Application of Chromosome Mapping to Understanding Evolutionary History of *Anopheles* Species

Maryam Kamali

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

Entomology

Igor V. Sharakhov, Chair

Allan W. Dickerman

Richard D. Fell

Maria V. Sharakhova

Zhijian (Jake) Tu

8 May 2013

Blacksburg, VA

Keywords:

Anopheles spp., chromosomal and molecular phylogeny, microsatellites

Application of Chromosome Mapping to Understanding Evolutionary History of *Anopheles* species

Maryam Kamali

ABSTRACT

Malaria is the main cause of approximately one million deaths every year that mostly affect children in south of Sub-Saharan Africa. The *Anopheles gambiae* complex consists of seven morphologically indistinguishable sibling species. However, their behavior, ecological adaptations, vectorial capacity, and geographical distribution differ. Studying the phylogenetic relationships among the members of the complex is crucial to understanding the genomic changes that underlie evolving traits. These evolutionary changes can be related to the gain or loss of human blood choice or to other epidemiologically important traits. In order to understand the phylogenetic relationships and evolutionary history of the members of the *An. gambiae* complex, breakpoints of the 2Ro and 2Rp inversions in *An. merus* and their homologous sequence in the outgroup species were analyzed using fluorescent *in situ* hybridization (FISH), library screening, whole-genome mate-paired sequencing and bioinformatics analysis. Molecular phylogenies of breakpoint genes were constructed afterwards. In addition, multigene phylogenetic analyses of African malaria vectors were performed. Our findings revised the chromosomal phylogeny, and demonstrated the ancestry of 2Ro, $2R+^{p}$ and 2La arrangements. Our new chromosomal phylogeny strongly suggests that vectorial capacity evolved repeatedly in members of the *An. gambiae* complex, and the most important vector of malaria in the world, *An. gambiae*, is more closely related to ancestral species than was previously thought. Our molecular phylogeny data were in agreement with chromosomal phylogeny, indicating that the position of the genetic markers with respect to chromosomal inversion is important for interpretation of the

phylogenetic trees. Multigene phylogenetic analysis revealed that a malaria mosquito from humid savannah and degraded rainforest areas, *An. nili*, belongs to the basal clade and is more distantly related to other major African malaria vectors than was assumed previously. Finally, for the first time a physical map of 12 microsatellite markers for the Asian malaria vector *An. stephensi* was developed. Knowledge about the chromosomal position of microsatellites was shown to be important for a proper estimation of population genetic parameters. In conclusion, our study improved understanding of genetics and evolution of some of the major malaria vectors in Africa and Asia.

Dedication

I dedicate this dissertation to my parents, Karim Kamali and Shahrbanou Razizadeh.

Acknowledgements

I would like to take this opportunity to thank all the people who helped and inspired me during my studies in these years. First of all I would like to thank my advisor Dr. Igor Sharakhov. I have to say that it is not possible to properly thank him in a few sentences. I am truly grateful that I had this opportunity to do research under his mentorship. He always helped and encouraged me toward my goals. I thank him for all his patient, endless support and academic advice. He sets an example of a great scientist and mentor that I would like to follow.

I would also like to thank my committee members. I specifically want to thank Dr. Maria Sharakhova for all her kindness, help, support, and advice during my PhD. For me she is truly an example of a passionate scientist who enthusiastically cares about her student's research. I have learned a lot from her. I would like to thank her both for her academic help and all the kindness throughout these years. I would like to thank Dr. Allan Dickerman for his valuable advice and help on the molecular phylogeny section of my project. I thank Dr. Richard Fell for all his help and support. He is truly a great teacher and I have learned a lot about insects in his class. I would like to thank Dr. Zhijian Tu for all his help and guidance. Without his help and valuable contributions, some parts of my project wouldn't have been accomplished. I also want to thank Dr. Loke Kok, the department head, for all his support during these years. I really enjoyed all his advisement during the student meetings. I would also like to thank Vice President and Dean for Graduate Education, Dr. Karen DePauw, for her support during my graduate studies.

I would like to thank my lab members for being such a supportive group and for all the enjoyable moments we had over the years. I am so glad to get to know all of them. I thank Ashley Peery for her constant support, encouragement and kindness, Philip George for all his help, Atashi Sharma for her kindness and bringing energy to our lab, Fan Yang for his kind support, and our former

lab mate Ai Xia for her help. My special thanks to Lisa Moore, Vladimir Timoshevsky and Anastasia Naumenko. I would also like to thank members of the Adelman lab, especially my dear friend Azadeh Aryan, who has been a great support for me.

I thank our collaborators Elina Baricheva, Dmitrii Karagodin, Christophe Antonio-Nkondjio, Cyrille Ndo, Frederic Simard. The department of Entomology, Kathy Shelor, Sarah Kenley, Robin Williams, Sandra Gabbert. Many thanks to Dr. Sally Paulson, J. Reese Voshell and Stephen Hiner. Being a teaching assistant in their classes was a great opportunity for me.

I would like to thank my dear friends in Blacksburg, Zahra Mashhadi, Hojat Ghandi, Leyla Nazhandali, Masoud Agah, Somayesadat Badieyan and Reza Sohrabi. Moreover, I thank Reyhane Ojani, Hoda Koushyar, Nasibeh Azadehfard, and Fatemeh Hashemi. I feel so lucky to get to know each one of you and you were like a family for me during all these years. Thank you for all your kindness and support.

I would like to thank my family for their continuous support and encouragement. I thank my husband, Mohammad Saied Dehghani for his kind support during my studies. I appreciate all his help and encouragement; without his constant support this would have not been accomplished. He truly encouraged me and supported me during all our years together and I feel lucky to have him. I also thank my siblings Masoud and Kaivan and their families for their kind support. At the end I would like to thank my parents, Karim Kamali and Shahrbanou Razizadeh, for their endless sacrifice, support, kindness and encouragements. Without their help this could have not been possible. They always supported, inspired and encouraged me throughout my life. I dedicate this dissertation to them.

Attribution

Several colleagues aided in the writing and research behind two of my chapters presented as part of this dissertation. A brief description of their contributions is included here.

Chapter 2: A New Chromosomal Phylogeny Supports the Repeated Origin of Vectorial Capacity in Malaria Mosquitoes of the *Anopheles gambiae* Complex.

Chapter 2 was published in the PLoS Pathogens journal.

Igor V. Sharakhov, PhD (Institute of Cytology and Genetics, Novosibirsk, Russia) is currently a professor in Entomology at Virginia Tech. Dr. Sharakhov was a co-author on this paper, principal investigator for the grants supporting the research, helped with the experimental design, and provided editorial comments.

Ai Xia, PhD (Department of Entomology, Virginia Tech) is currently an associate professor in Entomology at Nanjing Agricultural University, China. Dr. Xia was a co-author on this paper, and helped with fluorescent *in situ* hybridization experiments.

Zhijan (Jake) Tu, PhD (Department of Entomology, University of Arizona) is currently a professor in Biochemistry at Virginia Tech. Dr. Tu was a co-author on this paper and helped with analyzing mate-paired sequencing data, novel transposable element detection, and provided *An. stephensi* genomic and BAC DNA sequences as well as editorial comments.

Chapter 5: An Integrated Chromosome Map of Microsatellite Markers and Inversion Breakpoints for an Asian Malaria Mosquito, *Anopheles stephensi*

Chapter 5 was published in the Journal of Heredity.

Igor V. Sharakhov, PhD (Institute of Cytology and Genetics, Novosibirsk, Russia) is currently a professor in Entomology at Virginia Tech. Dr. Sharakhov was a co-author on this paper, principal investigator for the grants supporting the research, helped with the experimental design, and provided editorial comments.

Maria V. Sharakhova, PhD (Institute of Cytology and Genetics, Novosibirsk, Russia) is currently a research scientist in Entomology at Virginia Tech. Dr. Sharakhova was a co-author on this paper, helped with the experiments, and provided editorial comments.

Zhijan (Jake) Tu, PhD (Department of Entomology, Arizona State University) is currently a professor in Biochemistry at Virginia Tech. Dr. Tu was a co-author on this paper and provided *An. stephensi* genomic sequences and editorial comments.

Elina Baricheva, PhD (Institute of Cytology and Genetics, Novosibirsk, Russia) is currently the head of the laboratory of evolutionary cell biology, Institute of Cytology and Genetics, the Siberian Branch of the Russian Academy of Sciences, Russia. Dr. Baricheva helped with cloning experiments.

