Identification and Characterization of Y Chromosome and M Locus Genes in *Anopheles* and *Aedes* Mosquitoes Using the Chromosome Quotient Method

Andrew Brantley Hall

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)
Zach Adelman
Igor Sharakhov
Liqing Zhang

March 2nd, 2016
Blacksburg, VA

Keywords: Y chromosome, M locus, sex-determination, male-determining factor, chromosome quotient

Identification and Characterization of Y Chromosome and M Locus Genes in *Anopheles* and *Aedes* Mosquitoes Using the Chromosome Quotient Method

Andrew Brantley Hall

ABSTRACT

In mosquitoes, sex determination is initiated by a dominant male-determining factor located on the Y chromosome in *Anopheles* mosquitoes or in a small Y-like region called the M locus in *Aedes* mosquitoes. Before my research, not a single gene from the *Anopheles* Y or *Aedes* M locus had ever been discovered.

During the course of my undergraduate research in the Tu lab, I developed the chromosome quotient (CQ) method which identifies Y chromosome/M locus sequences by comparing the ratio of alignments from separate pools of female and male Illumina sequencing data. The focus of my dissertation is using the CQ method to identify potential male-determining factors in *Aedes* and *Anopheles* mosquitoes.

First, we identified a novel gene tightly-linked to the M locus in *Aedes aegypti* called *myo-sex*. *Myo-sex* encodes a myosin heavy chain protein that is highly expressed in the pupa and adult male. *Myo-sex* is generally only found in males, but can sporadically be found in females due to a rare recombination. The fact that *myo-sex* can be found in females combined with a lack of early-embryonic expression suggests that *myo-sex* is not the male-determining factor.

Next, we identified a gene in *Aedes aegypti*, *Nix*, which appeared to be persistently linked to the M locus and was expressed in the early embryo. *Nix* shows distant similarity at the amino acid level to *Transformer2*, a gene involved in the sex determination pathway of *Drosophila melanogaster*. *Nix* knockout with CRISPR/Cas9 resulted in feminization of genetic males and the production of the female isoforms of *doublesex* and *fruitless*, two key regulators of downstream sexual differentiation. Ectopic expression of *Nix* resulted in masculinization of genetic females. Based on these results, we concluded that *Nix* is a male-determining factor in *Aedes aegypti*.

We also characterized large portions of the *Anopheles gambiae* Y chromosome using PacBio sequencing and the CQ method. We discovered that 92.3 percent of predicted Y sequences fell into two classes, the *zanzibar* amplified region (ZAR) and the satellite amplified region (SAR). This analysis fills in a large piece of the *Anopheles gambiae* genome missing since 2002.

Public Abstract

Female mosquitoes feed on vertebrate blood to obtain protein necessary for egg development. In the process of blood-feeding, mosquitoes transmit pathogens that cause diseases including: malaria, dengue fever, Chikungunya, and Zika. On the other hand, male mosquitoes do not blood feed and therefore do not transmit disease.

Mosquito control strategies that use harmless males to control the populations of deadly females have been proposed but have been infeasible due to the lack of basic understanding of sex determination. In mosquitoes, the master gene responsible for beginning the sex determination cascade is located on the Y chromosome or a Y-like region called the M locus. Y chromosome or M locus sequences are rarely assembled due to the massive numbers of repeats present in these regions. Here, we use a newly-developed method to find genes from the Y chromosome and M locus. We identified the first M locus genes from the yellow fever mosquito, *myo-sex* and *Nix*. Removing *Nix* from male mosquitoes resulted in feminization and adding *Nix* to female mosquitoes resulted in masculinization. Thus, we concluded that *Nix* is the male-determining factor in the yellow fever mosquito. *Nix* is the first male-determining factor to be identified in any insect. The discovery of *Nix* paves the way for mosquito strategies that would reduce transmission of mosquito-borne pathogens by converting deadly females into harmless males.

To address the lack of an assembly of the Y chromosome of the African malaria mosquito *Anopheles gambiae*, we constructed a database of redundant PacBio reads representative of the non-recombining region of the *Anopheles gambiae* Y chromosome called Ydb. Ydb represents a major improvement compared to the previous reference sequences available for the *Anopheles gambiae* Y. Using Ydb, we determined the gene content and repeat structure of the *Anopheles gambiae* Y chromosome.

Acknowledgements

First I would like to thank my parents who have been incredibly supportive and encouraging through both my undergraduate and graduate degrees. I especially want to thank my wife Xiaofang Jiang who has helped in almost every aspect of my research and helped me become a much better scientist. I would also like to thank my advisor Jake Tu for giving me unparalleled freedom to pursue my research interests and supporting my ideas to completion. Without Dr. Tu's insights, none of this research would have been possible. I would like to thank all the other members of my committee Zach Adelman, Igor Sharakhov, and Liqing Zhang for their input, ideas, and support. I especially want to thank Zach Adelman for believing that Nix could be the male-determining factor and going forward with knocking it out. I would like to thank Igor Sharakhov for getting me involved in the *Anopheles* 16 genomes project and reading all my papers and providing prompt feedback. I would like to specifically thank Maria Sharakhova the Sharakhov lab in whole for all their FISH to support that the genes I found are Y/M-linked – nothing drives home the message like one of their FISH images. I would like to thank my lab mates and collaborators at Virginia Tech People who have helped along the way including: Frank Criscione, James Biedler, Yumin Qi, Wanqi Hu, Randy Saunders, Atashi Sharma, Vladimir Timoshevskiy, Ashley Peery, Sanjay Basu, and Michelle Anderson. I would like to thank Dr. Bevan and Dennie Munson who do an excellent job of running the GBCB program. The NSF East Asian and Summer Institutes (EAPSI) which was a wonderful experience. I would like to think my EPASI host, Xiao-Guang Chen, for the welcoming environment in Guangzhou. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE-1519168) and National Institutes of Health Grants No. AI113643 and AI105575.

Attributions

Chapters 1 and 5 were written by the candidate to provide background and conclusions for the three manuscripts that make up the body of the dissertation. Chapters 2 and 3 were the primary project of the candidate but other members contributed their own efforts and need to be acknowledged. Chapter 4 was a large consortium where the candidate contributed large portions of the final results.

Chapter 2

ABH – Identified *myo-sex* and M-linked BAC using the CQ method, performed PCR to verify male-specific amplification, sequenced the full-length myo-sex gene, drafted the manuscript and made the figures. VT and MS – Mapped *myo-sex* and the BAC using FISH. XJ – Performed bioinformatic analysis of *myo-sex* to construct the phylogeny. SB and MA – Contributed and maintained transgenic mosquito lines with transgenes inserted around the M locus. WH – Performed quantitative PCR on myo-sex in recombinant individuals. IS – Initiated the project, coordinated FISH analysis, critically reviewed the manuscript and figures. ZA– Initiated the project and coordinated the genetic crosses, helped to revise the manuscript and figures. ZT– Initiated and coordinated the project, helped to write and revise the manuscript and figures.

Chapter 3

ABH identified and characterized Nix. SB performed the CRISPR/Cas9 and ectopic expression experiments. XJ performed RNAseq analysis. YQ and JKB designed and performed RT-PCR and ddPCR assays. VT, MVS and IVS. performed FISH and interpreted the data. JKB and RE performed PCR and cloned Nix in nix- individuals. MAEA provided double-marked *A. aegypti* and materials for *A. albopictus* analysis. XC sequenced male *A. albopictus*. ZT and ZNA

initiated and designed the study. ABH, ZT, and ZNA wrote the manuscript with input from SB and IVS.

Chapter 4

Conceived project: NJB, SJE, IVS, ZT; Coordinated project: NJB; Genome and BAC sequencing: NHB, JH, DR, AMP, ABH, ZT, NJB; PacBio read correction and assembly: SK, AMP; RNA-Seq, RT-PCR and gene validation: PAP, CC, OSA, ACa, TD, EF, RG, TN, NW, ACr; Computational analysis of Y linkage: ABH, PAP, CC, LA, XJ, ASt, SZ, MWH, SJE, ZT; Cytogenetics and fluorescence *in situ* hybridization: ASh, MVS, VAT, IVS; Phylogeny reconstruction and simulations: CC, MWH. Wrote paper: NJB, ZT, ABH, PAP with input from other authors. Specific contributions of ABH in computational analysis: Identified Y-linked PacBio reads and created Ydb, a database of Y sequences in Anopheles gambiae, identified Y-linked BACs using the CQ method, identified Y genes using the CQ method and caldera pipeline.

Table of Contents

List of figures

List of tables

Chapter 1: Introduction

1.1 Sexual reproduction and sex determination

In mosquitoes there are two sexes: males and females (*1*, *2*). Both male and female mosquitoes feed on nectar, but for females the diet of nectar does not contain the necessary nutrients for egg production (*1–3*). Therefore, female mosquitoes feed on blood to obtain these nutrients while males do not (*1–3*). Blood-feeding females serve as a potent vector for blood-borne pathogens while the non-blood-feeding males do not serve as a vector (*2*). In terms of disease burden, the most important pathogens vectored by mosquitoes include: Plasmodium, the causative agent of malaria, yellow fever virus, and dengue fever virus (*4*). These pathogens infect hundreds of millions annually, so research on methods to control these diseases is of critical importance (*4*). Methods harnessing the genetic differences between male and female mosquitoes to reduce mosquito populations have been proposed but due to a lack of basic understanding of mosquito sex determination, these methods were infeasible (*1*).

1.1.1 Advantages of sexual reproduction

Mosquitoes reproduce by sexual reproduction of sexually dimorphic male and female individuals (*2*). However, this is just one of many variations of reproduction present in living organisms (*5*, *6*). The simplest form of reproduction is asexual where a single individual duplicates its genetic material and divides in two (*5*). Almost all bacteria and archaea reproduce by some variation of this method (*5*). It is alluring to think that due to the simplicity of asexual reproduction it should predominate among most multicellular forms of life as well. However, because genomes of the offspring resulting from asexual reproduction differ only in the few errors induced by DNA polymerase during genome replication, there are limited opportunities on which natural selection can act (*5*). Organisms limited to asexual reproduction may fail to adapt

to changing environmental conditions (*5*, *7*). Furthermore, it is difficult for organisms limited to asexual reproduction to eliminate deleterious genes from their genomes (*7*). Bacteria and archaea can partially overcome this problem through lateral transfer of genes through conjugation and random horizontal transfer (*8*).

Sexual reproduction, which joins alleles from different individuals through fusion of gametes, can combine genes in new combinations that prove to be advantageous and can also separate beneficial and deleterious genes (*5*, *7*). The genetic recombination resulting from sexual reproduction has huge advantages over asexual reproduction (*7*). In fact, almost all animals produce offspring by sexual reproduction at some time in their lives (*5*).

1.1.2 Hermaphroditism and Dioecy

Separate sexes are not necessary for sexual reproduction as hermaphroditism, where individuals act as both "males" and "females", is common in many animal taxa (*9*, *10*). However, dioecy, where males and females are distinct entities where males produce small mobile gametes (sperm) and females produce large sessile gametes (eggs), is more common than hermaphroditism in animals and has evolved multiple independent times (*5*).

The genetics of hermaphroditism are conceptually simple because every individual of a species carries out the same program for the development of sexual organs and behavior (*5*, *9*). Dioecy significantly complicates the matter because about half of the individuals need to develop as males and the other half needs to develop as females (*5*). Dioecy requires a switch to determine whether an individual will develop into a male or female (*5*). These switches can be either environmental or genetic (*5*, *11*). Temperature is the most prevalent environmental switch, and intuitively sex determination by temperature is called temperature-dependent sex determination (*11*, *12*). Temperature-dependent sex determination is prevalent in turtles and

other reptiles like alligators and crocodiles where the temperature experienced by the embryo during development determines whether the individual embryos develop into males or females (*11*).

1.1.3 Genetic sex determination and sex chromosomes

In animals with genetic sex determination, a genetic switch inherited from one of the parents is responsible for leading the sex determination cascade down the male or female program (*5, 12, 13*). The genetic switch is generally a dominant gene that initiates either the male or female program of sex determination when present, but exceptions exist (*5, 12, 13*). The lack of the dominant gene generally results in the opposite development plan (*5, 12, 13*). A dominant sex determining gene must be inherited by only half of offspring to maintain a balanced sex ratio, but most genes do not show this pattern of inheritance (*13*). At least initially, a dominant sex determining gene could exist as a single locus with dominant (functional) and recessive (non-functional) alleles (*13*). Thus, half of the offspring would receive the functional sex determining gene while the other half would not (*13*).

However, in the examples so far characterized the most common configuration is a dominant sex-determining gene located on a sex-limited chromosome (*5, 12, 13*). For a balanced sex ratio to be maintained, only one copy of the sex-limited chromosome must be present. For proper meiosis, the sex-limited chromosome must pair with another chromosome during meiosis. One common configuration that can satisfy these requirements in an XY/XX arrangement (*12, 14*). Here, the Y is the sex-limited chromosome with the dominant male-determining gene. Males have one Y and one X and the Y pairs with the X during meiosis. Females have two X chromosomes. In the XY/XX configuration, males are called heterogametic sex because half of gametes produced by these individuals carry the Y chromosome while the other half carry the X

chromosome (*15*). Females are called homogametic because all eggs contain the same complement of chromosomes (*15*). The X and Y are called sex chromosomes (*5*, *16*, *17*). Sex chromosomes differ from autosomes in the fact that they are not homologous across their entire length (*5*, *18*). X and Y chromosomes are found in many animal taxa including therian mammals, flies, beetles, and true bugs (*5*, *6*, *19–21*).

An alternative configuration is possible when females are the heterogametic sex. The chromosomes in this configuration are called ZW/ZZ (*5*, *6*). In this case, a dominant female-determining gene is located on the W chromosome and females have one W and one Z (*5*, *6*). Males have two Z chromosomes. Birds, snakes, monotremes, and butterflies tend to have Z and W chromosomes (*5*, *6*, *11*, *22*, *23*).

One feature that varies among sex chromosomes is the amount of differentiation between the X and Y (or Z and W) (*13*, *19–21*, *24*, *25*). Sex chromosomes range from entirely differentiated as in *Drosophila* species to only marginally differentiated in Medaka fish (*26–28*). In *Drosophila* species, there is no recombination in males which has caused the X and Y to become completely differentiated (*26*, *29*). In contrast, the Medaka Y contains only a 280 kb differentiated region (*27*, *28*). The differentiated region of a Y or W chromosome are called the male-specific region and female-specific region respectively (*13*, *19*). The undifferentiated parts of the Y or W that mediate pairing with the X or Z are called the pseudoautosomal regions (*13*, *19*). The details of the degeneration of sex chromosomes that lead to these differences will be discussed in detail later in this chapter. There are also cases when the X/Y and Z/W chromosomes are indistinguishable by morphological markers (*6*, *13*, *22*). Due to the inability to see the differentiated region of the sex chromosome, these cases are generally referred to as

homomorphic sex chromosomes (*13*). However, many times a differentiated region is present just not visible (*30*).

1.1.4 Sex determination in humans

Sex determination in human males is initiated by the presence of the testis-determining gene *Sry* carried on the Y chromosome (*31–34*). In the canonical experiment that proved involvement of *Sry* in testis development, XY individuals that had developed as females were shown to have mutations in *Sry* (*32*).

Sry is a SOX (*Sry*-like box) transcription factor that likely arose from a duplication of *SOX3* to the Y around 180 million years ago (*20*, *34–36*). Transcription of *Sry* begins around 6-8 weeks after zygote formation (*36*). The *Sry* protein interacts with the steroidogenic factor 1 protein to activate the transcription of *SOX9* (*33*, *35*, *36*). SOX9 transcription activates a cascade of genes that result in the bipotential gonad developing into testes which then produce testosterone putting the embryo on the path to male development (*35*).

1.1.5 *Drosophila* sex determination

In 1916, Calvin Bridges observed that XXY flies develop into females and X0 flies develop into sterile males (*16*). These results contrast with mammals where XXY individuals develop into males and X0 individuals develop into females (*34*). Based on this observation, Bridges concluded that the Y is not male-determining in *Drosophila* (*16*, *17*). Thus, sex in *Drosophila* species is determined by the number of X chromosomes (*37*). The question is how does a *Drosophila* embryo know how many X chromosomes it has?

X counting involves "numerator proteins", *Sisterless-a* and *Sisterless-b*, produced by genes on the X chromosome and autosomally-encoded "denominator proteins" *Deadpan* and *Extramacrochaetae* (*37*, *38*). The numerator and denominator gene products interact in a dosage-

dependent fashion (*38*). The "denominator proteins" are always present in two copies because they are located on autosomes (*38*). The amount of "numerator protein" differs depending on the number of X chromosomes present (*38*). If only one X chromosome is present as in males, the "denominator proteins" are in excess and block the effects of the "numerator proteins" (*38*). If two X chromosomes are present as in females, excess "numerator proteins" are free to act on their downstream target *Sex-lethal* (*Sxl*) (*38*). The *Sxl* gene has two promoters, an early promoter and a late promoter. *Sisterless-a* and *Sisterless-b* act to promote the transcription of *Sxl* from the early promoter (*37*, *38*). Thus, *Sxl* is transcribed from the early promoter only in female flies (*37*). When *Sxl* is transcribed from its late promoter, the product of early *Sxl* transcription in females splices out a premature stop codon in the late *Sxl* primary transcript producing the female isoform of the late *Sxl* (*37, 38*). In males where no early *Sxl* product is present, the premature stop codon is not spliced out resulting in a truncated non-functional late *Sxl* protein (*37*, *38*). Functional *Sxl* protein in females then regulates the splicing of *transformer*, the next gene in the sex determination cascade (*37–39*). Mirroring the splicing of the *Sxl* mRNA, functional *Sxl* in females splices out a premature stop codon from the *transformer* mRNA (*38, 39*). Thus, females produce a functional *transformer* protein while males produce a truncated non-functional version (*38*, *39*). Functional *transformer* then goes on to splice the *doublesex* and *fruitless* which are transcription factor that regulate somatic sexual differential and sex-specific nervous system development respectively (*37–42*). *Doublesex* and *fruitless* differ between males and females in a single female-specific exon (*41*, *42*). Functional transformer in females interacts with a series of proteins including *transformer2* to activate the weak 3' splice side of this female exon in *doublesex* and a similar activity to promote the inclusion of the female-specific exon in

the *fru* transcript (*37*, *38*, *41*). Thus, a difference in the number of X chromosomes leads to male and female flies (*37*, *38*).

1.1.6 Other characterized sex determination mechanisms

Genes involved in sex determination have been found with varying degrees of evidence in several other species. Here, I will briefly review these characterized sex-determining genes and the interesting role of DM-domain proteins in sex determination.

In the Medaka fish, *Oryzias latipes*, the male specific region of the Y chromosome is tiny, encompassing only 280,000 bp and containing a single gene (*27*, *28*). This gene is a duplicated copy of DMRT1 called DM-Y (*27*). Interestingly, DMRT1 is related to *doublesex* in *Drosophila* (*27*). Deletion experiments combined with the fact that DM-Y is the only gene in the male-specific region of the Y indicate DM-Y is likely the male-determining factor in Medaka (*27*, *28*).

In the African clawed frog, *Xenopus laevis*, a DMRT1 homolog is also implicated in sex determination (*43*). *Xenopus* species are female-heterogametic ZZ/ZW (*43*). A truncated copy of DMRT1 called DM-W was found on the W chromosome (*43*). Interestingly, DM-W has the DNA-binding domain but lacks the transactivation domain characteristic of DM-domain containing proteins (*43*). Therefore, DM-W could act as a dominant negative repressor. The gonads of transgenic male (ZZ) tadpoles with DM-W were partially feminized implicating DM-W as the female-determining factor in *Xenopus laevis* (*43*).

A DMRT1 homolog is also implicated in sex determination in chickens (*44*). Birds are female-heterogametic ZZ/ZW (*44*). In chickens, DMRT1 is Z-linked and may act in a dosage-dependent manner (*44*). When DMRT1 was knocked down with RNAi male gonads exhibited feminization (*44*). Thus, two copies of DMRT1 appear to be necessary for testes development,

but this does not preclude the presence of a dominant female-determining factor located on the W (*44*).

Silkworms (*Bombyx mori)* are female-heterogametic ZZ/ZW (*45*). Recently, a piRNA located on the W called Fem was identified to be the dominant female-determining factor (*23*). Fem negatively regulates a gene called Masc which when present causes male-specific splicing of *doublesex* (*23*). When Fem is present and negatively regulates Masc, the female-specific isoform of *doublesex* predominates (*23*). Conversion phenotypes were not observed because Masc is also involved in dosage compensation which proved fatal when misregulated (*23*).

1.2 The canonical theory of the degeneration of Y chromosomes

Dominant sex determining genes are often present on degenerate sex chromosomes (*13*). Interestingly, much of the difficulty in the identification of dominant male or female determining genes arise because they are located on these degenerate sex chromosomes (*13, 24, 46, 47*). The canonical theory of Y chromosome evolution presented by Brian and Deborah Charlesworth in their 2000 and 2005 papers "*The Degeneration of Y chromosomes*" and "*Steps in the Evolution of Heteromorphic Sex Chromosomes*" describe how a normal pair of autosomes degenerate into a heteromorphic X and Y (*24*, *46*). Briefly, after one of the autosomes acquires a sex-determining gene, linkage between sexually antagonistic genes and the sex-determining gene causes recombination to stop in a small region (*24*, *46*). Cessation of recombination is advantageous because it limits the sex-determining gene and the surrounding sexually antagonistic genes to the sex they benefit, shielding the opposite sex from their deleterious effects. At this point, the chromosome with the sex-determining gene is called the proto-Y. The non-recombining region usually progressively expands to encompass most of the Y chromosome due to the advantageous effect of limiting sexually antagonistic genes to the sex they benefit (*13, 46*).

The canonical theory of Y chromosome evolution accurately describes the evolution of the mammalian Y chromosome. The mammalian Y stopped recombining with the X in discreet stages creating evolutionary strata with different levels of differentiation between the X and the Y (*19*). There is clear homology between the genes on the X and Y in the evolutionary strata (*19*, *20*). There are a few other examples of Y chromosomes that have clear homology with the X (*13*, *48*). However, most of these fall in the new-Y category (*13*).

In insects the origin and evolution of Y chromosomes is not so clear. Insect Y chromosomes originated independently multiple times shrouding their origins in mystery (*49*). Several studies have discussed the possibility of Y-replacement in *Drosophila* (*26*, *50*). *D. melanogaster* and most other *Drosophila* species have old Y chromosomes where the origin is unknown. No homology (besides rRNA) has ever been found between the *D. melanogaster* X and Y (*25*, *26*). Due to the lack of homology between the X and Y, it is unclear whether the canonical theory of Y chromosome evolution can be directly applied to the *Drosophila* Y (*26*, *50*).

Rapid Y chromosome replacement in *Drosophila* is facilitated by the fact that no recombination occurs in male *Drosophila* and the lack of a sex-determining gene on the *Drosophila* Y (*25*, *26*, *29*, *50*). These facts allow for the formation of neo-Y chromosomes in *Drosophila* species (*25*, *29*). The formation of a neo-Y occurs when a fusion of an X and autosome is fixed in a population (*25*, *29*). In this scenario the old Y is either lost or absorbed by an autosome. Examples of neo-Y chromosomes are found in *D. pseudoobscura* and *D. miranda* (*25*, *29*). These examples of non-canonical Y chromosome origin caused extensive speculation about the origin of the *D. melanogaster* Y (*26*).

Non-canonical origin of the Y chromosome was recently documented in Homoptera (*26*). Most Homoptera have XX/X0 sex-determination where females have two X chromosomes while

males have just one (*26*). Independently, two species of Homoptera recently gained a Y chromosome from a supernumerary B chromosome (*26*). This is a surprising origin for a Y chromosome (*26*).

1.3 The *Anopheles* Y chromosome

Anopheles mosquitoes have well-differentiated X and Y chromosomes (*51*). The long arm of the X and Y tend to appear similar in many species, but the short arms of the X and Y appear highly divergent (*52*). The *Anopheles* Y chromosome is thought to have a male-determining function because a XXY triploid individual was observed to be male (*51*). Very few sequences totaling only 180kb from the *Anopheles* gambiae Y had been identified before CQ analysis (*53*, *54*). These sequences were almost entirely repetitive (*54*). Here we review what little is known about the *Anopheles gambiae* Y based on cytogenetics.

Cytogenetic literature teems with examples of heterochromatin variation on the X and Y chromosomes of *Anopheles* mosquitoes. These studies document the apparent morphological similarity between the heterochromatic arms of the *Anopheles* X and Y chromosomes and demonstrate the potential for recombination between the heterochromatic arms (*52*, *55*).

1.3.1 Morphological similarity between the heterochromatic arms of the X and Y

Many researchers observed morphological similarity between the long heterochromatic arms of the X and Y chromosomes of *Anopheles* mosquitoes. Four papers note the apparent similarity in size and banding pattern between the heterochromatic arms of the X and Y chromosomes in *An. atroparvus* (*55–58*). Further similarity between the heterochromatic arms of the X and Y chromosomes can be found in *An. albimanus* where the long arms of the X and Y were declared homologous based on appearance (*59*). Two more examples of similarity in size between the heterochromatic arms of the X and Y are found in *An. willmori* and *An. culicifacies*

(*52*, *60*). As an interesting side note, all evidence for a male-determining gene on the *Anopheles* comes from *An. culicifacies* (*51*). In *An. stephensi* the long arms of the X and Y are similar in size and the satellite Ast72A hybridizes equally to both the X and Y (*61*).

1.3.2 The long arms of the *Anopheles* X and Y can recombine

Papers from three separate *Anopheles* species report recombination or the potential for recombination between the long arms of the X and Y (*52*, *56*, *62*). Experiments in *An. culicifacies* have shown that the long heterochromatic arms of the X and Y, which appear morphologically similar, can recombine (*52*, *63*). The same paper also shows a lack of recombination between the short, presumably differentiated, arms of the X and Y (*63*). In *An. atroparuvs* it was shown that the X and Y pair during meiosis and form chiasmata (*55*). Evidence for the potential of recombination in *An. quadrimaculatus* Species A was found using a male-linked translocation strain. Recombination was presumed based on circumstantial evidence, but the possibility for recombination was confirmed based on the pairing of the X and Y during meiosis (*62*).

A recent study on rDNA also provides indirect evidence for recombination between the rDNA loci on the long arms of the X and Y (*64*). This study was performed on four species of the *An. albitarsis* complex. In these species the rDNA arrays exist on both the X and Y, the inferred ancestral location of the rDNA loci (*65*). The authors cloned the ITS2 spacer between the 5.8s and 28s rDNA genes from both males and females of the four species (*64*). In total, they sequenced 217 clones (*64*). Interestingly, no differences were found between the sequence of ITS2 in males and females of the four species (*64*). This serves as indirect evidence of X-Y recombination because there should be differences between the X and Y ITS2 if no recombination occurred (*64*).

1.3.3 *Anopheles* Y chromosome polymorphism

The size of the *Anopheles* X and Y chromosomes varies drastically between species and even within species due to variation in the amount of constitutive heterochromatin (*60*). In fact, the amount of heterochromatin on the X and Y was used to differentiate cryptic *Anopheles* species before molecular markers were developed. Intra and interspecies variation in the amount of constitutive heterochromatin is a common occurrence in many organisms (*60*).

Why are the X and Y so amenable to gain or loss of blocks of constitutive heterochromatin? Compared to the autosomes, heterochromatin occupies an enormous portion of the *Anopheles* X and Y chromosomes. Besides the rDNA genes, no known genes reside within these blocks of constitutive heterochromatin on the X and Y (*66*). Large duplications and deletions of constitutive heterochromatin may cause no apparent phenotypic effects so there may not be strong selection for or against the amount of constitutive heterochromatin in *Anopheles* (*60*). Heterochromatin variation plays unknown roles in evolution, but may factor into reproductive isolation and speciation is some *Anopheles* species (*67*).

Starting with mosquitoes in the *An. gambiae* complex, dramatic size differences in the *An. gambiae* Y chromosome have been observed (*68*). In a recent study, the first molecular evidence of polymorphism was found in a strain of *An. gambiae*. The *An. gambiae* Y chromosome usually lacks the rDNA locus (*65*). However, the ASEMBO strain of *An. gambiae* was found to have rDNA on the Y chromosome (*69*). This could mean recent recombination with the X, or it could be an ancestral remnant because a recent ancestor of *An. gambiae*, *An. merus* has rDNA located on the Y (*65*). Overall, compared to other *Anopheles* species, the *An. gambiae* Y chromosome is small and shares little or no morphological similarity with the X (*70*).

A thorough analysis of the metaphase karyotypes of numerous *Anopheles* mosquitoes has been performed by Visut Baimai. The ideograms produced in these papers illustrates a stunning

amount intra and interspecies X-Y polymorphism (*60*). Baimai was also the first to note that the Y chromosome increases as well as decreases in size. Considerable variation in the size of the Y was noted as sometimes the long heterochromatic arm of the *An. atroparvus* Y occasionally appeared longer the long arm of the X (*55*). There are clear examples of huge Y chromosomes (e.g. *An. karwari* B) that are much bigger than their X counterpart and the Y of ancestral species, indicating that the Y does in fact gain blocks of heterochromatin (*71*). On the opposite end of the spectrum, the Y can apparently shrink to less than a quarter of the size of the X (e.g. *An.* dirus E) (*60*).

Unequal crossing over has been suggested as the mechanism for the observed Y chromosome polymorphism. Baimai proposes a mechanism for the polymorphism of the *An. willimori* X chromosome using exchange of heterochromatin blocks from the Y resulting from chromosomal breakage (*60*). Baimai also proposed both unequal crossing over and chromosome breakage as potential mechanisms (*71*). There is one example of heterochromatin variation caused by unequal crossing over in the deer mouse, *Peromyscus beatae* (*72*). Just like in *Anopheles*, the X and Y heterochromatin varies between individuals. The authors observed unequal exchange of meiotic X-Y heterochromatin that caused the variation (*72*). Thus, this hypothesis is applicable to the *Anopheles* Y.

The cytogenetic literature reviewed here strongly suggests that the *Anopheles* Y chromosome is a degenerate X based on morphological similarity, recombination, and the presence of the rDNA locus on the Y in most species.

1.3.4 The mystery of homomorphic sex chromosomes

Though Charlesworth's theory about the degeneration of Y chromosomes explains the origin and evolution of many Y chromosomes there appear to be many exceptions (*13*). The biggest

exception is the fact that many species appear to have a dominant sex determining gene located on an almost undifferentiated homomorphic sex chromosome (*13*). It is unclear if a differentiated region exists or how large such a region is in most of these cases (*13*). While more studies need to be undertaken to understand the size of the undifferentiated regions of these chromosomes, Charlesworth's theory predicts that the small undifferentiated region should rapidly expand to encompass the majority of the chromosome (*13*).

A pertinent example of homomorphic sex chromosomes that appear to be old is chromosome 1 of *Aedes* mosquitoes (*73–75*). Male sexual development is initiated by a dominant male-determining factor in *Aedes* mosquitoes (*76–79*). This predicted male-determining factor has been mapped to chromosome 1 position 1q21, a region called the M locus for its role in initiating male development (*78, 80, 81*). Therefore, there are two types of chromosome 1 – those with the M locus containing the male-determining factor which are called the M chromosome, and those without the male-determining factor called the m chromosome (*30*). Shared markers around the M-locus between *Aedes* and *Culex* mosquitoes indicate that the M-locus shares a common origin between these mosquitoes suggesting it is not a newly-formed M-locus (*74*, *75*).

1.4 Difficulties in assembling Y, W, and M locus sequences

A huge challenge to the identification of Y chromosome, W chromosome, and M locus genes and sequences is the fact that they rarely appear in genome assemblies (*50*, *53*, *54*, *82*). For the rest of this section I will generalize Y/W/M locus sequences to Y for simplicity.

1.4.1 Biases against Y chromosomes in Sanger sequencing

A general goal for Sanger sequencing is to sequence autosomes to around 8x coverage but this can differ by project (*83*, *66*). If genomic libraries are made from mixed male and female

samples, for every four autosomes there is one Y chromosome (*26*, *84*). Therefore, if autosomes are sequenced to 8x coverage, the Y should only be sequenced to 2x coverage (*84*). Practically, even if the Y chromosome was not almost entirely repetitive this coverage would not be enough for a good assembly. Further compounding this problem is the fact that there is a systematic bias against heterchromatic sequences in Sanger sequencing further reducing the effective coverage of the Y (*85*). Combined with the fact that the Y chromosome is enormously repetitive it is unsurprising that Y chromosomes remain unassembled and unannotated in most traditional genome assemblies.

1.4.2 Y chromosome assembly with SHIMS

A special approach to sequence and assemble the human Y chromosome was developed by David Page and has been successfully applied to many other Y chromosomes including the human, chimpanzee, rhesus macaque, rat, bull, and opossum (*19*, *21*, *86*, *87*). The method is called single-haplotype iterative mapping and sequencing, abbreviated SHIMS (*21*). SHIMS sequences Y-linked BAC clones one at time and then finds overlapping regions of these BACs to generate an assembly (*19*). Notably, only the euchromatic region of the Y chromosomes tends to be assembled with SHIMS (*19*). This method was unable to assemble the heterochromatic portion of the human Y, but did sample the repeats and the various junctions of repeats (*19*). The fact that SHIMS works best with the euchromatic portion of the Y is relevant here because the *Anopheles* Y chromosome is fully heterochromatic (*52*). Another downside of SHIMS is that it is extremely expensive and extremely slow, especially in the era of next-generation sequencing.

