consequently, it is reasonable to assume that all *Anopheles* mosquitoes have dosage compensation and implement dosage compensation by the same or similar mechanisms because all *Anopheles* species share the same X chromosome (Neafsey et al. 2015). However, since the X chromosomes of *Anopheles* and *Drosophila* evolved independently (Toups and Hahn 2010), it is likely that their mechanisms for dosage compensation also evolved independently. In *Drosophila* species, dosage compensation is implemented by doubling the expression level of genes on the X chromosome in males (Conrad and Akhtar 2011). Dosage compensation in *Anopheles* could result from either the doubling of the expression level of X-linked genes in only males, or by doubling the expression level of X-linked genes in both males and females and subsequently silencing the expression of one X chromosome in females.

In *Drosophila*, the sex-lethal (*Sxl*) gene controls dosage compensation via male specific lethal-2 (MLS2) through the (Male-specific lethal) MSL-complex (Penalva and Sanchez 2003). Besides MSL-2, the MSL-complex includes males absent on the first (MOF), MSL1, MSL3, maleless (MLE), and the roX1 or roX2 non-coding RNAs (Conrad and Akhtar 2011). Even though some orthologs of these genes exist in some of the *Anopheles* genomes, these genes may perform different functions in *Anopheles* (Zdobnov 2002; Behura et al. 2011). For example, the ortholog of sex-lethal in *Anopheles* mosquitoes is not sex-specifically alternatively spliced and doesn’t function in sex determination or the initiation of dosage compensation (Traut et al. 2006). However, because dosage compensation may evolve by modifying existing epigenetic regulation systems, research focused on conserved proteins involved in epigenetic networks may provide insights into the mechanism of dosage compensation in *Anopheles* (Graves 2014). Genes involved in the sex-determination pathway are often also involved in dosage compensation, such as in *Drosophila* species and *Bombyx mori* (Penalva and Sanchez 2003; Kiuchi et al. 2014).

Drosophila and *Anopheles* appear to have acquired dosage compensation independently. This example of convergent evolution may be attributed to shared features between these two families. First, both X chromosomes evolved from the same pair of autosomes, meaning that the same dosage-sensitive genes may have caused the need for dosage compensation in both species. Second, both families have relatively large effective population sizes. Large effective population size provides high genetic diversity, which may have led to a quick adaptive dosage compensation mechanism. Lastly, both genera are male heterozygotic. Although dosage compensation exists in ZW species (Smith et al. 2014), it is relatively more common in XY species due to several potential reasons: mutations occur more frequently in males, the X chromosome effective population size increases due to sexual selection, and stronger natural selection acts on males (J E Mank 2013).

3.6.2 Sex-biased genes and their chromosomal distribution

Sex-biased genes are generally identified through comparing male and female samples. Therefore, the number of sex-biased genes detected is dependent on factors such as: species, sampled tissue and experimental and analytical methodology. Research on sex-biased genes in *An. gambiae* showed that because testes in males are proportionately smaller than ovaries in blood-fed females, testes-enriched genes were significantly underrepresented and ovary-enriched genes were highly overrepresented in the sex-biased genes detected from the whole body samples (Baker et al. 2011; Baker and Russell 2011). Here we compare transcriptomes from adult male and adult virgin non-blood-fed female mosquitoes, which have much smaller ovaries than blood-fed females. Thus, the effect of allometry is not as pronounced as in comparison between males and blood-fed females. Future transcriptomic analysis of sex-, stage-, and tissue-specific samples of *An. stephensi* will enable the distinction between sex-biased genes detected in this research that result from overall sex-biased expression and those caused by tissue-specific expression.

Nonrandom distributions of sex-biased genes have been observed in several species (Vicoso and Bachtrog 2015). However, in 2005, Hahn and Lanzaro did not identify non-random distributions of sex-biased genes in *An. gambiae* (Hahn and Lanzaro 2005). Conversely, Magnusson et. al reported a deficit of male-biased genes in the *An. gambiae* X chromosome (Magnusson et al. 2012). Our data shows that the X chromosome is depleted of genes with highly male-biased expression in *An. stephensi.* However, for genes with low sex-biased expression (less than 2-fold difference between the sexes), the trend does not exist. Highly male-biased genes are primarily expressed in the gonads, so the nonrandom distribution of highly male-biased genes may be attributed to specific tissues. This is consistent with GO terms associated with extremely male-biased genes (F/M > 16 or M/F >16). This explanation is also consistent with the previous evidence of underrepresentation of testis-expressed genes on the *An. gambiae* X chromosome (Baker and Russell 2011). Further experiments like those done in several flies (Baker et al. 2011; Vicoso and Bachtrog 2015), where transcriptomes of adult female and male soma are sequenced will help to evaluate the contribution of gonads to the chromosomal distribution of sex-biased genes in *An. stephensi*.

3.6.3 Gene duplications in the evolution of sex-biased genes

We have shown that gene duplication is one mechanism that leads to the formation of new sex-biased genes and we present three examples of orthologous groups where duplicated gene copies become specialized, each to a different sex. Although the examples are identified from whole adult bodies, the sex-biased expression is likely due to tissue dimorphism. The sex-specific venom allergens are likely expressed in salivary glands based on previous research on their orthologs (Arcà et al. 2005), while the sex-specific genes actin and myosin are muscle-specific (Vyazunova and Lan 2004; A B Hall et al. 2014; Labbé et al. 2012). This is consistent with the findings that most of the genes that are specialized to one sex also become tissue-specific (Gallach and Betrán 2011; Chen et al. 2012; Wyman, Cutter, and Rowe 2012). As sexually antagonistic conflicts can be tissue-specific and stage-specific, duplicated genes can evolve specific expression profiles to solve the conflicts. Therefore, future transcriptomic studies on tissue- and stage-specific transcription in both males and females will provide a high resolution view of how sexually antagonistic conflicts affect mosquito gene expression and duplication.

Actin-4 is expressed specifically in female flight muscles and has the highest level of expression during the pupal stage (Labbé et al. 2012). Our analysis shows that that *Actin-4* has female-biased expression in both adult *Ae. aegypti* and *An. stephensi*. The male-biased paralog of *Actin-4*, *Actin-3,* also has conserved expression in both species. *Actin-3* may perform a similar, but male-specific, function to *Actin-4*. Interestingly, despite the huge difference in the expression profiles, these two actin genes only differ by four amino acids, indicating that small mutations between these two genes may have large functional consequences. The interactions of actin and myosin are crucial for functions in the cell, including movement. Thus, actin and myosin genes may coevolve. The sex-specific sub-functionalization of the duplicated myosin genes in *Ae. aegypti* may have been partially triggered by the sex-specialization of actin genes. Nevertheless, more studies on the expression profiles and functional characterizations will further our understanding of these genes.

3.6.4 Potential vector control applications of dosage compensation and sex-specific genes

Mosquitoes transmit pathogens to humans and livestock. For example, *An. gambiae* is the primary malaria vector in Africa and *An. stephensi* is the key vector of urban malaria on the Indian subcontinent (Singh et al. 1999). *Ae. aegypti* is a major vector of dengue fever, yellow fever, and *chikungunya* (Nene et al. 2007). Only female mosquitoes feed on blood and transmit disease, while males are harmless. Dosage compensation functions on a sex-to-sex basis. Thus, manipulation of genes involved in dosage compensation can potentially result in female lethality by distorting X-linked gene expression dosage. In addition, the study of sex-biased genes will shed light on mosquito sex determination and sexual differentiation, processes that can be used in novel genetic approaches for vector control. For example, the female-specific promoter of Actin-4 has been used to create flightless female mosquitoes as a vector control strategy (Fu et al. 2010; Labbé et al. 2012). Other sex-biased genes such as the myosin genes identified in this study could be used in genetic vector control strategies.

3.7 Acknowledgements

This work is supported by the Fralin Life Science Institute and the Virginia Experimental Station, and by National Institutes of Health (NIH) grants AI77680, AI105575, and AI113643. We thank Janet Webster for critical reading of the manuscript and Thomas Randy Saunders for mosquito care.

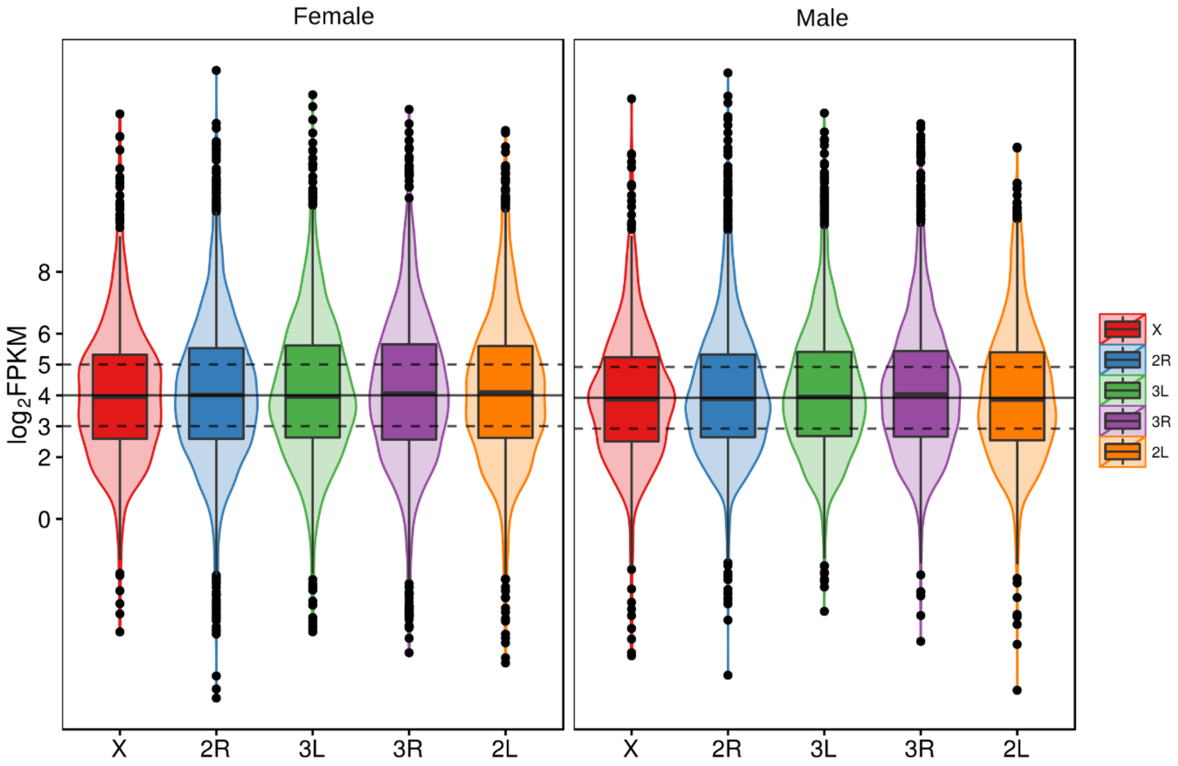

Figure 3.1 The distribution of log2 transformed RPKM values of genes on different autosomal arms and the X chromosome in males and females.

Inactive and low-expressed genes (genes with RPKM value less than two in one of the samples) were removed in this analysis. The width of the violin plots shows the density of genes at different log2 RPKM values. Boxplots are also shown in which the bottom and top of the box are the first and third quartiles, and the solid band inside the box is the median. The solid black horizontal line in each panel represents the median log2 RPKM value of autosomes in the corresponding sample. Dashed black horizontal lines above and below the black lines represent +1 and -1 of median log2 RPKM.

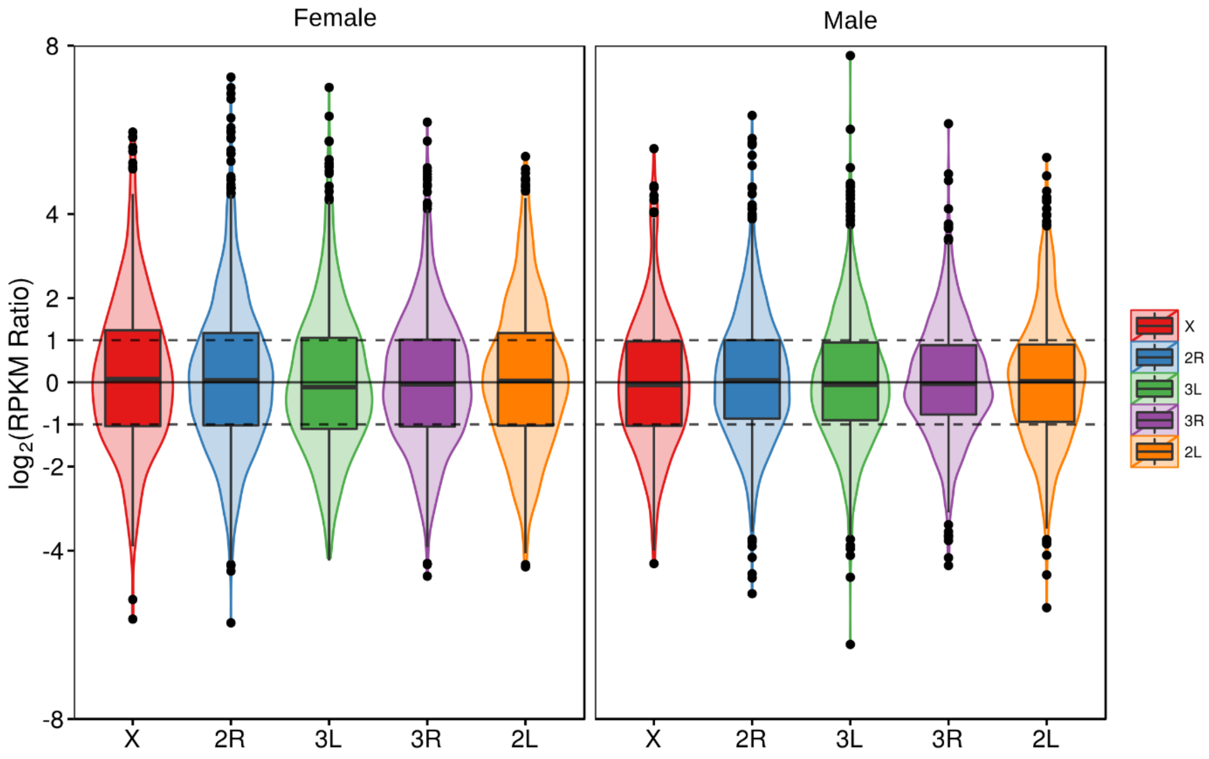

Figure 3.2 The distribution of the log2 normalized ratio of RPKM values in *An. stephensi* to their one-to-one orthologs in *A. aegypti* on different chromosome arms in males and females.