Dmitrii Karagodin, MS (Novosibirsk State University, Novosibirsk, Russia) is currently a senior laboratory technician in laboratory of evolutionary cell biology, Institute of Cytology and Genetics, the Siberian Branch of the Russian Academy of Sciences, Russia. Dmitrii Karagodin helped with cloning experiments.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 3

Chapter 4

Supplementary

Chapter 5

LIST OF FIGURES

Chapter 1

Chapter 2

Supplementary

Chapter 3

Chapter 4

Chapter 5

CHAPTER 1 Introduction

Malaria

Malaria is one of the most dangerous infectious diseases in the world. About half of the world population (almost 3.3 billion people) is at risk of malaria. According to the World Health Organization report 2010, 81% of cases of malaria and 91% of deaths occur in Africa, and most of them are children under five years of age in sub Saharan Africa [1]. Malaria is transferred by the bite of a female mosquito of genus *Anopheles* and is caused by a parasite from the genus *Plasmodium*. There are five species of *Plasmodium* which affect humans including *P. falciparum, P. vivax, P. ovale, P. malariae* and *P. knowlesi*. A significant majority of malaria infections are caused by *P. falciparum* and *P. vivax*. Being more dominant in Africa, *P. falciparum*, is responsible for the majority of deaths; however *P. vivax* is more widespread and less dangerous. *P. ovale , P. malariae* and *P. knowlesi* are found in lower frequencies [1].

Anopheles

Mosquitoes belong to order Diptera and family Culicidae and constitute an ancient monophyletic group dating back to the Jurassic period (about 260 Million years ago). The family Culicidae consists of more than 4500 species. The ancestral species gave rise to two subfamilies, Culicinae, which includes the genera *Aedes* and *Culex* and Anophelinae, which includes *Anopheles*. Anopheline mosquitoes are the only species that can transmit human malaria [2]. According to the analysis of mitochondrial DNA, it has been estimated that the Anophelinae and Culicinae subfamilies diverged about 145-200 million years ago. The divergence time between subgenera

Anopheles, *Cellia* and *Anopheles* is estimated at 90-106 million years ago, and between lineages within the subgenus *Anopheles* it is estimated to have occurred 70-85 million years ago [3].

The majority of Anopheline species belong to the genus *Anopheles* [4]. About 500 species of *Anopheles* have been described, however only about 30 *Anopheles* species are important vectors of malaria and are involved in majority of malaria transmission [5,6,7,8].

In Sub-Saharan Africa, which has 90% of the world malaria cases, malaria is transmitted by five major vectors, *An. gambiae*, *An. arabiensis, An. funestus, An. nili* and *An. moucheti* which are highly anthropophilic [6,9,10]. The Asian mosquito *An. stephensi* is the major vector of malaria in the south of Iran and the Persian Gulf [11], as well as urban areas of India [12], and rural areas of Pakistan and East Afghanistan [13]. Three morpho-ecological variants have been identified within *An. stephensi* populations—*type, intermediate,* and *mysorensis*—which can be identified by the number of ridges on the egg [14,15,16].

The *An. gambiae* complex

Complexes of sibling species are common among different organisms, including both vertebrates and non vertebrates [18,19]. They are found in fungi [20], Hymenoptera [21], fish [22], tree frogs [23], and *Anopheles* [24]. Sibling species of *Anopheles* are morphologically indistinguishable and are closely related to each other, however they have different capacities to transfer human malaria [25].

The *Anopheles gambiae* complex belongs to series Pyretophorus of subgenus *Cellia* and consists of seven sibling African malaria mosquito species that differ remarkably in ecological adaptation, geographical distribution, and host-seeking behavior. *Anopheles gambiae* sensu stricto Giles 1902, is one of the major vectors of malaria in Africa. It is highly anthropophilic,

prefers human blood and is naturally endophilic (tendency to rest inside the house after they take the blood) and endophagic (feeds inside). Oviposition takes placed on the temporal fresh water pools or water reservoirs made by humans. *Anopheles gambiae* has a wide distribution in Sub-Saharan Africa and is dominant species in humid savannas [26,27]. There is a zoophilic population of *An. gambiae* in the island of S˜ao Tom´e in the Gulf of Guinea in West Africa which differs from other populations of *An. gambiae* by being exophagic and endophilic [28]. Zoophilic populations of *An. gambiae* were also found in Madagascar [29].

Anopheles arabiensis Patton 1905, is also an anthropophilic species and is a major vector of malaria in Sub-Saharan Africa and has a wide distribution. It is found in sympatry with *An. gambiae* and breeds in temporal fresh water reservoirs. *Anopheles arabiensis* is mostly dominant in arid savanna and steppes. It is an endophilic and exophagic species [10,24] , however it rests outdoors more commonly than *An. gambiae* [30]. Although anthropophilic, *An. arabiensis* is considered a highly opportunistic species and even though it prefers to take blood from human, they shift to take blood from animals when the density of domestic animals increases [27]. In contrast to mainland Africa, there is a zoophilic population of *An. arabiensis* in Madagscar with the preference for cattle odors [29].

Anopheles merus Donitz 1902, and *An. melas* Theobald, 1903 are minor vectors of malaria in Africa and they breed in brackish water and have a narrow distribution. *Anopheles merus* is distributed in the east coast and *An. melas* on west coast of Africa [31]. They can complete their life cycle in fresh water, however their distribution is limited to coastal area, because they cannot compete with *An. gambiae* in the mainland [8]. The role of *An. merus* in transferring malaria in Madagascar has been reported [32].

An. bwambae White, 1985 is also a minor vector and is restricted to mineral water breeding sites. It has a narrow distribution and is found in thermal springs in the Semliki forest in Uganda [31]. This species can not complete its life cycle in fresh water and avoids oviposion in fresh water [8,33].

Anopheles merus, *An. melas* and *An. bwambae* are exophagic and bite humans in the absence of alternative hosts [31]. *Anopheles quadriannulatus* A Theobald 1911, and *An. quadriannulatus* B Hunt, Coetzee and Fettene 1998, are allopatric species which were named according to crossing experiments [34]. They breed in fresh water, are zoophilic, feed on animal blood and have a large number of hosts. *Anopheles quadriannulatus* A is found in the south of Africa and is solely exophilic, however, *An. quadriannulatus* B is only limited to Ethiopia and is endophilic in high altitudes in Ethiopia [24,31]. They are not natural malaria vectors, but are to some degree susceptible to *Plasmodium* infections [35,36,37]. *Anopheles quadriannulatus'* innate immune system has an important role in natural refractoriness to malaria. Silencing LRIM1, LRIM2, genes which encode leucine–rich repeat proteins, and TEP1 genes, which encode thioester-containing protein, will stop *Plasmodium berghei* melanization and turns *An. quadriannulatus* A into a vector [38]. According to chromosomal, cross-mating and molecular evidence, Ethiopian species of *An. quadriannulatus* B has been recently named *An. amharicus* Hunt, Wilkerson & Coetzee sp. n.[39].

All of the sibling species of *An. gambiae* complex may cross spontaneously in the laboratory, however the rate and frequency of hybrids are very low in nature (less than 0.1%) [40], and hybrids are rarely detected (0.02-0.76%) [41,42,43]. This indicates reproductive isolation among the species [40].

Chromosomal inversions in the *Anopheles gambiae* complex

Members of the *An. gambiae* complex are morphologically indistinguishable; however they can be distinguished base on fixed chromosomal differences. *Anopheles* mosquitoes have the mitotic karyotype of one pair of acrocentric sex chromosomes X (X and Y in males) and two pairs of submetacentric autosomes chromosome 2 and 3 [24,44]. Polytene chromosomes consist of five arms. In the *An. gambiae* complex, paracentric inversions are very abundant. For example, more than 120 polymorphic inversions have been detected in natural populations [24,31]. Polymorphic inversions segregate within a species population, however fixed inversions differentiate species.

There are 32 common polymorphic inversions in the *An. gambiae* complex [31]. Among members of the *An. gambiae* complex, *An. gambiae* and *An. arabiensis* have the highest number of polymorphic inversions and are also widely distributed in Sub-Saharn Africa [31]. *An. melas* which breeds in brackish water also has a high inversion polymorphism [45]. However, *An. bwambae* and *An. quadriannulatus* A, have few inversions, and no inversion polymorphism is observed in *An. merus* or *An. quadriannulatus* B [46].

There are ten fixed inversions in the *An. gambiae* complex. As compared with the standard chromosomal type, in the paracentric chromosomal inversion, a part of the chromosome has a reverse orientation. These inversions can be recognized based on the inverted banding pattern in the giant polytene chromosomes of *Anopheles* species [31]. The standard arrangements in all chromosomal arms are X+, 2L+, 2R+, 3L+, 3R+. Five inversions, Xbcd, Xag, are on the X chromosome; three on the right arm of chromosome 2: 2Ro, 2Rp, 2Rm; one on the left arm of chromosome 2: 2La; and one on the left arm of third chromosome: 3La.