Y chromosomes in genome assemblies generated from next-generation sequencing technologies seem to fare just as poorly as those from Sanger sequencing. At least in Sanger sequencing there are reads from the Y chromosome in the 400-1000 bp range. Finding a 1000 bp

Y-linked contig in an Illumina based assembly is rare (*82*). Due to the difficulties in identifying the small Y-linked fragments from genomes derived from next-generation sequencing technology, we developed a method called the chromosome quotient.

1.5 The chromosome quotient (CQ) method

1.5.1 Development of the CQ method

During my time as an undergraduate, I developed the chromosome quotient (CQ) method to identify Y chromosome sequences (Figure 1)(*82*). The CQ methods relies on the simple principle that Y chromosome sequences are present only in males (*82*). Therefore, Y chromosome sequences should have alignments from male sequence data, and no alignments from female sequence data (*82*). The CQ of a reference sequence is simply the ratio of female-to-male alignments to that reference sequence (*82*). For example, if a reference sequences has 10 alignments from female data and 11 alignments from male data it has a CQ of 0.909.

Using the CQ method, Y chromosome sequences can be differentiated from autosome and X sequences by their distinctive pattern of alignments from male and female Illumina sequencing data (*82*). Males and females have the same complement of autosomes, so autosomal sequences are expected to have roughly equal numbers of male and female alignments. Females have two X chromosomes while males have one, so X sequences are expected to have roughly twice as many female alignments as male alignments. The Y chromosome is present only in males, so Y chromosome sequences should have alignments from male data but no alignments from female data.

However, we found that Y sequences were often highly repetitive and therefore had closely-related copies present on the autosomes and X or were recently duplicated to the Y maintaining high nucleotide identity to autosome and X sequences (*82*). Thus, Y sequences can

plausibly have alignments from female Illumina data (*82*). A major innovation of the CQ method is to not ignore sequences with alignments from female Illumina data (*82*). Generally, we classify a sequence is Y-linked if it has a CQ less than 0.2 or 0.3 depending on the data available (*82, 88*). A major advantage of inferring Y-linkage based on the ratio of female-to-male alignments and not simple subtraction is that bacterial contamination and misassembled or poorly error-corrected sequences are not spuriously reported as Y linked (*82*). The CQ method requires a minimum number of male alignments which we generally set at 30 when using a genomic reference and 20 for an RNA-seq assembly (*82*, *88*). If the sequence was bacterial contamination or poorly assembled the male reads will fail to align and the sequence will not be reported as Y-linked (*82*). This proves to be invaluable in reducing the false positive rate when working with poor-quality reference assemblies like those produced from error-prone PacBio reads.

The CQ method utilizes cheap, abundant, short reads from the Illumina sequencing platform as the source for the male and female sequence data (*82*). The male and female Illumina reads are separately aligned to reference sequences using bowtie (*89*). Due to the fact that Y sequences are generally highly repetitive and therefore often have nearly-identical copies elsewhere in the genome, we require perfect alignment across the length of the read (*82, 88*). Another scenario that requires high-stringency alignments is genes that have been recently duplicated from the autosomes or X to the Y and retain high nucleotide identity with the progenitor gene. While requiring perfect stringency decreases the total number of reads aligned, this parameter dramatically increases the number of Y sequences identified. For example, using low-stringency alignments only 11 potential Y sequences can be identified from an assembly of the *An. stephensi* genome (*82*). In contrast, 317 Y sequences can be identified when only perfect

alignments are taken into account (*82*). We have shown that the CQ method can identify Y sequences maintaining up to 97 percent identity to autosomal or X sequences. We have implemented a script in Perl to calculate CQs for reference sequences which is available from GitHub: https://github.com/brantleyhall/Chromosome-Quotient.

We tested whether Y chromosome sequences could be differentiated from autosome and X sequences using two species with well-assembled Y chromosomes and male and female Illumina sequence data: humans and *Drosophila melanogaster*. Using the CQ method, we were successfully able to recover 89 percent of human Y sequences and 69 percent of *Drosophila melanogaster* Y sequences with a false positive rate of 2.44 percent and 1.85 percent for humans and *Drosophila melanogaster* respectively (*82*).

1.5.2 Generating reference sequences for the CQ method

Due to the fact that Y chromosome sequences are rarely found in reference genome assemblies, assemblies of some sort need to be undertaken to serve as the reference sequences for the CQ method. There are two good sources of these references assemblies – the male Illumina reads used in the CQ method, and RNA-seq (*88*). Because the CQ method requires male reads, these reads can also be used to assemble Y chromosome sequences. A rough Illumina assembly can be generated from the male Illumina reads using an assembler like AbYSS or Platanus (*90, 91*). These assemblies generally have N50 contig sizes less than 500 bp, but nevertheless often contain fragments of Y chromosome sequences. RNA-seq samples can be assembled with Trinity (*92*).

Using the CQ method, we have identified the first insect Y chromosome genes outside of *Drosophila* including six Y chromosome genes in *Anopheles* mosquitoes, three from *An. stephensi* and three from *An. gambiae* (*82*).

The CQ method only identifies Y chromosome sequences, not genes. Oftentimes, hundreds of candidate Y sequences are identified and in the case of *Aedes* mosquitoes, tens-of-thousands of candidate M locus sequences were identified. To find interesting Y genes in the sea of candidate Y sequences, we developed a pipeline we call caldera (https://github.com/brantleyhall/Chromosome-Quotient/blob/master/caldera.pl). Caldera takes into account many features to determine whether a Y sequence predicted by the CQ method is likely an interesting Y chromosome gene. First, we take into account the CQ which helps to estimate the probability of Y linkage. Next, we take into account of the number of male alignments. If there are too many male alignments, it is likely the sequence is highly repetitive on the Y, like a Y-enriched satellite or transposon. For example, the complete *Nix* transcript has 249 male alignments and is known to be a single copy gene. Therefore, if a sequence of equivalent length has 10,000 alignments it is likely present in many copies on the Y and should be excluded. Then caldera takes into account the relaxed CQ, which is the ratio of female-to-male alignments using blastn with relaxed parameters. Relaxed CQ is used to detect genes without close paralogs elsewhere in the genome. Several interesting Y genes, *GUY1*, *sYG2*, *gYG2*, and *Nix* all have zero female alignments even using the relaxed blastn parameters so an RCQ of zero is a good indication of linkage to the Y chromosome or M locus. Next, caldera takes into account the expression profile of the candidate genes using RNA-seq. Any sequences without RNA-seq alignments at the appropriate time points are discarded. An M factor needs to be expressed in the early embryo, but shouldn't be expressed in adult females or in embryos before the initiation of zygotic transcription 2 hours after embryo deposition, so these time points are of great interest for identifying whether a candidate gene could function as the M factor. Finally, caldera eliminates repeats. Repeats account for the vast majority of Y sequences predicted by the CQ method so we have to go to great

lengths to separate the repeats from interesting low-copy number genes. Caldera uses several databases to assess whether a sequence is a repeat including a database of known repeats for the species, a blastn against the genome to count how many distinct places the candidate gene aligns and with what percent identity, and finally a blastx against the NR database where the output is parsed and compared to a database of TE-related words like gag/pol/retrotransposon/LTR/transposase/viral. Using the caldera pipeline, it is possible to narrow down the hundreds of candidate Y sequences to a handful of interesting Y chromosome genes.

1.5.3 CQ vs. YGS

Another method to identify Y sequences was implemented using a kmer-based approach to identify Y sequences based on a lack of non-repetitive female-derived kmers called YGS (*93*). The major problem with this method is there are a number of sequences besides Y sequences that may not have any female-derived kmers. First, bacterial contamination is often present in genome assemblies. If the females sequenced for the YGS method don't have the exact same bacterial species present, the bacterial sequences from the reference genome will be spuriously classified as Y-linked. Bacterial contamination isn't such a problem for the YGS method when using extremely high quality genome assemblies like those of humans and *Drosophila melanogaster*. However, bacterial contamination poses a huge problem for the YGS method in the poor-quality genome assemblies generated from Illumina sequencing or RNA-seq specifically for finding for Y sequences. YGS also struggles with poorly-assembled or error-corrected reference genomes like those generated from PacBio sequencing. We performed PacBio sequencing on *An. gambiae* (Chapter 4) and error-corrected the reads. We then ran the CQ method and YGS on these reads. YGS identified a huge fraction of the reads as Y-linked simply because they had so many errors there were no unique female kmers. We also ran the CQ

method and YGS on known Y chromosome sequences from *An. gambiae* which can be found in table S8 of the "Rapid remodeling of the Y chromosome in a recent radiation of malaria mosquitoes". Due to the fact that many sequences are not Y-specific but instead Y-enriched, the YGS method failed to classify any of the major Y-repeats as Y-linked. Thus, we conclude the CQ method is superior to the YGS method for the identification of Y sequences from non-model organisms.

Despite the publication of the *A. aegypti* genome, genes from the male-specific M-locus had not been identified before our studies described in Chapters 2 and 3 *(30, 88)*. The absence of M-locus genes in the *A. aegypti* genome assembly impeded identification of the dominant male-determining factor hypothesized to be located in the M-locus *(38)*.

Currently, only 180 kb of reference sequence is available from the *An. gambiae* Y chromosome *(38)*. Furthermore, this reference sequence likely contains significant misassemblies (Personal communication). Therefore, the repeat structure and gene content of the *An. gambiae* Y chromosome remains almost completely unknown *(38)*. In Chapter 4, we describe Ydb, a database Y-linked, error-corrected PacBio reads that are representative of the non-recombining region of the *An. gambiae* Y chromosome.

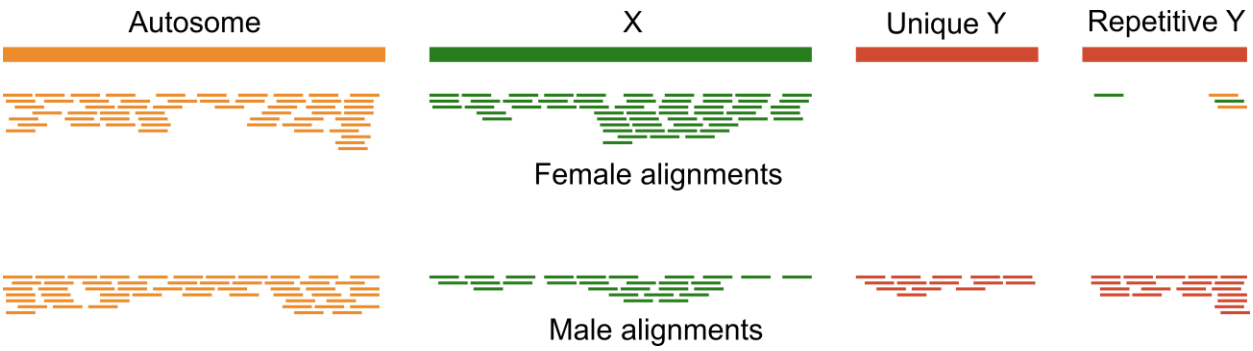

Figure 1.1: The CQ Method. The CQ Method. Y chromosome sequences have a distinctive pattern of alignments from male and female Illumina data. Autosomal sequences have roughly equal alignments from male and female data. X chromosome sequences have twice as many alignments from female as from male data. Y chromosome sequences have alignments from male data and alignments from female data only in repetitive regions.

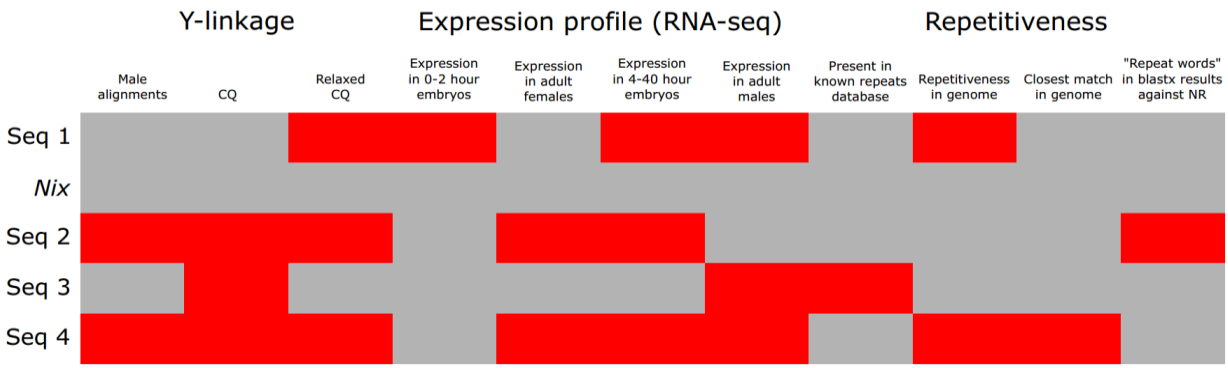

Figure 1.2: Caldera pipeline to identify the most interesting Y chromosome genes. Red boxes indicate negative marks against a Y sequence being an interesting Y gene.

1.6: References

1. P. A. Papathanos *et al.*, Sex separation strategies: past experience and new approaches. *Malar. J.* **8 Suppl 2**, S5 (2009).

2. A. N. Clements, The biology of mosquitoes. Development Volume 1, ed. nutrition and reproduction (1992).

3. A. N. CLEMENTS, The Sources of Energy for Flight in Mosquitoes. *J. Exp. Biol.* **32**, 547–554 (1955).

4. M. A. Tolle, Mosquito-borne diseases. *Curr. Probl. Pediatr. Adolesc. Health Care*. **39**, 97–140 (2009).

5. D. Bachtrog *et al.*, Sex Determination: Why So Many Ways of Doing It? *PLoS Biol*. **12**, e1001899 (2014).

6. Tree of Sex: A database of sexual systems. *Sci. Data*. **1** (2014) (available at http://dx.doi.org/10.1038/sdata.2014.15).

7. J. F. Crow, Advantages of sexual reproduction. *Dev. Genet.* **15**, 205–213 (1994).

8. C. M. Thomas, K. M. Nielsen, Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nat Rev Micro*. **3**, 711–721 (2005).

9. M. T. Ghiselin, The evolution of hermaphroditism among animals. *Q. Rev. Biol.*, 189–208 (1969).

10. E. L. Charnov, Simultaneous hermaphroditism and sexual selection. *Proc. Natl. Acad. Sci.* **76**, 2480–2484 (1979).

11. J. J. Bull, Sex determination in reptiles. *Q. Rev. Biol.*, 3–21 (1980).

12. J. J. Bull, *Evolution of sex determining mechanisms* (The Benjamin/Cummings Publishing Company, Inc., San Francisco, 1983).

13. D. Charlesworth, J. E. Mank, The birds and the bees and the flowers and the trees: lessons from genetic mapping of sex determination in plants and animals. *Genetics*. **186**, 9–31 (2010).

14. D. Bachtrog, A dynamic view of sex chromosome evolution. *Curr. Opin. Genet. Dev.* **16**, 578–585 (2006).

15. J. J. Bull, E. L. Charnov, Changes in the heterogametic mechanism of sex determination. *Heredity (Edinb).* **39**, 1–14 (1977).

16. C. B. Bridges, Non-Disjunction as Proof of the Chromosome Theory of Heredity (Concluded). *Genetics*. **1**, 107–163 (1916).

17. C. B. Bridges, Sex in relation to chromosomes and genes. *Am. Nat.* **59**, 127–137 (1925).

18. V. B. Kaiser, D. Bachtrog, Evolution of sex chromosomes in insects. *Annu Rev Genet*. **44**, 91–112 (2010).

19. H. Skaletsky *et al.*, The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*. **423**, 825–837 (2003).

20. D. Cortez *et al.*, Origins and functional evolution of Y chromosomes across mammals. *Nature*. **508**, 488–93 (2014).

21. J. F. Hughes *et al.*, Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature*. **437**, 100–103 (2005).

22. B. Vicoso, V. B. Kaiser, D. Bachtrog, Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. *Proc. Natl.*

Acad. Sci. **110**, 6453–6458 (2013).

23. T. Kiuchi *et al.*, A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature*. **509**, 633–636 (2014).

24. B. Charlesworth, D. Charlesworth, The degeneration of Y chromosomes. *Philos Trans R Soc L. B Biol Sci*. **355**, 1563–1572 (2000).

25. D. Bachtrog, E. Hom, K. M. Wong, X. Maside, P. de Jong, Genomic degradation of a young Y chromosome in Drosophila miranda. *Genome Biol*. **9**, R30 (2008).

26. A. Bernardo Carvalho, L. B. Koerich, A. G. Clark, Origin and evolution of Y chromosomes: Drosophila tales. *Trends Genet.* **25**, 270–277 (2009).

27. M. Matsuda *et al.*, DMY is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature*. **417**, 559–563 (2002).

28. M. Kondo, I. Nanda, U. Hornung, M. Schmid, M. Schartl, Evolutionary origin of the medaka Y chromosome. *Curr. Biol.* **14**, 1664–1669 (2004).

29. A. B. Carvalho, A. G. Clark, Y chromosome of D. pseudoobscura is not homologous to the ancestral Drosophila Y. *Science.* **307**, 108–110 (2005).

30. A. B. Hall *et al.*, A male-determining factor in the mosquito Aedes aegypti. *Science.* (2015), doi:10.1126/science.aaa2850.

31. P. Berta *et al.*, Genetic evidence equating SRY and the testis-determining factor. *Nature*. **348**, 448–450 (1990).

32. R. J. Jäger, M. Anvret, K. Hall, G. Scherer, A human XY female with a frame shift mutation in the candidate testis-determining gene SRY (1990).

33. N. A. Hanley *et al.*, SRY, SOX9, and DAX1 expression patterns during human sex determination and gonadal development. *Mech. Dev.* **91**, 403–407 (2000).

34. P. N. Goodfellow, R. Lovell-Badge, SRY and sex determination in mammals. *Annu. Rev. Genet.* **27**, 71–92 (1993).

35. R. Sekido, R. Lovell-Badge, Sex determination involves synergistic action of SRY and SF1 on a specific Sox9 enhancer. *Nature*. **453**, 930–934 (2008).

36. P. Koopman, Sry and Sox9: mammalian testis-determining genes. *Cell. Mol. Life Sci.* **55**, 839–856 (1999).

37. H. Salz, J. W. Erickson, Sex determination in Drosophila: The view from the top. *Fly (Austin).* **4**, 60–70 (2010).

38. S. F. Gilbert, Chromosomal sex determination in Drosophila (2000).

39. R. T. Boggs, P. Gregor, S. Idriss, J. M. Belote, M. McKeown, Regulation of sexual differentiation in D. melanogaster via alternative splicing of RNA from the transformer gene. *Cell*. **50**, 739–747 (1987).

40. K. Hoshijima, K. Inoue, I. Higuchi, H. Sakamoto, Y. Shimura, Control of doublesex alternative splicing by transformer and transformer-2 in Drosophila. *Science (80-.).* **252**, 833–836 (1991).

41. K. C. Burtis, B. S. Baker, Drosophila doublesex gene controls somatic sexual differentiation by producing alternatively spliced mRNAs encoding related sex-specific polypeptides. *Cell*. **56**, 997–1010 (1989).

42. E. Demir, B. J. Dickson, fruitless splicing specifies male courtship behavior in Drosophila. *Cell*. **121**, 785–94 (2005).

43. S. Yoshimoto *et al.*, A W-linked DM-domain gene, DM-W, participates in primary ovary development in Xenopus laevis. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 2469–2474 (2008).

44. C. A. Smith *et al.*, The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature*. **461**, 267–271 (2009).

45. K. P. Arunkumar, K. Mita, J. Nagaraju, The silkworm Z chromosome is enriched in testis-specific genes. *Genetics*. **182**, 493–501 (2009).

46. D. Charlesworth, B. Charlesworth, G. Marais, Steps in the evolution of heteromorphic sex chromosomes. *Hered.* **95**, 118–128 (2005).

47. D. Bachtrog, Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet*. **14**, 113–124 (2013).

48. Z. Liu *et al.*, A primitive Y chromosome in papaya marks incipient sex chromosome evolution. *Nature*. **427**, 348–352 (2004).

49. J. B. Pease, M. W. Hahn, Sex chromosomes evolved from independent ancestral linkage groups in winged insects. *Mol Biol Evol*. **29**, 1645–1653 (2012).

50. A. B. Carvalho, Origin and evolution of the Drosophila Y chromosome. *Curr. Opin. Genet. Dev.* **12**, 664–668 (2002).

51. R. H. Baker, R. K. Sakai, Triploids and male determination in the mosquito, Anopheles culicifacies. *J. Hered.* **70**, 345–346 (1979).

52. R. K. Sakai, R. H. Baker, K. Raana, M. Hassan, Crossing-over in the long arm of the X and Y chromosomes in Anopheles culicifacies. *Chromosoma*. **74**, 209–218 (1979).

53. J. Krzywinski, D. R. Nusskern, M. K. Kern, N. J. Besansky, Isolation and characterization of Y chromosome sequences from the African malaria mosquito Anopheles gambiae.

Genetics. **166**, 1291–1302 (2004).

54. J. Krzywinski, M. A. Chrystal, N. J. Besansky, Gene finding on the Y: fruitful strategy in Drosophila does not deliver in Anopheles. *Genetica*. **126**, 369–375 (2006).

55. L. Tiepolo, M. Fraccaro, U. Laudani, G. Diaz, Homologous bands on the long arms of the X and Y chromosomes of Anopheles atroparvus. *Chromosoma*. **49**, 371–4 (1975).

56. M. Fraccaro, U. Laudani, A. Marchi, L. Tiepolo, Karotype, DNA replication and origin of sex chromosomes in Anopheles atroparvus. *Chromosoma*. **55**, 27–36 (1976).

57. R. Mezzanotte, L. Ferrucci, A. Marchi, Y chromosome in the sibling speciesAnopheles atroparvus (van Thiel, 1927) andA. labranchiae (Falleroni, 1926) (Diptera: Culicidae): Differential behaviour of the short arm after acid-alkaline treatment and Coriphosphine-O staining. *Experientia*. **35**, 312–313 (1979).

58. A. Marchi, R. Mezzanotte, Inter- and intraspecific heterochromatin variation detected by restriction endonuclease digestion in two sibling species of the Anopheles maculipennis complex. *Heredity (Edinb).* **65 (Pt 1)**, 135–42 (1990).

59. J. A. Seawright, M. Q. Benedict, S. Narang, P. E. Kaiser, WHITE EYE AND CURLED, RECESSIVE MUTANTS ON THE X CHROMOSOME OF ANOPHELES ALBIMANUS. *Can. J. Genet. Cytol.* **24**, 661–665 (1982).

60. V. Baimai, A. Treesucon, U. Kijchalao, Heterochromatin variation in chromosome X in a natural population of Anopheles willmori (Diptera: Culicidae) of Thailand. *Genetica*. **97**, 235–9 (1996).

61. X. Jiang *et al.*, Genome analysis of a major urban malaria vector mosquito, Anopheles stephensi. *Genome Biol.* **15**, 459 (2014).

62. S. Mitchell, J. Seawright, Recombination between the X and Y Chromosomes in Anopheles quadrimaculatus Species A. *J. Hered.* **80**, 496–499 (1989).

63. R. H. Baker, R. K. Sakai, U. T. Saifuddin, A. Perveen, Induced chromosomal abberations in Anopheles culicifacies. *Mosq. News*. **38** (1978).

64. C. Li, R. C. Wilkerson, Intragenomic rDNA ITS2 variation in the neotropical Anopheles (Nyssorhynchus) albitarsis complex (Diptera: Culicidae). *J. Hered.* **98**, 51–9 (2007).

65. S. M. Paskewitz, D. M. Wesson, F. H. Collins, The internal transcribed spacers of ribosomal DNA in five members of the Anopheles gambiae species complex. *Insect Mol. Biol.* **2**, 247–57 (1993).

66. R. A. Holt *et al.*, The genome sequence of the malaria mosquito Anopheles gambiae. *Science (80-.).* **298**, 129–149 (2002).

67. M. Fraccaro, L. Tiepolo, U. Laudani, A. Marchi, S. D. Jayakar, Y chromosome controls mating behaviour on Anopheles mosquitoes (1977).

68. S. Bonaccorsi, G. Santini, M. Gatti, S. Pimpinelli, M. Colluzzi, Intraspecific polymorphism of sex chromosome heterochromatin in two species of the Anopheles gambiae complex. *Chromosoma*. **76**, 57–64 (1980).

69. E. E. Wilkins, P. I. Howell, M. Q. Benedict, X and Y chromosome inheritance and mixtures of rDNA intergenic spacer regions in Anopheles gambiae. *Insect Mol. Biol.* **16**, 735–41 (2007).

70. M. Gatti, G. Santini, S. Pimpinelli, M. Coluzz, Fluorescence banding techniques in the identification of sibling species of the anopheles gambiae complex. *Heredity (Edinb).* **38**, 105–8 (1977).

71. V. Baimai, Heterochromatin Accumulation and Karyotypic Evolution in Some Dipteran Insects. *Zool. Stud.* (1988), pp. 75–88.

72. P. D. Sudman, I. F. Greenbaum, Unequal crossing over and heterochromatin exchange in the X-Y bivalents of the deer mouse, Peromyscus beatae. *Chromosoma*. **99**, 183–9 (1990).

73. M. A. Toups, M. W. Hahn, Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. *Genetics*. **186**, 763–766 (2010).

74. C. A. Malcolm *et al.*, A sex-linked Ace gene, not linked to insensitive acetylcholinesterase-mediated insecticide resistance in Culex pipiens. *Insect Mol Biol*. **7**, 107–120 (1998).

75. A. Mori, D. W. Severson, B. M. Christensen, Comparative linkage maps for the mosquitoes (Culex pipiens and Aedes aegypti) based on common RFLP loci. *J. Hered.* **90**, 160–164 (1999).

76. G. A. H. McClelland, Sex-linkage in Aedes aegypti. *Trans roy Soc trop Med Hyg*. **56** (1962).

77. G. A. McClelland, Sex-linkage at two loci affecting eye pigment in the mosquito Aedes aegypti (diptera: culicidae). *Can J Genet Cytol*. **8**, 192–198 (1966).

78. M. E. Newton, D. I. Southern, R. J. Wood, X and Y chromosomes of Aedes aegypti (L.) distinguished by Giemsa C-banding. *Chromosoma*. **49**, 41–49 (1974).

79. M. E. Newton, R. J. Wood, D. I. Southern, Cytological mapping of the M and D loci in the mosquito, Aedes aegypti (L.). *Genetica*. **48**, 137–143 (1978).

80. V. A. Timoshevskiy, D. W. Severson, W. C. Black, I. V Sharakhov, M. V Sharakhova, An Integrated Linkage, Chromosome, and Genome Map for the Yellow Fever Mosquito

Aedes aegypti. *PLoS Negl. Trop. Dis.* **7**, e2052 (2013).

81. D. Shin, A. Mori, D. W. Severson, Genetic mapping a meiotic driver that causes sex ratio distortion in the mosquito Aedes aegypti. *J Hered*. **103**, 303–307 (2012).

82. A. B. Hall *et al.*, Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. *BMC Genomics*. **14**, 273 (2013).

83. V. Nene *et al.*, Genome sequence of Aedes aegypti, a major arbovirus vector. *Science.* **316**, 1718–1723 (2007).

84. A. B. Carvalho *et al.*, Y chromosome and other heterochromatic sequences of the Drosophila melanogaster genome: how far can we go? *Genetica*. **117**, 227–237 (2003).

85. M. D. Adams *et al.*, The genome sequence of Drosophila melanogaster. *Science.* **287**, 2185–2195 (2000).

86. J. F. Hughes *et al.*, Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature*. **483**, 82–86 (2012).

87. D. W. Bellott *et al.*, Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature*. **508**, 494–499 (2014).

88. A. B. Hall *et al.*, 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** , 179–191 (2014).

89. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. **10**, R25 (2009).

90. J. T. Simpson *et al.*, ABySS: A parallel assembler for short read sequence data. *Genome Res.* . **19** , 1117–1123 (2009).

91. R. Kajitani *et al.*, Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* (2014), doi:10.1101/gr.170720.113.

92. B. J. Haas *et al.*, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).

93. A. B. Carvalho, A. G. Clark, Efficient identification of Y chromosome sequences in the human and Drosophila genomes. *Genome Res.* **23**, 1894–1907 (2013).

Chapter 2: 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

Authors and affiliations:

A. Brantley Hall[1,2,4], Vladimir A. Timoshevskiy[3,4], Maria V. Sharakhova[3,4], Xiaofang Jiang[2,4], Sanjay Basu[3,4], Michelle A. E. Anderson[3,4], Wanqi Hu[1,4], Igor V. Sharakhov[2,3,4], Zach N. Adelman[2,3,4], and Zhijian Tu[1,2,4,*]

1. Department of Biochemistry, Virginia Tech

2. Program of Genetics, Bioinformatics, and Computational Biology, Virginia Tech

3. Department of Entomology, Virginia Tech

4. Fralin Life Science Institute, Virginia Tech

*Author for Correspondence: Dr. Zhijian Jake Tu, Department of Biochemistry, Virginia Tech, Blacksburg, VA 24061, USA, Tel: 540-231-8062, Fax: 540-231-9070, Email: jaketu@vt.edu

Data deposition: The sequence of *myo-sex* is available from GenBank (KF150020). The *Aedes aegypti* male and female Illumina sequencing data are available from the SRA (SRP023515).

Author contributions:

ABH – Identified *myo-sex* and M-linked BAC using the CQ method, performed PCR to verify male-specific amplification, sequenced the full-length myo-sex gene, drafted the manuscript and made the figures. VT and MS – Mapped *myo-sex* and the BAC using FISH. XJ – Performed

bioinformatic analysis of *myo-sex* to construct the phylogeny. SB and MA – Contributed and maintained transgenic mosquito lines with transgenes inserted around the M locus. WH – Performed quantitative PCR on myo-sex in recombinant individuals. IS – Initiated the project, coordinated FISH analysis, critically reviewed the manuscript and figures. ZA– Initiated the project and coordinated the genetic crosses, helped to revise the manuscript and figures. ZT– Initiated and coordinated the project, helped to write and revise the manuscript and figures.

Citation

Hall, Andrew Brantley, Vladimir A Timoshevskiy, Maria V Sharakhova, Xiaofang Jiang, Sanjay Basu, Michelle A E Anderson, Wanqi Hu, Igor V Sharakhov, Zach N Adelman, and Zhijian 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 Biology and Evolution* 6 (1): 179–91. doi:10.1093/gbe/evu002.

2.1: Abstract

The preservation of a homomorphic sex-determining chromosome in some organisms without transformation into a heteromorphic sex chromosome is a long-standing enigma in evolutionary biology. A dominant sex-determining locus (or M-locus) in an undifferentiated homomorphic chromosome confers the male phenotype in the yellow fever mosquito *Aedes aegypti*. Genetic evidence suggests that the M-locus is in a non-recombining region. However, the molecular nature of the M-locus has not been characterized. Using a recently developed approach based on Illumina sequencing of male and female genomic DNA, we identified a novel gene, *myo-sex*, that is present almost exclusively in the male genome, but can sporadically be found in the female genome due to recombination. For simplicity, we define sequences that are found primarily in the male genome as male-biased. Fluorescence *in situ* hybridization (FISH) on *A. aegypti* chromosomes demonstrated that the *myo-sex* probe localized to region 1q21, the established location of the M-locus. *Myo-sex* is a duplicated myosin heavy chain gene that is highly expressed in the pupa and adult male. *Myo-sex* shares 83 percent nucleotide identity and 97 percent amino-acid identity with its closest autosomal paralog, consistent with ancient duplication followed by strong purifying selection. Compared to males, *myo-sex* is expressed at very low levels in the females that acquired it, indicating that *myo-sex* may be sexually antagonistic. This study establishes a framework to discover male-biased sequences within a homomorphic sex-determining chromosome and offers new insights into the evolutionary forces that have impeded the expansion of the non-recombining M-locus in *A. aegypti*.

Keywords: sex-determination, sex chromosomes, chromosome quotient, *myo-sex*, non-recombining region, sexual antagonism

2.2: Introduction

In *Aedes* and *Culex* mosquitoes, male development is initiated by a dominant male-determining locus (M-locus) located on a homomorphic sex-determining chromosome (McClelland 1962; Newton, et al. 1974). On the other hand, *Anopheles* mosquitoes have fully differentiated heteromorphic sex chromosomes where the male-determining locus resides on the non-recombining Y chromosome (Clements 1992; Marın and Baker 1998). Evolutionary models suggest that homomorphic sex-determining chromosomes may eventually progress into heteromorphic sex chromosomes (Charlesworth, et al. 2005). After the acquisition of a sex-determining locus, linkage between sexually antagonistic genes and the sex-determining-locus is favored and may lead to an initial suppression of recombination followed by progressive expansion of the non-recombining region, transforming a homomorphic sex-determining chromosome into a heteromorphic sex chromosome (Bachtrog 2006; Bachtrog 2013; Bull 1983; Charlesworth and Charlesworth 2000; Charlesworth, et al. 2005; Charlesworth and Mank 2010). However, this fate is not inevitable because examples of old homomorphic sex-chromosomes have been observed (Charlesworth and Mank 2010).