The width of the violin plots shows the density of genes at different log2 RPKM ratios. Boxplots are also shown in which the bottom and top of the box are the first and third quartiles, and the solid band inside the box is the median. The solid black horizontal line in each panel represents 0 in the corresponding sample. The Dashed black horizontal lines above and below the black line represent +1 and -1.

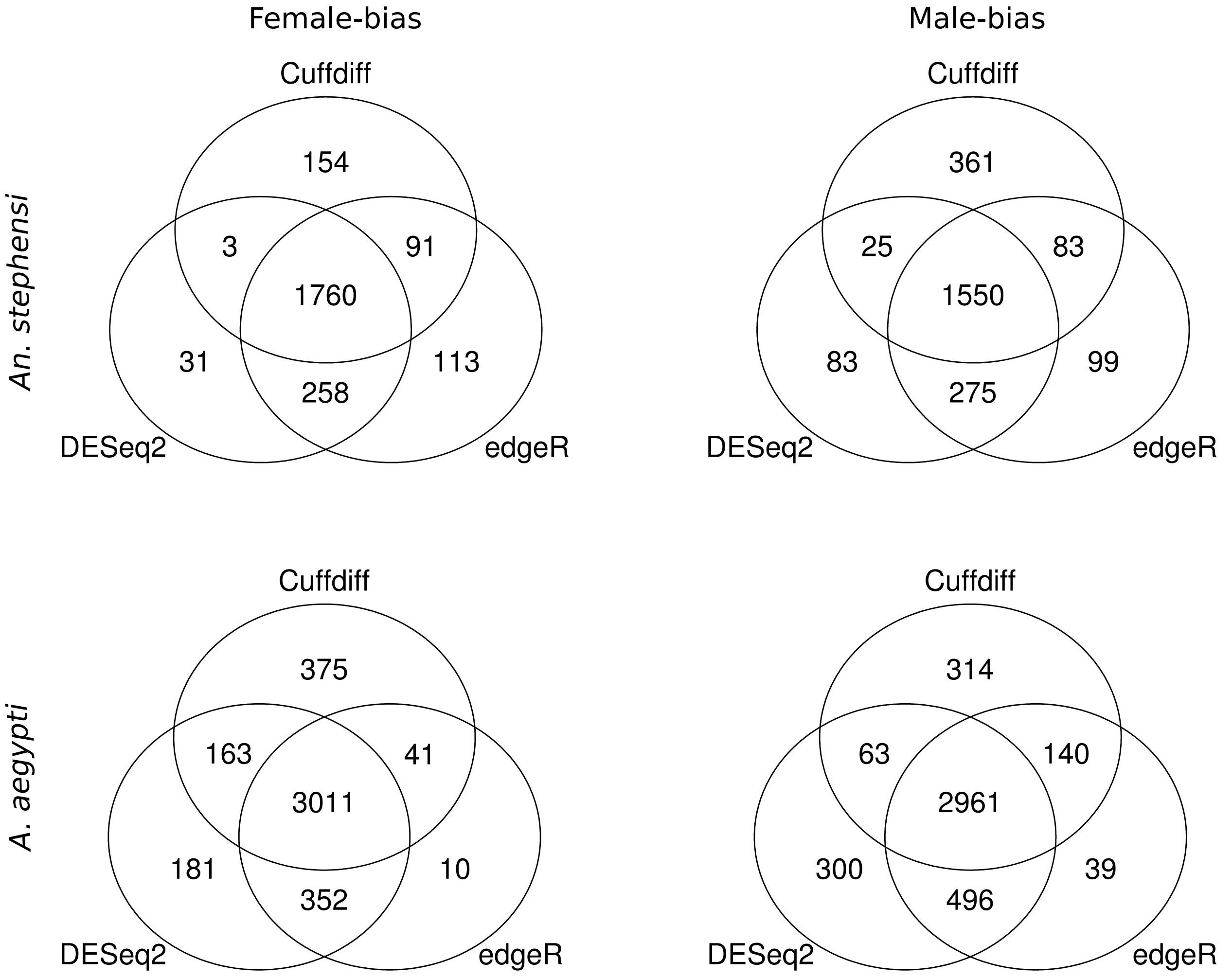

Figure 3.3 Venn diagrams of the overlap of differentially expressed genes based on Cufdiff, DESeq2 and edgeR.

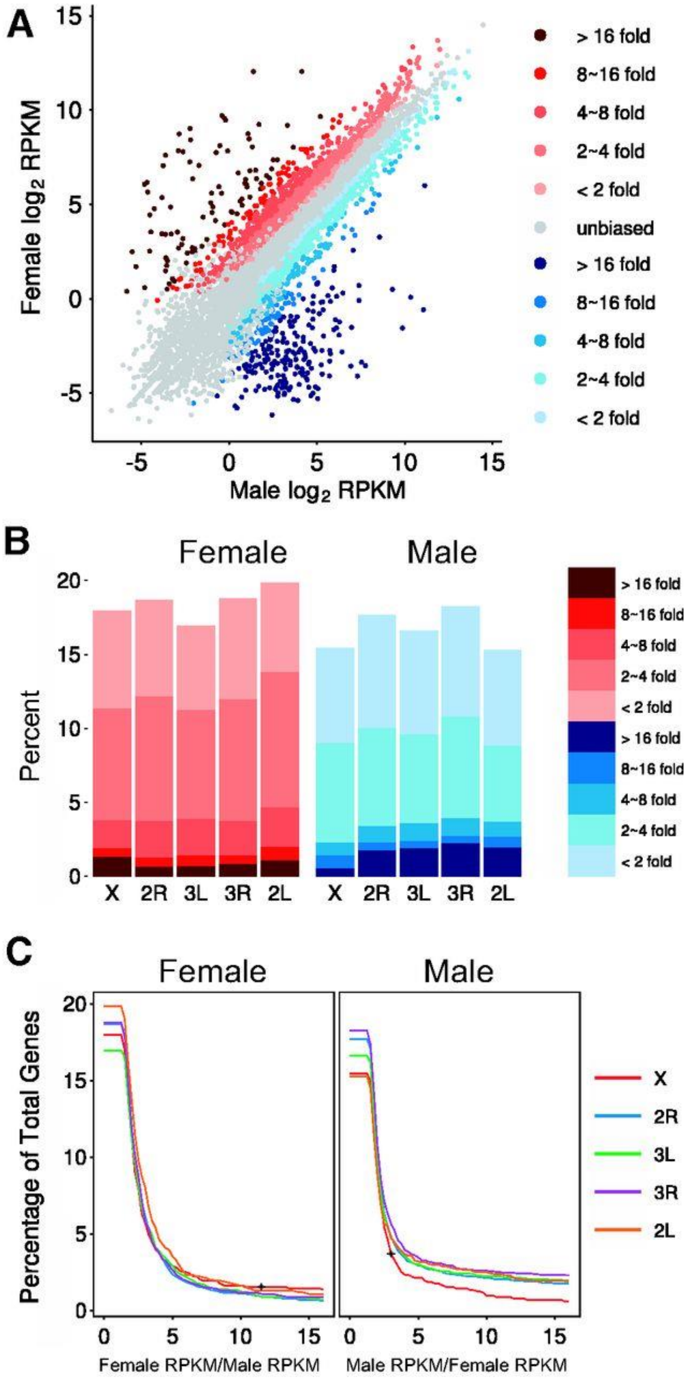

Figure 3.4 Distribution of sex-biased genes in *An. stephensi*

*(*A) Genome-wide sex-biased gene expression in *An. stephensi*. Darker shades of red represent greater female-biased expression. Darker shades of blue represent greater male-biased expression. (*B*) Percentage of sex-biased genes on five chromosomal arms. Left panel: female-biased genes;

Right panel: male-biased genes. Darker shades of red represent greater female-biased expression. Darker shades of blue represent greater male-biased expression. (*C*) Percentage of total genes identified as sex-biased at different magnitudes of sex-bias between the sexes on individual chromosome arms. The *x*axis indicates female to male ratio (left panel, panel female) or male to female ratio (right panel, panel male) of gene expression levels. The asterisk in Female panel indicates the cutoff ratio (11.5) above which permutation tests showed X chromosome feminization. The asterisk in Male panel indicates the cutoff ratio (3) above which permutation tests showed X chromosome demasculinization.

Table 3.1 Effect of the stringency of the expression level cutoff on the median gene expression of the X chromosome and autosomes.

	# of genes remained		Female				Male			
	X	Autosomes	X RPKM	Autosome RPKM	P-value*	XX:AA ratio	X RPKM	Autosome RPKM	P-value*	X:AA ratio
Original	1029	9933	10.67	11.49	0.48	0.93	10.91	11.23	0.10	0.97
Remove genes = 0 RPKM	1012	9719	11.11	12.04	0.32	0.92	11.29	11.79	0.05	0.96
Remove genes < 1 RPKM	927	8901	13.41	14.39	0.20	0.93	13.16	13.82	0.03	0.95
Remove genes < 2 RPKM	869	8440	15.42	16.03	0.33	0.96	14.61	15.18	0.08	0.96
Remove genes < 3 RPKM	826	8031	16.93	17.77	0.25	0.95	15.69	16.63	0.08	0.94
Remove genes < 4 RPKM	785	7680	18.78	19.23	0.30	0.98	16.91	17.81	0.12	0.95

*P-values were calculated based on Two-sample Wilcoxon rank sum tests.

3.8 References

Albritton, Sarah Elizabeth, Anna Lena Kranz, Prashant Rao, Maxwell Kramer, Christoph Dieterich, and Sevinç Ercan. 2014. “Sex-Biased Gene Expression and Evolution of the X Chromosome in Nematodes.” *Genetics* 197 (3): 865–83. doi:10.1534/genetics.114.163311.

Anders, S., P. T. Pyl, and W. Huber. 2014. “HTSeq A Python Framework to Work with High-Throughput Sequencing Data.” *bioRxiv*. doi:10.1101/002824.

Arcà, Bruno, Fabrizio Lombardo, Jesus G Valenzuela, Ivo M B Francischetti, Osvaldo Marinotti, Mario Coluzzi, and José M C Ribeiro. 2005. “An Updated Catalogue of Salivary Gland Transcripts in the Adult Female Mosquito, Anopheles Gambiae.” *The Journal of Experimental Biology* 208 (Pt 20): 3971–86. doi:10.1242/jeb.01849.

Baker, Dean A, Tony Nolan, Bettina Fischer, Alex Pinder, Andrea Crisanti, and Steven Russell. 2011. “A Comprehensive Gene Expression Atlas of Sex- and Tissue-Specificity in the Malaria Vector, Anopheles Gambiae.” *BMC Genomics* 12: 296. doi:10.1186/1471-2164-12-296.

Baker, Dean A, and Steven Russell. 2011. “Role of Testis-Specific Gene Expression in Sex-Chromosome Evolution of Anopheles Gambiae.” *Genetics* 189 (3): 1117–20. doi:10.1534/genetics.111.133157.

Behura, Susanta K., Morgan Haugen, Ellen Flannery, Joseph Sarro, Charles R. Tessier, David W. Severson, and Molly Duman-Scheel. 2011. “Comparative Genomic Analysis of Drosophila Melanogaster and Vector Mosquito Developmental Genes.” *PLoS ONE* 6 (7): e21504. doi:10.1371/journal.pone.0021504.

Beißbarth, Tim, and Terence P. Speed. 2004. “GOstat: Find Statistically Overrepresented Gene Ontologies with a Group of Genes.” *Bioinformatics* 20: 1464–65. doi:10.1093/bioinformatics/bth088.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20. doi:10.1093/bioinformatics/btu170.

Capella-Gutiérrez, Salvador, José M Silla-Martínez, and Toni Gabaldón. 2009. “trimAl Tutorial.” *Bioinformatics (Oxford, England)* 25: 1972–73. doi:10.1093/bioinformatics/btp348.

Charlesworth, B. 1996. “The Evolution of Chromosomal Sex Determination and Dosage Compensation.” *Current Biology : CB* 6 (2): 149–62. doi:10.1016/S0960-9822(02)00448-7.

Chen, Sidi, Xiaochun Ni, Benjamin H. Krinsky, Yong E. Zhang, Maria D. Vibranovski, Kevin P. White, and Manyuan Long. 2012. “Reshaping of Global Gene Expression Networks and Sex-Biased Gene Expression by Integration of a Young Gene.” *The EMBO Journal* 31 (12). Nature Publishing Group: 2798–2809. doi:10.1038/emboj.2012.108.

Conesa, Ana, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. 2005. “Blast2GO: A Universal Tool for Annotation, Visualization and Analysis in Functional Genomics Research.” *Bioinformatics (Oxford, England)* 21: 3674–76. doi:10.1093/bioinformatics/bti610.