Figure 1 shows the distribution of ten fixed inversions on chromosome arms of the *An. gambiae* complex [24]. The 2La inversion is fixed in *An. arabiensis* and *An. merus* and is polymorphic in *An. gambiae* [24,47]. 2Ro and 2Rp inversions are specific to *An. merus* and *An. gambiae* and *An. merus* share the Xag inversions. *An. arabiensis* possess Xbcd inversions. Finally, *An. bwambae* and *An. melas* share the 3La inversion and *An. melas* has an additional 2Rm inversion [24].

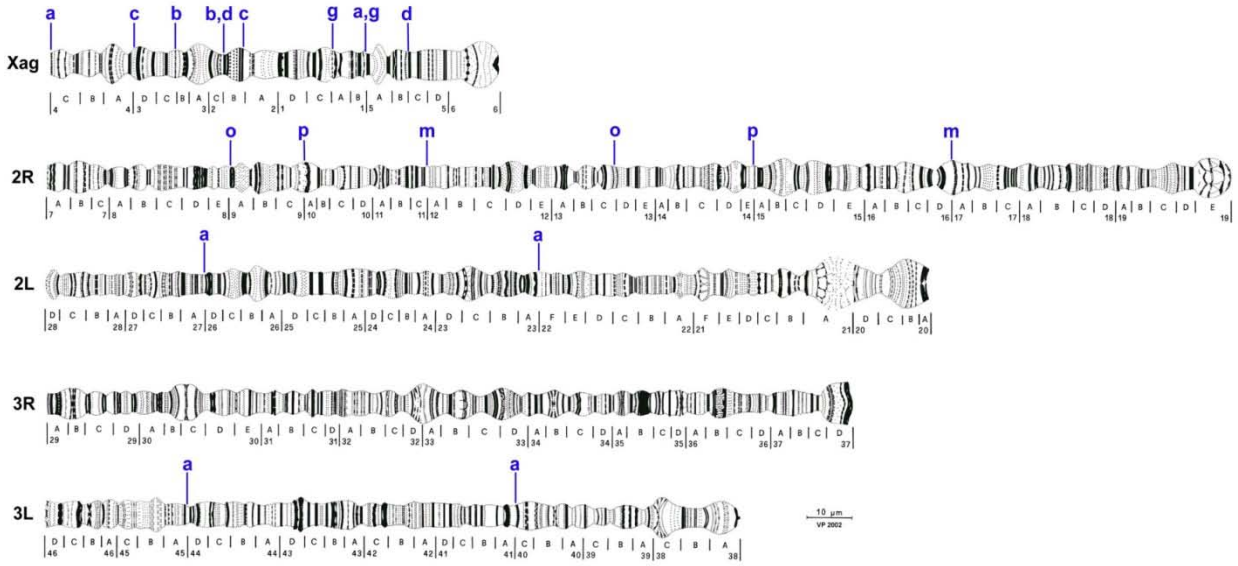

Figure 1. Distribution of 10 fixed inversion on polytene chromosome arms of *An. gambiae* complex [24, 65].

Significance of chromosomal inversions

Inversions occur when a chromosome breaks at two points and the segment is reinserted in reversed orientation. If the inverted region contains the centromere, it is called pericentric inversion. On the other hand, inversions that do not include the centromere are called paracentric inversions. The first evidence of chromosomal inversions was published by Alfred Sturtevant in 1921 [48]. Inversions have an important role in speciation and isolating between species and populations. Inversions can suppress recombination in heterozygotes and serve as a postzygotic

barrier by reducing the heteozygotes fecundity [49]. They can link and protect co- adapted alleles and protect them from recombination. For example, species specific loci that are mapped inside the inversions are suppressed from recombination [50,51,52]. Inversions can also be spread in natural populations because they protect favorable alleles from recombination. As a result inversions are involved in some important ecological adaptations, for example, an increase in tolerance to aridity [31]. Inversion 2La in *An. gambiae* larvae is associated with an increase in thermal resistance [53], and an increase in drought resistance in adults [54,55]. There is also a latitudinal cline in the frequency of inversions in *An. gambiae* and *An. arabiensis*. For example in Nigeria inversions are fixed in most northern dry regions, and standard arrangements are abundant in the humid climate of the south [31]. A study on molecular forms of *An. gambiae,* M and S forms and *An. arabiensis* in Cameroon revealed that inversions are involved in ecological specialization [56]. In *An. arabiensis* inversions are strongly associated with ecological clines. The range of chromosomal inversions 2Ra varies from zero in forest area, gradually increases to 40% in Geinea Savannah, and reaches 80% in the Sahel Savanas [31]. Moreover, inversion 2La is also associated with susceptibility to *Plasmodium* and significant differences have been observed among infection rates of *An. gambiae* populations which carry this inversion [57]. Two members of *An. gambiae* complex, *An. gambiae* and *An. arabiensis,* have the highest number of polymorphic inversions and occupy the widest geographical distribution and ecological conditions, especially as they relate to climate in Africa, moreover, they both carry the fixed 2La inversion which could help them to adapt to vast dry environments [31,55,58].

Chromosomal phylogeny in the *An. gambiae* complex

In 1936, Sturtevant and Dobzhansky, studied for the first time the inversion of the third chromosome of a wild race of *Drosophila pseudoobccura* for the purpose of finding the

historical relationship between populations [59]. It is possible to construct the phylogenetic relationship based on chromosomal inversions on those insects that have polytene chromosomes, and the phylogenetic relationship can be inferred from analyzing the distribution of fixed overlapping inversions [31,60,61]. This approach is based on two facts; fist, species-specific inversions do not introgress [62]; and second, inversions are mostly monophyletic in spite of the rare occurrence of breakpoint reuse [54]. As a result, whenever inversions happen, they are passed to the next generations and all the individuals that have the same inversion today, inherited it from the same ancestor [63]. Phylogenetic relationships in the *An. gambiae* complex can be inferred from distribution of 10 fixed inversions [24]. It was assumed that if a specific chromosomal arrangement is present in majority of species, that arrangement is ancestral. However, it was later shown that even rare arrangements can be ancestral only if that arrangement is also present in an outgroup species [44,47]. Therefore, genes found at inversion breakpoint regions in ingroup and outgroup species are expected to be in their ancestral order [64].

For a long time, *An. quadriannulatus* was considered ancestral because of its central position relative to other species, containing all the standard chromosomal arrangements, and some less specialized traits such as zoophily and a large number of hosts [24,31]. Based on this ancestry, *An. quadriannulatus* with standard arrangements gave rise to *An. arabiensis*, *An. gambiae* and *An. bwambae* by acquiring Xbcd and 2La, Xag, and 3La inversions (Figure 2). There is an introgression of 2La from *An. arabiensis* to *An. gambiae*. Later, *An. gambiae* gave rise to *An. merus* by acquiring the 2Rop inversion. Finally, *An. bwambae* gained a 2Rm inversion which lead to *An. melas* [24,31].

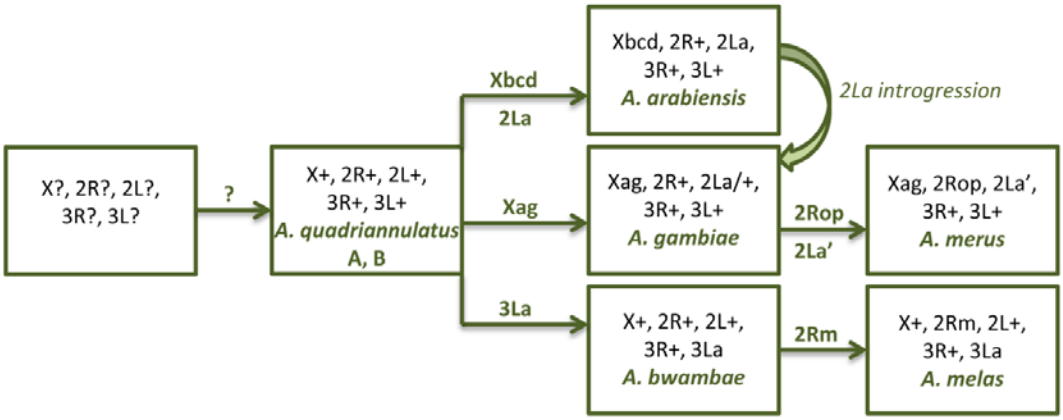

Figure 2. Phylogenetic relationship in *An. gambiae* complex based on ancestry of *An. quadriannulatus*
[24,31,65].