A. aegypti has three pairs of chromosomes, with chromosome 1 the shortest, chromosome 2 the longest, and chromosome 3 of medium length (Rai 1963; Sharakhova, et al. 2011; Timoshevskiy, et al. 2013). The M-locus has been mapped to chromosome 1, band 1q21 (McClelland 1962). For simplicity, we follow the nomenclature proposed by Matara and Rai (1978) and refer to the copy of chromosome 1 with the M-locus the M-chromosome, and the copy without the M-locus the m-chromosome. Genetic evidence suggests that there is a non-recombining M-locus in *A. aegypti* (Severson et al., 2002; Toups and Hahn, 2010). There are also cytological differences between the M-locus and the m-locus, consistent with clear

differentiation between the loci (Matara and Rai 1977; Matara and Rai 1978). The M-locus can be found in non-canonical locations in other *Culicinae* mosquitoes (Ferdig, et al. 1998; Venkatesan, et al. 2009), indicating either translocation or turn-over of the sex-determining gene. However, the M-locus of *A. aegypti* and *C. pipiens* are linked to the same markers suggesting that the *A. aegypti* M-locus likely shares a common origin with *C. pipiens* (Malcolm, et al. 1998; Mori, et al. 1999). Thus the sex-determining chromosome of the yellow fever and dengue fever mosquito *A. aegypti* may have remained stubbornly homomorphic since the evolutionary divergence between the *Aedes* and *Culex* lineages. Previous studies suggest that the homomorphic sex-determining chromosome of *A. aegypti* is ancestral within the mosquito lineage and they may even have remained homomorphic since the evolutionary divergence between the *Anopheles* and *Aedes/Culex* lineages approximately 150 MYA (Toups and Hahn 2010).

Despite the publication of the *A. aegypti* genome, sequences from the M-locus remain uncharacterized (Nene, et al. 2007). In species with well-defined heteromorphic sex chromosomes such as *Drosophila* and *Anopheles*, Y chromosome sequences have proven resistant to traditional methods of sequence assembly because of the repetitive nature of the Y (Carvalho 2002; Krzywinski, et al. 2006; Krzywinski, et al. 2004). In fact, Y chromosome sequences are rarely annotated in published genomes of non-model organisms (Carvalho, et al. 2009). Recent advances in sequencing technologies have allowed for the identification of more Y chromosome sequences. We identified a single Y chromosome gene in *Anopheles* mosquitoes by comparing male Illumina data to female Illumina data (Criscione, et al. 2013). We also identified six more Y chromosome genes in *Anopheles* mosquitoes using a more efficient approach called the chromosome quotient (abbreviated CQ) (Hall, et al. 2013).

In this study, we applied the CQ method to *A. aegypti*, a species with a homomorphic sex-determining chromosome, and identified sequences that were present primarily in the male genome, but were sporadically present in the female genome due to recombination. In this study we call sequences present primarily in the male genome male-biased sequences, without implying patterns of gene expression. The male-biased sequences discovered by the CQ method include: a novel gene, *myo-sex*, and the full-length BAC clone NDL62N22. Based on homology, strong purifying selection, and differential expression between males and recombinant females, *myo-sex* appears to be a functionally important myosin heavy-chain gene with the possibility for sexually antagonistic effects. *Myo-sex* is a product of an ancient duplication event and it is tightly linked to the M-locus. We discuss alternative scenarios which may begin to explain why *myo-sex* has not been incorporated into the non-recombining M-locus and shed light on the maintenance of homomorphic M chromosome in *A. aegypti*.

2.3: Methods

CQ Analysis

The CQ method was originally devised as an approach to identify Y chromosome sequences (Hall et al., 2013) and it is adapted here to identify male-specific or male-biased sequences in *A. aegypti*. The CQ method is based on the principle that male-specific or male-biased sequences should be present in male sequence data and absent from female sequence data. This simple principle is complicated by the presence of nearly identical autosomal paralogs and repetitive sequences. Thus, instead of searching for sequences exclusive to the male sequence data, female alignments are permitted as long as there are at least five times more alignments from the male ($CQ \leq 0.2$). The selection of this CQ cutoff is described below and at the beginning of the Results section. Interference from repetitive sequences is reduced by using extremely strict

alignment parameters. For an alignment to be valid, it must align with 100 percent nucleotide over the entire extent of the Illumina (Illumina, San Diego, California, USA) read. Bowtie, the ultrafast short read aligner, is used for alignment (Langmead, et al. 2009). To reduce the rate of false positives, a threshold for the number of male alignments is used. For this study, the threshold was set at 30.

Male and female Illumina sequence data was generated to perform the CQ method in *A. aegypti*. Genomic DNA was isolated separately from 10 male, and 6 virgin female Liverpool strain *A. aegypti*. The Qiagen (Hilden, Germany) DNeasy Blood and Tissue kit was used to isolate the DNA following the manufacturer suggested protocols. The male and female genomic DNA samples were each subjected to two lanes of Illumina sequencing on a HiSeq 1000 producing 99 bp paired-end reads. The coverage of the resulting reads was approximately 63x for males and 65x for females. The male and female sequence data was deposited to the SRA (SRP023515).

A control for the CQ method was performed on known *A. aegypti* autosomal sequences for normalization and to quantify the rate of false positives. Known autosomal supercontigs (Nene et al., 2007) were retrieved and repetitive sequences annotated by RepeatMasker, and gaps represented by Ns were removed generating 7,713 sequences. CQs were calculated for the 7,713 sequences with the male and female Illumina sequence data. These autosomal sequences had a median CQ of 1.274, slightly higher than the expected CQ of 1, indicating there was more female sequence data than male sequence data. Subsequently, we normalized all CQs by 1.274.

We then calculated the CQs for the *A. aegypti* IB12 strain supercontigs (version AaegL1), contigs (version AaegL1), transcripts (version AaegL1.3), ESTs (Aedes-aegypti_EST-CLIPPED_2012-12.fa.gz), and BAC-ends (Aedes-aegypti-Liverpool_BAC-ENDS_2012-

12.fa.gz). These sequences were all downloaded from VectorBase (Vectorbase.org). We noticed that there was a relatively high rate of apparent false positive sequences, so we set CQ 0.2 as the cutoff for further analysis of male-specific or male-biased candidates.

Further bioinformatics analysis of candidate male-biased sequences

Two additional steps were implemented to identify male-specific or male-biased genes among the sequences mentioned above with CQs less than 0.2. First, these sequences were compared to RNA-seq data spanning developmental time points from embryo to adult, including separate adult male and female sequence data (SRA SRP009679) (Biedler, et al. 2012). The transcriptome sequence data was compared to all the sequences with blastn requiring 100 percent nucleotide identity with an e-value less than $1x10^{-5}$. Second, to further ensure male-specificity or male-bias, candidate sequences were compared to the male and female Illumina sequence data, using blastn (E-value cutoff set at $1x10^{-5}$). This step employs relaxed search parameters and thus allows us to filter sequences that only have slight variations between males and females.

Obtaining the sequence of the full-length *myo-sex* transcript

RT-PCR was used to connect the five myosin ESTs. Phire II DNA polymerase (Thermo Scientific, Pittsburgh, Pennsylvania, USA) was used for RT-PCR following the manufacturer suggested protocol. cDNA was made from Liverpool *A. aegypti* pupa. RNA was isolated using the mirVana RNA isolation kit (Life Technologies, Carlsbad, California, USA) following the manufacturer protocol for total RNA isolation. cDNA was then synthesized with SuperScript RT (Life Technologies, Carlsbad, California, USA). The PCR products were run on 1 percent agarose gels, and gel-purified with the GE gel purification kit (GE Healthcare, United Kingdom) following the manufacturer suggested protocol. The RT-PCR products were cloned into the pGEM-T Easy vector (Promega, Madison, Wisconsin, USA) or CloneJET vector (Thermo

Scientific, Pittsburgh, Pennsylvania, USA) and sequenced. 5' and 3' RACE was performed to obtain the terminal ends of the *myo-sex* transcript. The SMARTer RACE cDNA Amplification kit (Clonetech Laboratories, Inc., Mountain View, California, USA) was used to perform RACE following the manufacturer suggested protocol. RACE-ready cDNA was synthesized from pupa cDNA. The RACE PCR products were cloned into the pGEM-T EASY vector and sequenced. Vector sequences were removed from the sequence results using NCBI VecScreen. Assembly was performed with Cap3 (Huang and Madan 1999). The sequence was corrected with the consensus of pupa RNA-seq using CLC Genomics Workbench (www.clcbio.com). The assembled *myo-sex* transcript was submitted to GenBank (KF150020) and is available in the supplemental materials (supplemental file S1). The primers used for RT-PCR and RACE are available in the supplemental materials (supplemental table S1).

Generating transgenic lines with transgenes flanking the M-locus

The methods for the generation of the sensor transgenic strain are detailed in Adelman, et al. (2008). To generate transgenic line J2, *A. aegypti* Liverpool strain embryos (n=663) were injected with 300ng/µl pGL3-PUb-Mos1 and 500ng/µl PUb-DsRED construct as previously described (Carpenetti, et al. 2012). G_0 injection survivors with DsRED somatic transient expression (10.7 percent) were crossed to Liverpool strain of the opposite gender resulting in the establishment of five pools designated J1-5 (two female, three male). Eight positive individuals out of 400 total screened were identified from line J2. Mosquitoes for this study were reared using techniques and conditions as described in Aryan, et al. (2013). Prior to crossing with the 3xP3-sensor strain, the J2 transgene insertion was moved into the kh^w genetic background.

Fluorescence *in situ* hybridization (FISH)

FISH was performed on mitotic and polytene chromosomes using the methods described in Timoshevskiy, et al. (2012). Polytene chromosomes were prepared using salivary glands of the fourth instar larvae and mitotic chromosomes were prepared using imaginal discs from the fourth instar larvae. Slides were placed in 2× SSC for 30 minutes at 37°C, pretreated with 0.1 mg/ml of pepsin for 5 minutes at 37°C, denaturated in 70 percent formamide in 2× SSC at 72°C for 2 min, and dehydrated in a series of cold (−20°C) ethanol (70 percent, 80 percent, 100 percent) for 3–5 minutes each. Hybridization mix contained: 50 percent formamide, 10 percent dextran sulfate, 100 ng of each probe per slide, and 3 µg of unlabeled repetitive DNA fractions per probe. DNA/probe mix was precipitated by adding 1/10 volume of sodium acetate and 2 volumes of 100 percent ethanol. The DNA pellet was dissolved in "master mix" (10 µl per slide) that contained 50 percent formamide, 10 percent dextransulfate, and 1.2× SSC. After that, DNA was denatured at 96°C for 7 minutes. Denatured DNA was placed on ice for 1 minute and incubated at 37°C for 30 minutes for pre-hybridization with unlabeled repetitive DNA fractions (C_0t3 DNA). Repetitive DNA fractions were isolated from *A. aegypti* genomic DNA. DNA was denatured and allowed to re-associate at 60°C in 1.2× SSC for 15-150 min depending on concentration. Single stranded DNA was digested using S_1 nuclease (Invitrogen Corporation, Carlsbad, California, USA). Ten µl of hybridization mix was placed on a slide, which had been preheated to 37°C, under a 22×22 mm coverslip, and glued by rubber cement. Slides were hybridized at 37°C in a dark humid chamber overnight. After hybridization, slides were dipped for washing in a Coplin jar with 0.4× SSC, 0.3 percent Nanodept-40 at 72°C for 2 minutes, and then in 2× SSC, 0.1 percent Nanodept-40 at room temperature for 5 minutes. Thereafter, slides were counterstained using Prolong with DAPI (Invitrogen Corporation, USA) or incubated with 1 µM YOYO-1 solution in 1× PBS for 10 minutes in the dark, rinsed in 1× PBS, and then

enclosed in antifade Prolong Gold (Invitrogen Corporation, Carlsbad, California, USA) under a cover slip. Slides were analyzed using a Zeiss LSM 510 Laser Scanning Microscope (Carl Zeiss Microimaging, Inc., Thornwood, New York, USA) at 1000× magnification.

***Myo-sex* expression analysis**

The RNA used for the expression profile of *myo-sex* was isolated using the mirVana RNA isolation kit (Life Technologies, Carlsbad California USA) following the manufacturer protocols for total RNA isolation. cDNA was then synthesized with the SuperScript III RT kit (Life Technologies, Carlsbad California USA) following the manufacturer suggested protocols. Using cDNA spanning developmental time points, RT-PCR was performed on *myo-sex* using Phire II DNA polymerase (Thermo Scientific, Pittsburgh, Pennsylvania, USA). The PCR products were verified to be *myo-sex* by sequencing. For the expression profile, 27 cycles were used with a melting temperature of 63^{o}C. To further assess *myo-sex* expression in Liverpool adult females, cDNA samples from virgin females and blood-fed females were amplified for 32 cycles (figure s1). To assess the expression of *myo-sex* in recombinant females that acquired *myo-sex*, both RT-PCR and quantitative RT-PCR were performed. qRT-PCR were performed using the SYBR Green based GoTaq qPCR kit from Promega (Madison, Wisconsin, USA) on a ABI 7300 real time PCR system (Life Technologies, Carlsbad California USA). Three biological replicates were included and a ribosomal protein gene RPS7 was used to normalize expression. Relative expression levels were quantified using ΔCt relative quantification method with RPS7 as the endogenous control (Sengul and Tu 2010). All primers are available in supplemental table S1.

***Myo-sex* evolutionary analysis**

The dN/dS ratio of *myo-sex* was calculated using JCoDA using a sliding window size of 200, and a jump size of 20. The Yang and Neilson dN/dS substitution model was used. The *myo-sex*

phylogeny was generated with MrBayes (Huelsenbeck and Ronquist 2001) using MUSCLE (Edgar 2004) for the alignments. The alignments and parameters used for phylogenetic inference are provided in supplemental file S2. For the synteny figure, orthologs were assigned by OrthoDB (Waterhouse, et al. 2011) and relative positions assigned by VectorBase.

An extremely male-biased BAC clone

CQ analysis of BAC-ends identified several BACs as male-specific or highly male-biased candidates and one such BAC was sequenced after PCR verification with male and female genomic DNA as templates. DNA was isolated from BAC NDL62N22 and was sequenced with PacBio sequencing. A single SMRT cell of PacBio sequencing (Pacific Biosciences, California, USA) was performed on the *A. aegypti* BAC clone NDL62N22 along with five other BACs from different species for a different project. The resulting PacBio sequences were assembled by Russell Durrett at the Weill Cornell Medical College using the Hierarchical Genome Assembly Process (HGAP) from PacBio. The resulting contig contained the BAC cloning vector, which was subsequently removed. The resulting BAC was 94,552 bp long. The sequence of BAC NDL62N22 is provided in the supplemental materials (supplemental file S1).

Male specific amplification of *myo-sex* and BAC clone NDL62N22

DNA was isolated from adult mosquitoes using the DNAzol (Life Technologies, Carlsbad California USA) reagent following the manufacturer protocols. Genomic DNA samples were isolated from pools of five individuals. PCR was performed in 25 males and 25 females from both the Liverpool and kh^{w} strains using 30 cycles with a melting temperature of 63^{o}C. The PCR products were verified to be the expected product by sequencing. Phire II DNA polymerase (Thermo Scientific, Pittsburgh, Pennsylvania, USA) was used for the PCR following the

manufacturer specified protocol. All primer sequences used in this study are available in supplemental table S2.

2.4: Results

Identification of candidate male-biased genomic sequences in *A. aegypti*

Separate male and female Illumina sequence data was generated from adults of the Liverpool strain of *A. aegypti,* resulting in coverage of 65x and 63x, respectively (SRA: SRP023515). To select a cutoff CQ for screening candidate male-specific sequences, we calculated CQs for supercontigs with known autosomal positions (Nene, et al. 2007). Each known autosomal supercontig was split at gaps denoted by Ns and bases masked by repeat masker. Chromosome quotients were calculated for 7,713 of the resulting autosomal sequences (fig. 1). Only seven of the split supercontig sequences had chromosome quotients less than 0.2, and only one had a CQ of zero. Thus, 0.2 was chosen as the cutoff for further analysis of male-specific or male-biased candidates.

We searched for sequences with alignments only from the male sequence data using various reference datasets from *A. aegypti* including: the *A. aegypti* IB12 strain supercontigs, contigs, annotated transcripts, Rockefeller strain ESTs, and Liverpool strain BAC-ends (fig. 1) (Jiménez, et al. 2004; Krebs, et al. 2002). As expected, the overwhelming majority of sequences had CQs distributed around one, the CQ expected for sequences present in equal numbers in males and females. Each set of reference sequences had sequences with alignments only from the male sequence data or many more male alignments than female alignments (table 1). In the current study, our focus was the identification of male-specific or male-biased genes, not slight sequence variations that may exist between the male and female samples nor highly-repetitive sequences. Thus, putative genes were identified by comparing the sequences with CQs less than

0.2 to the NCBI non-redundant protein database and transcriptome sequence data spanning developmental time points from embryo to adult, including separate adult males and females (SRA SRP009679) (Biedler, et al. 2012). No genes were identified in the supercontigs or contigs with CQs less than 0.2. We identified 28 transcripts with CQs less than 0.2. However, further analysis showed that none of the transcripts had characteristics of male-biased genes; 22 had either alignments from female transcriptome data, or do not appear to be transcribed due to the lack of alignments from any of our transcriptome data. The six remaining transcripts derive from contigs that do not appear to be male-biased based on CQ analysis. None of the 28 transcripts were male-biased when relaxed blastn parameters were used for comparison to the male and female Illumina data. Due to our lack of success finding male-biased genes in the published genome sequences, we focused on the unassembled ESTs and BACs: the ESTs because they were automatically candidates for genes, and the BACs because sequencing full-length male-biased BAC clones could result in the identification of genes.

Candidate male-biased EST and BAC sequences were further narrowed down by using relaxed alignment parameters to map the male and female Illumina data to the potential male-biased sequences with blastn. Even with the relaxed alignment parameters, 15 of the ESTs and 17 BAC-ends appeared unique or highly biased to the male sequence data. Analysis of the BAC sequences is described in a later section. Eight of the ESTs were removed because they had alignments from the adult female transcriptome data. The seven remaining ESTs were compared to the genome with blastn. One of the seven aligned with 97 percent nucleotide identity to a long contig that is not male biased. Blastx was used to compare the remaining six ESTs to the NCBI non-redundant protein database. One EST was eliminated as bacterial contamination. The

remaining five ESTs aligned to myosin heavy chain genes, with e-values less than $1x10^{-40}$ and had CQs of zero (table 2).

Discovery of *myo*-sex, an extremely male-biased myosin heavy-chain gene

When compared to the *A. aegypti* transcripts with blastn and blastx, the five remaining ESTs all aligned to the *A. aegypti* gene *AAEL005656* with approximately 83 percent nucleotide identity and 97 percent amino acid identity. RT-PCR and subsequent sequencing confirmed that all five ESTs derived from the same novel male-biased gene. Primers were also designed for 5' and 3' RACE to identify the terminal portions of the transcript. The full-length 5,990 bp myosin-gene transcript was assembled using the sequencing results of the RT-PCR and RACE products. We call this gene "*myo-sex*" because it is homologous to a myosin heavy chain gene and is male-biased. *Myo-sex* is not represented in the current *A. aegypti* genome assembly. However, a blastn search of the *A. aegypti* trace database used for genome assembly found 20 alignments with greater than 98 percent identity, indicating that its absence from the current assembly is likely a reflection of poor assembly or low coverage near the M-locus. Fluorescence *in situ* hybridization (FISH) experiments described below are consistent with our interpretation.

To verify that *myo-sex* is a male-specific gene, PCR was performed on five pools of male (n=5 per pool) and five pools of female (n=5 per pool) genomic DNA from both the Liverpool and kh^w strains of *A. aegypti*. Primers for *myo-sex* amplified a PCR product in male-genomic DNA from both the Liverpool and kh^w strains of *A. aegypti* but not from female genomic DNA from either strain (fig. 2). Amplicons were cloned and sequenced, and verified to be *myo-sex*. A ribosomal protein gene (*RPS7)* was amplified in both males and females to validate the integrity of the genomic DNA. The CQ of the assembled *myo-sex* transcript is zero, and even with relaxed

blastn parameters, few female reads align (table 2) further asserting that *myo-sex* is male-specific in our sequencing data.

Myo-sex hybridizes to the M-locus and can undergo recombination with the m-chromosome

To further verify that *myo-sex* was truly male-specific, we sought to determine its physical location within the *A. aegypti* genome. We took advantage of several transgenic strains in our possession. Adelman et al. (2008) described a transposon insertion strain (referred to here as "sensor") marked with a green fluorescent protein that resides 0.40cM from the sex-locus on the m-chromosome. The 0.40cM recombination distance is an average of three replicates that range from 0.12 to 0.64 and the variability may be related to the location of the “sensor” transgene in the rDNA repeats. During the course of other experiments, we obtained a second transgenic strain (J2, marked with DsRED) that appeared to be sex-linked as well. The J2 insertion recombined from the m-chromosome to the M-chromosome at a frequency of approximately 1.24 percent (fig. 3A). Subsequent crossing with sensor mosquitoes allowed us to establish both M-chromosome and m-chromosome strains containing both transgenes (fig. 3A). Calculated recombination distances between sex/sensor, sex/J2, and sensor/J2 established that these two transgenes flank the M-locus (fig. 3B). There are fewer J2-M and more M-sensor recombinants in the F3 than expected according to their respective average recombination frequencies, which may in part result from the previously observed variability in recombination distance (Adelman *et al*., 2008). FISH was performed on mitotic chromosomes of both $M^{J2sensor}$/m and m/$m^{J2sensor}$ individuals (fig. 3C-D). In both cases, *myo-sex* and the J2/sensor transgene colocalized to only one copy of the q-arm of chromosome 1, band 1q21, the established location of the sex-determining locus. The fact that *myo-sex* hybridized to a single $m^{J2sensor}$ chromosome suggests

that this gene recombined alongside the J2 transgene during the creation of the double-marked strain. In contrast, if *myo-sex* was truly autosomal, the *myo-sex* probe should have hybridized to both copies of chromosome 1. Hybridization to polytene chromosomes confirmed the colocalization of the sex-linked transgenes and *myo-sex* (fig. 3E). We conclude that *myo-sex* is a novel *A. aegypti* gene that is proximal to the M-locus and therefore present primarily in the male genome, but can sporadically appear in the female genome due to recombination.

***Myo-sex* expression profile, evolutionary origin, and selective pressures**

The expression profile of *myo-sex* was generated with cDNA spanning developmental time points (fig. 4A). RT-PCR was performed using total RNA isolated from: 1-12 hour embryos, 12-24 hour embryos, 24-36 hour embryos, 36-48 hour embryos, 48-60 hour embryos, 0-2 day larva, 2-4 day larva, 0-2 day pupa, adult females, adult males, male heads, male thorax, and male abdomens. The highest level of expression from RT-PCR appears in the pupa, with dark bands also appearing in the adult male, and male thorax. No bands were observed in the female samples where *myo-sex* was absent (fig. 4A and supplemental figure S1). The general expression profile shown in fig. 4A was consistent with RNA-seq analysis (supplemental table S2). As expected, genomic DNA PCR indicated that *myo-sex* is present in all five of the recombinant $m/m^{J2sensor}$ females tested (data not shown). In the recombinant females, however, RT-PCR showed only a very faint *myo-sex* band (fig. 4B). Similarly, qRT-PCR analysis indicated that *myo-sex* transcript level in the recombinant females that acquired *myo-sex* was equivalent to the background level observed in females without the *myo-sex* gene, both of which were hundreds of fold lower than that of the males (fig. 4C).

Phylogenetic analysis suggests that *myo-sex* is a close paralog of *AAEL005656* (fig. 5). We also examined the gene synteny of its paralogs: *AAEL005656* and the slightly more distant

AAEL005733. Based on the assignment of scaffolds to chromosomes in the *A. aegypti* genome, *AAEL005733* is located on chromosome 2, while the position of *AAEL005656* is unknown (Nene et al. 2007). *AAEL005656* appears to be an insertion that occurred after the divergence of *Aedes* and *Culex* mosquitoes. The gene synteny around *AAEL005733* is conserved in *A. aegypti, Anopheles gambiae* and *Culex quinquefasciatus* (fig. 6). These results suggest that *AAEL005733* represents the ancestral gene, and that a duplication of *AAEL005733* produced *AAEL005656*/*myo-sex.* Although it is not clear whether *AAEL005656* or *myo-sex* came first, it is likely that the duplication happened after the divergence between *Aedes* and *Culex* (fig. 5, fig 6).

Myo-sex and *AAEL005656* align well along their entire open reading frames and have a nucleotide identity of 83 percent. The predicted amino acid sequences of *AAEL005656* and *myo-sex* have an amino-acid identity of 97 percent. Using the coding sequences of *AAEL005656* and *myo-sex*, we calculated the ratio of nonsynonymous to synonymous mutations (dN/dS). The dN/dS ratio of *myo-sex* to *AAEL005656* was 0.0107, indicating strong purifying selection.

The sequence of BAC-clone NDL62N22 is male-biased in the Liverpool strain but not the *kh*^{w} strain of *A. aegypti*.

Nine of the 17 BACs that had male-specific or highly male-biased BAC-end sequences were from a library that is no longer available. Four of the remaining eight had BAC-end sequences that were nearly identical to multiple other sequences in the genome, which makes specific PCR for these BAC-ends very difficult. We tested all 4 BACs for which we could design specific PCR primers. One primer set designed to amplify the BAC clone NDL62N22 (CC867386.1) amplified a male-specific PCR product from Liverpool genomic DNA (fig. 2), while the other three BAC-ends did not produce male-specific PCR products. Genomic DNA was isolated from BAC NDL62N22 and PacBio sequencing was performed. The PacBio sequencing reads were

assembled using the PacBio HGAP assembler resulting in a single 94,552 bp contig. The T7 and SP6 BAC-ends align to the beginning and end of the contig respectively. No genes were found in the 94,500 bp contig. Although primers for the BAC NDL62N22 amplified a male-specific PCR product in the Liverpool strain of *A. aegypti*, the same primers amplified a PCR product in both males and females in the kh^w strain of *A. aegypti* (fig. 2). To further verify that BAC NDL62N22 is male-biased in the Liverpool strain, the full-length BAC sequence was cut into 1,000 bp fragments and CQs were calculated for each of these fragments. The CQs for the 1,000 bp pieces clearly show male-biased segments throughout the length of the BAC with Liverpool sequence data (supplemental figure S2). Using FISH on mitotic chromosomes of transgenic kh^w mosquitoes, the BAC hybridized to 1q21, the established location of the sex-determining locus (fig. 7). The BAC NDL62N22 hybridized to both the M-locus and the m-locus, confirming that it is not male-biased in kh^w strain of *A. aegypti*, consistent with the PCR results (fig. 2B).

2.5: Discussion

Discovery of the *myo-sex* gene and implications to finding the M-factor in a homomorphic sex-determining chromosomes

We identified extremely male-biased sequences around the *A. aegypti* M-locus using the CQ method, which was originally designed to identify Y chromosome sequences. Thus, we have shown that such a differential genomics approach can be used to identify male-biased sequences in species with a homomorphic sex-determining chromosome.

In this study, we identified two extremely male-biased sequences: *myo-sex*, a novel myosin heavy chain gene, and a full-length BAC clone NDL62N22. *Myo-sex* is the first *A. aegypti* gene that is passed primarily from fathers to sons and not inherited equally between males and females in the manner of a typical autosomal gene. However, *myo-sex* is not male-

limited because recombination can still occur. Earlier work suggests that there is a non-recombining M-locus in *A. aegypti* (Severson et al., 2002; Toups and Hahn, 2010). Thus, neither *myo-sex* nor BAC NDL62N22 are located within the *A. aegypti* M-locus. Genes in the non-recombining M-locus such as the M-factor are more likely to have male-specific sequences than sequences in the recombining regions of the genome. In theory, genes in the non-recombining M-locus should be easier to detect with the CQ method than *myo-sex* because they contain male-specific rather than male-biased sequences. However, we did not identify the dominant male-determining gene from the M-locus of *A. aegypti*.

The genome of *A. aegypti* was sequenced with Sanger technology, which has a well-known bias against heterochromatic sequences. Sequencing the M-locus is further complicated by the fact that the genomic DNA used for genome sequencing was derived from both male and female genomic DNA, relegating the M-locus to one-quarter the coverage of autosomal sequences (Carvalho et al. 2003). The estimated coverage of the *A. aegypti* genome sequencing used in genome assembly was only 7.63x meaning that the M-locus has less than 2x coverage. Combined with the bias against heterochromatic sequences the actual coverage of the M-locus may be considerably lower. Low coverage of the M-locus is probably a contributing factor as to why we did not identify male-biased genes from the supercontigs, contigs, or transcripts of the *A. aegyp*ti genome. Thus, the current assembly of the *A. aegypti* genome is uninformative when looking for candidates for the M-factor.

CQ analysis of unassembled sequences, in this case ESTs and BAC-ends, proved more successful than CQ analysis of the assembled genome for the identification of a male-biased sequences. However, future analysis undertaken on new datasets may be more effective than ESTs or BAC-ends. Given that the M-factor is likely expressed in early embryos, CQ analysis of

assembled transcripts from deep-coverage early embryonic RNA-seq data may lead to candidates for this sought-after male-determining factor. CQ analysis of genomic assemblies obtained from deep coverage, clone-independent methods such as Illumina sequencing may also lead to identification of gene fragments from the M-locus (Hall, et al. 2013).

We cannot rule out that the *A. aegypti* M-locus was recently derived because there is evidence that the M-locus can be found in non-canonical locations in other *Culicinae* mosquitoes, a subfamily that includes *Aedes* and *Culex* (Ferdig, et al. 1998; Venkatesan, et al. 2009), indicating either translocation or turn-over of the sex-determining gene. If so, the M-factor may be an allelic gene variation that is unique to the males or it is duplicated from a nearly identical autosomal paralog, either of which will be difficult to detect with the CQ method. However, the two scenarios described above are less likely in *A. aegypti* because previous studies suggest that the M-locus of *A. aegypti* and *C. pipiens* are linked to the same markers (Malcolm, et al. 1998; Mori, et al. 1999). There are also cytological differences between the M-locus and the m-locus, consistent with clear differentiation between the loci (Matara and Rai, 1977, 1978). Given sufficient genomic coverage and transcriptome sequence data, the M-factor is likely to be discovered with the CQ method.

Is *myo-sex* an example of a sexually antagonistic gene near the non-recombining M-locus?

Our findings suggest that *myo-sex* is functionally important because it is under strong purifying selection. While the nucleotide identity between *myo-sex* and its closest paralog, *AAEL005656*, is only 83 percent, the amino acid identity is 97 percent. Another indication of the potential functional importance of *myo-sex* is its apparently high level of temporally regulated expression. *Myo-sex* is highly expressed in adult males but not in females as *myo-sex* is rarely found in the female genome. However, in the rare females that acquired *myo-sex* through

recombination, *myo-sex* is expressed at a much lower level than it is expressed in males (fig. 4). Such reduction of transcript level could result from a number of mechanisms including the loss of a distant *myo-sex* enhancer during recombination, repression of *myo-sex* expression near the m-locus, or simply that *myo-sex* had evolved a male-specific pattern of expression. Regardless the mechanism, such an intriguing expression pattern may suggest that *myo-sex* is sexually antagnostic, being advantageous to males and/or deleterious to females. The impact of loss of *myo-sex* function in males and gain of *myo-sex* function in females may or may not be easy to discern in the laboratory. It is also important to point out that the recombinant individuals shown in fig. 3 may not be good subjects to investigate the effect of gain or loss of *myo-sex* function because any observed phenotypic differences could not be solely attributed to *myo-sex*. As a significant portion of the chromosomal arm participated in the recombination event, other yet-to-be-discovered genes in the M region may also have been gained or lost. Thus, true gain-of-function experiments such as ectopic expression of the *myo-sex* transgene and true loss-of-function experiments such as site-specific knockout of *myo-sex* are needed to determine the importance of this gene to male-specific morphology or behavior. *Myo-sex* is tightly linked to the M-locus and rarely found in females (figs 2 and 3). The potential role of *myo-sex* in the expansion of the non-recombining sex-determining region is discussed below in the context of sexually antagonism (Charlesworth, et al. 2005; Charlesworth and Mank 2010).

Recombination dynamics near the M-locus and the preservation of homomorphic sex-determining chromosomes.

As discussed above, the *A. aegypti* M-locus shares a common ancestor with *C. pipiens* suggesting that the M-locus is ancient (Malcolm, et al. 1998; Mori, et al. 1999). The evolutionary forces that limit the expansion of the non-recombining region around the M-locus and thus

maintain chromosome homomorphy are unknown (Toups and Hahn, 2010). We have shown that *myo-sex* still undergoes recombination, and recombination was detected by screening several thousand individuals. Our transgene-aided analysis of recombination is much more sensitive than traditional linkage mapping analysis in which often less than a few hundred individuals are analyzed. Sexually antagonistic or not, *myo-sex* is extremely male-biased and drastically different allele frequencies between the sexes favors reduced recombination with the M-locus (Bull 1983; Charlesworth and Mank 2010; Patten and Haig 2009; Rice 1987, 1984). The duplication that produced *myo-sex* likely happened long ago because *myo-sex* has diverged 17 percent at the nucleotide level from its closest autosomal paralog despite strong purifying selection. It is thus fascinating that *myo-sex* has not been incorporated into the non-recombining M-locus in the intervening time. This could be caused by the lack of sexually antagonistic genes around the M-locus that would benefit from incorporation into a non-recombining region. Sexual antagonism could be resolved with sex-specific expression, thus alleviating the need for a large non-recombining region (Vicoso et al. 2013). The lack of significant expression of *myo-sex* in the females that acquired *myo-sex* is interesting in this regard. Even if *myo-sex* is sexually antagonistic, the selective pressure to remove *myo-sex* from the female genome by incorporating it into the non-recombining M-locus is low because *myo-sex* expression is hardly detectable when it occasionally found itself in the females. However, we cannot rule out the possibility that leaky expression of *myo-sex* when present in females may confer sufficient selective pressure to keep *myo-sex* male-biased or closely linked to the M-locus. It is also possible that once established, the close linkage between *myo-sex* and the M-locus and the male-bias of *myo-sex* can be maintained without the presence of constant selective pressure exerted by sexual

antagonism. For example, the rare occurrence of *myo-sex* in the m-chromosome may be lost due to genetic drift.