Connallon, Tim, and Andrew G Clark. 2011. “The Resolution of Sexual Antagonism by Gene Duplication.” *Genetics* 187 (3): 919–37. doi:10.1534/genetics.110.123729.

Conrad, Thomas, and Asifa Akhtar. 2011. “Dosage Compensation in Drosophila Melanogaster: Epigenetic Fine-Tuning of Chromosome-Wide Transcription.” *Nature Reviews. Genetics* 13 (2): 123–34. doi:10.1038/nrg3124.

Deng, X, J B Hiatt, D K Nguyen, S Ercan, D Sturgill, L W Hillier, F Schlesinger, et al. 2011. “Evidence for Compensatory Upregulation of Expressed X-Linked Genes in Mammals, Caenorhabditis Elegans and Drosophila Melanogaster.” *Nat Genet* 43 (12): 1179–85. doi:10.1038/ng.948.

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32: 1792–97. doi:10.1093/nar/gkh340.

Fu, Guoliang, Rosemary S Lees, Derric Nimmo, Diane Aw, Li Jin, Pam Gray, Thomas U Berendonk, et al. 2010. “Female-Specific Flightless Phenotype for Mosquito Control.” *Proceedings of the National Academy of Sciences of the United States of America* 107: 4550–54. doi:10.1073/pnas.1000251107.

Gallach, Miguel, and Esther Betrán. 2011. "Intralocus Sexual Conflict Resolved through Gene Duplication." *Trends in Ecology and Evolution* 26 (5): 222–28. doi:10.1016/j.tree.2011.02.004.

Gelbart, Marnie E, and Mitzi I Kuroda. 2009. "Drosophila Dosage Compensation: A Complex Voyage to the X Chromosome." *Development (Cambridge, England)* 136 (9): 1399–1410. doi:10.1242/dev.029645.

Graves, Jennifer A. Marshall. 2014. "The Epigenetic Sole of Sex and Dosage Compensation." *Nature Genetics* 46 (3): 215–17. doi:10.1038/ng.2903.

Hahn, Matthew W, and Gregory C Lanzaro. 2005. "Female-Biased Gene Expression in the Malaria Mosquito Anopheles Gambiae." *Current Biology : CB* 15 (6): R192–93. doi:10.1016/j.cub.2005.03.005.

Hall, A B, V A Timoshevskiy, M V Sharakhova, X Jiang, S Basu, M A Anderson, W Hu, I V Sharakhov, Z N Adelman, and Z Tu. 2014. "Insights into the Preservation of the Homomorphic Sex-Determining Chromosome of Aedes Aegypti from the Discovery of a Male-Biased Gene Tightly Linked to the M-Locus." *Genome Biol Evol* 6 (1): 179–91. doi:10.1093/gbe/evu002.

Hall, Andrew Brantley, Sanjay Basu, Xiaofang Jiang, Yumin Qi, Vladimir A Timoshevskiy, James K Biedler, Maria V Sharakhova, et al. 2015. "A Male-Determining Factor in the Mosquito Aedes Aegypti." *Science*, no. May: 1–7.

Harrison, Peter W, Judith E Mank, and Nina Wedell. 2012. "Incomplete Sex Chromosome Dosage Compensation in the Indian Meal Moth, Plodia Interpunctella, Based on de Novo Transcriptome Assembly." *Genome Biology and Evolution* 4 (11): 1118–26. doi:10.1093/gbe/evs086.

Huelsenbeck, John P. 2001. "MrBayes : A Program for the Bayesian Inference of Phylogeny." *Dna Sequenc* 17 (4): 1–12.

Hunter Jr, Robert D, and W Keith Hartberg. 1986. "Observations on the Mitotic Chromosomes of the Mosquito Toxorhynchites Amboinensis (Doleschall)." *Mosq. Syst* 18 (2): 119–24.

Jaquiéry, Julie, Claude Rispe, Denis Roze, Fabrice Legeai, Gaël Le Trionnaire, Solenn Stoeckel, Lucie Mieuzet, et al. 2013. “Masculinization of the X Chromosome in the Pea Aphid.” *PLoS Genetics* 9 (8): e1003690. doi:10.1371/journal.pgen.1003690.

Jiang, Xiaofang, Ashley Peery, A Hall, Atashi Sharma, Xiao-Guang Chen, Robert M Waterhouse, Aleksey Komissarov, et al. 2014. “Genome Analysis of a Major Urban Malaria Vector Mosquito, Anopheles Stephensi.” *Genome Biology* 15 (9): 459. doi:10.1186/s13059-014-0459-2.

Jue, Nathaniel K, Michael B Murphy, Seth D Kasowitz, Sohaib M Qureshi, Craig J Obergfell, Sahar Elsisi, Robert J Foley, Rachel J O’Neill, and Michael J O’Neill. 2013. “Determination of Dosage Compensation of the Mammalian X Chromosome by RNA-Seq Is Dependent on Analytical Approach.” *BMC Genomics* 14: 150. doi:10.1186/1471-2164-14-150.

Kasahara, Masanori, Jutta Gutknecht, Keith Brew, Nigel Spurr, and Peter N. Goodfellow. 1989. “Cloning and Mapping of a Testis-Specific Gene with Sequence Similarity to a Sperm-Coating Glycoprotein Gene.” *Genomics* 5 (3): 527–34. doi:10.1016/0888-7543(89)90019-0.

Kayserili, Melek A., Dave T. Gerrard, Pavel Tomancak, and Alex T. Kalinka. 2012. “An Excess of Gene Expression Divergence on the X Chromosome in Drosophila Embryos: Implications for the Faster-X Hypothesis.” *PLoS Genetics* 8. doi:10.1371/journal.pgen.1003200.

Kharchenko, Peter V., Ruibin Xi, and Peter J. Park. 2011. “Evidence for Dosage Compensation between the X Chromosome and Autosomes in Mammals.” *Nature Genetics* 43 (12): 1167–69.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. 2013. “TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions.” *Genome Biology* 14 (4): R36. doi:10.1186/gb-2013-14-4-r36.

Kitzmiller, James B. 1963. “Mosquito Cytogenetics.” *Bulletin of the World Health Organization* 29: 345–55.

Kiuchi, Takashi, Hikaru Koga, Munetaka Kawamoto, Keisuke Shoji, Hiroki Sakai, Yuji Arai, Genki Ishihara, et al. 2014. “A Single Female-Specific piRNA Is the Primary Determiner of Sex in the Silkworm.” *Nature* 509 (7502): 633–36. doi:10.1038/nature13315.

Labbé, Geneviève M C, Sarah Scaife, Siân A. Morgan, Zoë H. Curtis, and Luke Alphey. 2012. “Female-Specific Flightless (fsRIDL) Phenotype for Control of Aedes Albopictus.” *PLoS Neglected Tropical Diseases* 6. doi:10.1371/journal.pntd.0001724.

Larschan, Erica, Eric P. Bishop, Peter V. Kharchenko, Leighton J. Core, John T. Lis, Peter J. Park, and Mitzi I. Kuroda. 2011. “X Chromosome Dosage Compensation via Enhanced Transcriptional Elongation in Drosophila.” *Nature* 471 (7336): 115–18. doi:10.1038/nature09757.

Lin, F, K Xing, J Zhang, and X He. 2012. “Expression Reduction in Mammalian X Chromosome Evolution Refutes Ohno’s Hypothesis of Dosage Compensation.” *Proc Natl Acad Sci U S A* 109 (29): 11752–57. doi:10.1073/pnas.1201816109.

Love, M. I., W. Huber, and S. Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *bioRxiv*. doi:10.1101/002832.

Lu, G, M Villalba, M R Coscia, D R Hoffman, and T P King. 1993. “Sequence Analysis and Antigenic Cross-Reactivity of a Venom Allergen, Antigen 5, from Hornets, Wasps, and Yellow Jackets.” *Journal of Immunology (Baltimore, Md. : 1950)* 150: 2823–30.

Magnusson, Kalle, Gareth J Lycett, Antonio M Mendes, Amy Lynd, Philippos-Aris Papathanos, Andrea Crisanti, and Nikolai Windbichler. 2012. “Demasculinization of the Anopheles Gambiae X Chromosome.” *BMC Evol. Biol.* 12: 69. doi:10.1186/1471-2148-12-69.

Mahajan, S., and D. Bachtrog. 2015. “Partial Dosage Compensation in Strepsiptera, a Sister Group of Beetles.” *Genome Biology and Evolution* 7 (2): 591–600. doi:10.1093/gbe/evv008.

Mank, J E. 2013. “Sex Chromosome Dosage Compensation: Definitely Not for Everyone.” *Trends Genet* 29 (12): 677–83. doi:10.1016/j.tig.2013.07.005.

Mank, J. E., and H. Ellegren. 2008. “All Dosage Compensation Is Local: Gene-by-Gene Regulation of Sex-Biased Expression on the Chicken Z Chromosome.” *Heredity* 102 (3): 312–20. doi:10.1038/hdy.2008.116.

Mank, Judith E, David J Hosken, and Nina Wedell. 2011. “Some Inconvenient Truths about Sex Chromosome Dosage Compensation and the Potential Role of Sexual Conflict.” *Evolution; International Journal of Organic Evolution* 65 (8): 2133–44. doi:10.1111/j.1558-5646.2011.01316.x.

Mank, Judith E. 2009. “The W, X, Y and Z of Sex-Chromosome Dosage Compensation.” *Trends in Genetics*. doi:10.1016/j.tig.2009.03.005.

Meiklejohn, Colin D., Emily L. Landeen, Jodi M. Cook, Sarah B. Kingan, and Daven C. Presgraves. 2011. “Sex Chromosome-Specific Regulation in the Drosophila Male Germline But Little Evidence for Chromosomal Dosage Compensation or Meiotic Inactivation.” *PLoS Biology* 9 (8): e1001126. doi:10.1371/journal.pbio.1001126.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5: 621–28. doi:10.1038/nmeth.1226.

Muñoz, D., A. Jimenez, O. Marinotti, and A. A. James. 2004. “The AeAct-4 Gene Is Expressed in the Developing Flight Muscles of Female Aedes Aegypti.” *Insect Molecular Biology* 13: 563–68. doi:10.1111/j.0962-1075.2004.00519.x.

Neafsey, Daniel E., Robert M. Waterhouse, Mohammad R. Abai, Sergey S. Aganezov, Max A. Alekseyev, James E. Allen, James Amon, et al. 2015. “Highly Evolvable Malaria Vectors: The Genomes of 16 Anopheles Mosquitoes.” *Science* 347 (6217): 1258522–1258522. doi:10.1126/science.1258522.

Nene, Vishvanath, Jennifer R. Wortman, Daniel Lawson, Brian Haas, Chinnappa Kodira, Zhijian Jake Tu, Brendan Loftus, et al. 2007. “Genome Sequence of Aedes Aegypti, a Major Arbovirus Vector.” *Science (New York, N.Y.)* 316 (5832): 1718–23. doi:10.1126/science.1138878.

Nozawa, M., N. Fukuda, K. Ikeo, and T. Gojobori. 2014. “Tissue- and Stage-Dependent Dosage Compensation on the Neo-X Chromosome in Drosophila Pseudoobscura.” *Molecular Biology and Evolution* 31 (3): 614–24. doi:10.1093/molbev/mst239.

Ohno, S. 1967. “Sex Chromosomes and Sex-Linked Genes. In Monographs on Endocrinology.” *Springer-Verlag, Heidelberg- Berlin- New York* 1.

Parisi, Michael, Rachel Nuttall, Daniel Naiman, Gerard Bouffard, James Malley, Justen Andrews, Scott Eastman, and Brian Oliver. 2003. “Paucity of Genes on the Drosophila X Chromosome Showing Male-Biased Expression.” *Science* 299 (5607): 697–700. doi:10.1126/science.1079190.

Parsch, John, and Hans Ellegren. 2013. “The Evolutionary Causes and Consequences of Sex-Biased Gene Expression.” *Nature Reviews. Genetics* 14 (2): 83–87. doi:10.1038/nrg3376.

Pease, J. B., and M. W. Hahn. 2012. “Sex Chromosomes Evolved from Independent Ancestral Linkage Groups in Winged Insects.” *Molecular Biology and Evolution* 29 (6): 1645–53. doi:10.1093/molbev/mss010.

Penalva, L. O. F., and L. Sanchez. 2003. “RNA Binding Protein Sex-Lethal (Sxl) and Control of Drosophila Sex Determination and Dosage Compensation.” *Microbiology and Molecular Biology Reviews* 67 (3): 343–59. doi:10.1128/MMBR.67.3.343-359.2003.

Pessia, E., T. Makino, M. Bailly-Bechet, A. McLysaght, and G. A. B. Marais. 2012. “Mammalian X Chromosome Inactivation Evolved as a Dosage-Compensation Mechanism for Dosage-Sensitive Genes on the X Chromosome.” *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1116763109.

Prince, Eldon G, Donna Kirkland, and Jeffery P Demuth. 2010. “Hyperexpression of the X Chromosome in Both Sexes Results in Extensive Female Bias of X-Linked Genes in the Flour Beetle.” *Genome Biol Evol* 2: 336–46. doi:10.1093/gbe/evq024.

Rambaut, Andrew. 2009. “FigTree, a Graphical Viewer of Phylogenetic Trees.” *Institute of Evolutionary Biology University of Edinburgh*.

Robinson, Mark, Davis Mccarthy, Yunshun Chen, and Gordon K Smyth. 2011. “edgeR : Differential Expression Analysis of Digital Gene Expression Data User ’ S Guide.” *Most* 23: 1–77. doi:10.1093/bioinformatics/btp616.

Singh, N, A K Mishra, S K Chand, and V P Sharma. 1999. “Population Dynamics of Anopheles Culicifacies and Malaria in the Tribal Area of Central India.” *Journal of the American Mosquito Control Association* 15: 283–90.