Later, *An. arabiensis* was considered the most likely ancestral species, having originated in Middle East and reaching to Africa via the Arabian Peninsula [41,42,43]. This assumption was based on the presence of 2La inversion in *An. arabiensis*, that was also identified cytologically in an outgroup species *An. subpictus*. Moreover, *An. arabiensis* is the only member of *An. gambiae* complex that is present in the horn of Africa and in the Arabian Peninsula and might have been zoophilic, acquireing the anthropophilic trait later [66]. In this scenario, *An. arabiensis* gave rise to *An. quadriannulatus* by acquiring standard $X+^{bcd}$ and $2L+^{a}$ arrangements. Consequently, *An. quadriannulatus* gave rise to *An. gambiae* by obtaining the Xag inversion and to *An. bwambae* by gaining the 3La inversion. There is also an introgression of 2La from *An. arabiensis* to *An. gambiae* [63]. Lastly, *An. gambiae* gave rise to *An. merus* by acquiring 2Rop, and an independent 2La inversion. *An. bwambae* gave rise to *An. melas* by obtaining the 2Rm inversion (Figure 3). It was suggested that 2La inversion has a multiple origin and *An. merus* has a molecularly independent 2La` inversion in comparison to *An. gambiae* and *An. arabiensis* [63].

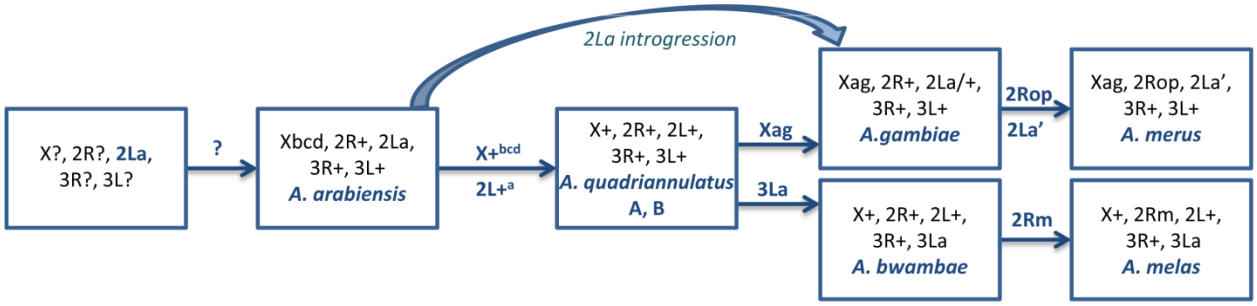

Figure 3. Phylogenetic relationship in *An. gambiae* complex based on ancestry of *An. arabiensis* [41,42,43,65].

The origin of 2La inversion and its monophyly was tested afterwards by detailed analyasis of breakpoint structure and it was revealed that *An. merus*, *An. gambiae* and *An. arabiensis* share the same 2La arrangement, and this inversion is monophyletic, has a unique origin, and is considered ancestral [47]. As a result, there is shift in phylogeny from nonvector, *An. quadriannulatus*, to vectors and either of species *An. gambiae*, *An. arabiensis* or *An. merus* which share this inversion and could be closer to the ancestral species. However the species closer to ancestral species was still unknown. Previously, a physical map of outgroup species, *An. funestus* and *An. stephensi*, were used to determine ancestral chromosome arrangements in the *An. gambiae* complex [44]. However the evolutionary history and phylogenetic relationship among the members of the complex remained unclear. In the following study, we will elucidate the evolutionary history and phylogenetic relationship among members of the *An. gambiae* complex.

Molecular phylogeny in the *An. gambiae* complex

It is important to understand the phylogenetic relationship among the members of *An. gambiae* complex to identify the specific genomic changes associated with the human blood choice, breeding site preference, and variations in vector competence. Molecular markers and the sequenced genome of *An. gambiae* are available [67]. However, phylogenetic relationship among the members of the complex is not resolved yet. Constructing phylogeny based on molecular markers in *An. gambiae* compelx is complicated due to the high degree of genetic similarity among sibling species. For example, in a study on the pattern of mitochondrial variation between these species an extensive gene flow was observed between *An. gambiae* and *An. arabiensis* [68]. Moreover, in a study of molecular evolution of four antimalaria immune genes in *An.gambiae* complex, introgression and the sharing of ancestral polymorphisms was observed [69].

In the *An. gambiae* complex, phylogenetic trees based on molecular data are not in agreement with inevrsion phylogeny data [70,71]. According to phylogeny trees obtained from rDNA, mt DNA, and satellite DNA (stDNA) of the Y chromosome, *An. gambiae* and *An. arabiensis* are sister taxa [71,72]. However, inversion phylogeny considers *An. gambiae* and *An. merus* as sister taxa [31,71,73]. Sequencing four DNA regions between or very close to 2La inversions in the *An. gambiae* complex, places *An. gambiae* and *An. merus* as sister taxa [63]. The mosaic genomic architecture of the *An. gambiae* complex was confirmed and it was approved that the phylogenetic relationship among members of *An. gambiae* complex varies widely between different genomic regions and that introgression happens in different genomic regions [74].

In the following study, we will use sequences near breakpoint regions to construct the molecular phylogeny among members of the *An. gambiae* complex. Breakpoint genes are good candidates

for this purpose because genes at the breakpoints are less subject to gene flow [75]. In a previous study on *Drosophila pseudoobscura*, a phylogenetic relationship was constructed based on the gene arrangements of the inversion breakpoints in the third chromosome [76]. It is important to consider the chromosomal location of a genetic marker with respect to breakpoints of fixed inversions when constructing phylogeny in closely related species.

Major African vectors of malaria

There are five major vectors of malaria in Africa, including *An. gambiae*, *An. arabiensis*, *An. funestus*, *An. nili* and *An. moucheti.* All of these species are anthropophilic and susceptible to *Plasmodium falciparum* [9]. *Anopheles nili* larvae breed in rivers and adults are considered to be a widespread vector in the humid savannas and forested areas of Africa [10,77,78]. *Anopheles moucheti* is restricted to forested areas of Central Africa and is mostly found close to rivers or slow moving streams [79], while *An. funestus* has a wide distribution in Africa and occupies a wide-range of ecological niches [80,81]. According to a study on rDNA and mtDNA sequences, Afrotropical Funestus and Afro-Oriental Minimus groups are considered sister taxa [82].

In a phylogenetic analysis of mtDNA and rDNA on anthropophilic *Plasmodium falciparum* malaria vectors within the subgenus *Cellia*, including *An. nili, An. gambiae, An. moucheti, An. arabiensis, An. funestus* and several outgroup species, *An. nili* is clustered separately from the *An. gambiae* complex [83]. Constructing phylogenetic relationship based on multiple gene sequences have been obtained in several organisms [84,85,86,87], however, these data are not available among African malaria vectors. In the following study, for the first time, we have used a multigene phylogeny approach to construct the phylogenetic relationship among major vectors of malaria in Africa.

Chromosomal mapping and microsatellites

Microsatellites are tandem repeats of sequence units that are less that 5 bp in length, and are widespread in eukaryotic genomes. Microsatellites are usually highly polymorphic in the number of their repeat units, and are variable in length [88]. They are informative genetic markers to infer the population structure of different organisms, and heterozygosity is detected according to genetic variation at microsatellite loci. Since microsatellites evolve neutrally, the degree of polymorphism is proportional to the rate of mutation. They can be used as genetic markers in population genetic studies and genome mapping. Microsatellites are also suitable markers to measure the amount of gene flow between populations [89].

Paracentric chromosomal inversions reduce recombination between alternative arrangements, specifically at the breakpoint regions [75,90]. As a result, they can capture and stabilize coadapted alleles [91]. Reduced recombination can affect loci within inversions and close to inversion breakpoints, therefore estimates of gene flow would be significantly different compared to loci that are located elsewhere in the genome [92,93].

It is important to study the population structure of major malaria vectors in order to better understand their epidemiology. The physical location of microsatellite markers with respect to polymorphic inversions is important for interpreting population genetic data. Microsatellites had been used to study the gene flow among malaria vectors, including *An. funestus, An. gambiae, An. nili*, and *An. stephensi* [94,95,96,97]. The position of microsatellites with respect to chromosomal inversions have been determined in *An. nili* and *An. funestus* [78,98]. However, the chromosomal map of microsatellite markers and inversion breakpoints has not been determined in *An. stephensi*. In this study, we aim to map microsatellite markers to polytene chromosomes of *An. stephensi*.