Another scenario that could contribute to the persistence of recombination between *myo-sex* and the M-locus is that the M-locus may be flanked by recombination hotspots, which may have a high intrinsic rate of recombination. Recombination hotspots are well known in yeast and humans (Gerton, et al. 2000; Myers, et al. 2005) and they are also recently characterized in detail in *Drosophila* (Chan, et al. 2012; Comeron, et al. 2012). Minisatellites are associated with such recombination hotspots in insects and spiders (Mita, et al. 1994; Sezutsu and Yukuhiro 2000). A genomic analysis of chromosomally mapped supercontigs demonstrated that the band 1q21, in which the *A. aegypti* M-locus resides, has an elevated coverage of minisatellites as compared with the neighboring regions (M. Sharakhova, unpublished data). Although recombination hotspots can change locations, the M-locus of both *A. aegypti* and *C. pipiens* are proximal to the tandem repeated ribosomal genes (Timoshevskiy, et al. 2013), which could promote recombination. Thus, it is possible that the non-recombining M-locus is near recombination hotspots, which sets a higher threshold against the expansion of the non-recombination region.

We also identified a BAC clone that is extremely male-biased in the Liverpool strain of *A. aegypti*, but not male-biased in the kh^{w} strain of *A. aegypti*. Even in kh^{w} where the BAC is not male-biased, the BAC is located directly adjacent to the M-locus. Inter-strain variation in male-biased sequences is noteworthy as it suggests ongoing plasticity near the M-locus. Comparative analysis between strains of *A. aegypti* will likely allow us to narrow in on the M-locus, which may be conserved between different strains.

Acknowledgements

We would like to thank David Severson for providing the BAC clones and Russell Durrett for running HGAP on the PacBio sequence data. We thank Frank Criscione for making the cDNA used for the expression profile. We thank Randy Saunders for mosquito care. This work was supported by National Institutes of Health Grants (AI 77680, AI105575, AI094289, and AI88035); and the Fralin Life Science Institute.

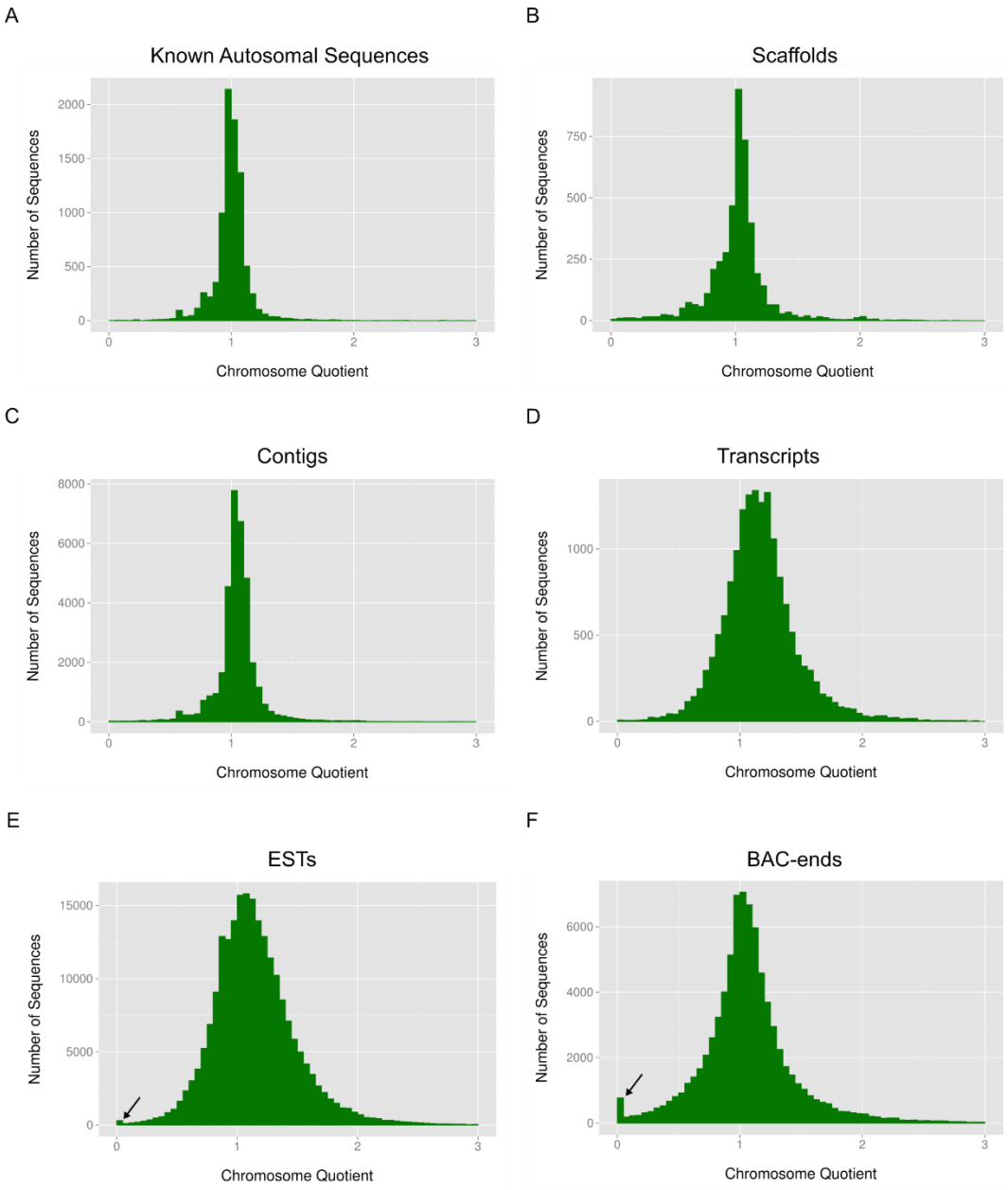

Figure 1. The distribution of chromosome quotients from *A. aegypti* including: (A) known autosomal sequences, (B) all scaffolds, (C) all contigs, (D) all transcripts, (E) ESTs, and (F) BAC-ends. Arrows in E and F indicate peaks in the distribution of CQs near zero, the CQ expected for male-biased sequences.

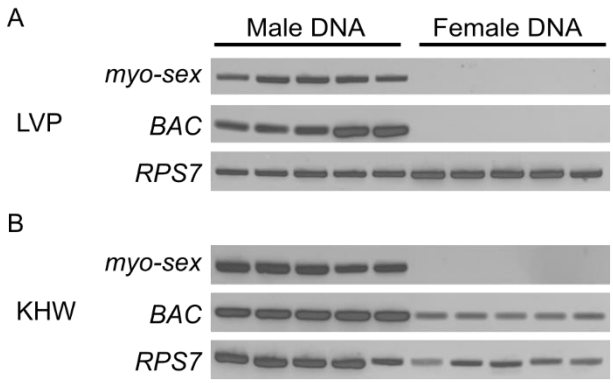

Figure 2. Genomic DNA amplification of *myo-sex* and BAC NDL62N22 from five pools of five male and female mosquitoes from the (A) Liverpool and (B) kh^{w} strains of *A. aegypti*. A ribosomal protein gene (*RPS7*) was amplified in both male and females samples to verify the integrity of the genomic DNA.

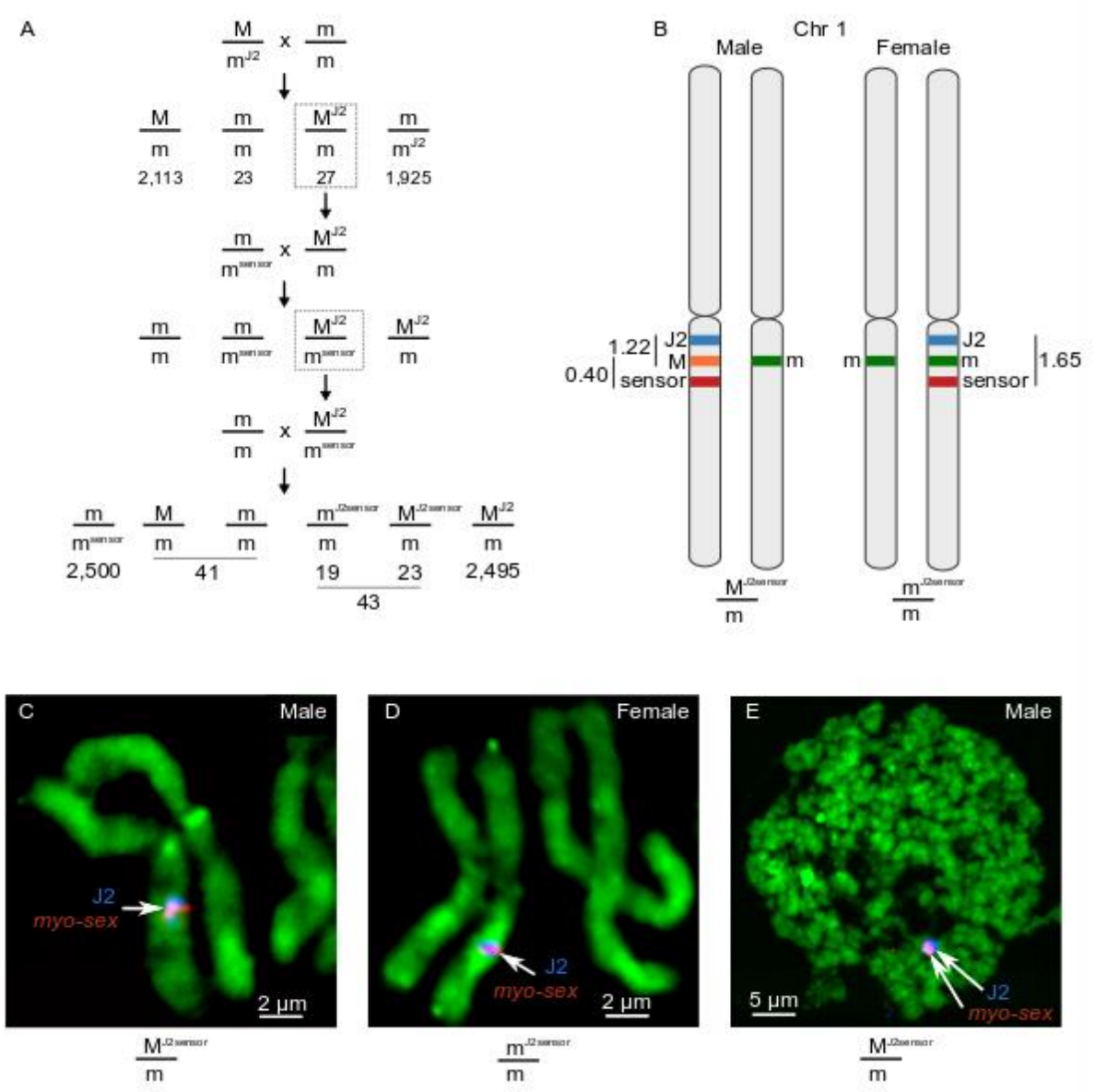

Figure 3. *Myo-sex* hybridizes to the location of the *A. aegypti* M-locus and can still recombine with the m-chromosome. (A) Using two transgenic strains of *A. aegypti* with sex-linked insertions, sensor and J2, a transgenic strain of *A. aegypti* was generated that has transgenes flanking the M-locus on the M-chromosome and a separate transgenic strain was generated with the two transgenes corresponding positions on the m-chromosome. Note that one of the 43 transgenic recombinants died before we can determine its sex. (B) A representation of the relative position of the transgenes with respect to the M-locus and the m-locus. Numbers

represent linkage in centimorgans. FISH on mitotic chromosomes from the *khw* strain of *A. aegypti* in males (C) and females (D) with probes for J2 and *myo-sex*. In (C) the blue J2 signal spans the entire presumed M-locus and the surrounding region. (E) FISH on male polytene chromosome shows that the *myo-sex* signal is fully within the J2 signal indicating that *myo-sex* is located between the J2 and sensor transgenes. *Myo-sex* was present in the transgenic strain with the transgenes on the m-chromosome. When J2 recombined with the m-chromosome to generate this transgenic line, *myo-sex* also moved to the m-chromosome with J2 indicating that *myo-sex* can still recombine.

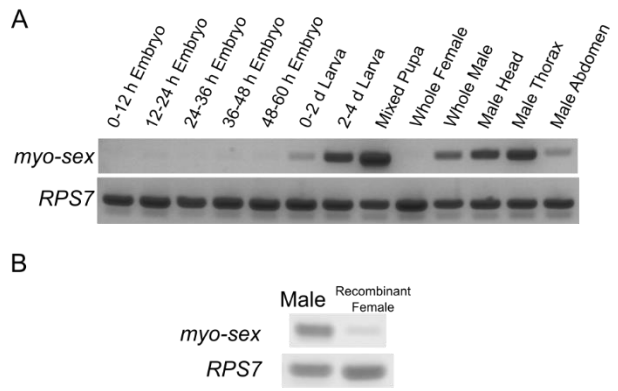

Figure 4. *Myo-sex* expreesion. (A) The expression profile of *myo-sex* from RT-PCR on cDNA spanning developmental time points. PCR products indicate expression of *myo-sex* in the larva, pupa, and adult male. A ribosomal protein gene (*RPS7*) was amplified in both male and females samples to verify the integrity of the cDNA. (B) RT-PCR on males and recombinant females. cDNA were synthesized from 10 pooled individuals in each sample. The recombinant females are the $m/m^{J2sensor}$ females shown in fig. 3A. Thirty-two cycles of amplification were performed. (C) Quantitative RT-PCR performed on males, normal females, and the $m/m^{J2sensor}$ recombinant females. Data are from three biological replicates of pools of three individuals for each test group. Relative expression levels were calculated using the ΔCt method with RPS7 as the endogenous control (Sengul and Tu 2010).

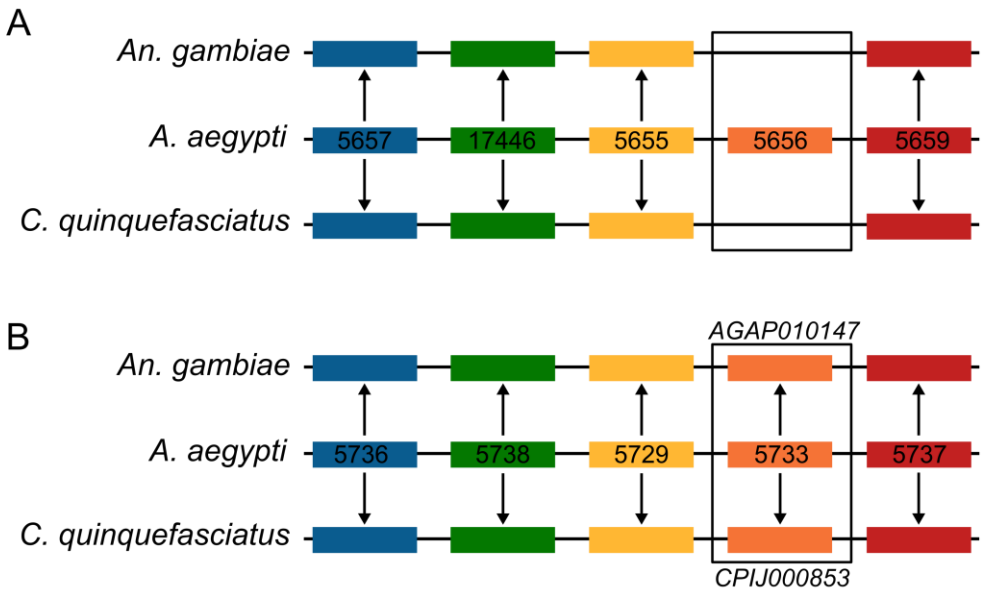

Figure 5. The synteny of the paralogs of *myo-sex*. The synteny of the closest paralog of *myo-sex*, *AAEL005656* (A) suggests that *AAEL005656* is inserted into a synteny block conserved among all three mosquitoes. The synteny of then next closest paralog of myo-sex, *AAEL005733* (B), is conserved in *A. aegypti*, *An. gambiae* and *C. quinquefasciatus*. *A. aegypti* gene names preceded by the VectorBase convention AAEL00. Genes of the same color within each panel are orthologues assigned by OrthoDB. The gene names of the orthologues of *AAEL005733* are shown in panel B.

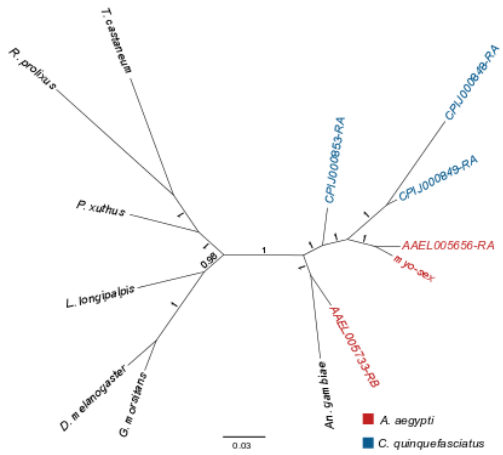

Figure 6. The phylogeny of *myo-sex* and other myosin heavy genes in insects. The phylogeny suggests that both *myo-sex* and *AAEL005656* originated after the evolutionary divergence of *Aedes* and *Culex* mosquitoes. The protein IDs for the genes represented in the phylogeny are as follows: *Aedes aegypti* (*AAEL005656-PA*), *Aedes aegypti* (*AAEL005733-PB*), *Culex quinquefasciatus* (*CPIJ000848-RA*), *Culex quinquefasciatus* (*CPIJ000849-PA*), *Culex quinquefasciatus* (*CPIJ000853-PA*), *Anopheles gambiae* (*AGAP010147-PA*), *Rhodnius prolixus* (*RPRC012274-PA*), *Tribolium castaneum* (*XP_001814139.1*), *Glossina morsitans* (*GMOY005703-PA*), *Drosophila melanogaster* (*FBpp0080463*), *Lutzomyia longipalpis* (*LLOTMP009501-PA*), and *Papilo xuthus* (*BAG30740.1*). We note that the current *CPIJ000853* protein model in Vectorbase lacks approximately 350 residues at the N-terminus and contains a 200 residue insertion compared to other proteins in the alignment (supplemental file S2). A semi-manual re-annotation of the *CPIJ000853* genomic sequence recovered a protein sequence that is 94% identical to *AAEL005733* (supplemental file 3). The re-annotated *CPIJ000853* grouped together with *AAEL005733* with credibility value of 1 in the new phylogeny (not shown).

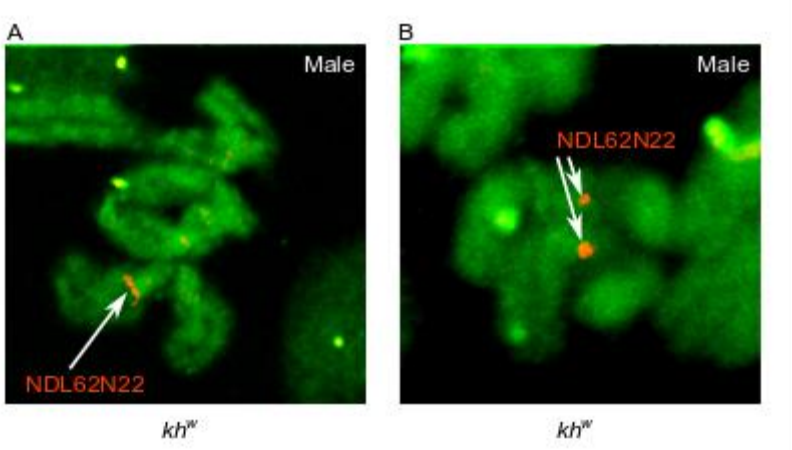

Figure 7. FISH on mitotic chromosomes with probes for BAC NDL62N22 in males hybridize to both the M-chromosome and the m-chromosome indicating that the BAC is not male-biased in the *kh*^{w} strain of *A. aegypti*. The signal is observed at band 1q21, the established location of the sex-determining locus.

Table 1. The total number of sequences, and the number of sequences with CQs less than 0.2 from the *A. aegypti* scaffolds, contigs, transcripts, ESTs, and BAC-ends. Where total sequences is the number of sequences for which CQs were calculated, with greater than or equal to 30 alignments from the male sequence data.

Sequences	Total sequences	CQs < 0.2
Known autosomal sequences	7,713	7
Supercontigs	4,505	40
Contigs	35,292	106
Transcripts	16,106	28
ESTs	221,753	747
BAC-ends	79,413	1,423

Table 2. The CQs and the ratio of alignments based on relaxed blastn parameters to the five male-biased myosin ESTs and the assembled *myo-sex* transcript (KF150020). Sequences starting with BQ and DV are ESTs from VectorBase.

Sequence	Female Alignments Bowtie	Male Alignments Bowtie	CQ	Female Alignments blastn	Male Alignments blastn	Ratio of alignments
BQ789600.1	0	34	0	0	125	0
BQ789612.1	0	63	0	5	175	0.023
BQ789634.1	0	74	0	36	147	0.19
BQ789633.1	0	79	0	5	177	0.022
DV248113.1	0	91	0	0	149	0
Myo-sex transcript	0	899	0	180	1820	.078

2.6: References

Adelman ZN, Anderson MA, Morazzani EM, Myles KM 2008. A transgenic sensor strain for monitoring the RNAi pathway in the yellow fever mosquito, *Aedes aegypti*. Insect biochemistry and molecular biology 38: 705-713.

Aryan A, Anderson MA, Myles KM, Adelman ZN 2013. Germline excision of transgenes in *Aedes aegypti* by homing endonucleases. Scientific reports 3.

Carvalho AB, Koerich LB, Clark AG 2009. Origin and evolution of Y chromosomes: *Drosophila* tales. Trends in Genetics 25: 270-277.

Biedler JK, Hu W, Tae H, Tu Z 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: e33933.

Bull JJ. 1983. Evolution of sex determining mechanisms: The Benjamin/Cummings Publishing Company, Inc.

Carpenetti TL, Aryan A, Myles KM, Adelman ZN 2012. Robust heat-inducible gene expression by two endogenous hsp70-derived promoters in transgenic *Aedes aegypti*. Insect Mol Biol 21: 97-106.

Carvalho AB 2002. Origin and evolution of the *Drosophila* Y chromosome. Current opinion in genetics & development 12: 664-668.

Carvalho AB, Koerich LB, Clark AG 2009. Origin and evolution of Y chromosomes: *Drosophila* tales. Trends in Genetics 25: 270-277.

Carvalho AB, Vibranovski MD, Carlson JW, Celniker SE, Hoskins RA, Rubin GM, Sutton GG, Adams MD, Myers EW, Clark AG 2003. Y chromosome and other heterochromatic

sequences of the Drosophila melanogaster genome: how far can we go? Genetica 117: 227-237.

Chan AH, Jenkins PA, Song YS 2012. Genome-wide fine-scale recombination rate variation in Drosophila melanogaster. PLoS genetics 8: e1003090.

Charlesworth D, Charlesworth B, Marais G 2005. Steps in the evolution of heteromorphic sex chromosomes. Heredity (Edinb) 95: 118-128. doi: 10.1038/sj.hdy.6800697

Charlesworth D, Mank JE 2010. The birds and the bees and the flowers and the trees: lessons from genetic mapping of sex determination in plants and animals. Genetics 186: 9-31. doi: 10.1534/genetics.110.117697

Clements A. 1992. The biology of mosquitoes. Development Volume 1, ed. nutrition and reproduction. In: London: Chapman & Hall.

Comeron JM, Ratnappan R, Bailin S 2012. The many landscapes of recombination in Drosophila melanogaster. PLoS genetics 8: e1002905.

Criscione F, Qi Y, Saunders R, Hall B, Tu Z 2013. 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.

Edgar RC 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics 5: 113.

Ferdig MT, Taft AS, Severson DW, Christensen BM 1998. Development of a comparative genetic linkage map for *Armigeres subalbatus* using *Aedes aegypti* RFLP markers. Genome research 8: 41-47.

Gerton JL, et al. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences 97: 11383-11390.

Hall AB, et al. 2013. Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. BMC genomics 14: 273.

Huang X, Madan A 1999. CAP3: A DNA sequence assembly program. Genome research 9: 868-877.

Huelsenbeck JP, Ronquist F 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754-755.

Jiménez L, Kang BK, DeBruyn B, Lovin D, Severson D 2004. Characterization of an *Aedes aegypti* bacterial artificial chromosome (BAC) library and chromosomal assignment of BAC clones for physical mapping quantitative trait loci that influence *Plasmodium* susceptibility. Insect Mol Biol 13: 37-44.

Krebs KC, Brzoza KL, Lan Q 2002. Use of subtracted libraries and macroarray to isolate developmentally specific genes from the mosquito, *Aedes aegypti*. Insect biochemistry and molecular biology 32: 1757-1767.

Krzywinski J, Chrystal MA, Besansky NJ 2006. Gene finding on the Y: fruitful strategy in *Drosophila* does not deliver in *Anopheles*. Genetica 126: 369-375.

Krzywinski J, Nusskern DR, Kern MK, Besansky NJ 2004. Isolation and characterization of Y chromosome sequences from the African malaria mosquito *Anopheles gambiae*. Genetics 166: 1291-1302.

Langmead B, Trapnell C, Pop M, Salzberg SL 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

Malcolm CA, et al. 1998. A sex-linked *Ace* gene, not linked to insensitive acetylcholinesterase-mediated insecticide resistance in *Culex pipiens*. Insect Mol Biol 7: 107-120.

Marın I, Baker BS 1998. The evolutionary dynamics of sex determination. Science 281: 1990-1994.

McClelland G 1962. Sex-linkage in *Aedes aegypti*. Trans roy Soc trop Med Hyg 56.

Mita K, Ichimura S, James TC 1994. Highly repetitive structure and its organization of the silk fibroin gene. Journal of molecular evolution 38: 583-592.

Mori A, Severson D, Christensen B 1999. Comparative linkage maps for the mosquitoes (*Culex pipiens* and *Aedes aegypti*) based on common RFLP loci. Journal of Heredity 90: 160-164.

Motara M, Rai K 1978. Giemsa C-banding patterns in *Aedes* (Stegomyia) mosquitoes. Chromosoma 70:51-58.

Motara MA, Rai KS 1977. Chromosomal differentiation in two species of *Aedes* and their hybrids revealed by Giemsa C-banding. Chromosoma 64: 125-132.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321-324.

Nene V, et al. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 316: 1718-1723. doi: 10.1126/science.1138878

Newton ME, Southern DI, Wood RJ 1974. X and Y chromosomes of *Aedes aegypti* (L.) distinguished by Giemsa C-banding. Chromosoma 49: 41-49.

Patten MM, Haig D 2009. Maintenance or loss of genetic variation under sexual and parental antagonism at a sex-linked locus. Evolution 63: 2888-2895.

Perrin N 2009. Sex reversal: a fountain of youth for sex chromosomes? Evolution 63: 3043-3049.

Rai K 1963. A comparative study of mosquito karyotypes. Annals of the Entomological Society of America 56: 160-170.

Rice WR 1987. The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. Evolution 41: 911-914.

Rice WR 1984. Sex chromosomes and the evolution of sexual dimorphism. Evolution: 735-742.Sengul, MS, Tu Z 2010. Expression analysis and knockdown of two antennal odorant-binding protein genes in *Aedes aegypti*. Journal of Insect Science 10: 171.

Severson D, Meece J, Lovin D, Saha G, Morlais I 2002. Linkage map organization of expressed sequence tags and sequence tagged sites in the mosquito, *Aedes aegypti*. Insect Mol Biol 11: 371-378.

Sezutsu H, Yukuhiro K 2000. Dynamic rearrangement within the Antheraea pernyi silk fibroin gene is associated with four types of repetitive units. Journal of molecular evolution 51: 329-338.

Sharakhova MV, et al. 2011. Imaginal discs–a new source of chromosomes for genome mapping of the yellow fever mosquito *Aedes aegypti*. PLoS neglected tropical diseases 5: e1335.

Shin D, Mori A, Severson DW 2012. Genetic mapping a meiotic driver that causes sex ratio distortion in the mosquito *Aedes aegypti*. J Hered 103: 303-307. doi: 10.1093/jhered/esr134

Stöck M, et al. 2011. Ever-young sex chromosomes in European tree frogs. PLoS Biology 9: e1001062.

Stöck M, et al. 2013. Low rates of X-Y recombination, not turnovers, account for homomorphic sex chromosomes in several diploid species of alearctic green toads (*Bufo viridis* subgroup). Journal of evolutionary biology.

Timoshevskiy VA, Severson DW, Black WC, Sharakhov IV, Sharakhova MV 2013. An Integrated Linkage, Chromosome, and Genome Map for the Yellow Fever Mosquito Aedes aegypti. PLoS neglected tropical diseases 7: e2052.

Timoshevskiy VA, Sharma A, Sharakhov IV, Sharakhova MV 2012. Fluorescent in situ Hybridization on Mitotic Chromosomes of Mosquitoes. J Vis Exp: e4215. doi: doi:10.3791/4215

Toups MA, Hahn MW 2010. Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. Genetics 186: 763-766.

Venkatesan M, Broman KW, Sellers M, Rasgon JL 2009. An initial linkage map of the West Nile Virus vector *Culex tarsalis*. Insect Mol Biol 18: 453-463. doi: 10.1111/j.1365-2583.2009.00885.x

Vicoso B, Kaiser VB, Bachtrog D 2013. Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. Proceedings of the National Academy of Sciences 110: 6453-6458.

Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV 2011. OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. Nucleic acids research 39: D283-D288.

Chapter 3: A male-determining factor in the mosquito *Aedes aegypti*

Andrew Brantley Hall[1,2,3†], Sanjay Basu[3,4†], Xiaofang Jiang[1,2,3], Yumin Qi[2,3], Vladimir A. Timoshevskiy[3,4], James K. Biedler[2,3], Maria V. Sharakhova[3,4], Rubayet Elahi[2], Michelle A. E. Anderson[3,4], Xiao-Guang Chen[5], Igor V. Sharakhov[1,3,4], Zach N. Adelman[1,3,4*], and Zhijian Tu[1,2,3*]

Affiliations:

[1]Interdisciplinary PhD Program in Genetics, Bioinformatics, and Computational Biology, Virginia Tech.

[2]Department of Biochemistry, Virginia Tech, United States of America.

[3]Fralin Life Science Institute, Virginia Tech United States of America.

[4]Department of Entomology, Virginia Tech United States of America.

[5]School of Public Health and Tropical Medicine, Southern Medical University, Guangdong, People's Republic of China

*Correspondence to: jaketu@vt.edu or zachadel@vt.edu

†Equal contribution

Author contribution:

ABH identified and characterized Nix. SB performed the CRISPR/Cas9 and ectopic expression experiments. XJ performed RNAseq analysis. YQ and JKB designed and performed RT-PCR and ddPCR assays. VT, MVS and IVS. performed FISH and interpreted the data. JKB and RE performed PCR and cloned Nix in nix- individuals. MAEA provided double-marked *A. aegypti* and materials for *A. albopictus* analysis. XC sequenced male *A. albopictus*. ZT and ZNA

initiated and designed the study. ABH, ZT, and ZNA wrote the manuscript with input from SB and IVS.

Citation

Hall, A. B., S. Basu, X. Jiang, Y. Qi, V. A. Timoshevskiy, J. K. Biedler, M. V. Sharakhova, et al. 2015. “A Male-Determining Factor in the Mosquito Aedes Aegypti.” *Science*, May. doi:10.1126/science.aaa2850.

3.1: Abstract

Sex determination in the mosquito *Aedes aegypti* is governed by a dominant male-determining factor (M factor) located within a Y chromosome–like region called the M locus. Here, we show that an M-locus gene, *Nix*, functions as an M factor in *A. aegypti*. *Nix* exhibits persistent M linkage and early embryonic expression, two characteristics required of an M factor. *Nix* knockout with clustered regularly interspaced short palindromic repeats (CRISPR)–Cas9 resulted in largely feminized genetic males and the production of female isoforms of two key regulators of sexual differentiation: doublesex and fruitless. Ectopic expression of *Nix* resulted in genetic females with nearly complete male genitalia. Thus, *Nix* is both required and sufficient to initiate male development. This study provides a foundation for mosquito control strategies that convert female mosquitoes into harmless males.

3.2: Main text

Insects employ diverse molecular mechanisms to determine sex (1–3). Sex is determined by X chromosome dosage in fruit flies (4), heterozygosity of the complementary sex determiner locus in honeybees (5), and a female-specific Piwi-interacting RNA in the silkworm Bombyx mori (6). Similar to mammals, sex determination in many insects is governed by an M factor located on a Y chromosome or homomorphic sex-determining chromosome (1). Despite the availability of genomic resources, no M factor has yet been characterized in any insect due to the difficulties of identifying genes in repeat-rich regions (1–3). Sex determination in mosquitoes is of particular interest because only adult females transmit pathogens responsible for dengue and yellow fever (7, 8). Consequently, a mosquito M factor would be useful in implementing vector control strategies where female mosquitoes are converted into harmless males (7).