Smith, G, Y R Chen, G W Blissard, and A D Briscoe. 2014. “Complete Dosage Compensation and Sex-Biased Gene Expression in the Moth Manduca Sexta.” *Genome Biol Evol* 6 (3): 526–37. doi:10.1093/gbe/evu035.

Straub, Tobias, and Peter B. Becker. 2007. “Dosage Compensation: The Beginning and End of Generalization.” *Nature Reviews Genetics* 8 (1): 47–57. doi:10.1038/nrg2013.

Sturgill, D, Y Zhang, M Parisi, and B Oliver. 2007. “Demasculinization of X Chromosomes in the Drosophila Genus.” *Nature* 450 (7167): 238–41. doi:10.1038/nature06330.

Toups, Melissa A.;, and Matthew W. Hahn. 2010. “Retrogenes Reveal the Direction of Sex-Chromosome Evolution in Mosquitoes.” *Genetics* 186 (2). Genetics Society of America: 763–66. doi:10.1534/genetics.110.118794.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. 2012. “Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks.” *Nature Protocols* 7: 562–78. doi:10.1038/nprot.2012.016.

Traut, Walther, Teruyuki Niimi, Kazuho Ikeo, and Ken Sahara. 2006. “Phylogeny of the Sex-Determining Gene Sex-Lethal in Insects.” *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada* 49: 254–62. doi:10.1139/G05-107.

Uebbing, S, A Kunstner, H Makinen, and H Ellegren. 2013. “Transcriptome Sequencing Reveals the Character of Incomplete Dosage Compensation across Multiple Tissues in Flycatchers.” *Genome Biol Evol* 5 (8): 1555–66. doi:10.1093/gbe/evt114.

Vicoso, Beatriz, and Doris Bachtrog. 2011. “Lack of Global Dosage Compensation in Schistosoma Mansoni, a Female-Heterogametic Parasite.” *Genome Biology and Evolution* 3: 230–35. doi:10.1093/gbe/evr010.

———. 2015. “Numerous Transitions of Sex Chromosomes in Diptera.” *PLoS Biol* 13 (4): e1002078. doi:10.1371/journal.pbio.1002078.

Vicoso, Beatriz, and Brian Charlesworth. 2009. “The Deficit of Male-Biased Genes on the D. Melanogaster X Chromosome Is Expression-Dependent: A Consequence of Dosage Compensation?” *Journal of Molecular Evolution* 68 (5): 576–83. doi:10.1007/s00239-009-9235-4.

Vicoso, Beatriz, J. J. Emerson, Yulia Zektser, Shivani Mahajan, and Doris Bachtrog. 2013. “Comparative Sex Chromosome Genomics in Snakes: Differentiation, Evolutionary Strata, and Lack of Global Dosage Compensation.” *PLoS Biol* 11 (8): e1001643. doi:10.1371/journal.pbio.1001643.

Vyazunova, I, and Q Lan. 2004. “Stage-Specific Expression of Two Actin Genes in the Yellow Fever Mosquito, Aedes Aegypti.” *Insect Molecular Biology* 13 (3): 241–49. doi:10.1111/j.0962-1075.2004.00481.x.

Waterhouse, Robert M., Fredrik Tegenfeldt, Jia Li, Evgeny M. Zdobnov, and Evgenia V. Kriventseva. 2013. “OrthoDB: A Hierarchical Catalog of Animal, Fungal and Bacterial Orthologs.” *Nucleic Acids Research* 41 (D1): D358–65. doi:10.1093/nar/gks1116.

Wright, A. E., and J. E. Mank. 2012. “Battle of the Sexes: Conflict over Dosage-Sensitive Genes and the Origin of X Chromosome Inactivation.” *Proceedings of the National Academy of Sciences* 109 (14): 5144–45. doi:10.1073/pnas.1202905109.

Wright, A. E., H. K. Moghadam, and J. E. Mank. 2012. “Trade-off Between Selection for Dosage Compensation and Masculinization on the Avian Z Chromosome.” *Genetics* 192 (4): 1433–45. doi:10.1534/genetics.112.145102.

Wyman, Minyoung J., Asher D. Cutter, and Locke Rowe. 2012. “Gene Duplication in The Evolution of Sexual Dimorphism: Duplicates and Sex-Biased Gene Expression.” *Evolution* 66 (5): 1556–66. doi:10.1111/j.1558-5646.2011.01525.x.

Xiong, Y, X Chen, Z Chen, X Wang, S Shi, J Zhang, and X He. 2010. “RNA Sequencing Shows No Dosage Compensation of the Active X-Chromosome.” *Nat Genet* 42 (12): 1043–47. doi:10.1038/ng.711.

Zdobnov, E. M. 2002. “Comparative Genome and Proteome Analysis of Anopheles Gambiae and Drosophila Melanogaster.” *Science* 298 (5591): 149–59. doi:10.1126/science.1077061.

Chapter 4: Single Molecule RNA Sequencing uncovers *trans*-splicing and updates annotations in *Anopheles stephensi*

4.1 Abstract

Here, we used PacBio Iso-Seq to sequence a cDNA library from the Asian malaria mosquito *Anopheles stephensi*. More than 600,000 full length cDNAs, referred to as reads of insert, were identified. Due to the inherently high error-rate of PacBio sequencing, we tested different approaches for error-correction. We found that error-correction using Illumina RNA-Seq generated more data than using the default SMRT pipeline. The full-length error-corrected PacBio reads greatly improved the gene annotation of *Anopheles stephensi*: 4,867 gene models were updated and 1,785 alternatively-spliced isoforms were added to the annotation. In addition, six inter-chromosomal *trans*-splicing events were identified in *An. stephensi*. All six *trans*-splicing events appear to be conserved in *Culicidae*, as they are also found in *An. gambiae* and *Aedes aegypti*. The proteins encoded by *trans*-splicing events are also highly conserved and the orthologs of these proteins are *cis*-spliced in outgroup species, implying *trans*-splicing may arise as a mechanism to rescue genes that broke-up during evolution.

4.2 Introduction

RNA splicing is the process by which introns are removed from the pre-messenger RNA (pre-mRNA), and exons are joined to form a mature messenger RNA (mRNA). RNA splicing is an essential step to form mature eukaryotic because the majority of eukaryotic mRNAs have introns. For the majority of eukaryotic genes, splicing is mediated in *cis* by the spliceosome. The spliceosome brings the exons on both side of an intron into close proximity and then cleaves the 5’ splice site and ligates the 5’ splice site to the branch point on the intron. This produces a lariat structured RNA. The spliceosome then cuts the 3’ splice site, ligates exons, and releases the lariat.

Splicing can also occur *in trans*, where exons from multiple separate pre-mRNAs are joined. *Trans*-splicing is well studied in trypanosomes and nematodes, where a spliced leader RNA is spliced to the 5' ends of the first exon on many pre-mRNAs (Douris, Telford, and Averof 2010). The reaction is similar to nuclear *cis*-splicing, but generates a Y-shaped RNA instead of a lariat. In higher eukaryotes, *trans*-spicing did not involve spliced leaders. *Trans*-splicing has been observed in fruit flies, rodents, humans, and many other organisms (Shao et al. 2012; Lasda and Blumenthal 2011; Horiuchi, Giniger, and Aigaki 2003; Caudevilla et al. 1998; Dorn, Reuter, and Loewendorf 2001; Herai and Yamagishi 2010). Based on the relationship of the two pre-mRNAs joined in *trans*-splicing, *trans*-splicing can be grouped into three categories: inter-allelic, intragenic and intergenic. A well-known example of inter-allelic *trans*-splicing is the *lola* gene in *Drosophila* (Horiuchi, Giniger, and Aigaki 2003). *Lola* is essential for nervous system development. *Trans*-splicing of *lola* was inferred from interallelic complementation tests on lethal mutations in *lola* exons and verified by allelic SNP makers in *Drosophila* hybrids. Later, a study utilizing RNA-Seq data from *Drosophila* hybrids identified more *trans*-splicing between homologous alleles, suggesting inter-allelic *trans*-splicing occurs commonly. Intragenic *trans*-splicing is the scenario where splicing occurs between two pre-mRNAs from the same genetic loci. The two pre-mRNAs can come from the same strand, and the examples are the Carnitine O-octanoyltransferase gene in the rat liver where the exons are duplicated in the mRNA (Caudevilla et al. 1998). They can also come from the opposite strand, like the fruit fly *mod* (*modifier of mdg4*) genes (Dorn, Reuter, and Loewendorf 2001). Intergenic *trans*-splicing occurs when the pre-mRNAs come from different genes. These genes can be located at distant genomic loci such as different chromosomes. For example, the *bursicon* gene in *Anopheles gambiae* is *trans*-spliced from three exons on 2L chromosome arm and one exons on 2R chromosome arm (Robertson et al. 2006).

The understanding of *trans*-splicing has been significantly improved by the advent of next generation sequencing technology. *Trans*-splicing events are generally identified by finding non-co-linear transcripts, which are RNA-Seq sequences that fail to align to the corresponding DNA sequences in the reference genome in a linear pattern. Although this approach cannot detect inter-allelic and other trans-splicing that generate co-linear transcripts, a significant number of *trans*-splicing have been detected (Davidson, Majewski, and Oshlack 2015; Liu et al. 2015). For example, a recent study on eight insect species across five orders detected 1,627 *trans*-splicing events (Kong et al. 2015). Some of the *trans*-splicing events are conserved across species, indicating that *trans*-splicing is not transcriptional noise and is likely to be functionally significant (Kong et al. 2015). Besides, the previous notion that fusion transcripts are the markers of tumor cells has been called into question, as several students and the ENCODE project demonstrated that chimeric RNAs are common in both normal tissues and cell lines (Gingeras 2009). Fusion transcripts do not necessary imply oncogenic chromosomal rearrangements, because *trans*-splicing can likely contribute as well.

The SMRT isoform sequencing (Iso-Seq) technology from Pacific Biosciences has also been applied to discover fusion genes (Weirather et al. 2015). Iso-Seq can generate full-length transcript sequences from the polyA-tail to the 5′ end, providing isoform-level resolution of transcriptome data. Iso-Seq have already been applied in a wide variety of organisms to improve the genome annotation and to discover new genes and isoforms. When it comes to detecting *trans*-splicing, RNA-Seq data by Illumina can only provide information on the small segment around the *trans*-spliced site. The structure of full *trans*-spliced mRNA is hard to infer from RNA-Seq data, due to the fact that majority of the reads generated from the *trans*-spliced mRNAs cannot be differentiated from the ones from *cis*-spliced mRNA. This will not be an issue for Iso-seq data, which provide reads representing full-length transcripts.

In this research, we use both SMRT isoform sequencing (Iso-Seq) data and Illumina RNA-Seq data to detect *trans*-splicing events in Asian malaria mosquito *Anopheles stephensi*. To eliminate false positive discoveries due to PCR chimeras and transcriptional noise, only *trans*-splicing events supported by both are used. In total, we identified six *trans*-splicing events in *Anopheles stephensi*, all of which are also found and conserved *Aedes aegypti*. The proteins encoded by the *trans*-spliced mRNAs are also highly conserved and their orthologs are co-linearly transcribed in *Culicidae* out-groups. This finding indicates the need to preserve the mRNA completeness and protein function of genes broken-up during the course of evolution may be the driving force behind *trans*-splicing. We have also used the Iso-seq data to improve the *An. stephensi* annotation.

4.3 Methods

4.3.1 Library preparation and data sequencing

Adult male *Anopheles stephensi* were obtained. Total RNA was isolated using DNAeasy blood and tissue kit. The RNA was then reverse transcribed using the SMRTer PCR cDNA synthesis kit and amplified. Three sequencing libraries (1-2kb, 2-3kb, 3-6kb) were prepared according to the PacBio Iso-Seq protocol. The sequencing was performed on the PacBio RS II using P4-C2 chemistry. Four SMRT cells were run from each of the three libraries. RNA-Seq libraries were prepared and sequenced as described before (Jiang et al. 2015).

4.3.2 SMRT pipeline analysis for Iso-Seq data

Analysis was performed using the PacBio SMRT-Analysis package v2.3 (http://www.pacb.com/devnet/). Analysis was run on the three libraries separately. The Iso-Seq bioinformatics pipeline consists of two major modules: classify and cluster. Reads of insert were obtained by identifying the adapter separator and then merging subreads into consensus sequence reads. Reads of insert were then classified into full-length, non-artificial-concatemer reads and non-full-length reads. Full-length, non-artificial-concatemer reads from the same isoform were clustered using the ICE algorithm and consensus isoforms were predicted. The consensus isoforms were then polished by Quiver utilizing the non-full-length reads. Default parameters were used when running Quiver, which means only consensus with more than 99% accuracy were binned into high-quality isoforms by Quiver. The low-quality isoforms binned by Quiver are generally the ones with low transcription level or low sequencing depth. Although less accurate, the low-quality isoforms also contain useful information and are used together with high-quality isoforms for our analysis.

4.3.3 Error correction of Iso-Seq data

The reads of insert of the three libraries were combined and subjected to error correction. RNA-Seq data was processed as shown in Figure 1 to achieve best performance for error correction. First, raw RNA-Seq reads were trimmed with Trimmomatic with parameter "2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36" (Bolger, Lohse, and Usadel 2014). The resulting trimmed paired reads were merged with FLASH with default parameters (Magoč and Salzberg 2011). The merged reads along with reads that failed to merge and unpaired reads from Trimmomatic were combined into one fastq file. This fastq file was of 26 Gb in size and used as short reads to correct the reads of insert with proovread (Hackl et al. 2014). Proovread is a high accuracy PacBio correction tool, which works via iterative alignment of short reads to produce consensus sequences. Proovread outputted high accuracy PacBio reads with low quality regions trimmed as well as complete corrected PacBio reads including poorly corrected regions. Only the high accuracy trimmed Iso-Seq reads were used for further analysis.