References

1. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.
2. Besansky NJ (2008) Genome Analysis Of Vectorial Capacity In Major Anopheles Vectors Of Malaria Parasites.
3. Krzywinski J, Grushko OG, Besansky NJ (2006) Analysis of the complete mitochondrial DNA from Anopheles funestus: An improved dipteran mitochondrial genome annotation and a temporal dimension of mosquito evolution. Molecular phylogenetics and evolution 39: 417-423.
4. Harbach R (2004) The classification of genus Anopheles (Diptera: Culicidae): a working hypothesis of phylogenetic relationships. Bulletin of Entomological Research 94: 537-554.
5. White BJ, Hahn MW, Pombi M, Cassone BJ, Lobo NF, et al. (2007) Localization of candidate regions maintaining a common polymorphic inversion (2La) in Anopheles gambiae. Plos Genetics 3: e217.
6. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596.
7. Collins FH, Paskewitz SM (1995) Malaria: current and future prospects for control. Annual review of entomology 40: 195-219.
8. 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.
9. Fontenille D, Simard F (2004) Unravelling complexities in human malaria transmission dynamics in Africa through a comprehensive knowledge of vector populations. Comparative immunology, microbiology and infectious diseases 27: 357-375.
10. Sinka ME, Bangs MJ, Manguin S, Coetzee M, Mbogo CM, et al. (2010) The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. Parasit Vectors 3: 117.
11. Manouchehri A, Javadian E, Eshighy N, Motabar M (1976) Ecology of Anopheles stephensi Liston in southern Iran. Trop Geogr Med 28: 228-232.
12. Kamali M, Sharakhova MV, Baricheva E, Karagodin D, Tu Z, et al. (2011) An Integrated Chromosome Map of Microsatellite Markers and Inversion Breakpoints for an Asian Malaria Mosquito, Anopheles stephensi. Journal of Heredity 102: 719-726.
13. Grouchy J (1987) Chromosome phylogenies of man, great apes, and Old World monkeys. Genetica 73: 37-52.
14. Sweet W, Rao B (1937) Races of Anopheles stephensi Liston, 1901. Indian Medical Gazette 72: 665-674.
15. Subbarao S, Vasantha K, Adak T, Sharma V, Curtis C (1987) Egg-float ridge number in Anopheles stephensi: ecological variation and genetic analysis. Medical and Veterinary Entomology 1: 265-271.
16. Rao BA, Sweet WC, Subbarao AM (1938) Ova measurements of A. stephensi type and A. stephensi var. mysorensis. J Malar Inst India 1: 261-266.
17. Sinka ME, Bangs MJ, Manguin S, Rubio-Palis Y, Chareonviriyaphap T, et al. (2012) A global map of dominant malaria vectors. Parasit Vectors 5: 69.
18. Mayr E (1942) Systematics and the origin of species, from the viewpoint of a zoologist: Harvard Univ Pr.
19. Yoder AD, Olson LE, Hanley C, Heckman KL, Rasoloarison R, et al. (2005) A multidimensional approach for detecting species patterns in Malagasy vertebrates. Proceedings of the National Academy of Sciences of the United States of America 102: 6587.
20. De Vienne D, Refrégier G, Hood M, Guigue A, Devier B, et al. (2009) Hybrid sterility and inviability in the parasitic fungal species complex Microbotryum. Journal of evolutionary biology 22: 683-698.

21. Campbell B, Steffen-Campbell J, Werren J (1994) Phylogeny of the Nasonia species complex (Hymenoptera: Pteromalidae) inferred from an internal transcribed spacer (ITS2) and 28S rDNA sequences. Insect Molecular Biology 2: 225-237.
22. Taylor E, Dodson J (1994) A molecular analysis of relationships and biogeography within a species complex of Holarctic fish (genus Osmerus). Molecular Ecology 3: 235-248.
23. Maxson L, Pepper E, Maxson RD (1977) Immunological resolution of a diploid-tetraploid species complex of tree frogs. Science 197: 1012-1013.
24. Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V (2002) A polytene chromosome analysis of the Anopheles gambiae species complex. Science 298: 1415-1418.
25. White G (1974) Anopheles gambiae complex and disease transmission in Africa. Transactions of the Royal Society of Tropical Medicine and Hygiene 68: 278-298.
26. Pates H, Curtis C (2005) Mosquito behavior and vector control. Annu Rev Entomol 50: 53-70.
27. Besansky NJ, Hill CA, Costantini C (2004) No accounting for taste: host preference in malaria vectors. Trends in parasitology 20: 249-251.
28. Charlwood JD, Alcântara J, Pinto J, Sousa CA, Rompão H, et al. (2005) Do bednets reduce malaria transmission by exophagic mosquitoes? Transactions of the Royal Society of Tropical Medicine and Hygiene 99: 901-904.
29. Duchemin JB, Tsy JM, Rabarison P, Roux J, Coluzzi M, et al. (2001) Zoophily of Anopheles arabiensis and An. gambiae in Madagascar demonstrated by odour-baited entry traps. Medical and veterinary entomology 15: 50-57.
30. Highton R, Bryan JH, Boreham P, Chandler J (1979) Studies on the sibling species Anopheles gambiae Giles and Anopheles arabiensis Patton (Diptera: Culicidae) in the Kisumu area, Kenya. Bulletin of entomological research 69: 43-53.
31. Coluzzi M, Sabatini A, Petrarca V, Di Deco M (1979) Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. Transactions of the Royal Society of Tropical Medicine and Hygiene 73: 483-497.
32. Tsy JMP, Duchemin JB, Marrama L, Rabarison P, Le Goff G, et al. (2003) Distribution of the species of the Anopheles gambiae complex and first evidence of Anopheles merus as a malaria vector in Madagascar. Malaria Journal 2: 33.
33. White G, Magayuka S, Boreham P (1972) Comparative studies on sibling species of the Anopheles gambiae Giles complex (Dipt., Culicidae): bionomics and vectorial activity of species A and species B at Segera, Tanzania. Bull Entomol Res 62: 295-317.
34. Hunt RH, Coetzee M, Fettene M (1998) The Anopheles gambiae complex: a new species from Ethiopia. Transactions of the Royal Society of Tropical Medicine and Hygiene 92: 231-235.
35. Takken W, Eling W, Hooghof J, Dekker T, Hunt R, et al. (1999) Susceptibility of Anopheles quadriannulatus Theobald (Diptera: Culicidae) to Plasmodium falciparum. Trans R Soc Trop Med Hyg 93: 578-580.
36. Habtewold T, Povelones M, Blagborough AM, Christophides GK (2008) Transmission blocking immunity in the malaria non-vector mosquito Anopheles quadriannulatus species A. PLoS Pathog 4: e1000070.
37. Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V (2002) A polytene chromosome analysis of the Anopheles gambiae species complex. Science 298: 1415-1418.
38. Habtewold T, Povelones M, Blagborough AM, Christophides GK (2008) Transmission blocking immunity in the malaria non-vector mosquito Anopheles quadriannulatus species A. PLoS pathogens 4: e1000070.
39. Coetzee M, Hunt RH, Wilkerson R (2013) Anopheles coluzzii and Anopheles amharicus, new members of the Anopheles gambiae complex. Zootaxa 3619: 246-274.
40. Coluzzi M, Sabatini A, Petrarca V, Di Deco M (1979) Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. Transactions of the Royal Society of Tropical Medicine and Hygiene 73: 483-497.

41. Temu E, Hunt R, Coetzee M, Minjas J, Shiff C (1997) Detection of hybrids in natural populations of the Anopheles gambiae complex by the rDNA-based, PCR method. Annals of tropical medicine and parasitology 91: 963-965.
42. Touré Y, Petrarca V, Traore S, Coulibaly A, Maiga H, et al. (1998) The distribution and inversion polymorphism of chromosomally recognized taxa of the Anopheles gambiae complex in Mali, West Africa. Parassitologia 40: 477.
43. Ayala FJ, Coluzzi M (2005) Chromosome speciation: humans, Drosophila, and mosquitoes. Proceedings of the National Academy of Sciences of the United States of America 102: 6535-6542.
44. Xia A, Sharakhova MV, Sharakhov IV (2008) Reconstructing ancestral autosomal arrangements in the Anopheles gambiae complex. Journal of Computational Biology 15: 965-980.
45. Bryan J, Petrarca V, Di Deco M, Coluzzi M (1987) Adult behaviour of members of the Anopheles gambiae complex in the Gambia with special reference to An. melas and its chromosomal variants. Parassitologia 29: 221.
46. Petrarca V, Carrara G, Di Deco M, Petrangeli G (1984) Cytogenetic and biometric observations on members of the Anopheles gambiae complex in Mozambique]. Parassitologia 26: 247.
47. Sharakhov I, White B, Sharakhova M, Kayondo J, Lobo N, et al. Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in the Anopheles gambiae complex; 2006. National Acad Sciences.
48. Sturtevant A (1921) A case of rearrangement of genes in Drosophila. Proceedings of the National Academy of Sciences of the United States of America 7: 235.
49. Kirkpatrick M (2010) How and why chromosome inversions evolve. PLoS biology 8: e1000501.
50. Noor MAF, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and the reproductive isolation of species. Proceedings of the National Academy of Sciences 98: 12084.
51. Noor MAF, Grams KL, Bertucci LA, Almendarez Y, Reiland J, et al. (2001) The genetics of reproductive isolation and the potential for gene exchange between Drosophila pseudoobscura and D. persimilis via backcross hybrid males. Evolution 55: 512-521.
52. Rieseberg LH (2001) Chromosomal rearrangements and speciation. Trends in Ecology & Evolution 16: 351-358.
53. Rocca K, Gray EM, Costantini C, Besansky NJ (2009) 2La chromosomal inversion enhances thermal tolerance of Anopheles gambiae larvae. Malar J 8: 147.
54. Gray EM, Rocca K, Costantini C, Besansky NJ (2009) Inversion 2La is associated with enhanced desiccation resistance in Anopheles gambiae. Malar J 8: 215.
55. Fouet C, Gray E, Besansky NJ, Costantini C (2012) Adaptation to aridity in the malaria mosquito Anopheles gambiae: chromosomal inversion polymorphism and body size influence resistance to desiccation. PLoS One 7: e34841.
56. Simard F, Ayala D, Kamdem GC, Pombi M, Etouna J, et al. (2009) Ecological niche partitioning between Anopheles gambiae molecular forms in Cameroon: the ecological side of speciation. BMC ecology 9: 17.
57. Petrarca V, Beier JC (1992) Intraspecific chromosomal polymorphism in the Anopheles gambiae complex as a factor affecting malaria transmission in the Kisumu area of Kenya. The American journal of tropical medicine and hygiene 46: 229.
58. Coetzee M, Craig M, Le Sueur D (2000) Distribution of African malaria mosquitoes belonging to the Anopheles gambiae complex. Parasitology today 16: 74-77.
59. Sturtevant A, Dobzhansky T (1936) Inversions in the third chromosome of wild races of Drosophila pseudoobscura, and their use in the study of the history of the species. Proceedings of the National Academy of Sciences of the United States of America 22: 448.
60. Coluzzi M, Di Deco M, Cancrini G (1973) Chromosomal inversions in Anopheles stephensi. Parassitologia 15: 129.
61. Bhutkar A, Gelbart WM, Smith TF (2007) Inferring genome-scale rearrangement phylogeny and ancestral gene order: a Drosophila case study. Genome biology 8: R236.