Male development in *Aedes aegypti* is initiated by an M factor located on the homomorphic sex determining chromosome within a Y chromosome–like region called the M locus (9–11). The highly repetitive nature of the *Aedes aegypti* M locus has impeded the discovery of an M factor (3, 12, 13). To overcome this bottleneck, we developed the chromosome quotient method to find male specific (M-linked) genomic sequences by comparing the ratio of female to male alignments to reference sequences (12, 13). First, we separately sequenced the genomes of males and females from two strains of *A. aegypti*: Liverpool and khw. Then, we generated a rudimentary assembly using the male khw strain data because repeat rich regions like the M locus are often underrepresented in Sanger-based genome assemblies (14). Next, we aligned the male and female Illumina data to this assembly and identified 164 contigs that were potentially M linked (defined as more than 5 times as many alignments from male data as from female data in both strains) (table S1). Of the 164 sequences, 140 were either absent

from RNA sequencing (RNA-seq) data altogether, absent from early embryo RNA-seq samples, or present in female-derived RNA-seq samples. Within the 24 remaining sequences, we identified only one new gene that is a distant homolog of transformer-2 (table S2), which is involved in the splicing of *doublesex* (*dsx*) and *fruitless* (*fru*), two key regulators of sexual differentiation in Drosophila melanogaster (4). We named this gene *Nix*. Because of the tantalizing link to sex determination, we hypothesized that *Nix* may function as an M factor in *A. aegypti*.

The *Nix* cDNA spans 985 base pairs and encodes a 288–amino acid polypeptide containing two RNA recognition motifs (GenBank KF732822) (fig. S1 and tables S2 and S3). Primers for *Nix* amplified a polymerase chain reaction (PCR) product exclusively in male genomic DNA (Fig. 1A). We previously described two transgenes (J2 and sensor) that closely flank the M locus (13). Fluorescence *in situ* hybridization (FISH) to mitotic chromosomes using MJ2sensor/m males confirmed that the *Nix* signal localizes to only one homologous copy of chromosome 1 at position 1q21, the location of the M locus (Fig. 1D and fig. S2) (11). Digital droplet PCR indicated that one haploid copy of *Nix* is present in males and zero copies of *Nix* are present in females (Fig. 1E). Next, we analyzed whether recombination could separate *Nix* from the M locus. By screening 5000 individuals, we identified 19 recombinants where the J2 transgene was separated from the M locus (13, 15). In females from a colony established from these individuals, we could not identify *Nix* by PCR, supporting the conclusion that *Nix* is located within the M locus (Fig. 1B). Transcription of *Nix* was first detected 3 to 4 hours after oviposition (Fig. 1C and fig. S3), corresponding to the beginning of the syncytial blastoderm stage before sex is established (16). Thus, *Nix* exhibits two necessary characteristics of an M factor: persistent M linkage and early embryonic expression.

To investigate the role of *Nix* in mosquito development, we generated somatic loss-of-function mutants by injecting clustered regularly interspaced short palindromic repeats–Cas9 (CRISPR-Cas9) (17, 18) and synthetic guide RNAs (sgRNAs) targeting *Nix* into embryos oviposited by females that had mated with double-markedMJ2sensor/m transgenic males (fig. S4). In the absence of any phenotypic changes, virtually all males resulting from this cross would be double-marked, whereas all females would be unmarked. Genetic lesions in *Nix* were confirmed by RNA-seq and DNA sequencing and were associated specifically with *Nix* guide RNA target sites (fig. S5 and table S4).

Somatic knockout of *Nix* resulted in feminization or deformities in sexually dimorphic organs in more than two-thirds (55 of 79) of double-marked males (designated hereafter as *Nix*– males), whereas unmarked females (control) were morphologically typical (Fig. 2A and figs. S6 to S8). As somatic mosaics, levels of feminization were variable among *Nix*– individuals. The phenotype of each *Nix*– male was scored for the extent of feminization (table S5 and figs. S6 to S8). A common morphological feminization that appeared in 53% (42 of 79) of *Nix*– males was the absence of one or both gonocoxites and gonostyli, features specific to male genitals used to grasp the female during mating (Fig. 2E and figs. S6 and S7) (19). We also observed feminized antennae with fewer and shorter setae than normal males in 44% (35 of 79) of *Nix*– males (Fig. 2D and figs. S6 and S8).

We further investigated the molecular mechanism of the feminization of *Nix*– males. *Dsx* and *fru* are essential genes in the sex-determination pathway of many insects, and differential splicing of each results in a downstream cascade that programs the development of sexually dimorphic traits (20–23).We confirmed that *Nix*– males produced female splice variants of both *dsx* and *fru* at 0.47 and 1.44 times the amounts in wild-type females, respectively (Fig. 2, B and

C; table S6; and fig. S9). Using RNA-seq to examine the expression of sex-biased genes in *Nix*– males, we found a global feminization of sex-biased gene expression consistent with the observed morphological feminization and the key regulatory functions of *dsx* and *fru* (Fig. 2, F and G, and fig. S10). Thus, *Nix* is required to initiate male development and functions upstream of the two master regulators of sexual differentiation.

To determine whether *Nix* was sufficient for male determination, we investigated the effect of ectopic expression of *Nix* in genetic females. Embryos oviposited by females that had mated with double-marked M/mJ2sensor transgenic males were injected with a plasmid expressing *Nix* under the control of the *A. aegypti* polyubiquitin promoter (fig. S4) (24). In this case, virtually all genetic females would be double-marked, whereas genetic males would be unmarked. In our first experiment, more than two-thirds (16 of 23) of the double-marked females (designated hereafter as *Nix*+ females) exhibited extensive masculinization or deformities of the genitalia (Fig. 3A, table S7, and fig. S11). Two male-specific structures of the external genitalia, the gonocoxites and gonostyli (19), were clearly visible in 43% (10 of 23) of *Nix*+ females, whereas a further 26% (6 of 23) had deformed genitalia (Fig. 3, A and C, and table S7). Testes were identified in 34% (8 of 23) of *Nix*+ females and accessory glands; vasa deferentia were identified in 60% (14 of 23) of *Nix*+ females (Fig. 3, B to D; table S7; and fig. S11). In a second experiment, 27% (5 of 18) of *Nix*+ females exhibited masculinized or deformed genitalia (Fig. 3A and table S7). Thus, *Nix* is sufficient to initiate male development.

Using Illumina sequences from male genomic DNA and male RNA-seq, we identified a homolog of *Nix* in the Asian tiger mosquito, *A. albopictus*, with 52%identity at the amino acid level (e-value = 10−71) (GenBank KP765684 and figs. S12 and S13). This gene is only found in male genomic DNA and is expressed in adult males and early embryos of *A. albopictus*,

suggesting that *Nix* may be a conserved M factor in these Aedes mosquitoes. We also searched for *Nix* in other mosquito genera, including Culex and Anopheles, but found only autosomal or X-linked genes with RNA recognition motifs.

Here, we demonstrate that an M-locus gene, *Nix*, is an M factor in *A. aegypti* because it is both required and sufficient to initiate male development, although complete sex conversion has not been achieved in our transient assays. *Nix* encodes a potential splicing factor, and the absence of *Nix* shifts the alternative splicing of *dsx* and *fru* toward their female isoforms. The discovery of *Nix* provides an opportunity to characterize the remaining genes and interactions in the *A. aegypti* sex-determination pathway, which may be informative in unraveling the sex determination cascades of mosquitoes in general.

Aedes aegypti is a major vector for dengue, yellow fever, and chikungunya viruses, and only female mosquitoes feed on blood and transmit these pathogens. Thus, genetic control methods that introduce a male bias to reduce mosquito populations are attractive and potentially effective measures to reduce the incidence of mosquito borne disease (7, 8). When dosage compensation and sex determination are linked, as in the silkworm, manipulation of the sex-determination pathway results in sex-specific embryonic lethality due to misregulation of dosage compensation (6). In contrast, we have obtained partial sex change phenotypes from both *Nix* knockout and ectopic expression, presumably because *A. aegypti* does not require dosage compensation. Thus, this study provides the foundation for developing mosquito control strategies by converting females into harmless males or selectively eliminating deadly females.

Acknowledgments

We thank A. James and C. Barillas-Mury for critical review of the manuscript. We thank R. Saunders for mosquito care. Sequencing data used in this study has been submitted to the NCBI under the accessions: KF732822, KP765684, KP842989-KP843003, SRP044709, SRP046160, SRP047470, SRP047401, SRP034669, SRP055126, and SRP055127. The authors declare no competing financial interests. This research was supported by the National Institute of Health grant AI113643, the National Science Foundation Graduate Research Fellowship grant DGE-1519168, the National Natural Science Foundation grant 81420108024, the Fralin Life Science Institute, and the Virginia Agriculture Experimental Station. This study was considered by the Virginia Tech Institutional Biosafety Committee and was determined not to fall under Dual Use Research of Concern (DURC).

Supplementary Materials

www.sciencemag.org/content/348/6240/1268/suppl/DC1

Materials and Methods

Figs. S1 to S14

Tables S1 to S9

References (25–38)

12 November 2014; accepted 7 May 2015

Published online 21 May 2015;

10.1126/science.aaa2850

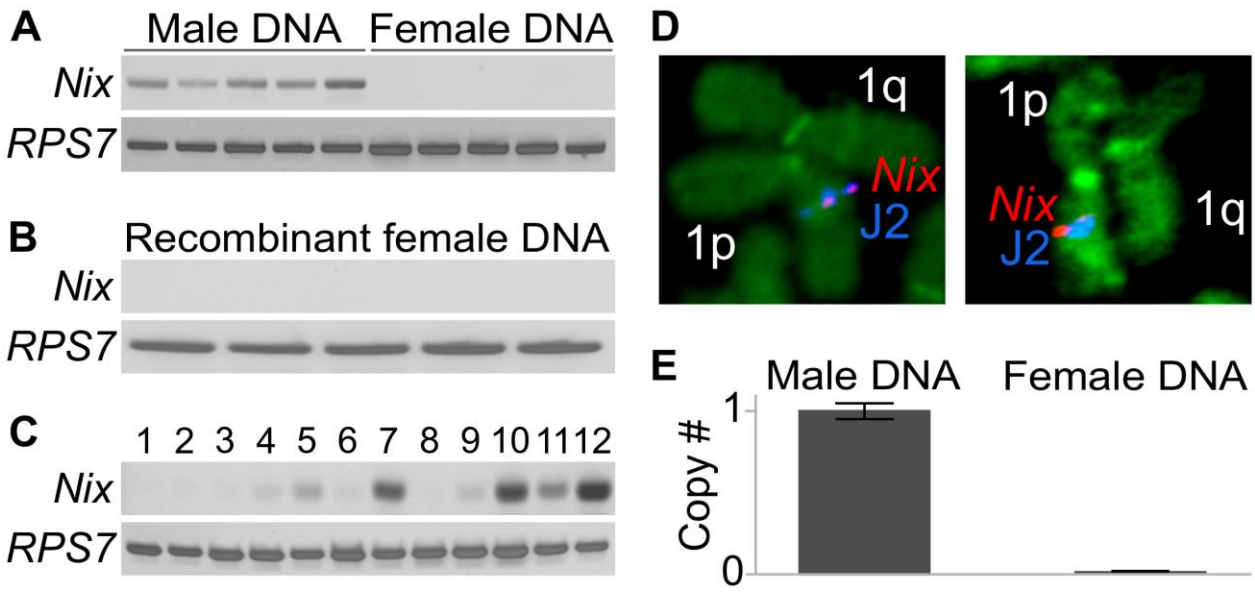

Fig. 1. *Nix* is located within the M-locus. **(A)** PCR for *Nix* in male and female genomic DNA. **(B)** PCR for *Nix* in genomic DNA from recombinant female $m^{J2sensor}$/m *A. aegypti*. A ribosomal protein, *RPS7*, was used as a positive control. **(C)** RT-PCR expression profile of *Nix* from 0-12 hour embryo cDNA starting at 0-1 hours in 1 hour increments. **(D)** FISH with a probe for *Nix* and the *J2* transgene in mitotic chromosomes of *J2* transgenic males. **(E)** *Nix* copy number as determined by digital droplet PCR (ddPCR) on male and female genomic DNA. Error bars represent s.e.m.

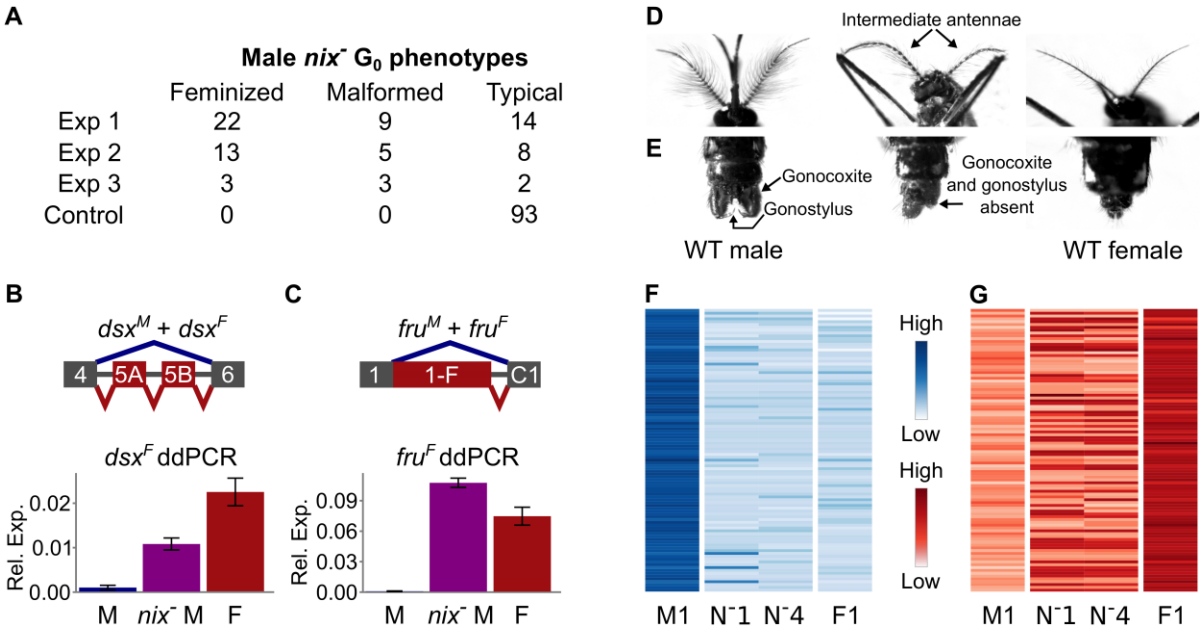

Fig. 2. Knockout with CRISPR/Cas9 demonstrates that *Nix* is required for male development. **(A)** Phenotypes of injected individuals. **(B)** The female isoform of *dsx* (dsx^F) is present *nix*- *A. aegypti* detected by ddPCR. Error bars represent s.e.m. **(C)** The female isoform of *fru* (fru^F) is present in the *nix*- *A. aegypti* detected by ddPCR. Rel. Exp. – Relative expression. Error bars represent s.e.m. **(D)** Feminization of the antennae in a *nix*- male individual. **(E)** Feminization in the genitals of a *nix*- male individual. **(F & G)** The log2 RPKM expression level heat map of the top 100 male-biased **(F)** and female-biased **(G)** genes in wild-type males, *nix*- male individuals, and wild-type females. Two heat maps from *nix*- male are shown here. All other heat maps are shown in Fig. S10.

A **Female *nix*$^{+}$ G_0 external phenotypes**

	Masculinized	Malformed	Normal
Exp 1	10	6	6
Exp 2	2	3	13
Control	0	0	134

B **Female *nix*$^{+}$ G_0 male features**

	Testes	Acessory glands	Claspers
Exp 1	8	14	9

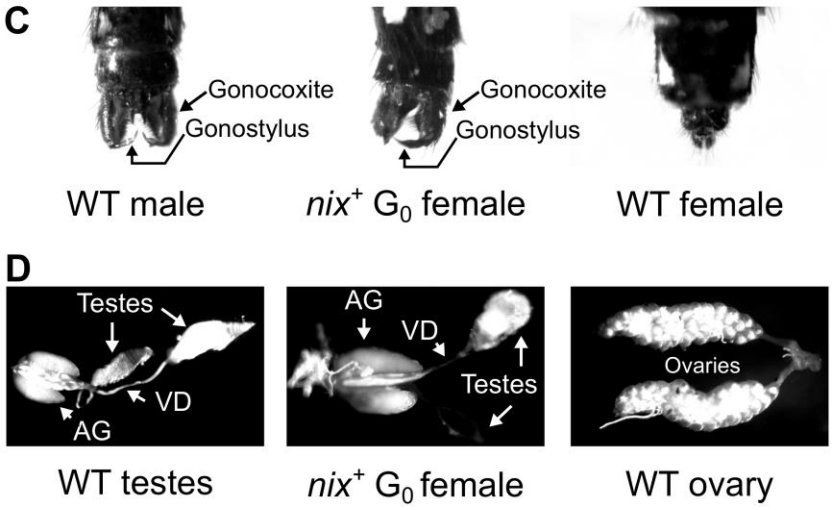

Fig. 3. Ectopic expression demonstrates that *Nix* is sufficient to initiate male development. **(A)** The phenotypes of *nix*$^{+}$ females. Thirty individuals were sacrificed at the larval stage to examine gene expression and therefore have an undetermined phenotype. **(B)** The number of *nix*$^{+}$ females with male-specific features from experiment 1. **(C)** Wild-type genitals compared to the genitals of *nix*$^{+}$ females which have gonocoxites and gonostyli. **(D)** Wild-type testes and ovaries compared to gonads of a *nix*$^{+}$ female which had testes and accessory glands. Wild-type images and *nix*$^{+}$ images are viewed under x55 and x80 respectively. AG: accessory glands, VD: vas deferens.

3.3: References

1. D. Bachtrog *et al.*, Sex Determination: Why So Many Ways of Doing It? *PLoS Biol*. **12**, e1001899 (2014).

2. Tree of Sex: A database of sexual systems. *Sci. Data*. **1** (2014) (available at http://dx.doi.org/10.1038/sdata.2014.15).

3. D. Charlesworth, J. E. Mank, The birds and the bees and the flowers and the trees: lessons from genetic mapping of sex determination in plants and animals. *Genetics*. **186**, 9–31 (2010).

4. H. Salz, J. W. Erickson, Sex determination in *Drosophila*: The view from the top. *Fly (Austin).* **4**, 60–70 (2010).

5. M. Hasselmann *et al.*, Evidence for the evolutionary nascence of a novel sex determination pathway in honeybees. *Nature*. **454**, 519–522 (2008).

6. T. Kiuchi *et al.*, A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature*. **509**, 633–636 (2014).

7. P. A. Papathanos *et al.*, Sex separation strategies: past experience and new approaches. *Malar J*. **8**, S5 (2009).

8. M. R. Wise de Valdez *et al.*, Genetic elimination of dengue vector mosquitoes. *Proc. Natl. Acad. Sci.* . **108** , 4772–4775 (2011).

9. B. M. Gilchrist, J. B. S. Haldane, Sex linkage and sex determination in a mosquito, *Culex molestus*. *Hereditas*. **33**, 175–190 (1947).

10. G. A. H. McClelland, Sex-linkage in *Aedes aegypti*. *Trans roy Soc trop Med Hyg*. **56** (1962).

11. M. E. Newton, R. J. Wood, D. I. Southern, Cytological mapping of the M and D loci in the mosquito, *Aedes aegypti* (L.). *Genetica*. **48**, 137–143 (1978).

12. A. B. Hall *et al.*, Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. *BMC Genomics*. **14**, 273 (2013).

13. A. B. Hall *et al.*, 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** , 179–191 (2014).

14. R. Hoskins *et al.*, Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. *Genome Biol.* **3**, research0085.1–0085.16 (2002).

15. Z. N. Adelman, M. a E. Anderson, E. M. Morazzani, K. M. Myles, A transgenic sensor strain for monitoring the RNAi pathway in the yellow fever mosquito, *Aedes aegypti*. *Insect Biochem. Mol. Biol.* **38**, 705–713 (2008).

16. J. K. Biedler, W. Hu, H. Tae, Z. Tu, 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**, e33933 (2012).

17. M. Jinek *et al.*, A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Sci.* . **337** , 816–821 (2012).

18. L. Cong *et al.*, Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science.* **339**, 819–823 (2013).

19. N. Becker, *Mosquitoes and Their Control* (Kluwer Academic / Plenum Publishers, 2003).

20. K. C. Burtis, B. S. Baker, *Drosophila doublesex* gene controls somatic sexual differentiation by producing alternatively spliced mRNAs encoding related sex-specific polypeptides. *Cell*. **56**, 997–1010 (1989).

21. M. Salvemini *et al.*, Genomic organization and splicing evolution of the *doublesex* gene, a Drosophila regulator of sexual differentiation, in the dengue and yellow fever mosquito *Aedes aegypti*. *BMC Evol. Biol.* **11**, 41 (2011).

22. M. Salvemini *et al.*, The Orthologue of the Fruitfly Sex Behaviour Gene *Fruitless* in the Mosquito *Aedes aegypti* Evolution of Genomic Organisation and Alternative Splicing. *PLoS One*. **8**, e48554 (2013).

23. S. Whyard *et al.*, Silencing the buzz: a new approach to population suppression of mosquitoes by feeding larvae double-stranded RNAs. *Parasit. Vectors*. **8**, 96 (2015).

24. M. A. E. Anderson, T. L. Gross, K. M. Myles, Z. N. Adelman, Validation of novel promoter sequences derived from two endogenous ubiquitin genes in transgenic Aedes aegypti. *Insect Mol. Biol.* **19**, 441–449 (2010).

Chapter 4: Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes

Andrew B. Hall[a,1], Philippos-Aris Papathanos[b,c,1], Atashi Sharma[d,1], Changde Cheng[e,f,1,2], Omar S. Akbari[g], Lauren Assour[h], Nicholas H. Bergman[i], Alessia Cagnetti[b], Andrea Crisanti[b,c], Tania Dottorini[c], Elisa Fiorentini[c], Roberto Galizi[c], Jonathan Hnath[i], Xiaofang Jiang[a], Sergey Koren[j], Tony Nolan[c], Diana Radune[i], Maria V. Sharakhova[d,k], Aaron Steele[h], Vladimir A. Timoshevskiy[d], Nikolai Windbichler[c], Simo V. Zhang[l], Matthew W. Hahn[l,m], Adam M. Phillippy[j], Scott J. Emrich[e,h], Igor V. Sharakhov[a,d,k,3], Zhijian Tu[a,n,3], Nora J. Besansky[e,f,3]

[a]The Interdisciplinary PhD Program in Genetics, Bioinformatics, and Computational Biology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA; [b]Section of Genomics and Genetics, Department of Experimental Medicine, University of Perugia, 06132 Perugia, Italy; [c]Department of Life Sciences, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom; [d]Department of Entomology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA; [e]Eck Institute for Global Health, University of Notre Dame, Notre Dame, Indiana 46556, USA; [f]Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana 46556, USA; [g]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California 91125, USA & Department of Entomology, University of California, Riverside Center for Disease Vector Research, Institute for Integrative Genome Biology, University of California, Riverside, CA, 92521, USA; [h]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, Indiana 46556, USA; [i]National Biodefense Analysis and Countermeasures Center, Frederick, MD 21702, USA; [j]Genome Informatics Section,

Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA; [k]Laboratory of Evolutionary Cytogenetics, Tomsk State University, Tomsk 634050, Russia; [l]School of Informatics and Computing, Indiana University, Bloomington, Indiana 47405, USA; [m]Department of Biology, Indiana University, Bloomington, Indiana 47405, USA; [n]Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA.

[1]These authors contributed equally to this work.

[2]Present address: Department of Integrative Biology, University of Texas, Austin, Texas 78712, USA

[3]Corresponding author. E-mail: Igor@vt.edu (I.V.S.); Jaketu@vt.edu (Z.T.); nbesansk@nd.edu (N.J.B)

Short title: The Y chromosome of *Anopheles*

Key words: Anopheles gambiae, PacBio, RNA-Seq, tandem repetitive DNA, Y-chromosome

Author Contributions:

Conceived project: NJB, SJE, IVS, ZT; Coordinated project: NJB; Genome and BAC sequencing: NHB, JH, DR, AMP, ABH, ZT, NJB; PacBio read correction and assembly: SK, AMP; RNA-Seq, RT-PCR and gene validation: PAP, CC, OSA, ACa, TD, EF, RG, TN, NW, ACr; Computational analysis of Y linkage: ABH, PAP, CC, LA, XJ, ASt, SZ, MWH, SJE, ZT; Cytogenetics and fluorescence *in situ* hybridization: ASh, MVS, VAT, IVS; Phylogeny reconstruction and simulations: CC, MWH. Wrote paper: NJB, ZT, ABH, PAP with input from other authors. Specific contributions of ABH in computational analysis: Identified Y-linked

PacBio reads and created Ydb, a database of Y sequences in Anopheles gambiae, identified Y-linked BACs using the CQ method, identified Y genes using the CQ method and caldera pipeline.

Submitted to PNAS on December 20th 2015.

4.1: Abstract

Y chromosomes control essential male functions in many species, including sex determination and fertility. However, due to obstacles posed by repeat-rich heterochromatin, knowledge of Y chromosome sequences is limited to a handful of model organisms, constraining our understanding of Y biology across the tree of life. Here, we leverage long single-molecule sequencing to determine the content and structure of the non-recombining Y (NRY) chromosome of the primary African malaria mosquito, *Anopheles gambiae*. We find that the *An. gambiae* Y consists almost entirely of a few massively amplified, tandemly arrayed repeats, some of which can recombine with similar repeats on the X chromosome. Sex-specific genome re-sequencing in a recent species radiation, the *An. gambiae* complex, revealed rapid sequence turnover within *An. gambiae* and among species. Exploiting 52 sex-specific *An. gambiae* RNA-Seq datasets representing all developmental stages, we identified a small repertoire of Y-linked genes that lack X gametologs and are not Y-linked in any other species except *An. gambiae*—with the notable exception of *YG2*, a candidate male-determining gene. *YG2* is the only gene conserved and exclusive to the Y in all species examined, yet sequence similarity to *YG2* is not detectable in the genome of a more distant mosquito relative, suggesting rapid evolution of Y chromosome genes in this highly dynamic genus of malaria vectors. The extensive characterization of the *An. gambiae* Y provides a long-awaited foundation for studying male mosquito biology, and will inform novel mosquito control strategies based on the manipulation of Y chromosomes.

Significance Statement

Interest in male mosquitoes has been motivated by the potential to develop novel vector control strategies, exploiting the fact that males do not feed on blood and do not transmit diseases such as malaria. However, genetic studies of male *Anopheles* mosquitoes have been impeded by the lack of molecular characterization of the Y chromosome. Here we show that the *An. gambiae* Y chromosome contains a very small repertoire of genes, with massively amplified tandem arrays of a small number of satellites and transposable elements constituting the vast majority of the sequence. These genes and repeats evolve rapidly, bringing about remodeling of the Y even among closely related species. Our study provides a long-awaited foundation for studying mosquito Y chromosome biology and evolution.

4.2: Introduction

Sex chromosomes carry a master switch gene responsible for sex determination (1). They are thought to derive from an ordinary pair of autosomes, and have multiple independent origins across the tree of life (2, 3). In animals with morphologically distinct (heterogametic) sex chromosomes, the Y has ceased crossing over with the X across some or all of its length and the non-recombining region is transmitted clonally by males (4, 5). The absence of recombination initiates progressive genetic decay—gene loss and accumulation of repetitive sequences—but there is increasing recognition that even relatively old and otherwise highly degenerate Y chromosomes retain functional importance not only for sexual reproduction but for their contributions to global gene regulation affecting health and survival (6-10). Notwithstanding these critical roles, the Y chromosome remains one of the most recalcitrant and poorly characterized portions of any genome more than a decade into the post-genomic era, with current knowledge resting largely on only two animal groups: mammals and *Drosophila* (2, 11).

Mosquitoes in the genus *Anopheles* are the exclusive vectors of human malaria, a disease that claimed nearly 600,000 lives globally in 2013—the majority in sub-Saharan Africa (12). Although fifteen years of intensified vector control efforts (mainly insecticide-impregnated bed nets) have successfully averted an estimated 663 million clinical cases of malaria (13), further progress toward elimination in the most malarious regions will depend upon the development of novel methods of vector control complementary to existing approaches (14). One method currently under development entails genetic modification of the mosquito to bias the population sex ratio toward males (which do not bite), with the goal of local population reduction or elimination (15-17). Modeling has shown that the most efficient means toward this end is the

engineering of a driving Y chromosome (18). A molecular-level understanding of the *Anopheles* Y chromosome is important to inform and optimize such a strategy.

Historical cytogenetic studies established that the *Anopheles* Y chromosome is entirely heterochromatic, but also suggested that, contrary to *Drosophila* and in common with mammals, it bears partial homology to the X chromosome and plays a male-determining role (19-24). However, efforts to characterize Y chromosome sequences in *Anopheles* have been thwarted by a lack of directed resources and effective tools. The unsurpassed medical importance of the African malaria mosquito *Anopheles gambiae* motivated its early selection for whole genome sequencing (25), making it second only to the model organism *D. melanogaster* as a fully sequenced insect genome. Yet because of formidable obstacles to assembling repeat-dense Y chromosome sequences (26), efforts to assemble or even to assign Y chromosome sequences in the framework of the *An. gambiae* genome project were largely unsuccessful despite separate sequencing of males and females (27), leaving its content and organization obscure. Here, we leverage the increased length and reduced bias of single-molecule sequencing (28), along with sex-specific transcriptional profiling, to identify an extensive data set of Y chromosome sequences and explore their organization and evolution in the young species radiation known as the *An. gambiae* complex (29), which contains some of the most important vectors of human malaria. We find massive remodeling of the repeat-dense Y chromosome and remarkably few genes, none of which have counterparts on the X. Only the *YG2* gene—the earliest to be expressed in 3-hour male embryos—is conserved and exclusive to the Y across species in the complex, and thus is a possible male-determining factor. Yet no sequence similarity to *YG2* can be detected in the genome of an Asian malaria vector from the same subgenus (30), underscoring the rapid evolution of Y-linked genes in the evolutionarily dynamic genus *Anopheles* (31).

Contrary to the species branching order, the *YG2* gene tree from the *An. gambiae* complex supported the grouping of two major malaria vectors with a history of substantial autosomal introgression (32), consistent with the hypothesis that even the Y chromosome may have crossed species boundaries. Although a Y chromosome assembly awaits further technological developments, the compilation and comparative analysis of Y chromosome sequences in the *An. gambiae* complex substantially advances our understanding of the composition, organization and evolution of the *Anopheles* Y chromosome and lays the groundwork for exploiting the Y chromosome to control disease transmission.

4.3: Results

Identification of *An. gambiae* Y chromosome sequences. Gross cytological estimates suggest that the *An. gambiae* Y chromosome constitutes ~10% of the 264 Mb *An. gambiae* genome (27, 33), yet a mere 0.18 Mb of unordered sequences have previously been assigned to the Y in the PEST reference genome assembly [www.vectorbase.org; (34)]. Similarly, a recent *An. stephensi* genome project identified only 57 short unordered Y sequences spanning ~50 kb from genomic reads, and 11 contigs spanning ~200 kb from BAC clones that were assigned to the Y chromosome (35). To overcome this impediment, we developed a strategy based on long-read, PacBio single molecule real-time (SMRT) sequencing (36). Template genomic DNA was extracted from male siblings of a single-pair mating that inherited the same paternal *An. gambiae* Y chromosome and was sequenced to 70X autosomal (35X heterosomal) coverage with PacBio SMRT sequencing [*Supporting Information* (*SI*) *Appendix*, Text S1.1]. Consensus-based error correction with PBcR (37) resulted in 40X autosomal (20X heterosomal) coverage of PacBio corrected reads with an N50 size of 2,799 bp (*SI Appendix*, Text S1.2). A whole-genome assembly of the entire PacBio dataset was performed with Celera Assembler (38, 39), resulting in a 294 Mb assembly with an N50 contig size of 101,465 bp (*SI Appendix*, Text S1.2). However, the average raw read length (2,479 bp) was insufficient for *de novo* reconstruction of the Y chromosome, and concerns about potentially misassembled Y chromosome contigs or scaffolds led us to focus exclusively on individual (unassembled) PacBio corrected reads from genomic DNA for all subsequent Y chromosome analysis.

As a complementary strategy to investigate the organization of large (100-kb) contiguous pieces of the *Anopheles* Y chromosome, PacBio sequencing of individual *An. gambiae* BAC clones was performed (*SI Appendix*, Text S1.3). These BACs were deemed potentially Y-linked

based on initial computational analysis of available BAC-end sequences (*SI Appendix*, Text S1.3). Directed, high-coverage PacBio sequencing yielded sufficient information to completely assemble each BAC computationally without manual finishing, and detailed computational analysis supported the majority as originating from the Y chromosome (*SI Appendix*, Text S1.3, Figs. S1-S2).