4.3.4 Fusion transcripts detection with both Iso-Seq data and RNA-Seq data

Anopheles stephensi Indian strain genome version2 was downloaded from Vectorbase (Giraldo-Calderon et al. 2015). Based on our previous research (data not shown), we are able to include the majority of the *Anopheles stephensi* Indian genome in five FASTA sequences, with each of the sequences representing one chromosomal arm. The five fasta sequences were used as the genome sequence in the fusion transcripts detection analysis. The high accuracy trimmed Iso-Seq reads were aligned to the genome and then processed by the fusion_finder.py script of PacBio pbtranscript-tofu package (https://github.com/PacificBiosciences/cDNA_primer.git). Proovread could potentially remove *trans*-spliced transcripts when it removed PCR chimeric reads during correcting. Therefore, the unpolished consensus isoforms were added to the above analysis to provide more reads. MapSplice, a splice junction discovery software was used to predict fusion genes in the RNA-Seq data (Wang et al. 2010). With the "—fusion" option, MapSplice performed canonical and semi-canonical fusion junction detection after the RNA-Seq reads were aligned to the genome. Only fusion junctions supported by both Iso-Seq and RNA-Seq data were used. In total, six splice junctions were identified.

4.3.5 Genome annotation updates

Genome version2 and annotation version2.2 of the *Anopheles stephensi* Indian strain were downloaded from Vectorbase (Giraldo-Calderon et al. 2015). PASA (Program to Assemble Spliced Alignments) Release r20140417 (http://pasapipeline.github.io/) was used to update the existing annotations using evidence generated from the high accuracy trimmed Iso-Seq reads, and we then compared the updated annotation to existing gene structure annotations (Haas et al. 2003).

As the high accuracy trimmed Iso-Seq reads kept the transcribed orientation, option "--transcribed_is_aligned_orient" were added when the PASA pipeline were launched.

4.4 Results

4.4.1 Error correction outperforms SMRT pipeline in both data accuracy and data quantity

Each library was composed of four cells and each cell produced around 50,000 to 60,000 reads of insert. Full-length transcripts were defined by the presence of 5' primer, 3' primer, and the polyA tail in the reads of insert. Approximately 38%, 31%, and 9% of reads of insert were identified as full length and non-chimeric reads for library of 1-2kb insert size, 2-3kb insert size, and 3-6kb insert size, respectively (Table 4.1). During the clustering process of the SMRT pipeline, on average 2 to 3 full length and non-chimeric reads can be clustered as one consensus isoforms for library with small reads length size. For the 3-6kb size library, most consensus isoforms are composed of only one full length and non-chimeric reads. Therefore, in order to obtain sufficient number of long reads for analysis, libraries with large reads length require deep sequencing depth. Only less than 17% of total consensus isoforms were polished as high-quality isoforms by Quiver. This indicates that in order to obtain enough high accuracy data through the SMRT pipeline alone, the number of cells sequenced for each library should be higher to provide enough coverage, particularly for long reads library. Both all polished isoforms include both high-quality ones and low-quality ones were used for the further analysis to avoid discarding useful information.

Alternatively, we used RNA-Seq data to error-correct Iso-Seq data. Of 1,321 million base pairs of reads of insert, 668 million base pairs were corrected by Proovread with a high accuracy (Table 4.2). The mean of the average quality scores of each read of insert improved from 14.73 to 36. 4 after correction, indicating a significant improvement of accuracy. Compared with polished high-quality isoforms from the SMRT pipeline, although the mean value of median quality scores of high accuracy corrected reads was slightly lower, 30 times more base pairs were corrected. In addition, the mean value of median quality scores of high accuracy corrected reads was 37.85, equivalent to an accuracy above 99.98%. This result showed that it is favorable to use high-quality short reads to correct erroneous Iso-Seq reads. This is also more economical as the RNA-Seq data needed for the analysis is significantly cheaper than additional Iso-Seq data needed.

4.4.2 Proteins encoded by *trans*-splicing are conserved

490 *trans*-splicing events were detected based on RNA-Seq processed by MapSplice. 3,359 *trans*-splicing events were found by PacBio pbtranscript-tofu package. In both RNA-Seq and Iso-Seq technology, PCR chimeras could cause a large number of false positive results. Therefore, we set a criterion that splice junctions must be supported by both to be considered as valid. In the end, six pairs of splice junctions were identified. All these six *trans*-splicing events are inter-chromosomal.

Trans-spliced mRNA 1 (Tm1) is one mRNA created from two *trans*-splicing events (Figure 3.2 A). The Pre-mRNAs of Tm1 are located in chromosome elements 1, 2 and 3. This mRNA has five exons: two shared with gene ASTEI07024, one shared with gene ASTEI02601, one shared with the intron of gene ASTEI04882. The mRNA encoded a 475 aa peptide with two domains. The first domain encodes MiT/TFE transcription factors, N-terminal (IPR031867), which is shared with gene ASTEI07024. The second domain is Myc-type, basic helix-loop-helix domain (IPR011598), shared with gene ASTEI02601. ASTEI07024 is a mosquito specific gene. Alignment of peptide sequences of ASTEI02601 to its *Drosophila* ortholog FBgn0041164 revealed that the exon utilized by the Tm1 *trans*-splicing event contributes to amino acid sequences that do not exist in their *Drosophila* orthologs. This is also the case for the exon shared between Tm1 and ASTEI04882 when we aligned the peptide sequence of ASTEI04882 to that of the ortholog FBgn0034176. No obvious *Drosophila* ortholog for the complete Tm1 protein has been observed. There is some similarity between the protein of FBgn0263112 to the peptide sequences coded by the first three exons, particularly the third exon of Tm1 (28.87% identify). Interestingly, the complete 474 aa peptide sequence coded by Tm1 is highly homologous to some dipteran out-groups. The examples include gene XP_011304746 in *Fopius arisanus* (33.17% identity) and XP_012252483 in *Athalia rosae* (31.53% identity). In these genes, the mRNAs are co-linear to the genome, and thus likely to be *cis*-spliced.

The exons of Tm2 come from ASTEI01093 and ASTEI00334 (Figure 3.2 B). Both genes do not have orthologs outside of mosquitos. The protein encoded by Tm2 consists of 515 amino acids, which belongs to the neurotransmitter-gated ion-channel (IPR006201) family. This protein is orthologous to the *Drosophila* gene FBgn0037950 with high similarity (83.57% identity). This protein is conserved in *Insecta*. All of its non-mosquito orthologs appear to be *cis*-spliced.

The donor sites of the *trans*-splicing event of Tm3 and Tm4 are identical (Figure 3.2 C and D). The genes on the acceptor site of these two events are paralogous to each other. The paralogs are in close proximity but of different orientation, probably due to a tandem duplication. The coding sequences of the two paralogs were identical, and consequently Tm3 and Tm4 encode identical protein. *Trans*-splicing exists in both ASTEI004497 and ASTEI004495, as supported by full length transcripts covering the 3' UTR in both genes. The encoded protein is a neurotransmitter symporter (IPR000175). The *Drosophila* gene FBgn0181657 is annotated as an ortholog to ASTEI02036 but in fact, sequence alignment showed that this is only a partial match, and FBgn0181657 more fully aligns over its full length to the fusion protein of ASTEI02036 and ASTEI004497/ASTEI004495. This protein is highly conserved across *Insecta*. Like Tm2, Tm3 and Tm4 orthologs outside mosquitoes are *cis*-spliced.

The longest read we obtained of Tm5 is 2,142 bp (Figure 3.2 E). This read likely represents incomplete mRNA with five prime end trimmed, because the start code is missing. Nevertheless, this read covers the *trans*-splicing site between chromosome element 1 and 3. Tm5 joins exons from ASTEI06203 and ASTEI00378. The protein Tm5 encodes is uncoordinated protein 13 (IPR027080). It is conserved across *Insecta*. FBpp0300963 is annotated as ortholog to ASTEI06203. Interestingly, the last 53 amino acids encoded by FBpp0300963 is 76% identical to the 54 amino acids encoded by the last exon of ASTEI00378, only 7% identical to amino acid encoded by the last exon of ASTEI06203. This implies that the fusion protein is ancestral and *trans*-splicing between exons of ASTEI006203 and ASTEI00378 is a way to keep the protein intact.

4.4.3 Trans-splicing is highly conserved in *Culicidae*

To investigate whether the above *trans*-splicing events are *An. stephensi* specific or are conserved, we checked the transcriptome data of *Anopheles gambiae* and *Aedes aegypti*. We predicted *trans*-splicing sites using MapSplice (Wang et al. 2010) with RNA-Seq data as described in the methods. All the *trans*-splicing events in *An. stephensi* also exist in *An. gambiae* (Table 4.3). In addition, the chromosomal assignment of the orthologs involved in *trans*-splicing are the same and the sequences around the splice sites are highly identical between these two species. In *Aedes aegypti*, the supercontigs are not assigned to chromosomes and thus chromosomal position cannot be inferred. Based on supercontigs, the orthologs of the *trans*-splicing *An. stephensi* events are observed with a few differences. First, the *trans*-spliced gene may share the exons with a different *cis*-spliced gene. For example, the ASTEI02601 orthologs AAEL010693 and AAEL010696 do not share exons with Tm1 in *Aedes aegypti*. Instead, the shared exon is in their neighboring gene AAEL010700. Second, duplication events are different between *Anophelinae* and *Culicinae*: the

ASTEI07024 orthologs were duplicated and located on different supercontigs in *Aedes aegypti*, while the ortholog of gene ASTEI004495/ASTEI004497 is a single gene AAEL012596 in *Aedes aegypti*. Interestingly, *trans*-splicing was maintained during duplications of these genes.

4.4.4 Genome annotation improvement

Comparisons of the existing gene annotation of *Anopheles stephensi* and the updated annotation by PASA with error-corrected high-accuracy Iso-Seq data can be seen in the link http://tu07.fralin.vt.edu/cgi-bin/PASA_r20140417/cgi-bin/status_report.cgi?db=ECRItr. The existing gene annotation is based on gene annotation software Maker (Holt and Yandell 2011), where protein homology based and *ab inito* prediction were applied. Transcriptomes were not used by this Maker annotation. In the existing annotation, 11,789 protein coding genes were annotated. Each gene has only one isoform, which indicates that alternative splicing is largely ignored. In addition, genes were mostly annotated with UTRs missing. The updated annotation enhanced the existing one by adding UTRs, identifying alternative spliced isoforms, and adjusting exon boundaries. In total, 3,323 genes were updated with the addition of UTRs, 1,785 genes were updated with alternatively spliced isoforms, and 1,923 genes were updated with exons adjusted or genes merged. These structural changes of genes altered 1,878 protein sequences (Table 4.4).

One example demonstrating the improvement of gene annotation can be reflected in the annotation of the gene *doublesex* (Suzuki et al. 2001). *doublesex* is a gene essential for sexual dimorphism and it contains male-specific and female-specific isoforms. In our analysis, the Iso-Seq data was obtained from males only, so we would expect to observe only the male isoform. The gene *doublesex* in mosquitos spans a region of 90,000 base pairs with another gene inserted in one of its introns. As a result, in the majority of *Anophelinae*, this gene is mis-annotated as two genes. After the annotation updating by PASA (Figure 3.3), the two parts of *doublesex* ASTEI07080 and

ASTEI07082 were merged into one complete model. This model is the complete male isoform of *doublesex* as expected.

4.5 Discussion

4.5.1 The evolution of the six *trans*-splicing events

Two separate gene break-ups are necessary to form the *trans*-splicing of gene of Tm1. The first one which separated the third and four exons of Tm1 happened before the formation of *Diptera. Culicidae* adopted *trans-splicing* to join these two separated exons, while *Drosophila* either did not used *trans*-splicing or aborted it at a later time point. The second gene breakup which separated exon2 and exon3 happened only to *Culicidae* and they adopted *trans*-splicing to form a functional protein. In *Aedes,* the region transcribing the first pre-mRNA of Tm1 was duplicated and both copies kept their capability to be *trans*-spliced. The breakup of the ancestor genes of other *trans*-spliced mRNAs happened after formation of *Diptera* but before *Culicidae*. All their *Drosophila* orthologs remained as canonical genes that can be created from *cis*-splicing, while the formation of the complete mRNAs dependent on *trans*-splicing in *Anophelinae* and *Aedes*. The high conservation of the *trans-splicing* sites across three species indicates the single origin for each *trans*-splicing event.

4.5.2 Speculation on the evolution of *trans*-splicing

Although *trans*-splicing has been observed in many higher eukaryotes, its mechanism remains largely unclear. One well known model is *trans*-splicing through mutually complementary intron sequences. The introns of two separate pre-mRNAs will pair, bringing the two molecules together, promoting *trans*-splicing (Wally, Murauer, and Bauer 2012). A recent study on *Drosophila* showed that two intronic RNA sequences are critical to initiate *trans*-splicing in the *mod* gene (Gao et al. 2015). In both models, the nucleotide sequences of pre-mRNAs effect the conformation of the RNA- spliceosome complex and then influence splicing, which is essentially the same as *cis*-splicing. It is reasonable to assume that the splicing machinery is no different for *trans*-splicing. *Trans*-splicing is observed across a wide range of eukaryotes and likely exists in all eukaryotes, just like *cis*-splicing (Douris, Telford, and Averof 2010). In addition, the only factor differentiating *cis*-splicing and *trans*-splicing is whether there are more than one pre-mRNAs. As a process in three dimensions, splicing requires spatial proximity of splice sites (Hiller et al. 2007; Warf and Berglund 2010). No matter whether the two separate pre-mRNA interact with each other through base pairing or binding to the spliceosome using some motif, as long as the splice sites are spatially close and accessible, splicing reactions likely carry on just as it will in *cis*.