62. Torre A, Merzagora L, Powell J, Coluzzi M (1997) Selective introgression of paracentric inversions between two sibling species of the Anopheles gambiae complex. Genetics 146: 239.
63. Caccone A, Min GS, Powell JR (1998) Multiple origins of cytologically identical chromosome inversions in the Anopheles gambiae complex. Genetics 150: 807.
64. Bhutkar A, Gelbart WM, Smith TF (2007) Inferring genome-scale rearrangement phylogeny and ancestral gene order: a Drosophila case study. Genome Biol 8: R236.
65. Kamali M, Xia A, Tu Z, Sharakhov IV (2012) A New Chromosomal Phylogeny Supports the Repeated Origin of Vectorial Capacity in Malaria Mosquitoes of the Anopheles gambiae Complex. PLoS pathogens 8: e1002960.
66. Ayala FJ, Coluzzi M (2005) Chromosome speciation: humans, Drosophila, and mosquitoes. Proceedings of the National Academy of Sciences of the United States of America 102: 6535.
67. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science 298: 129-149.
68. Besansky NJ, Lehmann T, Fahey GT, Fontenille D, Braack LEO, et al. (1997) Patterns of mitochondrial variation within and between African malaria vectors, Anopheles gambiae and An. arabiensis, suggest extensive gene flow. Genetics 147: 1817-1828.
69. Parmakelis A, Slotman M, Marshall J, Awono-Ambene P, Antonio-Nkondjio C, et al. (2008) The molecular evolution of four anti-malarial immune genes in the Anopheles gambiae species complex. BMC Evolutionary Biology 8: 79.
70. Mathiopoulos KD, Powell JD, McCutchan TF (1995) An anchored restriction-mapping approach applied to the genetic analysis of the Anopheles gambiae malaria vector complex 1. Molecular Biology and Evolution 12: 103-112.
71. Besansky NJ, Powell JR, Caccone A, Hamm DM, Scott JA, et al. (1994) Molecular phylogeny of the Anopheles gambiae complex suggests genetic introgression between principal malaria vectors. Proceedings of the National Academy of Sciences 91: 6885.
72. Krzywinski J, Sangaré D, Besansky NJ (2005) Satellite DNA from the Y chromosome of the malaria vector Anopheles gambiae. Genetics 169: 185-196.
73. Pape T (1992) Cladistic analyses of mosquito chromosome data in Anopheles subgenus Cellia (Diptera: Culicidae). Mosquito Systematics 24: 1-11.
74. Wang-Sattler R, Blandin S, Ning Y, Blass C, Dolo G, et al. (2007) Mosaic genome architecture of the Anopheles gambiae species complex. PLoS One 2: e1249.
75. Navarro A, Betran E, Barbadilla A, Ruiz A (1997) Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. Genetics 146: 695.
76. Wallace AG, Detweiler D, Schaeffer SW (2011) Evolutionary History of the Third Chromosome Gene Arrangements of Drosophila pseudoobscura Inferred from Inversion Breakpoints. Molecular Biology and Evolution.
77. Gillies M, De Meillon B (1968) The Anophelinae of Africa south of the Sahara (Ethiopian zoogeographical region). The Anophelinae of Africa south of the Sahara (Ethiopian Zoogeographical Region).
78. Peery A, Sharakhova MV, Antonio-Nkondjio C, Ndo C, Weill M, et al. (2011) Improving the population genetics toolbox for the study of the African malaria vector Anopheles nili: microsatellite mapping to chromosomes. Parasites & vectors 4: 1-10.
79. Kengne P, Antonio-Nkondjio C, Awono-Ambene H, Simard F, Awolola T, et al. (2007) Molecular differentiation of three closely related members of the mosquito species complex, Anopheles moucheti, by mitochondrial and ribosomal DNA polymorphism. Medical and veterinary entomology 21: 177-182.
80. Hay SI, Guerra CA, Gething PW, Patil AP, Tatem AJ, et al. (2009) A world malaria map: Plasmodium falciparum endemicity in 2007. PLoS Medicine 6: e1000048.
81. Sharakhov IV, Sharakhova MV, Mbogo CM, Koekemoer LL, Yan G (2001) Linear and spatial organization of polytene chromosomes of the African malaria mosquito Anopheles funestus. Genetics 159: 211-218.

82. Garros C, Harbach RE, Manguin S (2005) Morphological assessment and molecular phylogenetics of the Funestus and Minimus Groups of Anopheles (Cellia). Journal of medical entomology 42: 522-536.
83. Marshall JC, Powl JR, Caccone A (2005) Phylogenetic relationships of the anthropophilic Plasmodium falciparum malaria vectors in Africa. The American journal of tropical medicine and hygiene 73: 749-752.
84. Shalchian-Tabrizi K, Minge MA, Espelund M, Orr R, Ruden T, et al. (2008) Multigene phylogeny of choanozoa and the origin of animals. PLOS ONE 3: e2098.
85. Koepfli K-P, Deere K, Slater G, Begg C, Begg K, et al. (2008) Multigene phylogeny of the Mustelidae: resolving relationships, tempo and biogeographic history of a mammalian adaptive radiation. BMC biology 6: 10.
86. Gao F, Katz LA, Song W (2013) Multigene-based analyses on evolutionary phylogeny of two controversial ciliate orders: Pleuronematida and Loxocephalida (Protista, Ciliophora, Oligohymenophorea). Molecular Phylogenetics and Evolution.
87. Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798-804.
88. Bruford MW, Wayne RK (1993) Microsatellites and their application to population genetic studies. Current opinion in genetics & development 3: 939-943.
89. Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics 5: 435-445.
90. Sturtevant A, Beadle G (1936) The relations of inversions in the X chromosome of Drosophila melanogaster to crossing over and disjunction. Genetics 21: 554.
91. Dobzhansky T, Dobzhansky TG (1970) Genetics of the evolutionary process: Columbia University Press.
92. Lanzaro GC, Touré YT, Carnahan J, Zheng L, Dolo G, et al. (1998) Complexities in the genetic structure of Anopheles gambiae populations in west Africa as revealed by microsatellite DNA analysis. Proceedings of the National Academy of Sciences 95: 14260-14265.
93. Tripet F, Dolo G, Lanzaro GC (2005) Multilevel analyses of genetic differentiation in Anopheles gambiae ss reveal patterns of gene flow important for malaria-fighting mosquito projects. Genetics 169: 313-324.
94. Cohuet A, Dia I, Simard F, Raymond M, Rousset F, et al. (2005) Gene flow between chromosomal forms of the malaria vector Anopheles funestus in Cameroon, Central Africa, and its relevance in malaria fighting. Genetics 169: 301-311.
95. Lehmann T, Hawley WA, Kamau L, Fontenille D, Simard F, et al. (1996) Genetic differentiation of Anopheles gambiae populations from East and West Africa: comparison of microsatellite and allozyme loci. Heredity 77: 192-200.
96. Ndo C, Antonio-Nkondjio C, Cohuet A, Ayala D, Kengne P, et al. (2010) Population genetic structure of the malaria vector Anopheles nili in sub-Saharan Africa. Malaria journal 9: 161.
97. Dube M, Gakhar S (2010) Genetic differentiation between three ecological variants (‘type’,‘mysorensis’ and ‘intermediate’) of malaria vector Anopheles stephensi (Diptera: Culicidae). Insect Science 17: 335-343.
98. Sharakhov I, Braginets O, Grushko O, Cohuet A, Guelbeogo W, et al. (2004) A microsatellite map of the African human malaria vector Anopheles funestus. Journal of Heredity 95: 29-34.