To identify presumptive Y-linked sequences among the unassembled genomic PacBio reads, we implemented two recent computational approaches (*SI Appendix*, Text S2) that exploit short-read (Illumina) genomic sequencing from sex-specific DNA pools (*SI Appendix*, Text S1.4, Table S1). The Y chromosome genome scan (YGS) approach was designed to operate on scaffolds from a genome assembly derived from mixed sexes or males; after identification and masking of identical repeats, scaffolds are classified as Y-linked if they have few or no kmer-length matches to female Illumina sequences (40). As applied to *An. gambiae* male PacBio corrected reads, which were treated as "scaffolds," YGS failed to unambiguously classify Y-linked sequences, apparently due to the extremely small fraction of Y chromosome sequence that is exclusive to the Y in *An. gambiae* as opposed to highly enriched there (see below; *SI Appendix*, Table S8). A second approach, the chromosome quotient (CQ) method (30), infers Y-linkage based on the female-to-male ratio of sequence alignments to a reference—in this case *An. gambiae* female-to-male Illumina sequences aligned to PacBio reads. At a conservative threshold value ($CQ \leq 0.2$) imposed across the length of a PacBio read, the CQ method classified 79,475 unassembled reads (246 Mb) as presumptive Y chromosome sequences, which populate a database that we denote Ydb (*SI Appendix*, Text S2.2, Tables S4-S7; Other Supporting File 1). Although the rate of false positives in Ydb should be low (30) (*SI Appendix*, Text S2.1), the conservative CQ threshold necessarily means that Ydb is incomplete with respect to possible Y

chromosome sequences that share extended sequence identity with other chromosomes, as would be expected for pseudoautosomal regions or sequences recently acquired from elsewhere in the genome. However, it is likely that Ydb represents much of the non-recombining (male-limited) Y chromosome (NRY) (*SI Appendix*, Table S4). Greater than 94% of sequence classes comprising Ydb were validated as Y-linked in *An. gambiae* (Fig. 1; *SI Appendix* Text S3, Table S9) through genomic PCR and physical mapping by fluorescent *in situ* hybridization (FISH) of representative sequences to mitotic chromosomes of male *An. gambiae* larvae (*SI Appendix*, Text S4).

The *An. gambiae* Y contains massively amplified satellites and retrotransposons. We conducted a detailed computational assessment of Y chromosome repeat content based on analysis of Ydb. Our inferences should be minimally affected by redundant and overlapping reads, as they are based on proportional content (relative abundance), and PacBio coverage has limited bias (28, 38). Initially, both PacBio Ydb reads and assembled BAC sequences were screened for interspersed repeats and low complexity DNA with RepeatMasker 4.0.3 (41), using the *An. gambiae* PEST RepeatMasker library augmented with previously characterized Y chromosome satellite and retrotransposon sequences (42, 43) (*SI Appendix*, Text S3). Anticipating that the *An. gambiae* Y chromosome contains previously unknown repeats, or repeats whose structures differ from those represented in the reference repeat library, we characterized both annotated and unclassified output from RepeatMasker through iterative clustering and consensus building of sequences in Ydb and the Y-linked BACs. This strategy ultimately revealed that ~98% of bases in Ydb belong to a very few repetitive sequence classes, amplified extensively (Fig. 2A; *SI Appendix*, Text S3, Table S9).

Satellite DNA accounts for ~49% of all bases in Ydb (*SI Appendix*, Table S9). Yet only six different satellite monomers were identified and two—AgY477 and AgY373—predominate, comprising 93% of all satellite DNA sequence in Ydb. Moreover, the satellite sequences are found as long tandem arrays in Ydb, largely devoid of transposable elements (TEs). These data suggest that satellite DNA is an abundant and homogeneous constituent of the NRY, a major Y chromosome sequence feature that we refer to as the SAR, for Satellite Amplified Region (Fig. 2A). The absence of other repetitive sequence classes interspersed within the SAR suggests limited genetic exchange between satellites and other repeats, but we find evidence of recombination and higher-order repeat structures among satellites within the SAR. Different satellite monomers that share extensive sequence similarity (AgY477 and AgY373; AgY280 and AgY53D) frequently co-occur on the same PacBio read, often as interspersed or even chimeric monomers, indicative of sister chromatid or intrachromatid exchange (Fig. 2C; *SI Appendix,* Text S3.1, Figs. S5-S8).

Another 43.5% of bases in Ydb are TE-related, of which only eight distinct TE types, mainly retrotransposons, were identified (*SI Appendix,* Text 3.2, Table S9). Remarkably, one particular element alone—a 6.9 kb Ty3/Gypsy LTR retrotransposon that we designate *zanzibar*—comprises almost 27% of bases in Ydb as a whole, and accounts for more than 61% of the bases classified as TEs (Fig. 2, *SI Appendix,* Table S9). Unlike the typical interspersed arrangement of TEs in genomic euchromatin, *zanzibar* is arranged on the Y in massively amplified head-to-tail tandem arrays, in which one LTR followed by one gag/pol region are repeated in succession like beads on a string: $(gag/pol+LTR)_n$. From its abundance and organization, we infer that this *Zanzibar* Amplified Region (designated ZAR; Fig. 2A,B)—like the SAR—is another prominent organizational feature of the NRY. *Zanzibar* monomers

(gag/pol+LTR) in the array may carry insertions of a variety of other TEs or TE fragments. Remarkably, every copy of *zanzibar* carrying a particular TE type (*e.g.*, *mtanga*) contains precisely the same TE sequence inserted into precisely the same *zanzibar* sites (Fig. 2, *SI Appendix,* Text S3.2), as though replica insertions in different *zanzibar* copies are not the result of independent transposition events. Taken together, these data—the precise tandem organization of the ZAR and the peculiar clonal nature of insertions—strongly suggest that *zanzibar* retrotransposons no longer function in the manner expected of autonomous transposable elements. Instead, *zanzibar* sequences appear to have been the substrate for illegitimate recombination and megabase-spanning tandem amplifications on the Y chromosome, analogous to the process described for the genesis of centromeric repeats in maize (44). Despite its evident origin as an autonomous transposable element, the present ZAR structure most closely resembles satellite sequence and thus may reflect a general pattern of sequence amplification and evolution on the *Anopheles* Y chromosome.

Of the remaining ~7.5% of Ydb bases, we were able to classify ~5.5% as repetitive, but the last ~2% could not be clustered (*SI Appendix,* Table S9). This small unclustered fraction contains a heterogeneous mixture of less abundant types of repetitive sequences, degenerate copies of repeats categorized above, and Y chromosome genes, many of which are also appear to be multicopy (see below; *SI Appendix,* Text S7, Fig. S11). Only four Ydb reads were classified by a metagenomic analysis as originating from another organism (*SI Appendix,* Text S2.2), suggesting that nearly all of Ydb is legitimate *An. gambiae* sequence.

Extensive structural dynamism of the Y chromosome in a young species radiation. Cytological observations conducted in the 1970s revealed striking differences in sex

chromosome heterochromatin among populations and between species in this complex (45, 46). Not only did the staining intensity and pattern vary, but also length of the Y chromosome, ranging from less than half the length of the X in one *An. gambiae* population to almost the same length as the X in others (45). However, a mechanistic understanding of the phenomenon was lacking. Our finding that ~98% of the bases in Ydb constitute highly repetitive sequence organized into tandem arrays suggests that the cytological observations may have their basis in rapid expansion and contraction of tandem repeats on the Y chromosome, through unequal crossover and a variety of other mechanisms (42, 43, 47-49). We applied computational and cytogenetic methods to assess the nature and degree of Y chromosome remodeling within *An. gambiae* and among sibling species during the relatively brief (2 MY) evolutionary history of the species complex.

Intraspecific variation in the SAR and ZAR of *An. gambiae* was assessed computationally among three laboratory colonies and among 85 individuals sampled from a natural population in Cameroon (*SI Appendix*, Text S1.4-1.5, Table S1-S2). From the *An. gambiae* Pimperena colony—our reference—and two additional colonies (G3 and Asembo), we generated Illumina sequences from sex-specific genomic DNA pools, and aligned them to the consensus sequences of *An. gambiae* monomer repeat units compiled from Ydb. Alignments were performed twice, using either a strict read-mapping protocol (for CQ calculations) or a less stringent mapping protocol used to produce a metric analogous to CQ, termed "relaxed CQ" (RCQ) (*SI Appendix*, Text S5, Tables S6-S7). Presence/absence and relative abundance of each major repeat was estimated from the number of mapped reads; male-bias was estimated from the female-to-male ratio of sequences reflected by CQ and RCQ (*SI Appendix*, Tables S6-S7). Using similar strategies, we also interrogated individually sequenced wild-caught *An. gambiae*

mosquitoes of both sexes (40 males; 45 females) from Cameroon (*SI Appendix*, Figs. S11-13, Tables S12-14).

Overall, the pattern of Y-linkage as inferred by CQ and RCQ was qualitatively similar among *An. gambiae* samples. However, the corresponding copy number of Y-linked sequences was much more labile (Fig. 3; *SI Appendix*, Text S5, Figs. S11-S13, Tables S6-7, S13-S15). The most dramatic copy number variation of any SAR or ZAR component was displayed by satellite sequences AgY53D and AgY280 in male samples (Fig. 3). Although normalized numbers of read alignments are not precise reflections of copy number, they can convey a rough approximation of relative abundance if copy number varies across orders of magnitude. Indeed, counts of male alignments to AgY53D and AgY280 spanned four to five orders of magnitude from the Asembo to the Pimperena colony (*SI Appendix*, Tables S6-S7), and even within the natural population from Cameroon, alignment counts among individual males (normalized to number per million for each sample) spanned three orders of magnitude (*SI Appendix*, Table S13), suggesting major expansions or contractions in array length. Copy number variation of this magnitude on the Y chromosome would be expected to affect its length; our cytogenetic length estimates indeed varied among individual males from the *An. gambiae* Pimperena colony (from ~25.9 Mb to ~47.8 Mb; *SI Appendix*, Text S4.2, Table S11), consistent with prior studies (50).

The sibling species complex to which *An. gambiae* belongs radiated rapidly and recently, within the last 2 MY (32). To examine the extent of structural divergence of the Y chromosome between species over this relatively short time frame, we generated Illumina sequences from sex-specific pools of three additional members of the complex (*An. arabiensis*, *An. quadriannulatus*, *An. merus*) (*SI Appendix*, Table S1), and assessed male bias and relative abundance as described

above for *An. gambiae* (*SI Appendix*, Text S5.2, Tables S6-S7, S15), and by FISH (Fig. 4). Rapid and extensive remodeling of the Y chromosome between species is evidenced by dramatic examples of turnover in both the SAR and ZAR. Satellite AgY477, heavily male-biased and a major component of the Y in *An. gambiae*, is abundant but not strongly sex-biased in *An. merus*, and is not detected in *An. arabiensis* or *An. quadriannulatus* of either sex (Fig. 1; SI Appendix Text S5.2, Table S15). Similarly, *zanzibar* is neither strongly sex-biased nor abundant in *An. arabiensis* and *An. merus*, yet in *An. quadriannulatus* (not a sister species of *An. gambiae*), this retrotransposon has an *An. gambiae*-like pattern of expanded tandem arrays on the Y chromosome (Fig. 1, Fig. 4, *SI Appendix*, Text S5.2, table S15).

The *Anopheles* Y recombines with the X chromosome. Meiotic pairing, chiasma formation, and crossing-over between the sex chromosomes have been reported for three anopheline species in two different subgenera (19, 21, 24). Indirect evidence in *An. stephensi*—the presence of rDNA and a major satellite sequence on both sex chromosomes (20, 35)—also hints at possible genetic exchange, although not necessarily crossing-over. In *An. gambiae*, the apparent stability of X- and Y-linked translocations (51) suggested that crossing-over between the X and Y chromosomes does not occur. If a pseudoautosomal region exists on the *An. gambiae* Y, our methods do not allow its unambiguous detection. However, several observations are consistent with some form of genetic exchange between the X and NRY. First, our computational results with CQ and YGS suggest that *An. gambiae* Y chromosome sequences are only rarely exclusive to the Y, although they may be greatly enriched there. Indeed, physical (FISH) mapping of individual components of the SAR (e.g., AgY53B; Fig. 4, *SI Appendix*, Fig. S10), or of fluorescently labeled sequences derived from the entire microdissected Y chromosome (Fig. 5A),

reveals extensive cross-hybridization of Y repeats with X chromosome heterochromatin due to sequence similarity between satellite monomers (*SI Appendix*, Text 4.1.2). Additional support for occasional X-Y genetic exchange emerged from individually sequenced male and female *An. gambiae* from Cameroon (Fig. 5B; *SI Appendix*, Text S6). Normalized counts of individual female Illumina sequence reads mapping to consensus satellite monomer sequences normally found on the Y chromosome yielded unexpectedly high counts in 6 of 45 females for AgY477, and in another (mostly non-overlapping) 6 of 45 females in the case of AgY53D (*SI Appendix*, Text S6, Table S14). As these same females lacked correspondingly high copies of other Y chromosome sequences, contamination by male genomic DNA (whether in the laboratory or via sperm stored in the spermatheca) is an unlikely alternative explanation. Finally, support for X-Y recombination comes from individual *An. gambiae* PacBio reads containing AgX367 monomers [a satellite from the X chromosome; (42)] together with AgY477 and AgY373 monomers (Fig. 5C), confirming previous evidence from PCR amplicon sequencing (42).

Rapid turnover of the small Y chromosome genic repertoire. Only three Y-linked genes had been identified previously in *An. gambiae* (30), designated *gYG1* to *gYG3* (hereafter, *YG1* to *YG3*). Aiming for comprehensive gene discovery on the *An. gambiae* Y chromosome, we performed extensive transcriptional profiling of developmentally staged *An. gambiae* embryos (nine time points), sexed larvae (three time points), and adults (whole and dissected males and females) through mRNA sequencing (RNA-Seq)—52 data sets in total (*SI Appendix*, Text S1.6, Table S3)—and integrated complementary approaches to gene finding (*SI Appendix*, Text S7). Gene candidates bearing significant similarity to known TEs or bacterial sequences were

discounted. Arising from the combined approaches were eight presumptive genes (*YG1-8*), including the three previously identified (Fig. 1; *SI Appendix*, Text S7).

To be considered valid Y genes, we required that they exhibit male-biased or male-specific expression from RNA-Seq as well as male-specific amplification by genomic PCR, conditions met by *YG1-5*. With the exception of *YG4*, this validated set was further confirmed by male-specific RT-PCR. Moreover, we were able to physically localize *YG5* to the Y chromosome by FISH (a homolog was also detected on chromosome 3; Fig. 4, SI Appendix, Text S7, Fig. S19). As we could not identify male-specific SNPs distinguishing *YG6-8* from their autosomal homologs, these genes could not be validated despite exclusive expression of *YG6* in male accessory glands (*SI Appendix*, Fig. S20), and elevated numbers of normalized read alignments to *YG8* from individually sequenced males versus females in our population sample from Cameroon (*SI Appendix*, Fig. S11, Tables S12-S14).

Beyond the strikingly small total number of genes identified as Y-linked in *An. gambiae* following this intensive search, it is noteworthy that gene number varies even between strains. Both *YG3* and *YG4* are Y-linked exclusively in the G3 strain of *An. gambiae*, not in Asembo or Pimperena (*SI Appendix*, Tables S6-S7). None of these genes have recognizable gametologs on the X, yet all have partial or complete homologs on the autosomes (*SI Appendix*, Text S7, Fig. S14), suggesting that they have been gained on the Y chromosome since its divergence from the X (see below).

To screen for candidate Y-linked genes in three *An. gambiae* sibling species (*An. arabiensis*, *An. merus*, and *An. quadriannulatus*), we used the male and female Illumina sequences from each species in conjunction with corresponding genome assemblies, mixed-sex transcript sets and *de novo* RNA-Seq assemblies (31), as well as *An. gambiae* genomic resources

(*SI Appendix*, Text S7). After merging the results of these approaches, we made two surprising observations. First, among all the Y-linked genes identified in *An. gambiae*, only one—*YG2*—was computationally detected and confirmed (by male-specific genomic PCR) as Y-linked in each of the other three very closely related species (Fig. 1). None of the other *An. gambiae* Y genes, with the sole exception of *YG1*, was Y-linked in any other species examined. For *YG1*, Y-linkage was validated in *An. arabiensis* and *An. quadriannulatus*, but in *An. merus* the ratio of female-to-male alignments was inconclusive and male-specific genomic PCR amplification was not possible owing to highly similar sequence elsewhere in the genome (Fig. 1, *SI Appendix*, Text S7, Tables S6-S7, Fig. S14). Thus *YG2* is the only gene both conserved on, and exclusive to, the Y chromosome in all four species examined. As it is the earliest to be expressed, at 3 hours of male embryonic development (*YG1* is not expressed until 4 hours), *YG2* is a possible male determining gene. Surprisingly, *YG2* is not a single-copy gene. The first suggestion that this might be the case was hinted by the number of read alignments to *YG2* from individual males in the Cameroon sample (Fig. 3). However, we have more definitive evidence for multiple, nearly identical copies of *YG2* in *An. gambiae*. Four distinct haplotypes were sampled repeatedly in Ydb PacBio reads (which derived from the same paternal Y chromosome; *SI Appendix*, Text S7.1.2, Fig. S17). Variant positions among *YG2* copies were not only validated by sequencing of genomic PCR amplicons from individual male *An. gambiae* derived from natural populations, but also through RNA-Seq data, which further indicates that multiple *YG2* copies are expressed (*SI Appendix*, Table S17).

With the exception of *YG2*, the near-complete absence of conserved Y-linkage between *An. gambiae* genes and corresponding genes in the sibling species was reinforced by the converse result, our second surprising observation: all genes identified as Y-linked in any one

sibling species could not be assigned to the Y chromosome in any of the other species (*SI Appendix*, Text S7.2, Tables S6-S7). In *An. quadriannulatus*, we found three novel Y-linked candidate genes (*SI Appendix*, Text S7, Table S20). In *An. arabiensis*, no genes other than *YG1* and *YG2* were detected on the Y chromosome. In *An. merus*, we found evidence supporting the duplication of a multi-gene segment from chromosome 3R onto the Y since its split from other *An. gambiae* complex lineages (*SI Appendix*, Text S7, Table S19). Among seven sequential genes on 3R in this segment (corresponding to AGAP009631-37 in *An. gambiae*), the first three (AGAP009631-33) and last two (AGAP009636-37) have detectable copies on the Y chromosome in *An. merus* (based initially on the relative number of normalized read alignments in females and males, later validated by male-specific PCR). Our data are consistent with the two intervening genes (corresponding to AGAP009634-35 on 3R) having been lost from the Y, and the five flanking genes becoming amplified, although further experimental evidence will be required to confidently reconstruct these events.

Possible Y chromosome introgression between hybridizing malaria vectors. The *YG2* gene potentially encodes a short, 56-aa peptide whose possible role in determining maleness is under investigation. In the Asian malaria vector *An. stephensi*,a Y-linked gene (*Guy1*) implicated in male determination also encodes a 56-aa sequence whose predicted secondary structure resembles that of the putative *YG2* peptide (52) even though primary sequence similarity is not detectable (30) over the relatively short evolutionary span since these lineages separated, ~30 MYA (53). The fact that *YG2* expression is detected in early embryos before any other *An. gambiae* Y-linked gene, taken together with its uniquely conserved Y chromosome location in all four *An. gambiae* complex species—in the face of otherwise rampant structural dynamism and

genic turnover on the Y—is consistent with a primary role in male determination in this group. For this reason, we predicted that a gene tree reconstructed from *YG2* would reflect the known species branching order (32). Although sibling species in the *An. gambiae* complex are not completely reproductively isolated, contemporary interspecific gene flow is possible only though female F1 hybrids; as their brothers are sterile, the Y chromosome cannot introgress (54). Contrary to this expectation, a *YG2* tree built from sequences derived from population samples of the four species considered in this study supported *An. gambiae* and *An. arabiensis* as most closely related (Fig. 6; *SI Appendix*, Text S8), an arrangement previously shown to be the result of massive historical introgression between these two species that involved most of the autosomes and the proximal ~10 Mb of the X chromosome (32). The simplest explanation for a gene tree disagreeing with the species tree in a rapid radiation such as the *An. gambiae* complex is incomplete lineage sorting (ILS). We performed coalescent simulations to assess the likelihood that the grouping of *An. gambiae* with *An. arabiensis* in the *YG2* tree is due to ILS alone. In 62 out of 1000 simulations under the species tree, we recovered *An. gambiae* and *An. arabiensis* as sister lineages (i.e. P=0.062; *SI Appendix*, Text S8), indicating that non-introgressing lineages could produce the observed tree a small fraction of the time. Although we cannot formally reject the null hypothesis at the 0.05 significance level, these results certainly do not rule out Y chromosome introgression. Introgression of the Y chromosome between species is conventionally viewed as unlikely (55), but it is important to consider that the pair of malaria vectors in question have historically exchanged the vast majority of the rest of their genomes, including part of the X chromosome (32). In this context, introgression of the Y chromosome is possible if not likely, as long as the introgression event(s) predated the development of male F1 hybrid sterility barriers between this species pair.

4.5: Discussion

From studies of mammals (6, 7, 11, 56, 57) and *Drosophila* (58-60), it is known that the Y chromosomes in both groups have lost most of their ancestral gene repertoires and have acquired copious amounts of repetitive and ampliconic/palindromic DNA. In the ~250 MY since *Drosophila* and *Anopheles* last shared a common Dipteran ancestor, there has been parallel evolution of heteromorphic sex chromosomes from the same ancestral linkage group (61), implying that the *Anopheles* Y must have undergone a similar fate of massive ancestral gene loss and genomic degradation. In one main characteristic—its male determining role—the *Anopheles* Y resembles the mammalian Y more than it does *Drosophila*, in which XO flies are (sterile) males and the scant Y chromosome genes are crucial only for male fertility (2). Yet in other respects, the *Anopheles* and *Drosophila* Y are much more similar. More than one-third of the human Y chromosome and 99.9% of the mouse Y is euchromatic (56, 62), whereas *Drosophila* and *Anopheles* Y chromosomes are entirely heterochromatic. Although relatively few in number, some ancestral X-Y gene pairs have been conserved throughout mammalian evolution due to their vital role as dosage-sensitive regulators of global gene expression (6, 7). Crucially, although cases are known in which a mammalian Y chromosome has acquired autosomal genes [*e.g.*, (63)], most extant mammalian Y-linked genes have an X-linked gametolog. By contrast, what little gene content exists on the Y in *Drosophila* or *Anopheles* is not only poorly conserved between species, but there are no recognizable ancestral gametologs; all known Y-linked genes in *Drosophila* seem to have an autosomal origin (58). The recent DNA-based duplication of a gene from chromosome 3R to the Y chromosome in *D. melanogaster* following its split from *D. simulans* ~4 MYA (64) mirrors our finding of a similar event in *An. merus* since the radiation of the *An. gambiae* complex, < 2 MYA. We conclude that the most salient factor uniting *Anopheles*

and *Drosophila* Y chromosomes may be the continuous gain of genes and functions from the autosomes (64), in contrast to the conservation of remaining ancestral gametologs seen on the mammalian Y. However, the *Anopheles* NRY appears to stand apart from both *Drosophila* and mammalian Y chromosomes in the relative paucity of male-specific content.

One of our main findings was rapid turnover in quantity and type of repetitive DNA on the Y chromosome within and between species in the *An. gambiae* complex. It is known that both satellite and ampliconic DNA regions are prone to rapid divergence in length, structure and sequence, due to unequal sister chromatid exchange between out-of-register repeat units and other mechanisms (49). On the Y chromosome such regions may be subject to accelerated rates of divergence compared to the rest of the genome. Between humans and chimps whose lineages diverged ~5 MYA, orthologous satellite arrays in the X centromere are collinear and share 93% sequence identity, while collinearity declines and sequence conservation drops to 78% between orthologous satellites in Y centromeres (65). Additional evidence of rapid length, structure and sequence evolution of satellites and ampliconic structures on the Y chromosome has been reported in mice species 1-2 MY diverged and mice subspecies separated by only ~900,000 years (65); between *D. melanogaster* and *D. simulans* that split ~4 MYA (60, 66); and among human males worldwide (67). Despite such pervasive remodeling of the *Anopheles* Y chromosome over short evolutionary distances, a transgene randomly inserted onto the Y chromosome of an *An. gambiae* strain in 2014 is transcriptionally active and has been stably integrated ever since, establishing that the Y chromosome is amenable to the molecular manipulation required for Y-linked genetic vector control strategies (68).

The high level of satellite DNA polymorphism within species could have important phenotypic consequences for fitness related traits (8, 9). Moreover, the dramatic degree of

satellite DNA turnover on the Y between closely related species has been implicated in hybrid incompatibility in *Drosophila* (69-71). Intriguingly, two genes known to cause hybrid incompatibility between *D. melanogaster* and *D. simulans* (*Hmr* and *Lhr*) function within species to repress transcripts from satellite DNAs and TEs (71). These species differ drastically in satellite DNA content; *D. simulans* contains four-fold less satellite DNA overall (5% versus 20% of the genome), and is particularly depauperate of the two most abundant satellite types in *D. melanogaster* (72). In *An. gambiae*, we found that AgY477 and AgY373 are the most abundant satellites on the NRY, and they are both expressed exclusively in adult male reproductive tissues; these satellite sequences are absent or altered in the other sibling species investigated (Fig. 1). Whether hybridization leads to misregulation of satellite DNA remains to be explored in the *An. gambiae* complex.

Laborious single-haplotype iterative mapping and sequencing has previously revealed the structure of mammalian Y chromosomes (11). In contrast, single-molecule sequencing now provides individual reads tens of kilobases in length, promising a resource-efficient alternative for characterizing Y chromosomes. Here, we were able to determine the content and structural characteristics of the heterochromatic *An. gambiae* NRY using this approach. Single-molecule sequencing reads were able to reveal complex repeat structures from whole-genome data and completely assemble heterochromatic BACs without manual finishing. However, the complete reconstruction of Y chromosomes remains a challenging problem. A recent PacBio assembly of *D. melanogaster* failed to completely assemble the Y chromosome (73), although it did successfully resolve the complex regions Mst77Y (74) and FDY (64). These results suggest that continued single-molecule read length improvements may soon enable the complete reconstruction of Y chromosomes from whole-genome data alone.

Acknowledgements: We thank F. Catteruccia and S.N. Mitchell for sharing unpublished data, J. Pease for assistance with simulations, and M. Kern, M. Menichelli, M.K. Lawniczak, I. Antoshechkin, T. Persampieri, R. Carballar, and R. D'Amato for technical assistance and discussion. Sequencing data and assemblies have been submitted to NCBI under two umbrella BioProject IDs: PRJNA254152 and SRP044019. Genomic sequencing was funded in part by a grant from the Eck Institute for Global Health, University of Notre Dame. RNA-Seq was funded in part from a European Community Seventh Framework Programme (FP7/2007–2013) under grant agreement N° 228421 (INFRAVEC). Individual laboratories were funded by the NIH [R01AI076584 (NJB, MWH), R21AI112734 (NJB, SJE), R21AI101459 (NJB), R21AI094289 and R21AI099528 (IVS), R21AI105575 (ZT), HHSN272200900039C (SJE)], FNIH through the VCTR program of the Grand Challenges in Global Health Initiative (NJB, ACr, PAP, TN), European Commission and Regione Umbria Grant I-MOVE (RG, EF, PAP), Rita-Levi Montalcini Career Development Award (PAP), Marie Curie Intra-European Fellowship for Career Development (IEF) PIEFGA-273268 (TD), European Research Council Grant 335724 (NW), NSF Graduate Research Fellowship grant DGE-1519168 (ABH), Department of Education GAANN Fellowship (AS), and Fralin Life Science Institute of Virginia Tech (IVS, ZT). This research was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (SK, AMP). The contributions of NHB, JH, and DR were funded under Agreement No. HSHQDC-07-C-00020 awarded by the Department of Homeland Security Science and Technology Directorate (DHS/S&T) for the management and operation of the National Biodefense Analysis and Countermeasures Center (NBACC), a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily

representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. In no event shall the DHS, NBACC, or Battelle National Biodefense Institute (BNBI) have any responsibility or liability for any use, misuse, inability to use, or reliance upon the information contained herein. The Department of Homeland Security does not endorse any products or commercial services mentioned in this publication.

Author Contributions: Conceived project: NJB, SJE, IVS, ZT; Coordinated project: NJB; Genome and BAC sequencing: NHB, JH, DR, AMP, ABH, ZT, NJB; PacBio read correction and assembly: SK, AMP; RNA-Seq, RT-PCR and gene validation: PAP, CC, OSA, ACa, TD, EF, RG, TN, NW, ACr; Computational analysis of Y linkage: ABH, PAP, CC, LA, XJ, ASt, SZ, MWH, SJE, ZT; Cytogenetics and fluorescence *in situ* hybridization: ASh, MVS, VAT, IVS; Phylogeny reconstruction and simulations: CC, MWH. Wrote paper: NJB, ZT, ABH, PAP with input from other authors.

4.6: References

1. Bull JJ (1983) *The Evolution of Sex Determining Mechanisms* (Benjamin/Cummings, Menlo Park, CA) p 316.
2. Bachtrog D (2013) Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature reviews. Genetics* 14(2):113-124.
3. Bachtrog D, *et al.* (2014) Sex determination: why so many ways of doing it? *PLoS Biol* 12(7):e1001899.
4. Charlesworth B & Charlesworth D (2000) The degeneration of Y chromosomes. *Philos Trans R Soc Lond B* 355(1403):1563-1572.
5. Rice WR (1984) Sex chromosomes and the evolution of sexual dimorphism. *Evolution* 38:735–742.
6. Bellott DW, *et al.* (2014) Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* 508(7497):494-499.
7. Cortez D, *et al.* (2014) Origins and functional evolution of Y chromosomes across mammals. *Nature* 508(7497):488-493.
8. Lemos B, Araripe LO, & Hartl DL (2008) Polymorphic Y chromosomes harbor cryptic variation with manifold functional consequences. *Science* 319(5859):91-93.
9. Lemos B, Branco AT, & Hartl DL (2010) Epigenetic effects of polymorphic Y chromosomes modulate chromatin components, immune response, and sexual conflict. *Proc. Natl. Acad. Sci. USA* 107(36):15826-15831.
10. Sackton TB, Montenegro H, Hartl DL, & Lemos B (2011) Interspecific Y chromosome introgressions disrupt testis-specific gene expression and male reproductive phenotypes in Drosophila. *Proc. Natl. Acad. Sci. USA* 108(41):17046-17051.

11. Hughes JF & Page DC (2015) The Biology and Evolution of Mammalian Y Chromosomes. *Annu. Rev. Genet.* 49:507-527.

12. World Health Organisation (2014) World Malaria Report: 2014. http://www.who.int/malaria/publications/world_malaria_report_2014/en/.

13. Bhatt S, *et al.* (2015) The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature* 526(7572):207-211.

14. malERA Consultative Group on Vector Control (2011) A research agenda for malaria eradication: vector control. *PLoS Med* 8(1):e1000401.

15. Galizi R, *et al.* (2014) A synthetic sex ratio distortion system for the control of the human malaria mosquito. *Nat Commun* 5:3977.

16. Windbichler N, Papathanos PA, & Crisanti A (2008) Targeting the X chromosome during spermatogenesis induces Y chromosome transmission ratio distortion and early dominant embryo lethality in Anopheles gambiae. *PLoS Genet* 4(12):e1000291.

17. Burt A (2014) Heritable strategies for controlling insect vectors of disease. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 369(1645):20130432.

18. Deredec A, Godfray HC, & Burt A (2011) Requirements for effective malaria control with homing endonuclease genes. *Proc Natl Acad Sci U S A* 108(43):E874-880.

19. Sakai RK, Baker RH, Raana K, & Hassan M (1979) Crossing-over in the long arm of the X and Y chromosomes in *Anopheles culicifacies*. *Chromosoma* 74:209-218.

20. Redfern CP (1981) Satellite DNA of Anopheles stephensi Liston (Diptera: Culicidae). Chromosomal location and under-replication in polytene nuclei. *Chromosoma* 82(4):561-581.

21. Mitchell SE & Seawright JA (1989) Recombination between the X and Y chromosomes in *Anopheles quadrimaculatus* species A. *J Heredity* 80:496-499.

22. Marchi A & Mezzanotte R (1990) Inter- and intraspecific heterochromatin variation detected by restriction endonuclease digestion in two sibling species of the Anopheles maculipennis complex. *Heredity* 65(Pt 1):135-142.

23. White GB (1980) Academic and applied aspects of mosquito cytogenetics. *Insect Cytogenetics*, eds Blackman RL, Hewitt GM, & Ashburner M (Blackwell Scientific Publications, Oxford), pp 245-274.

24. Fraccaro M, Laudani U, Marchi A, & Tiepolo L (1976) Karotype, DNA replication and origin of sex chromosomes in Anopheles atroparvus. *Chromosoma* 55(1):27-36.

25. Holt RA, *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298(5591):129-149.

26. Carvalho AB, *et al.* (2003) Y chromosome and other heterochromatic sequences of the Drosophila melanogaster genome: how far can we go? *Genetica* 117(2-3):227-237.

27. Krzywinski J, Nusskern D, Kern M, & Besansky NJ (2004) Isolation and characterization of Y chromosome sequences from the African malaria mosquito *Anopheles gambiae*. *Genetics* 166:1291-1302.