Splicing greatly diversifies the proteome by promoting the formation of new genes through alternative splicing. Allelic *trans*-splicing creates new combinations of alleles in mRNA (Horiuchi, Giniger, and Aigaki 2003). Intragenic *trans*-splicing can generate new transcripts by exon reuse (Caudevilla et al. 1998). Intergenic *trans*-splicing is supposed to be able to produce new genes by joining exons from different genes. However, novel gene formation through intergenic *trans*-splicing appears not to be favorable, as the *trans*-splicing events observed are largely involved in ancient gene rescue rather than new gene creation. This indicates that intergenic trans-splicing event likely evolved from an ancestry *cis*-splicing or allelic *trans*-splicing; the scenario where two random unrelated distant segment on the genome acquire the capability to be *trans*-spliced together should be rare if any. In addition, all the proteins encoded by our *trans*-spliced events are highly conserved. It appears that strong purification selection pressure acts to keep *trans*-splicing. If the protein is less essential, or other alternative strategy adopted as in the case of Tm1 in *Drosophila*, intergenic *trans*-splicing may never be adopted or may be abandoned later during evolution. This may be attributed to some inconvenient facts about intergenic *trans*-splicing. First, unlike *cis*-

splicing or the other two types of *trans*-splicing, it is hard to have two pre-mRNAs have the same or at least overlapping temporal and spatial transcription due to the fact that their DNA is located at different loci and they will not share the same regulatory mechanism. Second, upon transcription the physical distance of pre-mRNA could hinder *trans*-splicing. Additionally, it is hard for the two pre-mRNAs to coevolve when their DNA templates were shaped by potentially different evolutionary forces. Therefore, although existing as a rescue mechanism for essential breakup genes in multiple organism, intergenic *trans*-splicing is uncommon.

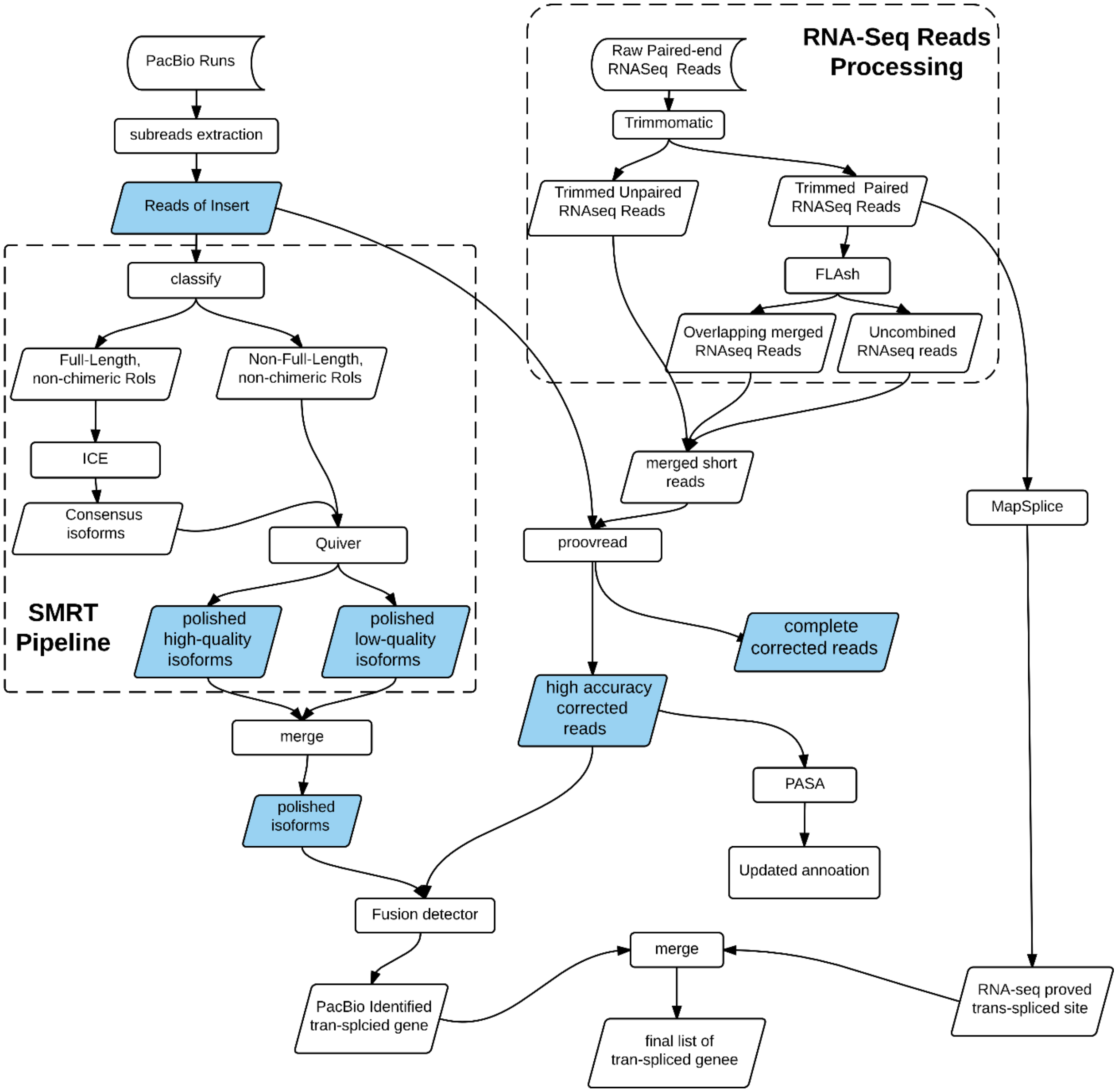
PacBio Runs
subreads extraction
Reads of Insert
classify
Full-Length, non-chimeric RoIs
Non-Full-Length, non-chimeric RoIs
ICE
Consensus isoforms
Quiver
SMRT Pipeline
polished high-quality isoforms
polished low-quality isoforms
merge
polished isoforms
Fusion detector
PacBio Identified tran-splcied gene
RNA-Seq Reads Processing
Raw Paired-end RNASeq Reads
Trimmomatic
Trimmed Unpaired RNAseq Reads
Trimmed Paired RNASeq Reads
FLAsh
Overlapping merged RNASeq Reads
Uncombined RNAseq reads
merged short reads
proovread
complete corrected reads
high accuracy corrected reads
PASA
Updated annoation
MapSplice
merge
final list of tran-spliced genee
RNA-seq proved trans-spliced site

Figure 4.1 Data processing and analysis pipelines for both RNA-seq data and Iso-Seq data. Processed Iso-Seq data highlighted in blue were compared in Table 4.2.

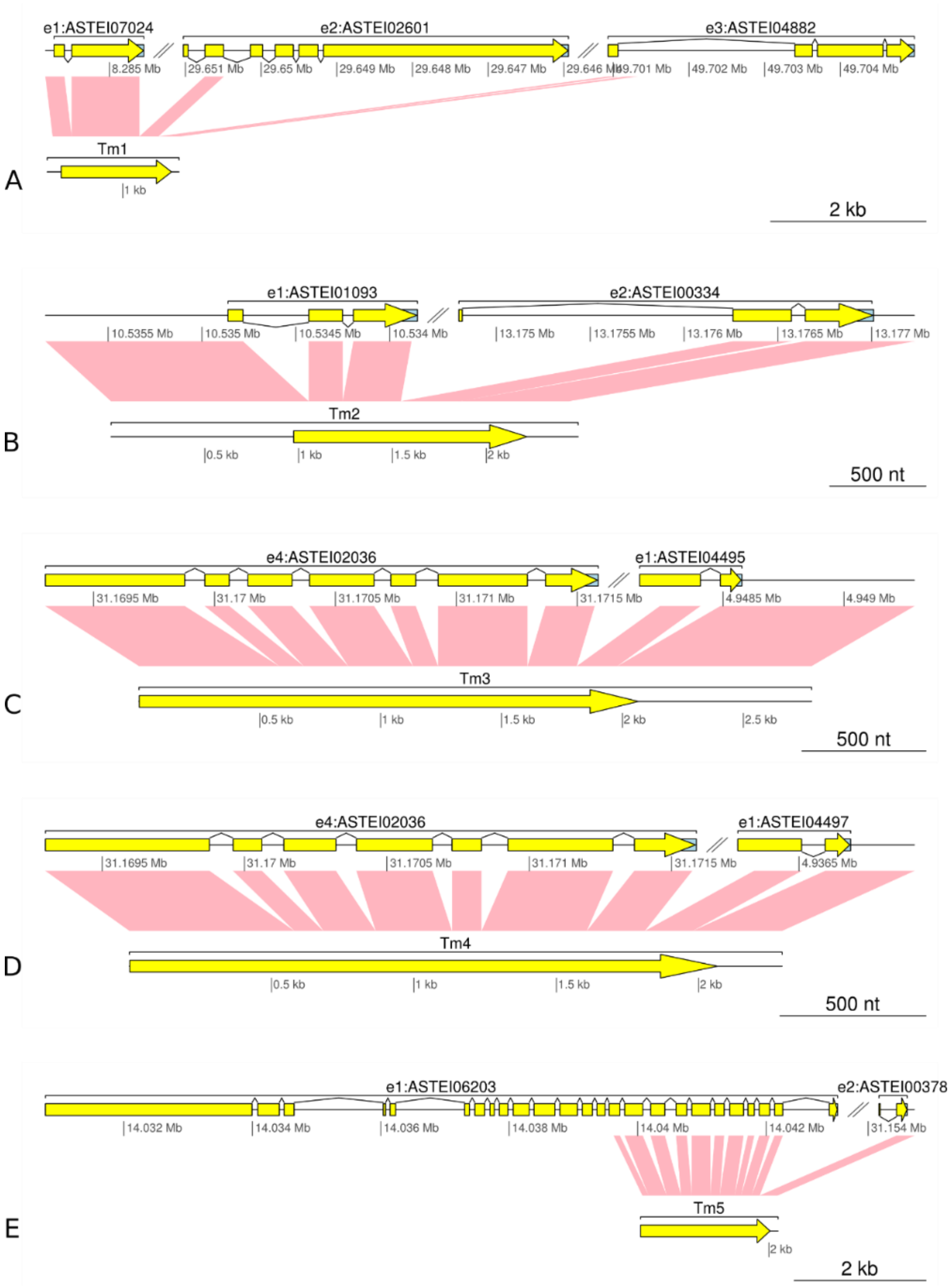

Figure 4.2 *Trans*-splicing events in *Anopheles stephensi*.

Each panel, the top part stands for the genomic region where the trans-spliced mRNA match to, with related genes annotated. The bottom part stands for the full-length mRNA sequence. Yellow bar presents coding region. Pink blocks represent matches between genomic sequence and mRNA.

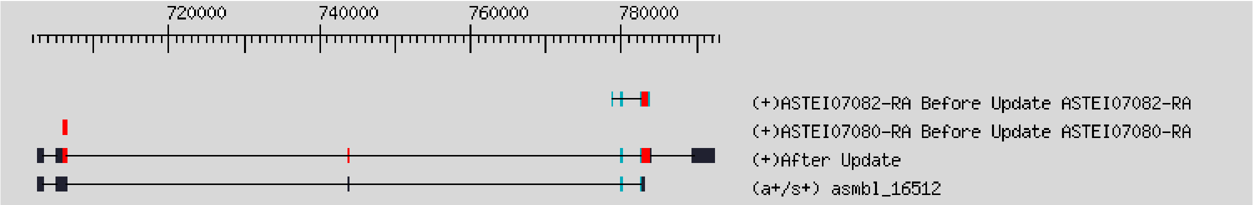
720000
740000
760000
780000
(+)ASTEI07082-RA Before Update ASTEI07082-RA
(+)ASTEI07080-RA Before Update ASTEI07080-RA
(+)After Update
(a+/s+) asmbl_16512

Figure 4.3 Updated annotation for *doublesex* gene in *Anopheles stephensi*.

The first row and second row represent gene ASTEI07082 and ASTI07080. The third row is an updated annotation from PASA which merges the two genes. The forth row is the evidence from transcripts that supports the updated annotation.