CHAPTER 2 A New Chromosomal Phylogeny Supports the Repeated Origin of Vectorial Capacity in Malaria Mosquitoes of the *Anopheles gambiae* Complex

Kamali M, Xia A, Tu Z, Sharakhov IV (2012) A New Chromosomal Phylogeny Supports the Repeated Origin of Vectorial Capacity in Malaria Mosquitoes of the *Anopheles gambiae* Complex. PLoS Pathog 8(10): e1002960. doi:10.1371/journal.ppat.1002960

New Chromosomal Phylogeny Supports the Repeated Origin of Vectorial Capacity in Malaria Mosquitoes of the *Anopheles gambiae* Complex

Maryam Kamali[1], Ai Xia[1]†, Zhijian Tu[2], Igor V. Sharakhov[1]*

[1]Department of Entomology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, United States of America.

[2]Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, United States of America.

*Correspondence to: Igor V. Sharakhov, Department of Entomology, 203 Fralin Life Science Institute, West Campus Drive, MC 0346, Virginia Tech, Blacksburg, Virginia, United States of America, 24061; Office phone: (540) 231-7316; Lab phone: (540) 231-0731; Fax: (540) 231-7126; email: igor@vt.edu.

†Current address: Department of Entomology, College of Plant Protection, Nanjing Agricultural University, No.1 Weigang, Nanjing, Jiangsu Province, 210095, P. R. China.

Abstract

Understanding phylogenetic relationships within species complexes of disease vectors is crucial for identifying genomic changes associated with the evolution of epidemiologically important traits. However, the high degree of genetic similarity among sibling species confounds the ability to determine phylogenetic relationships using molecular markers. The goal of this study was to infer the ancestral–descendant relationships among malaria vectors and nonvectors of the *Anopheles gambiae* species complex by analyzing breakpoints of fixed chromosomal inversions in ingroup and several outgroup species. We identified genes at breakpoints of fixed overlapping chromosomal inversions 2Ro and 2Rp of *An. merus* using fluorescence *in situ* hybridization, a whole-genome mate-paired sequencing, and clone sequencing. We also mapped breakpoints of a chromosomal inversion 2La (common to *An. merus*, *An. gambiae*, and *An. arabiensis*) in outgroup species using a bioinformatics approach. We demonstrated that the "standard" $2R+^{p}$ arrangement and "inverted" 2Ro and 2La arrangements are present in outgroup species *Anopheles stephensi*, *Aedes aegypti*, and *Culex quinquefasciatus.* The data indicate that the ancestral species of the *An. gambiae* complex had the 2Ro, $2R+^{p}$, and 2La chromosomal arrangements. The "inverted" 2Ro arrangement uniquely characterizes a malaria vector *An. merus* as the basal species in the complex. The rooted chromosomal phylogeny implies that *An. merus* acquired the 2Rp inversion and that its sister species *An. gambiae* acquired the $2R+^{o}$ inversion from the ancestral species. The karyotype of nonvectors *An. quadriannulatus* A and B was derived from the karyotype of the major malaria vector *An. gambiae.* We conclude that the ability to effectively transmit human malaria had originated repeatedly in the complex. Our findings also suggest that saltwater tolerance originated first in *An. merus* and then independently

in *An. melas*. The new chromosomal phylogeny will facilitate identifying the association of evolutionary genomic changes with epidemiologically important phenotypes.

Author summary

Malaria causes more than one million deaths every year, mostly among children in Sub-Saharan Africa. *Anopheles* mosquitoes are exclusive vectors of human malaria. Many malaria vectors belong to species complexes, and members within these complexes can vary significantly in their ecological adaptations and ability to transmit the parasite. To better understand evolution of epidemiologically important traits, we studied relationships among nonvector and vector species of the African *Anopheles gambiae* complex. We analyzed gene orders at genomic regions where evolutionary breaks of chromosomal inversions occurred in members of the complex and compared them with gene orders in species outside the complex. This approach allowed us to identify ancient and recent gene orders for three chromosomal inversions. Surprisingly, the more ancestral chromosomal arrangements were found in mosquito species that are vectors of human malaria, while the more derived arrangements were found in both nonvectors and vectors. Our finding strongly suggests that the increased ability to transmit human malaria originated repeatedly during the recent evolution of these African mosquitoes. This knowledge can be used to identify specific genetic changes associated with the human blood choice and ecological adaptations.

Introduction

Complexes of sibling species are common among arthropod disease vectors [1-3]. Members of such complexes are morphologically similar and partially reproductively isolated from each other. The *Anopheles gambiae* complex consists of seven African malaria mosquito sibling species. *Anopheles gambiae* and *An. arabiensis*, the two major vectors of malaria in Africa, are

both anthropophilic and can breed in temporal freshwater pools. *Anopheles gambiae* occupies more humid areas, while *An. arabiensis* dominates in arid savannas and steppes. *Anopheles merus* and *An. melas* breed in saltwater, and the habitat of *An. bwambae* is restricted to mineral water breeding sites. These three species are relatively minor malaria vectors mainly due to narrow geographic distributions [4]. *Anopheles quadriannulatus* A and *An. quadriannulatus* B are freshwater breeders and, although to various degrees susceptible to *Plasmodium* infections, are not natural vectors of malaria mainly due to zoophilic behavior [5-7]. Inferring the evolutionary history of the *An. gambiae* complex could be crucial for identifying specific genomic changes associated with the human blood choice, breeding site preference, and variations in vector competence. However, the high degree of genetic similarity, caused by the ancestral polymorphism and introgression, complicates the use of molecular markers for the reconstruction of a sibling species phylogeny [8-10]. Even the most recent genome-wide transcriptome-based phylogeny reconstruction of multiple *Anophelinae* species could not unambiguously resolve the relationships among *An. gambiae, An. arabiensis,* and *An. quadriannulatus* [11].

An alternative approach to inferring the phylogenetic relationships among species is to analyze the distribution of fixed overlapping inversions [4,7,12]. This approach is based on the fact that species-specific inversions do not introgress [13] and that inversions are predominantly monophyletic, despite rare occurrences of breakpoint reuse [14]. In addition, chromosomal inversions are more rare events and more consistent characters as compared with nucleotide substitutions [12,15]. Phylogenies based on inversion data are highly congruent with phylogenies based on DNA sequence data and are shown to be more information rich than are nucleotide data [15]. Members of the *An. gambiae* complex carry 10 fixed inversions that can be used for a

phylogeny reconstruction [7]. Five fixed inversions are present on the X chromosome, three inversions are found on the 2R arm, and one is found on each of the 2L and 3L arms (Figure S1) [7]. The only nonvectors in the complex, *Anopheles quadriannulatus* A and B, had been traditionally considered the closest species to the ancestral lineage because they have a large number of hosts, feed on animal blood, tolerate temperate climates, exhibit disjunctive distribution, and possess a "standard" karyotype [4,7,16,17]. More recently, the *An. arabiensis* karyotype had been assumed ancestral because it has the fixed 2La inversion, which was also found in two outgroup species from the Middle Eastern *An. subpictus* complex [18]. Both chromosomal phylogenies assumed the most recent speciation of *An. merus* and an independent origin of the cytologically identical 2La' inversion in this species [19]. A phylogenetic status of an inversion can be determined more precisely when breakpoints are identified and gene orders across breakpoints are compared between ingroup and multiple outgroup species. The genes found across inversion breakpoints in ingroup and outgroup species are expected to be in their ancestral order [12]. For example, the molecular analysis of the 2La inversion breakpoints and physical mapping of the sequences adjacent to the breakpoints in outgroup species identified the shared 2La inversion in *An. gambiae*, *An. merus*, and *An. arabiensis* and determined the ancestral state of the 2La arrangement [20-22].