28. Ross MG, *et al.* (2013) Characterizing and measuring bias in sequence data. *Genome Biol* 14(5):R51.

29. White BJ, Collins FH, & Besansky NJ (2011) Evolution of *Anopheles gambiae* in relation to humans and malaria. *Annual Review of Ecology Evolution and Systematics* 42:111-132.

30. Hall AB, *et al.* (2013) Six novel Y chromosome genes in *Anopheles* mosquitoes discovered by independently sequencing males and females. *BMC Genomics* 14:273.

31. Neafsey DE, *et al.* (2015) Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* 347:1258522.

32. Fontaine MC, *et al.* (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347:1258524.

33. Sharakhova MV, *et al.* (2007) Update of the Anopheles gambiae PEST genome assembly. *Genome Biol* 8(1):R5.

34. Giraldo-Calderon GI, *et al.* (2015) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.* 43(Database issue):D707-713.

35. Jiang X, *et al.* (2014) Genome analysis of a major urban malaria vector mosquito, Anopheles stephensi. *Genome Biol* 15(9):459.

36. Eid J, *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910):133-138.

37. Koren S, *et al.* (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology* 14(9):R101.

38. Koren S, *et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30(7):693-700.

39. Myers EW, *et al.* (2000) A whole-genome assembly of Drosophila. *Science* 287(5461):2196-2204.

40. Carvalho AB & Clark AG (2013) Efficient identification of Y chromosome sequences in the human and Drosophila genomes. *Genome Res.* 23(11):1894-1907.

41. Smit AFA, Hubley R, & Green P (2013-2015) (RepeatMasker Open-4.0, www.repeatmasker.org).

42. Krzywinski J, Sangare D, & Besansky NJ (2005) Satellite DNA from the Y chromosome of the malaria vector Anopheles gambiae. *Genetics* 169(1):185-196.

43. Rohr CJ, Ranson H, Wang X, & Besansky NJ (2002) Structure and evolution of *mtanga*, a retrotransposon actively expressed on the Y chromosome of the African malaria vector *Anopheles gambiae*. *Mol. Biol. Evol.* 19(2):149-162.

44. Sharma A, Wolfgruber TK, & Presting GG (2013) Tandem repeats derived from centromeric retrotransposons. *BMC Genomics* 14:142.

45. Bonaccorsi S, Santini G, Gatti M, Pimpinelli S, & Colluzzi M (1980) Intraspecific polymorphism of sex chromosome heterochromatin in two species of the *Anopheles gambiae* complex. *Chromosoma* 76(1):57-64.

46. Gatti M, Santini G, Pimpinelli S, & Coluzzi M (1977) Fluorescence banding techniques in the identification of sibling species of the *Anopheles gambiae* complex. *Heredity* 38(1):105-108.

47. Cohen S, Agmon N, Sobol O, & Segal D (2010) Extrachromosomal circles of satellite repeats and 5S ribosomal DNA in human cells. *Mob DNA* 1(1):11.

48. Ma J & Jackson SA (2006) Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res.* 16(2):251-259.

49. Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. *Science* 191(4227):528-535.

50. Gatti M, Bonaccorsi S, Pimpinelli S, & Coluzzi M (1982) Polymorphism of sex chromosome heterochromatin in the *Anopheles gambiae* complex. *Recent developments in the genetics of insect disease vectors,* eds Steiner WWM, Tabachnick WJ, Rai KS, & Narang S (Stipes Publishing Co., Champaign, IL), pp 32-48.

51. Curtis CF, Akiyama J, & Davidson G (1976) Genetic sexing system in *Anopheles gambiae* species A. *Mosq. News* 36:492-298.

52. Criscione F, Qi Y, Saunders R, Hall B, & Tu Z (2013) 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.* 22(4):433-441.

53. Kamali M, *et al.* (2014) Multigene phylogenetics reveals temporal diversification of major African malaria vectors. *PLoS One* 9(4):e93580.

54. Davidson G, Paterson HE, Coluzzi M, Mason GF, & Micks DW (1967) The *Anopheles gambiae* complex. *Genetics of Insect Vectors of Disease,* eds Wright JW & Pal R (Elsevier Publishing Company, Amsterdam).

55. Payseur BA (2009) Y not introgress? Insights into the genetics of speciation in European rabbits. *Mol. Ecol.* 18(1):23-24.

56. Skaletsky H, *et al.* (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423(6942):825-837.

57. Hughes JF, *et al.* (2012) Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* 483(7387):82-86.

58. Koerich LB, Wang X, Clark AG, & Carvalho AB (2008) Low conservation of gene content in the Drosophila Y chromosome. *Nature* 456(7224):949-951.

59. Bachtrog D, Hom E, Wong KM, Maside X, & de Jong P (2008) Genomic degradation of a young Y chromosome in Drosophila miranda. *Genome Biol* 9(2):R30.

60. Mendez-Lago M, *et al.* (2011) A large palindrome with interchromosomal gene duplications in the pericentromeric region of the D. melanogaster Y chromosome. *Mol. Biol. Evol.* 28(7):1967-1971.

61. Toups MA & Hahn MW (2010) Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. *Genetics* 186(2):763-766.

62. Soh YQ, *et al.* (2014) Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* 159(4):800-813.

63. Saxena R, *et al.* (1996) The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. *Nat. Genet.* 14(3):292-299.

64. Carvalho AB, Vicoso B, Russo CA, Swenor B, & Clark AG (2015) Birth of a new gene on the Y chromosome of Drosophila melanogaster. *Proc Natl Acad Sci U S A* 112(40):12450-12455.

65. Pertile MD, Graham AN, Choo KH, & Kalitsis P (2009) Rapid evolution of mouse Y centromere repeat DNA belies recent sequence stability. *Genome Res.* 19(12):2202-2213.

66. Wei KH, Grenier JK, Barbash DA, & Clark AG (2014) Correlated variation and population differentiation in satellite DNA abundance among lines of Drosophila melanogaster. *Proc Natl Acad Sci U S A* 111(52):18793-18798.

67. Repping S, *et al.* (2006) High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat. Genet.* 38(4):463-467.

68. Bernardini F, *et al.* (2014) Site-specific genetic engineering of the Anopheles gambiae Y chromosome. *Proc Natl Acad Sci U S A* 111(21):7600-7605.

69. Ferree PM & Barbash DA (2009) Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. *PLoS Biol* 7(10):e1000234.

70. Bayes JJ & Malik HS (2009) Altered heterochromatin binding by a hybrid sterility protein in Drosophila sibling species. *Science* 326(5959):1538-1541.

71. Satyaki PR, *et al.* (2014) The Hmr and Lhr hybrid incompatibility genes suppress a broad range of heterochromatic repeats. *PLoS Genet* 10(3):e1004240.

72. Lohe AR & Brutlag DL (1987) Identical satellite DNA sequences in sibling species of Drosophila. *J. Mol. Biol.* 194(2):161-170.

73. Berlin K, *et al.* (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.*:in press.

74. Krsticevic FJ, Schrago CG, & Carvalho AB (2015) Long-Read Single Molecule Sequencing To Resolve Tandem Gene Copies: The Mst77Y Region on the Drosophila melanogaster Y Chromosome. *G3*:in press.

75. Chaisson M & Tesler G (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 13(1):238.

76. Treangen TJ, Sommer DD, Angly FE, Koren S, & Pop M (2002) Next Generation Sequence Assembly with AMOS. *Current Protocols in Bioinformatics*, (John Wiley & Sons, Inc.).

77. Chin CS, *et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10(6):563-569.

78. Wood D & Salzberg S (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 15(3):R46.

79. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.

80. Garrison E & Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]*

81. Lobo NF, *et al.* (2010) Breakpoint structure of the Anopheles gambiae 2Rb chromosomal inversion. *Malar J* 9:293.

82. Santolamazza F, Della Torre A, & Caccone A (2004) Short report: A new polymerase chain reaction-restriction fragment length polymorphism method to identify *Anopheles arabiensis* from *An. gambiae* and its two molecular forms from degraded DNA templates or museum samples. *Am. J. Trop. Med. Hyg.* 70(6):604-606.

83. Dobin A, *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15-21.

84. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357-359.

85. Anders S, Pyl PT, & Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2):166-169.

86. Anders S & Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.

87. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997v2 [q-bio.GN]*.

88. Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842.

89. Fu L, Niu B, Zhu Z, Wu S, & Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23):3150-3152.

90. Coulibaly MB, *et al.* (2007) Segmental duplication implicated in the genesis of inversion 2Rj of *Anopheles gambiae*. *PLoS ONE* 2(9):e849.

91. Timoshevskiy VA, Sharma A, Sharakhov IV, & Sharakhova MV (2012) Fluorescent in situ hybridization on mitotic chromosomes of mosquitoes. *JoVE* 67:e4215.

92. Rozen S & Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, eds Krawetz S & Misener S (Humana Press, Totowa, NJ), pp 365-386.

93. George P, Sharma A, & Sharakhov IV (2014) 2D and 3D chromosome painting in malaria mosquitoes. *JoVE* 83:e51173.

94. Besansky NJ, *et al.* (1995) Cloning and characterization of the white gene from Anopheles gambiae. *Insect Mol. Biol.* 4(4):217-231.

95. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27(2):573-580.

96. Grabherr MG, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29(7):644-652.

97. Trapnell C, *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* 7:562–578.

98. Rozen S, *et al.* (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423(6942):873-876.

99. Marçais G & Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6):764-770.

100. Cheng C, *et al.* (2012) Ecological genomics of *Anopheles gambiae* along a latitudinal cline: a population-resequencing approach. *Genetics* 190(4):1417-1432.

101. Collins FH, *et al.* (1988) Comparison of DNA-probe and isoenzyme methods for differentiating *Anopheles gambiae* and *Anopheles arabiensis* (Diptera: Culicidae). *J. Med. Entomol.* 25(2):116-120.

102. Gouy M, Guindon S, & Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27:221-224.

103. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306-314.

104. Guindon S, *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59(3):307-321.

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

Figures

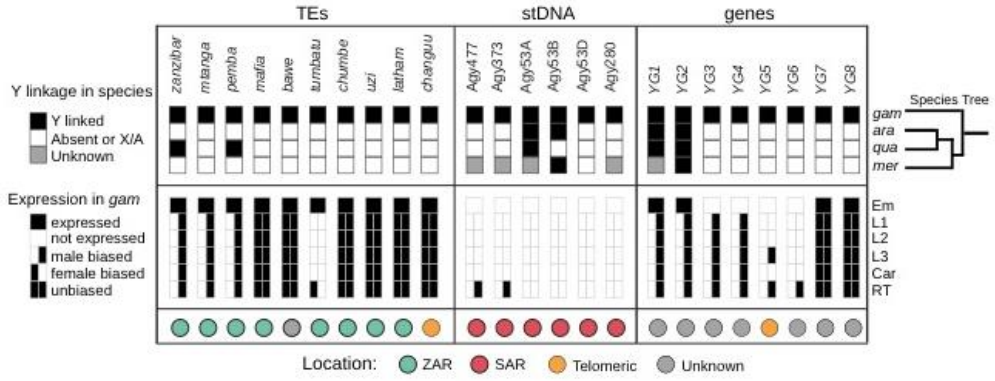

Fig. 1. Summary of major Y chromosome loci, showing rapid turnover of the Y chromosome content and expression patterns in the *Anopheles gambiae* species complex. In the **top panel**, black boxes indicate Y-linkage, white boxes indicate either total absence from the species or absence from its Y chromosome, and gray boxes indicate unknown status with regard to Y-linkage. Typically, sequences indicated by gray showed CQ or RCQ values of ~1, suggesting that they are either on both sex chromosomes, or on autosomes. Details are provided in *SI Appendix*, Tables S6-7, S15. At right, the species branching order provides an evolutionary context of the changes in Y chromosome content within the past 2 MY. Only *YG2* is conserved and exclusively on the Y chromosome in all four species of the *An. gambiae* complex. In the **middle panel**, sex-specific transcription in *An. gambiae* was assessed at different developmental stages and tissues, except for embryos (Em). The **bottom panel** shows the organization of the Y chromosome loci in *An. gambiae*, if known. TEs, transposable elements; stDNA, satellite DNA; *gam, An. gambiae; ara, An. arabiensis; qua, An. quadriannulatus; mer, An. merus*. Em, embryo; L1-L3, first to third instar larvae; RT, adult reproductive tissues; Car, adult carcass.

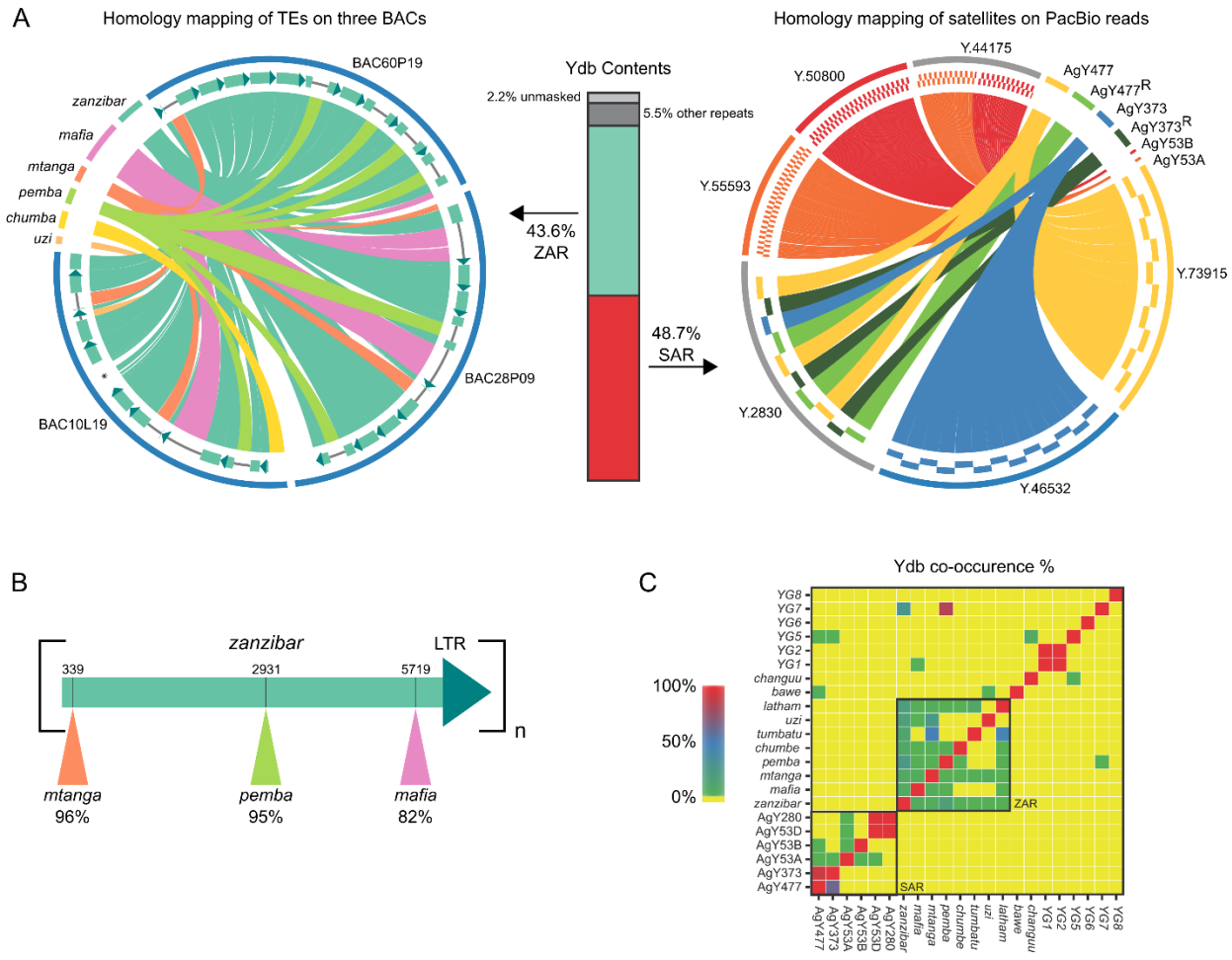

Fig. 2. The non-recombining Y (NRY) of *An. gambiae* mainly consists of massively amplified tandem arrays of a small number of satellites and transposable elements (TEs). A) Two major regions of the *An. gambiae* Y, the *zanzibar* amplified region (ZAR) and satellite amplified region (SAR), represent 92.3% of the sequences in Ydb (vertical bar plot). Ydb reflects the content of NRY in *An. gambiae*. Percentages were calculated by masking Ydb using annotated Y chromosome loci. The left circos plot, created by homology mapping of TEs on three Y chromosome BAC clones, shows the organization of the ZAR in the three BACs. As seen in these BACs, and as independently confirmed in PacBio reads, the ZAR consists of head-to-tail tandem arrays of *zanzibar* which sometimes have other transposons inserted. The arrays of *zanzibar* units inside each BAC are shown schematically directly inside the BAC ideograms

(blue semicircular lines) enclosing the circus plot. The dark green arrows of each *zanzibar* unit (shown enlarged in panel B) represent the single LTR; lines breaking *zanzibar* units indicate insertions of other TEs. A few small insertions (~200 bp) into *zanzibar* are too small to be visible in this plot. The asterisk in BAC10L19 corresponds to an atypical *zanzibar* unit that could result from recombination or misassembly. The circos plot at right, constructed by homology mapping of satellite monomers on PacBio reads from Ydb, shows the organization of the satellite amplified region (SAR). Shown are representative examples of the occurrence of homo-monomeric tandem arrays (Y73915, Y46532, Y55593), junctions between homo-monomeric tandem arrays (Y44175), and recombinant arrays (Y2830). The recombinant arrays are interspersed with recombinant and non-recombinant versions of AgY477 and AgY373 satellites (*SI Appendix*, Fig. S5). B) Schematic of a single *zanzibar* unit, consisting of a gag/pol domain and a single LTR; each unit is organized in a head-to-tail tandem array (see 2A, left circus plot). Shown by colored triangles are the canonical insertion sites of three other transposons (*mtanga*, *pemba*, *mafia*) into different *zanzibar* units. Percentages indicate the fraction of Ydb PacBio reads observed to carry TE insertions into *zanzibar* units at the precise insertion site illustrated (coordinates shown above the gag/pol domain). For example, we observed 243 of 256 (95%) PacBio reads in which *pemba* was inserted into *zanzibar* at position 2931. This phenomenon was independently confirmed in WGS Illumina reads. C) Co-occurrence matrix of Y chromosome loci in PacBio reads from Ydb. These results show that satellite sequences co-occur (in the SAR), as do TEs (in the ZAR), but that the ZAR and SAR regions are largely independent.

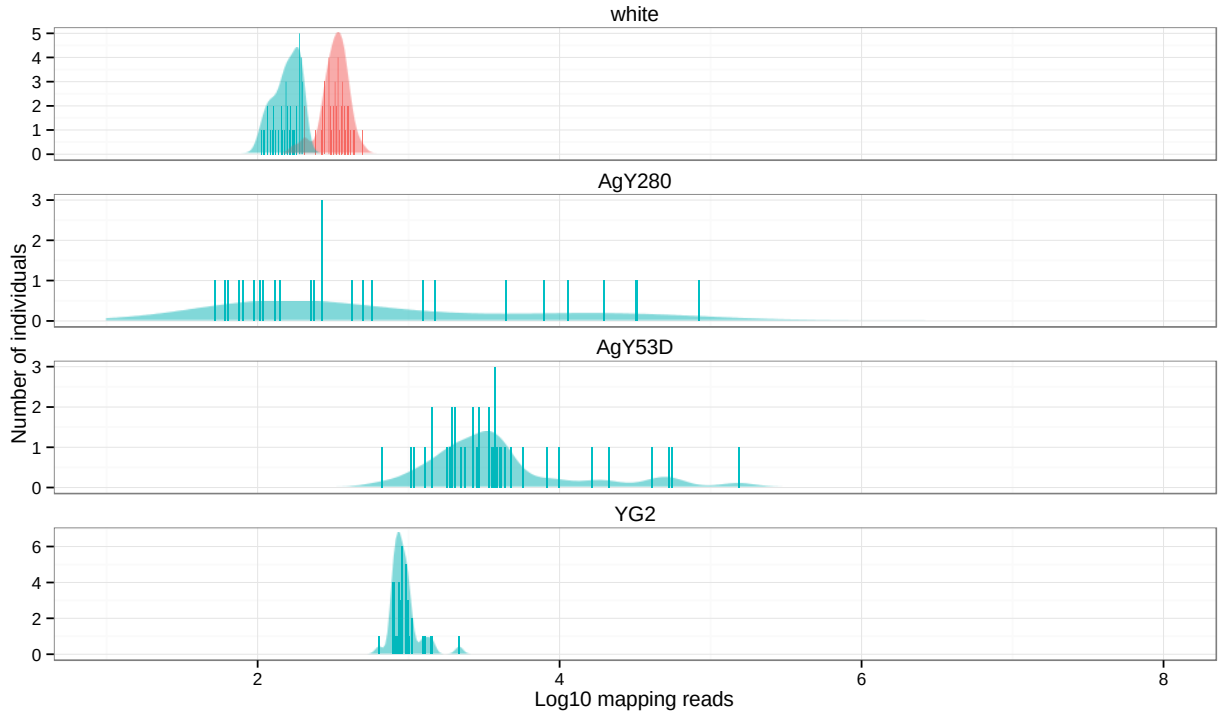

Fig. 3. Satellites AgY280 and AgY53D show extensive structural dynamism in males from a natural population of *An. gambiae*. Shown are histograms of the $\log_{10}$ numbers of read alignments from Illumina genomic sequence derived from 40 individual male mosquitoes from Cameroon, mapped to satellite monomers of AgY280 and AgY53D. For comparison are similar histograms of reads mapping to the presumptive male determining gene, *YG2*, and to the single-copy X-linked gene, *white* [in the latter case, reads come from 40 males (blue) and 45 females (pnk) from the Cameroon sample]. Numbers of read alignments to the satellite monomers varies drastically between individuals, in contrast to *YG2* and *white*, suggesting large within-population differences in satellite abundance on the Y.

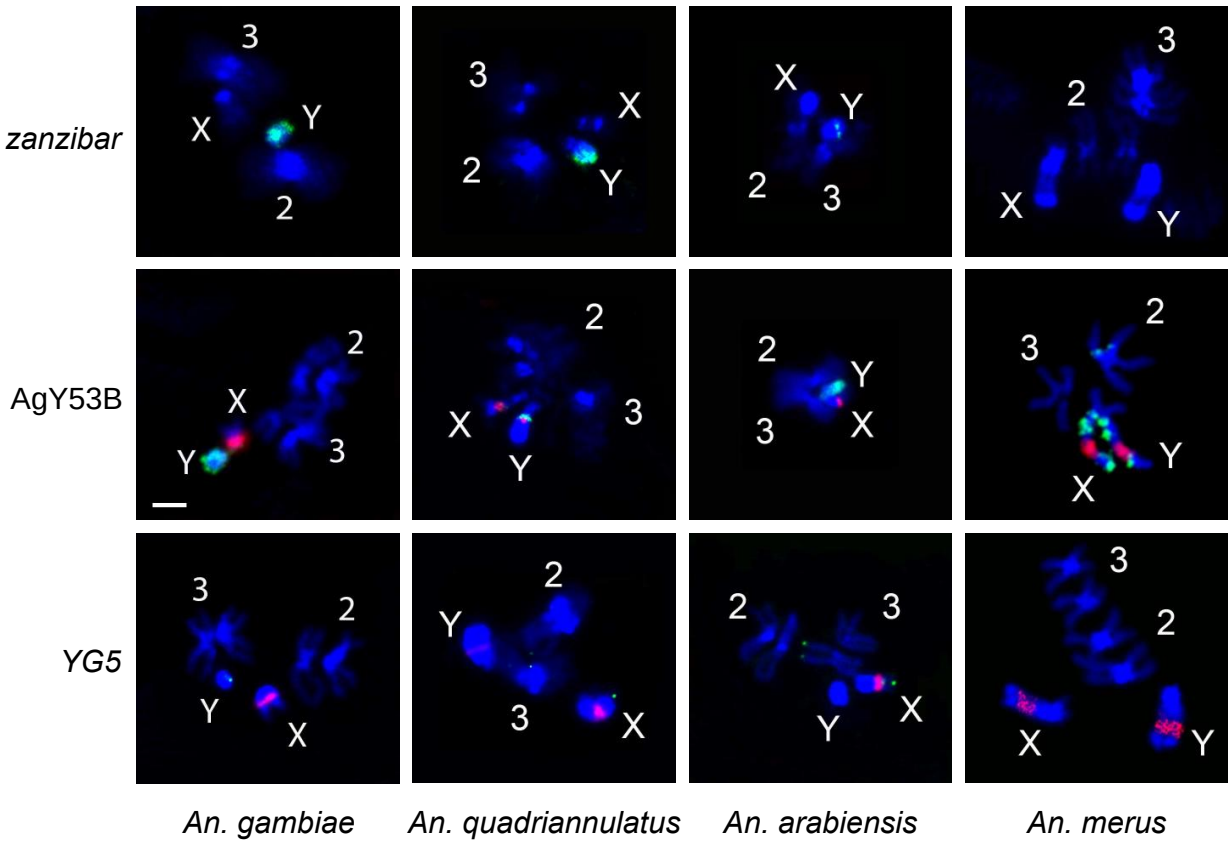

Fig. 4. Physical mapping supports structural dynamism of Y chromosome sequences in the *An. gambiae* complex. FISH of retrotransposon *zanzibar*, satellite AgY53B, and gene *YG5* (green signals) was performed on chromosomes of male *An. gambiae* Kisumu (*zanzibar*, *YG5*), *An. gambiae* Asembo (AgY53B), *An. quadriannnulatus* SANGWE, *An. arabiensis* Dongola, and *An. merus* MAF. Chromosomes were obtained from imaginal discs except for *An. merus* chromosomes hybridized to AgY53B, which were obtained from testes. The 18S rDNA probe (red signal) was used in all experiments except with *zanzibar*. Chromosomes were counterstained with DAPI (blue). The scale bar of 2 μm applies to all images.

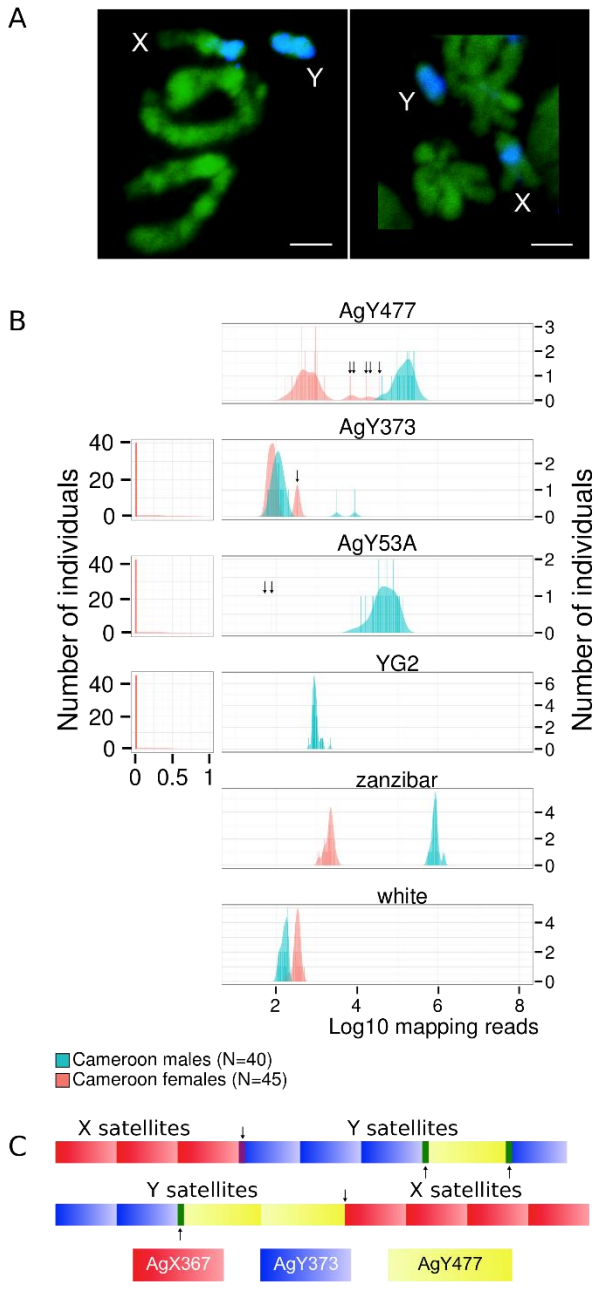

Fig. 5. The *An. gambiae* X and Y chromosomes are not genetically isolated. A) Painting of prometaphase (left panel) and metaphase (right panel) chromosomes from male larval imaginal discs of the *An. gambiae* Pimperena strain with a probe generated from microdissected Y chromosomes (labeled blue by the WGA3 kit with dNTP-Cy3). Chromosomes are counterstained with YOYO-1 (green). Scale bar: 2 μm. B) Histograms showing the $\log_{10}$ number of read

alignments from 40 individual *An. gambiae* males (blue) and 45 females (pink) from the Cameroon population, to satellite AgY477, AgY373, and AgY53A monomers, compared to numbers of reads from these sources aligning to *YG2* (exclusively Y-linked), *zanzibar* (heavily Y-biased), and the *white* gene (single-copy and X-linked). Arrows represent exceptionally high abundance in certain female *An. gambiae* individuals where X/Y recombination likely occurred. C) Two examples of single PacBio reads (pacbio_7224704_1 and pacbio_5551309_1) where predominantly X-linked (AgY367; shown iin red) and predominantly Y-linked (AgY373, shown in blue; AgY477, shown in yellow) satellites occur in the same PacBio read. Black arrows indicate the junction between the predominantly X-linked and predominantly Y-linked satellites. The purple and orange boxes indicate inferred recombination, between AgX367-AgY373 and AgX367-AgY477, respectively. Green boxes indicate recombination between AgY477-AgY373.

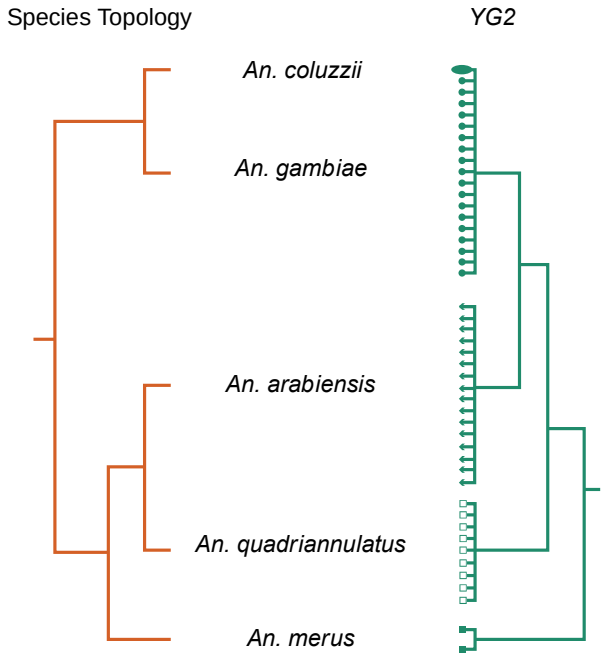

Fig. 6. Phylogeny inferred from a candidate male-determining gene on the Y chromosome, *YG2*, differs from the species branching order. Species topology (32) of five members of the *An. gambiae* complex (red branches) compared to a maximum likelihood phylogeny inferred from a Y chromosome-specific region of the *YG2* gene (green branches) sequenced from male *An. coluzzii* (ellipse), *An. gambiae* (filled circles), *An. arabiensis* (triangles), *An. quadriannulatus* (open squares) and *An. merus* (filled squares). Samples were drawn from colonies and natural populations (see text). The *YG2* tree was rooted at the midpoint; all nodes are supported by ≥ 95% bootstrap replicates. The topological disagreement involves *An. arabiensis;* in the species topology *An. arabiensis* is sister to *An. quadriannulatus*, while the *YG2* topology indicates a sister group relationship of *An. arabiensis* and *An. gambiae +An. coluzzii*.

Supporting Information Appendix:

Supplementary Text S1-S8
Figs. S1-S20
Tables S1-S22
References (74-105)

Chapter 5: Conclusions and future directions

5.1 The future of the CQ method

The CQ method has proven to be a reliable and robust method to identify Y chromosome sequences and has fared well compared to other approaches to identify Y chromosome sequences. The conceptual simplicity of the CQ method has proven to be one of its greatest advantages. Due to projects like the insect 5,000 genomes project (i5k) there are abundant opportunities in the future to use the CQ method to identify novel Y chromosome genes. For example, the i5k pilot project at Baylor sequenced males and females separately for stinkbugs and the sheep blowfly. We hope in the future more genome projects sequence males and females separately so that the same data used for genome assembly could also be used for CQ analysis.

How long into the future the CQ method will remain relevant? If the quality of genome assemblies suddenly increases to where chromosome-based assemblies are common, the CQ method would be irrelevant. Genomes assembled using the standard Illumina technology of today almost never reach that level of completion. Interestingly, mosquito genome assemblies have decreased in quality over time. The *Anopheles gambiae* genome, published in 2002, remains either the best or one of the best mosquito genome assemblies because most scaffolds are ordered on chromosomes and the assembly contains millions of base pairs of heterochromatic sequence, features that are quite rare in modern Illumina-based assemblies (*1*, *2*). The *Aedes aegypti* genome has better assembly metrics than the *An. gambiae* genome, but has a large number of misassembles (*3*–*5*). The recent publication of 16 *Anopheles* genomes continues the trend of declining quality (*6*, *7*). Several genomes of wild-caught mosquitoes like *An. christyi* and *An. epirotictus* have much smaller N50 scaffold sizes than *An. gambiae*. Finally, a recent paper on the genome of *A. albopictus* provided an “assembly” with an N50 scaffold size of 1,105

kb (*8*). As long as Illumina-based genome assemblies are being produced, the CQ method is likely to remain relevant.