Table 4.1 PacBio SMRT pipeline outputs metrics

Data type	1-2kb	2-3kb	3-6kb
Number of reads of insert	248903	210594	202440
Number of five prime reads	136440	100147	54463
Number of three prime reads	146668	109776	65571
Number of poly-A reads	142375	106746	61281
Number of filtered short reads	13209	8876	7165
Number of non-full-length reads	138909	135425	175629
Number of full-length reads	96785	66293	19646
Number of full-length non-chimeric reads	96170	65955	19094
Average full-length non-chimeric read length	1388	1948	4357
Number of consensus isoforms	35248	30793	17405
Average consensus isoforms read length	1465	2044	4376
Number of polished high-quality isoforms	7414	6075	636
Number of polished low-quality isoforms	27834	24718	16769

Table 4.2 Comparisons of processed Iso-Seq data

		polished isoforms	polished high-quality isoforms	polished low-quality isoforms	high accuracy corrected reads	complete corrected reads	reads of insert
Mean Quality Score	min	0	0.01	0	16.01	0.11	0.98
	max	17.36	40	12.74	39.81	37.36	27.31
	average	12.7	37.15	7.72	36.4	28.25	14.73
	median	13.67	39.84	8.34	37.85	30.87	15.45
Length (base pairs)	Total (million)	190.6	23.8	166.8	668.6	1254.9	1321.4
	min	330	567	330	70	249	11
	max	25502	5673	25502	8098	31415	31532
	average	2284.59	1686.06	2406.55	1284.73	2005.91	1996.32
	median	1829	1630	1891	1145	1636	1641
	N25	4325	2005	4404	1939	4074	4214
	N50	2244	1742	2471	1444	2234	2342
	N75	1720	1416	1780	1046	1565	1624
	N90	1358	1193	1407	727	1181	1206
	N95	1213	1113	1244	613	919	928

Table 4.3 Trans-splicing sites in three *Culicidae*

An. stephensi								
	doner				acceptor			
	contig	start	strand	seq	contig	start	strand	seq
Tm1.1	stl-e1	8284604	-	ATCAAGAAGGATAATCATAACTGCA	stl-e2	29650741	-	ATATTGAAACGACGACGTCGCTCAA
Tm1.2	stl-e3	29650494	-	CGCAATGCTGCTGAAGCGTGTTGCG	stl-e2	49700929	+	GAAAATCAACCTGATTTTGCAACTC
Tm2	stl-e1	10533884	-	ATCTGTTCGATGATGATCGAAAGTT	stl-e2	13176261	+	ACCAAATCTTGTACCGTGTGCGATA
Tm3	stl-e4	31171572	+	ATCACTACTCCTGCCATCTGTGTCG	stl-e1	4936658	-	TGCACGATGTTGAAGATAAATACGC
Tm4	stl-e4	31171572	+	ATCACTACTCCTGCCATCTGTGTCG	stl-e1	4948239	+	TGCACGATGTTGAAGATAAATACGC
Tm5	stl-e1	14042259	+	AAAGCTGAAAGATGTCGTTGATCAG	stl-e2	31153557	-	GTAGCCAGCAGGCGAAGGAACTTTG
An. gambiae								
	doner				acceptor			
	contig	start	strand	seq	contig	start	strand	seq
Tm1.1	X	14803808	-	ATCAAGAAGGACAATCACAACTGCA	2L	40170764	+	ATGTTGAATCGACGACGCCGCTCAA
Tm1.2	2L	40171011	+	CGCAATGCTGCAGAAGCGTGTCGCG	2R	56727316	+	GAAAGTCAACTTGGTTTTGCAACTC
Tm2	X	4334405	-	ATCTGCTCGATGATGATCGAAAGTT	2R	13216641	-	ACCAAATCCTGTACCGTGTGCGATA
Tm3	3R	42647718	+	ATCACCACTCCTGCCATCTGTGTCG	X	8704053	-	TGCACGATGTTGAAGATGAATACGC
Tm4	3R	42647718	+	ATCACTACTCCTGCCATCTGTGTCG	X	8717291	+	TGCACGATGTTGAAGATGAATACGC
Tm5	X	1025131	+	AAAGCTGAAAGATGTCGTTGATCAG	2R	27239887	+	GTAGCCAGCAGGCGAAGGAACTTTG
Ae. aegypti								
	doner				acceptor			
	contig	start	strand	seq	contig	start	strand	seq
Tm1.1	supercont1.30	2525719	+	ATCAAGAAGGATAACCATAACTGCA	supercont1.497	69990	-	ATATTGAAACGACGACGCCTCTCGA
	supercont1.322	325659	-	ATCAAGAAGGATAACCATAACTGCA	supercont1.497	69990	-	ATATTGAAACGACGACGCCTCTCGA
Tm1.2	supercont1.497	69743	-	CATCATTCTGCAGCACAAACTGGCG	supercont1.541	339301	-	AATAGTCGAGCTGCTTTTGCATTTC
Tm2	supercont1.75	2116294	-	GTTTGCTCGATGATGATCGAAAGTT	supercont1.187	287309	-	ACCAAATCCTGAACGGTGTGGGATA
Tm3/Tm4	supercont1.496	104246	-	TATCACGACTCCGACGATCTGTGTT	supercont1.715	64386	+	GCACCAGGTTGAAGATGAATACGCC
Tm5	supercont1.54	128027	-	GAAGCTGAAGGACGTAGTCGATCAG	supercont1.179	490627	-	GTAACCAGCAGGCAAAGGAACTTTG

Table 4.4 Annotation improvement in *Anopheles stephensi* using PASA

	Num Gene Model Updates	Num Alt Splice isoforms to Add
EST assembly extends UTRs.	3323	0
EST assembly alters protein sequence, passes validation.	697	0
EST assembly properly stitched into gene structure.	1065	0
EST assembly stitched into Gene model requires alternative splicing isoform.	0	1785
EST-assembly found capable of merging multiple genes.	161	0
Totals (some models in multiple classes)	4867	1785

4.6 Reference

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20. doi:10.1093/bioinformatics/btu170.

Caudevilla, C, D Serra, a Miliar, C Codony, G Asins, M Bach, and F G Hegardt. 1998. "Natural Trans-Splicing in Carnitine Octanoyltransferase Pre-mRNAs in Rat Liver." *Proceedings of the National Academy of Sciences of the United States of America* 95 (21): 12185–90. doi:10.1073/pnas.95.21.12185.

Davidson, Nadia M, Ian J Majewski, and Alicia Oshlack. 2015. "JAFFA: High Sensitivity Transcriptome-Focused Fusion Gene Detection." *Genome Medicine* 7 (1). ??? 43. doi:10.1186/s13073-015-0167-x.

Dorn, R, G Reuter, and a Loewendorf. 2001. "Transgene Analysis Proves mRNA Trans-Splicing at the Complex mod(mdg4) Locus in Drosophila." *Proceedings of the National Academy of Sciences of the United States of America* 98 (17): 9724–29. doi:10.1073/pnas.151268698.

Douris, Vassilis, Maximilian J Telford, and Michalis Averof. 2010. "Evidence for Multiple Independent Origins of Trans-Splicing in Metazoa." *Molecular Biology and Evolution* 27 (3): 684–93. doi:10.1093/molbev/msp286.

Gao, Jun-li, Yu-jie Fan, Xiu-ye Wang, Yu Zhang, Jia Pu, Liang Li, Wei Shao, Shuai Zhan, Jianjiang Hao, and Yong-zhen Xu. 2015. "A Conserved Intronic U1 snRNP-Binding Sequence Promotes Trans -Splicing in Drosophila." *Genes & Development* 29: 760–71. doi:10.1101/gad.258863.115.3.

Gingeras, Thomas R. 2009. "Implications of Chimaeric Non-Co-Linear Transcripts." *Nature* 461 (7261): 206–11. doi:10.1038/nature08452.

Giraldo-Calderon, G. I., S. J. Emrich, R. M. MacCallum, G. Maslen, E. Dialynas, P. Topalis, N. Ho, et al. 2015. "VectorBase: An Updated Bioinformatics Resource for Invertebrate Vectors and Other Organisms Related with Human Diseases." *Nucleic Acids Research* 43 (D1): D707–13. doi:10.1093/nar/gku1117.

Haas, Brian J., Arthur L. Delcher, Stephen M. Mount S.M., Jennifer R. Wortman, Roger K. Smith, Linda I. Hannick, Rama Maiti, et al. 2003. "Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies." *Nucleic Acids Research* 31 (19): 5654–66. doi:10.1093/nar/gkg770.

Hackl, Thomas, Rainer Hedrich, Jörg Schultz, and Frank Förster. 2014. "Proovread: Large-Scale High-Accuracy PacBio Correction through Iterative Short Read Consensus." *Bioinformatics (Oxford, England)* 30 (21): 1–8. doi:10.1093/bioinformatics/btu392.

Herai, Roberto Hirochi, and Michel E Beleza Yamagishi. 2010. "Detection of Human Interchromosomal Trans-Splicing in Sequence Databanks." *Briefings in Bioinformatics* 11 (2): 198–209. doi:10.1093/bib/bbp041.

Hiller, Michael, Zhaiyi Zhang, Rolf Backofen, and Stefan Stamm. 2007. "Pre-mRNA Secondary Structures Influence Exon Recognition." *PLoS Genetics* 3 (11): e204. doi:10.1371/journal.pgen.0030204.

Holt, Carson, and Mark Yandell. 2011. "MAKER2: An Annotation Pipeline and Genome-Database Management Tool for Second-Generation Genome Projects." *BMC Bioinformatics*. doi:10.1186/1471-2105-12-491.

Horiuchi, Takayuki, Edward Giniger, and Toshiro Aigaki. 2003. "Alternative Trans-Splicing of Constant and Variable Exons of a Drosophila Axon Guidance Gene, Lola." *Genes & Development* 17 (20): 2496–2501. doi:10.1101/gad.1137303.

Jiang, X., J. K. Biedler, Y. Qi, a. B. Hall, and Z. J. Tu. 2015. “Complete Dosage Compensation in Anopheles Stephensi and the Evolution of Sex-Biased Genes in Mosquitoes.” *Genome Biology and Evolution*, 1–42. doi:10.1093/gbe/evv115.

Kong, Yimeng, Hongxia Zhou, Yao Yu, Longxian Chen, Pei Hao, and Xuan Li. 2015. “The Evolutionary Landscape of Intergenic Trans-Splicing Events in Insects.” *Nature Communications* 6. Nature Publishing Group: 8734. doi:10.1038/ncomms9734.

Lasda, Erika L., and Thomas Blumenthal. 2011. “Trans-Splicing.” *Wiley Interdisciplinary Reviews: RNA* 2 (3): 417–34. doi:10.1002/wrna.71.

Liu, S., W.-H. Tsai, Y. Ding, R. Chen, Z. Fang, Z. Huo, S. Kim, et al. 2015. “Comprehensive Evaluation of Fusion Transcript Detection Algorithms and a Meta-Caller to Combine Top Performing Methods in Paired-End RNA-Seq Data.” *Nucleic Acids Research*, 1–15. doi:10.1093/nar/gkv1234.

Magoč, Tanja, and Steven L. Salzberg. 2011. “FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies.” *Bioinformatics* 27 (21): 2957–63. doi:10.1093/bioinformatics/btr507.

Robertson, H. M., J. A. Navik, K. K. O. Walden, and H.-W. Honegger. 2006. “The Bursicon Gene in Mosquitoes: An Unusual Example of mRNA Trans-Splicing.” *Genetics* 176 (2): 1351–53. doi:10.1534/genetics.107.070938.

Shao, W., Q.-Y. Zhao, X.-Y. Wang, X.-Y. Xu, Q. Tang, M. Li, X. Li, and Y.-Z. Xu. 2012. “Alternative Splicing and Trans-Splicing Events Revealed by Analysis of the Bombyx Mori Transcriptome.” *RNA* 18 (7): 1395–1407. doi:10.1261/rna.029751.111.

Suzuki, M. G., F. Ohbayashi, K. Mita, and T. Shimada. 2001. “The Mechanism of Sex-Specific Splicing at the Doublesex Gene Is Different between Drosophila Melanogaster and Bombyx Mori.” *Insect Biochemistry and Molecular Biology* 31: 1201–11. doi:10.1016/S0965-1748(01)00067-4.

Wally, Verena, Eva M Murauer, and Johann W Bauer. 2012. “Spliceosome-Mediated Trans-Splicing: The Therapeutic Cut and Paste.” *Journal of Investigative Dermatology* 132 (8). Nature Publishing Group: 1959–66. doi:10.1038/jid.2012.101.

Wang, Kai, Darshan Singh, Zheng Zeng, Stephen J Coleman, Yan Huang, Gleb L Savich, Xiaping He, et al. 2010. “MapSplice: Accurate Mapping of RNA-Seq Reads for Splice Junction Discovery.” *Nucleic Acids Research* 38 (18): e178. doi:10.1093/nar/gkq622.

Warf, M. Bryan, and J. Andrew Berglund. 2010. “Role of RNA Structure in Regulating Pre-mRNA Splicing.” *Trends in Biochemical Sciences* 35 (3): 169–78. doi:10.1016/j.tibs.2009.10.004.

Weirather, Jason L, Pegah Tootoonchi Afshar, Tyson A Clark, Elizabeth Tseng, Linda S Powers, Jason G Underwood, Joseph Zabner, Jonas Korlach, Wing Hung Wong, and Kin Fai Au. 2015. “Characterization of Fusion Genes and the Significantly Expressed Fusion Isoforms in Breast Cancer by Hybrid Sequencing.” *Nucleic Acids Research* 43 (18): gkv562 – . doi:10.1093/nar/gkv562.

Chapter 5: Conclusions and Perspectives

5.1 The future of genome sequencing and assembly

The *An. stephensi* genome was assembled with a unique combination of 454, Illumina, PacBio and Bac-end sequencing. This is likely the only genome to ever be assembled with this combination of data types because 454 sequencing is no longer cost effective.

Currently, the most cost effective way to sequence and assemble a genome is using Illumina sequencing, but third-generation sequencing technologies are maturing rapidly. Pacific Biosciences (PacBio) sequencing seems to be getting better with every chemistry revision, and PacBio has been consistently delivering new chemistries which continue to improve read length. In our experience the N50 length of PacBio reads improved dramatically from when we initially sequenced *An. stephensi* in 2012 to when we sequenced *An. gambiae* in 2014. PacBio reads are getting so long that one of the biggest challenges is preparing a library with long enough DNA fragments (Panayotova et al. 2014). The BluePippin, an instrument that can perform automated size selection of ultra-long DNA fragments, is now being used in PacBio library preparation and has helped to increase read lengths (Panayotova et al. 2014).

The error-rate of PacBio sequencing was initially a huge challenge for software developers. Aligning and assembling very long reads with 10-15 percent error-rates was a completely new problem. Software to effectively utilize PacBio data is now rapidly improving (Berlin et al. 2015; Koren and Phillippy 2015). Initially, Illumina or 454 reads were aligned to the PacBio reads to correct the errors (Koren et al. 2012). However, this method tends to be extremely slow due to the massive number of reads that need to be aligned allowing for such a high error rate. A redeeming feature of PacBio sequencing is that errors are distributed randomly across the read. Therefore, it is possible to error-correct by consensus. Newer methods now error-correct by consensus by aligning shorter PacBio reads to longer ones (Berlin et al. 2015). If the sequencing was done with high enough depth, this approach can correct most of the errors in PacBio reads (Berlin et al. 2015). A big advantage of assemblies using PacBio data is the ability to obtain heterochromatic regions in the assembled genomes (Berlin et al. 2015). Genomes sequenced with Sanger technology, like *D. melanogaster* and *An. gambiae,* often included heterochromatic regions (Holt et al. 2002; Adams et al. 2000). However, the current crop of mosquito genome assemblies from the *Anopheles* 16 genomes project which were sequenced on the Illumina platform contain almost no heterochromatin (Daniel E. Neafsey et al. 2015). The future is bright for PacBio sequencing as a second-generation sequencing instrument has been announced that promises to dramatically increase throughput without a commensurate increase in cost.

Nanopore sequencing has been mentioned frequently as the next big thing in sequencing, but it is currently not widely used. A company called Oxford Nanopore has released the MinION sequencer. The size, cost, and speed of the MinION are initially impressive. The MinION, as it is only slightly larger than a standard USB flash drive, costs 900 dollars and as little as 500 dollars with a volume discount, and only takes a few hours to run. These attributes are highly-competitive against Illumina sequencing which requires a large initial capital investment and requires days or weeks for sequencing. However, the initial results from the MinION have been less than stellar (Goodwin et al. 2015). Error rates from the MinION appear to be around 35 percent which could make de-novo assembly using Oxford Nanopore data as the sole source of data very difficult (Goodwin et al. 2015). The throughput of the MinION is also going to severely limit its application in de-novo assembly project. I think the biggest advantage of Oxford Nanopore sequencing may be speed, not *de novo* assembly. For example, MinION devices could be taken on expeditions to sequence samples as they are collected or could be used by agencies like the CDC for real-time diagnosis of contagious disease.

One of the biggest questions about the future of genome sequencing and assembly is the demand for sequencing new genomes in the future versus resequencing to sample population-level genomic diversity. There have already been some large resequencing projects, and even bigger projects have been announced. Some of these notable resequencing projects include the 1000 human genomes project (Consortium 2012), and the 1000 *An. gambiae* genomes project. Both Google's Calico subsidiary and J. Craig Venter have announced plans to sequence tens-of-thousands of human genomes using the newly-announced Illumina HiSeq X Ten system, which has a theoretical throughput of 18,000 human genomes at a price of 1,000 dollars per year. If most of the genome sequencing demand is locked-up in resequencing, the market for improving sequencing technologies for *de novo* genome assemblies may not be big enough to justify the billions in research and development required to develop a new sequencing platform.

5.2 Comparison of the Indian strain and SDA strain assemblies of *Anopheles stephensi*

Two genomes of two strains of *An. stephensi*, Indian wild-type, and SDA, have recently been sequenced and assembled. These two assemblies provide a unique opportunity to compare and contrast different assemblies of essentially the same genome. It also provides an interesting opportunity to compare and contrast how differences in sequencing technology affect the quality of genome assemblies. Chapter 2 of this dissertation is on the assembly of the Indian wild-type strain of *An. stephensi,* so I will not go into great detail about the methods used. The SDA strain of *An. stephensi* was sequenced as part of a larger effort to sequence and assemble the genomes of 16 *Anopheles* mosquitoes (Daniel E. Neafsey et al. 2015). The sequencing and assembly of all genomes in the *Anopheles* 16 genomes project were performed at The Broad institute using the Illumina platform and the ALLPATHS LG assembler (Gnerre et al. 2011; D. E. Neafsey et al. 2013). Due to the high heterozygosity in mosquito genomes, most of the DNA used was isolated from a single female individual (D. E. Neafsey et al. 2013). Long insert Fosill libraries were also generated with insert sizes around 38-40 kb and greatly improved scaffolding (D. E. Neafsey et al. 2013).

I will start by comparing the contigs of these two assemblies. There are 31,761 and 8,946 contigs respectively in the Indian and SDA strains of the *An. stephensi* genome assemblies (Jiang et al. 2014; Daniel E. Neafsey et al. 2015). The N50 contig size of the Indian strain is 36,511 while the N50 contig size of the SDA strain is 72,570 (Jiang et al. 2014; Daniel E. Neafsey et al. 2015). While the number of contigs and contig N50 size favor the SDA strain assembly, the Indian assembly contains more bases (Jiang et al. 2014; Daniel E. Neafsey et al. 2015). The Indian strain contigs contain 209 Mb of sequence while the SDA strain contigs contain only 196 Mb of sequence (Jiang et al. 2014; Daniel E. Neafsey et al. 2015). One explanation for the discrepancy in contig number can be explained because the distribution of the length of contigs has a long tail. This means that there are few contigs that are very long, but many more that are very short. These short sequences generally contain repetitive sequences that often can not be assembled. The GC percentages were 44.80 percent and 45.02 for the Indian and SDA contigs respectively (Jiang et al. 2014; Daniel E. Neafsey et al. 2015).

Next, I will compare the scaffolds of the two assemblies. There are 23,371 scaffolds in the Indian strain assembly and 1,114 scaffolds in the SDA assembly (Jiang et al. 2014; Daniel E. Neafsey et al. 2015). This approximately 20 fold difference in the number of scaffolds is indicative of the different strategies used by the assemblers. The vast majority of this discrepancy comes from the minimum sequence length. The minimum sequence length of the Indian strain scaffolds is 486 bp; while the minimum sequence length of the SDA scaffolds is 1,000 bp (Jiang et al. 2014; Daniel E. Neafsey et al. 2015). The 1,000 bp minimum length of the SDA scaffolds is an arbitrary cutoff of the ALLPATHS LG assembly algorithm. In contrast, the Newbler assembler carries over sequences less than 1,000 that do not contain a gap to preserve information. The fact that there are a huge number of short contigs that are carried over into the scaffolds explains the 20x difference in the number of scaffolds between the assemblies. We would argue that the inclusive approach used by Newbler is better because there is interest in short sequences. For example, many of the short sequences contain repeats that could contain heterochromatic sequences that could be annotated. For example, satellite sequences are often present in the shortest of contigs. Another example is Y chromosome sequences, which are mostly shorter than 1,000 bp.

The next metric I will compare is the N50 scaffold size. N50 scaffold size is viewed to be one of the most important metrics for a genome assembly because the scaffolds are often what are actually used in downstream analysis. The N50 scaffold size of the Indian strain assembly is 1,591,355 bp while the N50 scaffold size of the SDA strain is only 837,295 bp (Jiang et al. 2014; Daniel E. Neafsey et al. 2015). This is quite a big different in N50 scaffold size and is quite surprising because the longest mate-pairs used in the Indian strain assembly were 20 kb, while the SDA strain assembly used ultra-long mate pairs around 38-40 kb. The longest scaffold of the Indian strain genome is also longer than the longest scaffold of the SDA strain genome at 5,975,090 bp and 3,396,703 bp respectively (Jiang et al. 2014; Daniel E. Neafsey et al. 2015). Another big discrepancy between the scaffolds is the percentage of gaps (N's). The Indian strain assembly is 5.35 percent gaps while the SDA strain assembly is 12.95 percent gaps (Jiang et al. 2014; Daniel E. Neafsey et al. 2015). We hypothesize that the difference in the gaps comes from the very long mate-pair reads used to scaffold the SDA assembly. While the contig metrics tend to favor the SDA strain assembly, the scaffold metrics strongly favor the Indian strain assembly as the superior assembly.

Next, I will compare the annotations of the two assemblies. The Indian strain annotation had 11,789 genes with 49,269 total exons (Jiang et al. 2014; Daniel E. Neafsey et al. 2015). The SDA strain annotation had 13,113 genes with 56,153 total exons (Jiang et al. 2014; Daniel E. Neafsey et al. 2015). There are many potential reasons for the differences in genes, like genes split between multiple contigs annotated as more than one gene.

The *Anopheles* 16 genomes project choose to sequence female individuals due to coverage of the X (D. E. Neafsey et al. 2013). Males have one X and one Y chromosome while females have two X chromosomes. Therefore, if males were sequenced, the coverage of the X chromosome would be half that of the autosomes. If a pool of mixed males and females were sequenced, the coverage of the X chromosome would be three-fourths that of autosomes. The *Anopheles* 16 genomes project erred on the side of caution and only sequenced females to produce a better assembly of the X. The Indian strain assembly was sequenced from a pool of mixed males and females and therefore contains Y chromosome sequences (Hall et al. 2013). Choosing the individual or individuals to be sequenced is a huge decision for a genome project, and both choices can be justified.

5.3 Iso-Seq and the future of transcriptome sequencing

Microarrays and RNA-Seq have revolutionized our understanding of gene expression but a new technology, PacBio Iso-Seq is superior in many ways. The major advantage of Iso-Seq is that full length transcripts can be sequenced without the need for assembly. Assembly of RNA-Seq is a very computationally intensive process and often results in high numbers of chimeric contigs.

Furthermore, assessing the isoforms present and their relative abundance present an even bigger challenge compared to assembly. We used PacBio Iso-Seq in the very early stages of its development, so much so that we had to specifically contact a sequencing center to ask to have it performed. Using the PacBio Iso-Seq, we were able to identify trans-spliced mRNAs, which are extremely difficult to identify with RNA-Seq. In the future, PacBio Iso-Seq may lead to huge improvements in the annotations of the 5' and 3' UTRs of genes. I think that PacBio Iso-Seq will turn out to be a technological revolution in transcript sequencing, and I am very glad to have used it so early after its development.

5.4 Reference

Adams, Mark D, Susan E Celniker, Robert A Holt, Cheryl A Evans, Jeannine D Gocayne, Peter G Amanatides, Steven E Scherer, Peter W Li, Roger A Hoskins, and Richard F Galle. 2000. "The Genome Sequence of Drosophila Melanogaster." *Science* 287 (5461). American Association for the Advancement of Science: 2185–95.

Berlin, Konstantin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. 2015. "Assembling Large Genomes with Single-Molecule Sequencing and Locality-Sensitive Hashing." *Nat Biotech* 33 (6). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 623–30.

Consortium, 1000 Genomes Project. 2012. "An Integrated Map of Genetic Variation from 1,092 Human Genomes." *Nature* 491 (7422). Nature Publishing Group: 56–65.

Gnerre, Sante, Iain MacCallum, Dariusz Przybylski, Filipe J Ribeiro, Joshua N Burton, Bruce J Walker, Ted Sharpe, et al. 2011. "High-Quality Draft Assemblies of Mammalian Genomes from Massively Parallel Sequence Data." *Proceedings of the National Academy of Sciences* 108 (4): 1513–18. doi:10.1073/pnas.1017351108.

Goodwin, Sara, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael Schatz, and W Richard McCombie. 2015. "Oxford Nanopore Sequencing and de Novo Assembly of a Eukaryotic Genome." *bioRxiv*, January.

Hall, Andrew Brantley, Yumin Qi, Vladimir Timoshevskiy, Maria V Sharakhova, Igor V Sharakhov, and Zhijian Tu. 2013. "Six Novel Y Chromosome Genes in Anopheles Mosquitoes Discovered by Independently Sequencing Males and Females." *BMC Genomics* 14 (1): 273. doi:10.1186/1471-2164-14-273.

Holt, Robert A., G Mani Subramanian, Aaron Halpern, Granger G Sutton, Rosane Charlab, Deborah R Nusskern, Patrick Wincker, et al. 2002. "The Genome Sequence of the Malaria Mosquito Anopheles Gambiae." *Science* 298 (5591): 129–49. doi:10.1126/science.1076181.

Jiang, Xiaofang, Ashley Peery, A Hall, Atashi Sharma, Xiao-Guang Chen, Robert M Waterhouse, Aleksey Komissarov, et al. 2014. "Genome Analysis of a Major Urban Malaria Vector Mosquito, Anopheles Stephensi." *Genome Biology* 15 (9): 459. doi:10.1186/s13059-014-0459-2.

Koren, Sergey, and Adam M Phillippy. 2015. "One Chromosome, One Contig: Complete Microbial Genomes from Long-Read Sequencing and Assembly." *Current Opinion in Microbiology* 23 (February): 110–20. doi:http://dx.doi.org/10.1016/j.mib.2014.11.014.

Koren, Sergey, Michael C Schatz, Brian P Walenz, Jeffrey Martin, Jason T Howard, Ganeshkumar Ganapathy, Zhong Wang, et al. 2012. "Hybrid Error Correction and de Novo Assembly of Single-Molecule Sequencing Reads." *Nat Biotech* 30 (7). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 693–700.

Neafsey, D. E., G. K. Christophides, F. H. Collins, S. J. Emrich, M. C. Fontaine, W. Gelbart, M. W. Hahn, et al. 2013. "The Evolution of the Anopheles 16 Genomes Project." *Genes|Genomes|Genetics* 3 (7): 1191–94. doi:10.1534/g3.113.006247.

Neafsey, Daniel E., Robert M. Waterhouse, Mohammad R. Abai, Sergey S. Aganezov, Max A. Alekseyev, James E. Allen, James Amon, et al. 2015. "Highly Evolvable Malaria Vectors: The Genomes of 16 Anopheles Mosquitoes." *Science* 347 (6217): 1258522–1258522. doi:10.1126/science.1258522.

Panayotova, N G, X H Zhou, G Yuan, D A Moraga, and S Shanker. 2014. "Optimization of Library Construction Protocol to Sequence Large Fragment Libraries on PacBio." *Journal of Biomolecular Techniques : JBT* 25 (Suppl). Bethesda, MD: Association of Biomolecular Resource Facilities: S13–S13.