The long, inaccurate reads produced by PacBio sequencing are unlikely to solve the problem of Y assembly. Our attempt to assemble the *An. gambiae* Y with PacBio sequencing failed completely. The assembly of the PacBio reads simply could not be used due to a large number of chimeric contigs combining Y-linked and clearly non-Y-linked sequences. The promised super-long reads of nanopore sequencing may finally lead to good Y assemblies, but initial analysis of the data produced by the Oxford Nanopore MinION has been less promising than initially hoped (*9*). Even if nanopore sequencing lives up to its high expectations, single-contig assemblies of extremely long tandemly-repeated regions like the zanzibar amplified region are highly unlikely. Clearly, in the near-term the CQ method will be relevant in identifying Y sequences in the exceptionally fragmented genome assemblies currently being generated.

5.2 *Nix* future directions

We demonstrated that *Nix* is required for male development and sufficient to initiate testes development. The next logical step is to test whether *Nix* is sufficient for fertile and competitive males. Constructs with *Nix* driven by its native promoter will be integrated into the genomes of *A. aegypti* and the resulting transgenic individuals will be observed for complete masculinization, fertility, and competitiveness. An interesting experiment would be to inject this same construct containing *Nix* into other *Aedes* species to see if *A. aegypti Nix* can initiate male development in these species. If positive results are seen in different *Aedes* species, the same experiment could even be performed in *Culex* mosquitoes.

We also want to explore how widespread *Nix* is in mosquitoes using bioinformatics. We are planning to sequence more species of *Aedes* mosquitoes to see if *Nix* orthologs can be identified. We attempted to find an ortholog of *Nix* in *Culex* mosquitoes, but failed to find any similar sequences. It is possible that *Nix* is present in *Culex* mosquitoes, but is so far diverged that it is beyond the limit of detection by similarity-based methods. Because *Culex* mosquitoes also have homomorphic sex chromosomes with presumably no dosage compensation, experiments to test the function of a male-determining factor should be comparatively easier compared to *Anopheles*. Therefore, I advocate sequencing male and females from multiple strains and species of *Culex* mosquitoes to perform CQ analysis in an attempt to identify the *Culex* M factor. Other mosquitoes that have homomorphic sex chromosomes like *Toxorhynchites* are enticing targets to identify if a *Nix* ortholog or other gene acts as the male-determining factor.

5.3 The importance of *myo-sex*

Sexually-antagonistic genes are often found on a differentiated sex chromosome (Y or W) because if they are only present in the sex they benefit their sexually-antagonistic effects are ameliorated (*10*). Due to its tight linkage to *Nix,* coupled with its extremely high expression level, we suspect that *myo-sex* plays an important male-specific role. Understanding whether *myo-sex* has an important male-specific role is crucial if we are to successfully use *Nix* as a tool to convert deadly female mosquitoes into harmless males. Supporting its potential male-specific importance, *myo-sex* was found to have a CQ near zero in two additional samples – a Thai strain of *A. aegypti* and a strain of *A. aegypti* formosus (Unpublished data).

The only negative mark against the potential male-specific function of *myo-sex* is that it can recombine away from the M locus. *Myo-sex* probably recombined from the M chromosome to m chromosome via illegitimate recombination. Because there is typically no *myo-sex* on the m

chromosome, the most likely scenario is that repeats similar to both chromosomes initiated an unequal recombination event that took *myo-sex* along with it. Even though *myo-sex* can be separated from *Nix*, it does not mean it is not important. First, we did not observe what happened to the males that lost *myo-sex*. Second, in females where *myo-sex* was present, its expression was highly down-regulated which could indicate sexual antagonism (*11*). Third, even *Sry*, the mammalian male-determining factor, can be unlinked from the Y, generally through recombination with the X (*12*). Thus, there is no reason to think that *myo-sex* is not highly important. I strongly advocate testing the function of *myo-sex* with CRISPR/Cas9.

5.4 The *Aedes* M chromosome

Chapters 2 and 3 lay out the first sequence-based knowledge of the *Aedes* M locus and lay the groundwork to further understand the evolution of homomorphic sex chromosomes. Due to the space restrictions of Science, we did not mention the evolutionary implications of the presence of *Nix* in the M locus of both *A. aegypti* and *A. albopictus*. The presence of *Nix* in the M locus of both these species indicates that the *Aedes* M locus is at the very least 40-100 million years old, the divergence time between *A. aegypti* and *A. albopictus*. We established that there is a differentiated region between the M and m chromosomes in *A. aegypti* containing *Nix* and most of the time *myo-sex*. While it is possible for *myo-sex* to recombine away, the smallest possible difference between the M and m chromosomes is the *Nix* genomic DNA due to its identity as the male-determining factor. Though strange, it is possible that a recombination event could move *Nix* from the M to m chromosome. However, because the male-determining potential of *Nix* defines the M chromosome, the old M chromosome without *Nix* will become an m chromosome and m chromosome that gained *Nix* the M chromosome. Due to the presence of a "non-recombining" region at least 40 million years old, the *Aedes* M chromosome is by definition a Y

chromosome where the male-specific (differentiated) region is small compared to the pseudoautosomal (non-differentiated) regions.

5.5 The *Anopheles gambiae* Y chromosome

Chapter 4 could essentially be called "The *Anopheles gambiae* Y chromosome project". The goal of this project was to sequence and assemble the Y chromosome of *An. gambiae*. Even though the *An. gambiae* genome was published in 2002, the Y chromosome was completely absent from this assembly. Further efforts only characterized approximately 180 kb of the Y chromosome in the intervening time (*13–15*). We view the Y chromosome as a big missing piece of the *An. gambiae* genome. First of all, the Y chromosome represents somewhere around 10 percent of the genome (*1*, *13*, *15*). Second, the *An. gambiae* Y chromosome is responsible for the initiation of male sex determination (*16*). Third, the *An. gambiae* Y chromosome seems to be modulate mating behaviors (*17*). Fourth, the Y chromosome has been observed to be incredibly polymorphic between different populations and individuals of *An. gambiae* (*18*).

The results described in Chapter 4 are useful in answering several basic questions about the *An. gambiae* Y chromosome. First, we identified the sequences that should constitute the majority of the non-recombining region of the *An. gambiae* Y chromosome. We found that four satellites and a tandemly-repeating transposable element and the subsequent insertions into this element represent more than 90 percent of known *An. gambiae* Y sequences. We were expecting to find a large number of somewhat repetitive repeats, but instead we found a very small number of highly repetitive repeats. Second, while we were unable to assemble the *An. gambiae* Y into a single chromosome or scaffolds spanning millions of base pairs, we were able to fulfill one major goal of a genome assembly - understanding the organization of these repeats on the Y chromosome. Third, through analysis of several members of the *An. gambiae* complex, we found

that only one gene, *gYG2* which I previously discovered (*19*), was conserved throughout all members of the *An. gambiae* complex. Combined with the fact that *gYG2* is expressed in 2-4 hour embryos (*19*), these results strongly suggest that *gYG2* is the most likely candidate to be the *An. gambiae* male-determining factor. Fourth, from analysis of ~80 individual genomes from the 1000 *Anopheles gambiae* genomes project, we identified a class of satellites that varied wildly in abundance between individuals that could help to explain some of the variation observed between individual *An. gambiae*. Fifth, we further solidified the relationship between the *An. gambiae* X and Y strongly suggesting that the *An. gambiae* Y originated as an X chromosome that has subsequently degenerated. The extremely abundant X-linked satellite AgX367 appears to share a common origin with the two most abundant Y-linked satellites AgY373 and AgY477 (*14*). We observed recombination between these satellite arrays finding AgY477 and AgY373 monomers in AgX367 arrays and vice versa. Therefore, these regions could serve as a "pseudo-pseudoautosomal" region that mediates pairing without the presence of a canonical pseutoautosomal region. However, a pseudoautosomal region 100 percent identical between the X and Y could not be detected with the CQ method, so we can't rule out this possibility. In *An. merus* where the long arms of the X and Y appear homologous in banding pattern, we found that the satellite AgY53B hybridized to both the X and Y in nearly equal amounts. This result is not likely due to cross-hybridization of a divergent satellite because with the strict alignment-parameters required by the CQ method, AgY53B had a CQ of ~1 in *An. merus*. These two results serve as the first molecular evidence of similarity between the X and Y chromosomes in *An. gambiae* and further supports the hypothesis that the *An. gambiae* Y is a degenerate X chromosome.

During the course of the research leading to Chapter 4, we exhaustively looked for a pseudoautosomal region in the assembled *An. gambiae* X chromosome. X chromosome sequences generally have a CQ around 2 while autosomal sequences generally have a CQ around 1 (*19*). A pseudoautosomal region of the X chromosome should also have a CQ around 1 because there should be two copies in both males and females. However, non-pseudoautosomal sequences on the X can also have CQs around 1 if they are highly repetitive with many more copies on the autosomes than the X. One of the best features of the *An. gambiae* genome assembly is that it extends millions of base pairs into heterochromatin (*2*). The assembled *An. gambiae* X chromosome is ~24 Mb (*1*). A heterochromatic region of the X extends from approximately 20 Mb to the end of the assembly at 24 Mb (*2*). This heterochromatic region is unsurprisingly much more repetitive than the rest of the euchromatic X chromosome. Therefore, when calculating CQs across the X chromosome, this region tends to have a depressed CQ because of the abundance of repeats in this region. Thus, one hypothesis could be that this region could be a pseudoautosomal region of the X or at the very least maintains some similarity with the Y. We have completely disproven this hypothesis. We found that the depression in CQ was mediated by highly repetitive sequences and a misassembly. This region contains a 400 kb scaffold of autosomal origin spuriously included in the X chromosome assembly of *An. gambiae*. This scaffolds contains genes: *AGAP001056*, *AGAP001057*, and *AGAP001058* which are all autosomal (unpublished data). After further inspection, we found no direct evidence of similar sequences between the X and Y throughout the entire 20-24 Mb region of the *An. gambiae* X. Therefore, we conclude that there is not a special relationship between the X and Y in the 20-24 Mb region of the *An. gambiae* X.

5.6 Final conclusions

The CQ method has enabled unparalleled studies of the gene and sequence content of the *Anopheles* Y chromosome and *Aedes* M locus. I hope it can be used in the future to discover more male-determining (or female-determining) genes and identify Y/W/M locus sequences from many more species.

5.7: References

1. R. A. Holt *et al.*, The genome sequence of the malaria mosquito Anopheles gambiae. *Science (80-.).* **298**, 129–149 (2002).
2. M. V Sharakhova *et al.*, Genome mapping and characterization of the Anopheles gambiae heterochromatin. *BMC Genomics*. **11**, 1–17 (2010).
3. P. Juneja *et al.*, Assembly of the genome of the disease vector Aedes aegypti onto a genetic linkage map allows mapping of genes affecting disease transmission. *PLoS Negl. Trop. Dis.* **8**, e2652 (2014).
4. V. A. Timoshevskiy, D. W. Severson, W. C. Black, I. V Sharakhov, M. V Sharakhova, An Integrated Linkage, Chromosome, and Genome Map for the Yellow Fever Mosquito Aedes aegypti. *PLoS Negl. Trop. Dis.* **7**, e2052 (2013).
5. V. Nene *et al.*, Genome sequence of Aedes aegypti, a major arbovirus vector. *Science (80-.).* **316**, 1718–1723 (2007).
6. D. E. Neafsey *et al.*, Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. *Sci. .* **347** (2015), doi:10.1126/science.1258522.
7. D. E. Neafsey *et al.*, The evolution of the Anopheles 16 genomes project. *G3*. **3**, 1191–1194 (2013).
8. V. Dritsou *et al.*, A draft genome sequence of an invasive mosquito: an Italian Aedes albopictus. *Pathog. Glob. Health*. **109**, 207–220 (2015).
9. S. Goodwin *et al.*, Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome. *bioRxiv* (2015) (available at http://biorxiv.org/content/early/2015/01/06/013490.abstract).
10. B. Vicoso, V. B. Kaiser, D. Bachtrog, Sex-biased gene expression at homomorphic sex

chromosomes in emus and its implication for sex chromosome evolution. *Proc. Natl. Acad. Sci.* **110**, 6453–6458 (2013).

11. A. B. Hall *et al.*, 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** , 179–191 (2014).
12. A. A. Rizvi, 46, XX man with SRY gene translocation: cytogenetic characteristics, clinical features and management. *Am. J. Med. Sci.* **335**, 307–309 (2008).
13. J. Krzywinski, D. R. Nusskern, M. K. Kern, N. J. Besansky, Isolation and characterization of Y chromosome sequences from the African malaria mosquito Anopheles gambiae. *Genetics*. **166**, 1291–1302 (2004).
14. J. Krzywinski, D. Sangaré, N. J. Besansky, Satellite DNA From the Y Chromosome of the Malaria Vector Anopheles gambiae. *Genetics*. **169**, 185–196 (2005).
15. J. Krzywinski, M. A. Chrystal, N. J. Besansky, Gene finding on the Y: fruitful strategy in Drosophila does not deliver in Anopheles. *Genetica*. **126**, 369–375 (2006).
16. R. H. Baker, R. K. Sakai, Triploids and male determination in the mosquito, Anopheles culicifacies. *J. Hered.* **70**, 345–346 (1979).
17. M. Fraccaro, L. Tiepolo, U. Laudani, A. Marchi, S. D. Jayakar, Y chromosome controls mating behaviour on Anopheles mosquitoes (1977).
18. S. Bonaccorsi, G. Santini, M. Gatti, S. Pimpinelli, M. Colluzzi, Intraspecific polymorphism of sex chromosome heterochromatin in two species of the Anopheles gambiae complex. *Chromosoma*. **76**, 57–64 (1980).
19. A. B. Hall *et al.*, Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. *BMC Genomics*. **14**, 273 (2013).

Appendix A: Finding Y genes in additional *Anopheles* species

The *Anopheles* 16 genomes project sequenced only female individuals (*1*, *2*). Therefore, the Y chromosome of these Anopheles species was completely ignored in the genome assemblies. The rationale for sequencing only females was to not halve the coverage of the X (*1*). However, we saw this as an opportunity to cheaply identify Y chromosome genes and potentially the M factor. We sequenced males (n=30) of *An. albimanus* STECLA, *An. minimus* MINIMUS1, *An. dirus* WRAIR2, and *An. atroparvus* EBRO using half a lane of Illumina sequencing each. All four species were obtained through BEI and were the same strains used for the *Anopheles* 16 genomes project. Female Illumina data from the Anopheles 16 genomes project for the CQ method was downloaded from the NCBI SRA.

New reference sequences were needed to perform the CQ method because the reference genomes are female-derived. We decided to use the RNA-seq from the *Anopheles* 16 genomes project. The RNA-seq was obtained from a single sample that was combined from samples spanning embryos to adults, including male embryos and adult males. Therefore, Y chromosome sequences should be included in the RNA-seq. We assembled this RNA-seq using Trinity (*3*). The assembly succeeded for *An. albimanus*, *An. minimus*, and *An. atroparvus*, but for an unknown reason failed for *An. dirus*. The reason for the failure was unknown after repeated attempts with Trinity failing at the butterfly stage.

We used the Trinity assemblies to perform CQ analysis. Because the male and female reads were not from the same source we needed to "balance" the number of male and female reads. Therefore, we performed the CQ method on the respective female-derived reference genomes and calculated the median CQ. Because sequencing was performed at different times with different methods, the percentage of alignments differed between the Anopheles 16

genomes samples and our own samples. Therefore, we added or subtracted female reads until the median CQ was close to one to even out this discrepancy. Therefore, autosome sequences have CQs near one, X sequences have CQs near two, and Y sequences have CQs near zero (*4*–*6*). If this step was not performed, the CQs would not fall within their normal ranges.

Next, we performed CQ analysis on the Trinity assemblies from the three species. Sequences with CQs less than 0.2 were identified from each sample. Note that all sequences identified here should be "genes" because they are from a RNA-seq de-novo assembly. 22, 253, and 591 candidate Y genes were identified from *An. albimanus, An. minimus,* and *An. atroparvus*, respectively. Note that the vast majority of these sequences are repetitive in nature. Interesting sequences in the NCBI NR database with alignments from potential Y genes in *An. albimanus* include: serpin and acyl-coa dehydrogenase. Interesting sequences in the NCBI NR database with alignments from potential Y genes in *An. minimus* include: dynein heavy chain (could be involved in sperm), apolipoprotein, D7, serine protease, and *AGAP012173*. See Table 1 for the names of the sequences.

Searches with blastn (word-size 7, evalue 1e-5) and tblastx against a database of known Y chromosome sequences from *An. stephensi* and *An. gambiae* did not identify any similarity (*7*). Furthermore, searches against Y "genes" from other species with the same blastn parameters here yielded no similarity. Therefore, either the repertoire of Y chromosome genes on the *Anopheles* Y changes rapidly on the timescale of *Anopheles* speciation, or have diverged sufficiently to preclude identification with similarity-based measures. This is consistent with our previous results that many Y genes like *sYG1*, *sYG3*, and *gYG5* have been recently duplicated (>85% nucleotide identity remains with the autosomal paralog) from an autosome to

the Y. The Y chromosome may be such a chaotic environment that genes Y genes do not tend to have a long lifespan.

The *GUY1* protein problem

The *GUY1*, *sYG2*, and *gYG2* proteins are roughly 56 amino acids long and fold into a helix loop helix secondary structures(*8*). Both *GUY1* and *sYG2* have been shown to kill females when transgenically inserted into the *An. stephensi* genome (unpublished results) strongly implicating that these genes are the male determining factor or involved in sex determination. *gYG2* is the only gene conserved in all members of the *An. gambiae* complex and it expressed at two hours after embryo deposition making it the most likely candidate for the M factor in Anopheles mosquitoes. The problem is that these proteins are so short that it is nearly impossible to find similarity between potentially orthologous sequences. There is slight detectable similarity between *GUY1* and *sYG2* with a high e-value but no detectable similarity between *GUY1-gYG2* or *sYG2-gYG2*. However, due to the similar structure one may hypothesize that *GUY1*/*sYG2* and *gYG2* may be orthologous. They also share other similar features in that they are the only "unique" genes on the respective *An. gambiae* and *An. stephensi* Y chromosomes, and both expressed in the early embryo. Because these genes are "unique" to the Y, they are likely the oldest known Y genes - younger genes likely retain similarity to their progenitor autosomal/X gene. However, a gene that moved to the Y instead of duplicated to the Y would also be "unique" to the Y. These similarities support the fact that there could be a conserved male-determining factor with a helix-loop-helix secondary structure. However, a conserved M factor would be extraordinarily hard to detect via similarity-based measures because of the length of the peptide sequences.

The phyre2 web server was used to predict the conformation of peptides encoded by potential Y chromosome genes (*9*). However, as the helix-loop-helix structure is so similar, many candidate Y genes as well as non-Y control sequences folded into helix-loop-helix structures. An expert in secondary structure may be able to perform a detailed analysis to narrow down the potential candidate *Guy1/sYG2/gYG2* homolog.

Another feature of *GUY1*/*sYG2*/*gYG2* is uniqueness to the Y which here is defined as no nucleotide or amino acid similarity to any other sequences on the autosomes or X using blastn or or bowtie2. We aligned the female Illumina reads using bowtie2 with the --sensitive-local (relaxed) parameter to all three Trinity assemblies. With these same parameters, *GUY1*/*sYG2*/*gYG2* have zero female alignments. Unfortunately, every sequence in all three trinity assembly had alignments from the female reads and were therefore not "unique" to the Y.

Problems and potential solutions

Male contamination in the female reads

Female mosquitoes store Y-chromosome containing sperm in their spermathecae (*10*). Therefore, sequencing mated females can result in low amounts of Y chromosome reads. It is unclear whether the *Anopheles* 16 genomes project sequenced virgin females, but our results indicate they did not. We downloaded the Anopheles 16 genomes female Illumina reads from *Anopheles stephensi* and aligned them to *GUY1*, a gene unique to the Y. We found alignments to *GUY1* indicating some sort of male contamination. There are not known Y chromosome genes in *An. albimanus*, *An. minimus*, *An. dirus*, and *An. atroparvus* so we can't test whether there was male-specific contamination in these supposedly female-specific reads. The performance of the CQ method is not negatively affected by small amounts of contamination. In these cases, unique

Y reads do not have CQ=0, instead CQ = 0.0X. However, when searching for "unique" Y chromosome genes like *GUY1*, this would eliminate all potential genes from contention. Unique Y chromosome genes can also be found using subtractive kmer-based approaches. However, any male contamination in the female samples will preclude subtraction-based approaches.

Genetic variation

Genetic variation between different strains of mosquitoes can causes the CQ method to output enormous numbers of false positive results. Here, the female mosquitoes and male mosquitoes should have been from the same ancestor, but the CQ method would perform better with mosquitoes from the same colony at the same time.

Bacterial contamination

CQ method works best when the mosquitoes are from the same colony and were raised together so that they share the same microbiome. To increase the percentage of mosquito-derived reads, larva could be raised in water with broad-spectrum antibiotics and the sugar water fed to adults could also contain antibiotics.

Low coverage of Y genes in RNA-seq

Important Y chromosome genes are not necessarily highly-expressed. *GUY1*, *sYG2*, and Nix all have less than 100 alignments from embryo-specific RNA-seq samples. In mixed RNA-seq samples, like those sequenced for the Anopheles 16 genomes project, it is highly unlikely there is enough RNA-seq reads for de-novo assembly of genes with low levels of expression.

Furthermore, even if important Y genes are identified from a different reference source, there may not be enough alignments from RNA-seq reads to indicate it is a gene.

Suggestions for future CQ analysis

Almost every case of sequencing male and females specifically for CQ analysis results in successful identification of Y chromosome sequences. On the contrary, piecing together male and female reads from different sources has rarely resulted in success. Therefore, the suggestion for future CQ analysis would be to sequence male and females from highly-inbred colonies or from paired-mating experiments, which have lived together as juveniles, but separated before sexual maturity (ensuring virgin females). The amount of sequence data needed is dependent on genome size, but a rough suggestion can be a single Illumina HiSeq 2000 lane. For identification of male-determining factors, RNA-seq of the early embryo is highly suggested.

Table 1

Species	Sequence	Alignment
An. albimanus	comp27_c0_seq1	acyl-coa dehydrogenase
An. albimanus	comp7283_c0_seq1	Serpin
An. minimus	comp14218_c0_seq1	*AGAP012173*
An. minimus	comp19605_c0_seq3	Dynein heavy chain
An. minimus	comp15128_c0_seq1	Apolipoprotein D
An. minimus	comp15874_c0_seq1	D7
An. minimus	comp11276_c0_seq1	Serine protease 14

An. albimanus sequences corresponding to table

>comp27_c0_seq1 len=174 path=[1:0-55 57:56-80 82:81-173]
GTCACTTTGTCCTTCTTTTGGCCCGTCTTGGGATCCGTCACCTCGGTCTGTGCGAACACCGTGAGTATGT
CGGCCAGACCACCTCCACTAATCCAGATCTTGGAACCATTCAACACGTAATGCGTACCGTCGGCCGATTT
CACCGCTCTGGTACGAATTGAACCGGCATCCGAA

>comp7283_c0_seq1 len=225 path=[203:0-224]
ACCACGAAAGTACAACGTGTTGATGATAAGCATAATCGTTTCCGGGCTTATGTGGGACAGGATGTCACCG
ATGCGACCATGGGTTTGATTGTAAACCCAACTGTTGATGGCTCGCACCGATTTCGCTTCGTTACCGACAA
AGTCCAGATACTGCAGCTCGCTCTTGTACAGTTTCTGAGCCAGCGCGTTGTAGCGTTGGTCAAAGTTGGT
ACCATCCTGGGCGAA

An. minimus sequences corresponding to table

>comp14218_c0_seq1
ATATATCACGTGCATCTGGCCATTGTGATTATGCTTCGGGCATACAAGATGCGTCAGCTGGACAACGAGT
TCGACTTTCTGGTCGCAACGGAAATGCCAACGGGTAGGAAATTGGATGACATACTGTACCATTACTCATC
GCCAAGACTACACACCGGCACTTTGTTCCTACAAGCGAAACATAAAAAGAACACGAGATCGAAGACAGAG
AGATATAACGAATTGCATCCTCATGAGGACTCCAAAACAAGAACAGATACAGAACATTTGAGGTTATTAT
ATCGATATGTTCATTTGGTGCAGTTCGTACCGAGTATCAACAAACGTACGATCGAAGAGACTGATGTAGC
TCGATGCGTATCGGTTCGTTGCTCGATAGAGTTGGAAAACCCCAAAAAAACCGAGCTGCCAATCAAGCAA
CTAGTTGGATTGTGCTTATTCCAAGAGAAGTTGAATCAATCAAATTTCGTGGTAGTTCTAGACGGGTATG
ATGAGATTGCACCATATTACAAATCTGTGGAGTTGAAGCGTTTCGCAGTACATGCTGGACTTTATGGGGT
GCAGACATAATATCGAAAGATATGCTCAAACATGCAAATTTTGACATACTTTGGTTGCAGTTAAATAAAT
ATTATTAAAACGTGAAAAGAAAGTAGTTTTTGGTGAAATGGAAAAAGAAATAAAAAAATTGTAAATTTCA
ATGTAAAATCGTAAATGCGATGTGGAATTCAGACTATTATTTGATCCCATATGTTGACACGGAATCATTT
TTGATTAACCATATTTTATCCGATAGAAAGAAGCATTCTATTTCACAATCTCTTCACTTGACGATTAAAG
TTAAAAAAAACCCGGTTTTTACAGAAGTTTATATTGGCTCAGAATTGCAACAGGATATTTTTAAGTGATA
GACGATTTTTTATTCAAATTCAACATTCAAATGAGGGGTCAAACAGCGAAAGTATAATCGAATGGATCAG
TTTCAGCCGAATAGTAATGTCATGCCGCTAAACAATCACTATGAACTTGTAGATTGTCGAAATATTTTAA
GATAGTAAATATAAGCTTATAAATTTAACTCGAAATAAATCATATGTATAAAAAATAAACATCCTGCT

>comp19605_c0_seq3

GTTTCGTGCGCTACTAGACATAGTGTAATCTATGATTTTATTTGGTTAAGTTCACTATTTTTGATTCACT
CACAACTCACATAGATCAGTAAACAACAACACTTTCTAAACTGGTGGATTCAAAACTGCATCACACACAT
AATTGTTTTGTTTTTATGGGCAACACTCACTGTGTTTAGAAGAAAACAATCTAATTGCACTACTAATTGT
ACAGCTTAAGAGCATGTTTTTGTGTGGATAAGGTTTCTAGAAAATTCTTATGCCCGCTAGCATTCATCTT
GTACCAATCGATGCTCTTAAGTTTTTTGATAAATTGTGCTGCAAGATCACTGCTCTAAGCAGCATGACCT
TTCAACATGATAGTGCACAAACGTGCAGATAAGGAAAGATGCGAAAATATCTCTTTGTTTATTTGATATC
ACACAGCAGTGCGACTCCACGCAACGTCCAGTGGCGCGGGTTAAAGTCCGTCTCGAAATCGATCGAACCG
ACGTACTTCTGATCGGTACGCTGCGGCTTCCGGTAGATGGGACATTCGTAGAGCCGCGGATCCTTACCAG
CAGTCGTGTTGATGGCGTAGATATAAATCACCGGCATCTGTTCGTACAGTATCTTTGGTTTCGATTCGAT
CAGCTTGCCACTGCGCCGGTCCAGCGATGCACCTTCCAGAAACAAACCGTACACGTAGACACCTTCCTGG
GGTGGATCGTGAATGTCTTCCTTGTTGTGGCGCGTAATTTGATTCTGCAACACAACCGAATCGAGGGCCC
AGCCTTTATGGGCTCGAGTTACCTCCTAGCGATGATGGAACCGTGCCCAAG

>comp15128_c0_seq1
TTTTTTTTTTTTTTTTTTTAATTTATAAATTTATTTTTATTTCAATCCTAAAACTAACATTTTATGATAC
GTTGGGATTCTTATTCATGGAATAAAATACACCGTAACATGTGCCATCGTTACTGACACGGTGGGTATGT
GGTAATTTCAGGATCAACTGAGCAGCAGAACTCTCCATCTTGAATGGTGGGACGAAGATCCGACTCGATC
AGGAACTCATCGACATACTGCTGCACCGCAGCCTCAGCCTGGGCGGATAGTTGGGGCGTACGGGACAATA
CCCAGGCACTTTCGGCACGCAGAGTGTCGGCCACTCCGAAGCAGCCCCAAACGACAGCATAGCTATCGTA
ATCTGTACCGAGTACCCAGTAGTTGGCCGAAACATCGATGGCACCGGCATCGAACGTAACGTTCAGCTTA
GCCTCGAGTGGAGACTCATCTGGGAAGGCAACCACTGCACGGCCAGTATCGGACGAACGGACCTGGGACG
GTGGTACCATCATCGAGTTGAAGACACGCACACTGCCATCATCGTTCAGCGAGTACTCGGCCGTAACGCA
TTCGCCACCACGCTGGAATACCTGCTCGTAGCGACGGATTTCGTACCACAGTCCCAGGTACCGGGCCACA
TCAAAGTTCACGTGCACCGGATAGTCCTGGCACTGGCCGGTCGAAACAACCTGTCCGCTGGCCAGCTGAA
CAAGGCCAAGCACGAACGCGGCTGCTAGCAATTGAATGGAGTTCATTTTGGAGGTTGTACTGTGAGTGGA
GACACT

>comp15874_c0_seq1
CGGTGAGATAGCATCATGTTGAAGAAACTCCTACTAAGCGTTGGACTGGTGTGGTGTCTGATCTCTTTGG
GACAGGCCCGAAAAGAATCAACGGTTGAGGAATGTGAGAAAAATATTGCCGCATCACTGAAAGATCGCGT
CTGTGAGCTGCGTCAATATACACCGGTTAGTACCGATGATATGGATAAGCACATGCAGTGTGTCCTGGAG
GTGGTCGGATTTGTAAACGACAATGGAGAAGTTAAGGAAAATGAGCTCCTGGGACTACTGCAGCGCGTTG
ATAGCAGTGTTCCCCATGCTACCAACATGAAGAAATGTGTAACCGAAGCATCCGGCGTGGGCAACGCAAA
GAAGGCAAACACCTTCTACACGTGCTTCTTGGGCACGAGTTCCTTGGACAGTTTCAAGTATGCGGTCGAC
TACAATGAGCTTTTGAGGGCCGGTAAGATGCAGCTAAGTGATCCATTCGATATGGGTCGTGTTTCGGCGT
TGATTAAGGAAATTGATGACGGTTTGTGTCAATAGAATGGAGGATCGTTCATTGTAATTTATATGGCGTT
ATGTGACAAACAGAAGTGAAATGTAGCATGAATGCAAACATGAAAA

>comp11276_c0_seq1
GCCAACGATACGATCCGACAGCTGAATGCCACATTGCGGTGATACGGGGAGGACATTGTTGGTGGTCTGT
TGTTCACCAAGATCGGTTACTGCACAGCATACCAAAGTACGGCGATCAATCTGCCCACAGCGGCTATTTG
CTAGAAACCTCGAATCATCAGGCGTTACGAGAAACTTGCTGTATATAGCCATCACTGGCGGACATTCTCG
AAAGGGAATACATTTACCCGGCTGTCCCAAAGGATTAATGCAGCTTTGTCCCTCTTCTAACGTTACAGCG
ACATCTACCACGGTTAGTACGCACAGCGACCACAGACAGCGATACAATTGACTTCTGTTCATTATTTTAC
CTGTTTACGGATCCGCGATCACAATTAGAATGACGACGGTGACACCACAACTGCCTTCTTTCGTCGGTTT

AAGTTAAACTGTTGGGGATCTGCGAGCAAAAAAAAAACCTAAAGATGCGTGTTTTACGAACTTGTTTGAC
AACACCAACACCAACTGCTACAGTGTGCGCAACGAGATGATGCTGATTTACCG

References

1. D. E. Neafsey *et al.*, The evolution of the Anopheles 16 genomes project. *G3*. **3**, 1191–1194 (2013).
2. D. E. Neafsey *et al.*, Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. *Sci.* . **347** (2015), doi:10.1126/science.1258522.
3. B. J. Haas *et al.*, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
4. A. B. Hall *et al.*, Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. *BMC Genomics*. **14**, 273 (2013).
5. A. B. Hall *et al.*, 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** , 179–191 (2014).
6. A. B. Hall *et al.*, A male-determining factor in the mosquito Aedes aegypti. *Science (80-.).* (2015), doi:10.1126/science.aaa2850.
7. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* . **25** , 3389–3402 (1997).
8. F. Criscione, Y. Qi, R. Saunders, B. Hall, Z. Tu, 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*. **22**, 433–441 (2013).
9. L. A. Kelley, M. J. E. Sternberg, Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4**, 363–371 (2009).
10. N. Becker, *Mosquitoes and Their Control* (Kluwer Academic / Plenum Publishers, 2003).