Based on the X chromosome fixed inversions, three species clades can be identified in the complex: (i) *An. bwambae*, *An. melas*, and *An. quadriannulatus* A and B (X+), (ii) *An. arabiensis* (Xbcd), and (iii) *An. merus* and *An. gambiae* (Xag) (Figure 1). The *An. gambiae–An. merus* and *An. bwambae–An. melas* sister taxa relationships have been supported by independent phylogenetic analyses of nuclear genes and mitochondrial DNA sequences [9,10,23]. Each clade has unique fixed inversions that can be used to unambiguously determine its phylogenetic status

if compared to gene arrangements in outgroup species: X+, 2Rm, 3La in the *An. bwambae–An. melas–An. quadriannulatus* clade, Xbcd in *An. arabiensis*, and Xag, 2Ro, 2Rp in the *An. gambiae–An. merus* clade. However, to efficiently pursue this research was not possible until recently when genome sequences of several outgroup mosquito species became available, including *An. stephensi* (series *Neocellia*, subgenus *Cellia*, subfamily *Anophelinae*) (this paper), and *Aedes aegypti* and *Culex quinquefasciatus* (both from subfamily *Culicinae*) [24,25]. In this study, we identified genes at the breakpoints of fixed overlapping inversions 2Ro and 2Rp of *An. merus* and homologous sequences in *An. stephensi*, *Ae. aegypti*, and *C. quinquefasciatus*. We demonstrated that the "inverted" 2Ro and the "standard" $2R+^{p}$ arrangements are ancestral in the complex. In addition, we found that the "inverted" 2La arrangement is present in evolutionary distant *Culicinae* species and, therefore, is ancestral. The inversion data support the basal position of the *An. gambiae–An. merus* clade and the terminal positions of the *An. arabiensis* and *An. melas* lineages. This rooted chromosomal phylogeny could be a means to examine specific genomic changes associated with evolution of traits relevant to vectorial capacity.

Results and discussion

To infer the ancestral-descendant relationships among chromosomal arrangements in the *An. gambiae* complex, we determined gene orders at the breakpoints of the *An. merus*-specific fixed overlapping inversions 2Ro and 2Rp in ingroup and several outgroup species, including *An. stephensi*, *Ae. aegypti*, and *C. quinquefasciatus*. In our first approach, we used *An. gambiae* DNA probes, which were identified at breakpoints of "standard" $2R+^{o}$ and $2R+^{p}$ arrangements, for the mapping to polytene chromosomes of *An. merus* and *An. stephensi* by fluorescence *in situ* hybridization (FISH). In our second approach, we performed mate-paired sequencing of the *An. merus* genome and mapped the read pairs to the *An. gambiae* AgamP3 genome assembly. The

inversion breakpoints of 2Ro and 2Rp in the *An. gambiae–An. merus* clade and their homologous sequences in the outgroup species were obtained and analyzed. This study reconstructed a rooted chromosomal phylogeny and revised evolutionary history of the *An. gambiae* complex.

Chromosome positions of the 2Ro and 2Rp inversion breakpoints in *An. merus*, *An. gambiae*, and *An. stephensi*

We mapped multiple *An. gambiae* DNA probes derived from the cytological breakpoints to the chromosomes of *An. merus* by FISH. *Anopheles gambiae* BAC clone 141A14 that spans the proximal $2R+^{o}$ breakpoint was identified by comparative mapping with *An. merus* in our previous study [21]. FISH of the BAC clone to *An. merus* chromosomes produced two separate signals on 2R indicating an inversion. Reiteration of this procedure with PCR fragments derived from the BAC clone allowed us to localize the breakpoint region within the BAC between genes AGAP002933 and AGAP002935. Further comparative mapping with *An. merus* demonstrated that the distal $2R+^{o}$ breakpoint in *An. gambiae* is located between genes AGAP001759 and AGAP001762 (Figure S2). We also performed FISH with polytene chromosomes of *An. merus* using multiple probes located near the $2R+^{p}$ cytological breakpoints of *An. gambiae*. The proximal $2R+^{p}$ breakpoint was found between genes AGAP003327 and AGAP003328, and the distal $2R+^{p}$ breakpoint was localized between AGAP001983 and AGAP001984 in *An. gambiae*. These gene pairs were neighboring in the genome of *An. gambiae*, but they were mapped in separate locations in *An. merus* (Figure S3). To determine gene arrangements in an outgroup species, we mapped genes at the $2R+^{o}$ and $2R+^{p}$ breakpoints to polytene chromosomes of *An. stephensi* (Figure S4 and Figure S5). The FISH results showed that the "inverted" 2Ro and "standard" $2R+^{p}$ arrangements are present in the outgroup species *An. stephensi* (Figure 2).

Structure of the 2Ro and 2R+^{o} inversion breakpoints in *An. merus* and *An. gambiae*

We performed mate-paired sequencing of the *An. merus* genome and mapped the read pairs to the *An. gambiae* AgamP3 genome assembly, which has all "standard" arrangements [26,27]. Mate-paired sequencing is the methodology that enables the generation of libraries with inserts from 2 to 5 kb in size. The 2 kb, 3 kb, and 5 kb DNA fragments were circularized, fragmented, purified, end-repaired, and ligated to Illumina paired-end sequencing adapters. The final libraries consisted of short fragments made up of two DNA segments that were originally separated by several kilobases. These genomic inserts were paired-end sequenced using an Illumina approach. Paired-read sequences that map far apart in the same orientation delineate inversions [28]. We executed a BLASTN search to find read pairs mapped to the putative breakpoint regions in the same orientation on chromosome 2 (Figure 3). Alignment of the read pairs to the genome of *An. gambiae* identified the 2Ro breakpoints at coordinates ~9.48 Mb and ~29.84 Mb. We also identified the 2La breakpoints at coordinates ~20.52 Mb and ~42.16 Mb, which confirmed a previous study and, thus, validated the approach [20]. However, the BLASTN search did not find the paired-read sequences that map at the opposite 2Rp breakpoints in the same orientation. This approach could not detect breakpoint regions longer than 5 kb. The 2Rp breakpoint regions in *An. merus* likely have larger sizes caused by accumulation of repetitive sequences. We also used the Bowtie program [29] to confirm the genomics positions of the 2Ro breakpoints. Both BLASTN and Bowtie results supported the position of the proximal 2Ro breakpoint to the region between genes AGAP001762 and AGAP002935, and they refined the position of the distal 2Ro breakpoint to the region between AGAP001760 and AGAP002933.

The genes adjacent to the 2Ro breakpoint were used as probes to screen the genomic phage library of *An. merus*. Positive *An. merus* phage clones were confirmed to span inversion breakpoints by FISH to polytene chromosomes of *An. gambiae, An. merus,* and *An. stephensi*. For example, hybridization of Phage 6D produced only one signal in the proximal 2Ro breakpoint in *An. merus* but two signals at both 2Ro breakpoints in *An. gambiae* (Figure S6). Phage 6D hybridized to only one locus in *An. stephensi*, confirming the 2Ro arrangement in this species. Confirmed phage clones were sequenced, and the exact breakpoint regions were identified by aligning the *An. merus* sequences and *An. gambiae* AgamP3, AgamM1, and AgamS1 genome assemblies available at VectorBase [26,30,31]. Thus, distal and proximal breakpoints were identified on a polytene chromosome map [7] and in the genome assembly of *An. gambiae* (Figure 4). In the AgamP3 assembly, the distal and proximal breakpoint regions span coordinates 9,485,167–9,486,712, and 29,838,366–29,839,163, respectively. The 2Ro breakpoint regions were 2.6 and 5.9 times smaller in *An. merus* as compared with the $2R+^{o}$ breakpoint regions in *An. gambiae* due to accumulation of transposable elements (TEs) in the latter species. The presence of TEs is a common signature of inversion breakpoints, as TEs usually mark breakpoints of derived arrangements [20,32]. Five various DNA transposons were found at the distal $2R+^{o}$ breakpoint, and one novel miniature inverted-repeat TE (MITE), Aga_m3bp_Ele1, was identified at the proximal $2R+^{o}$ breakpoint in *An. gambiae* (Figure 4). Smaller sizes of the breakpoint regions and the lack of TEs at the breakpoints of *An. merus* strongly suggest the ancestral state of the 2Ro arrangement.

Gene orders at the 2Ro, 2Rp, and 2La inversion breakpoints in outgroup species

We determined gene orders at the breakpoints of the *An. merus*-specific fixed overlapping inversions 2Ro and 2Rp in several outgroup species, including *An. stephensi, Ae. aegypti,* and *C.*

quinquefasciatus. The genes adjacent to the 2Ro and 2Rp breakpoint were used as probes to screen the genomic BAC library of the outgroup species *An. stephensi*. Sequences homologous to genes from the distal 2Ro breakpoint were found in the BAC clone AST044F8 of *An. stephensi*. In addition, we performed sequencing of the *An. stephensi* genome using 454 and Illumina platforms. Sequences homologous to genes from the proximal 2Ro breakpoint were identified in scaffold 03514 of the *An. stephensi* genome. We also detected homologous sequences in the genome assemblies of *Ae. aegypti* and *C. quinquefasciatus* available at VectorBase [27]. The analysis demonstrated that all studied outgroup species had the gene arrangement identical to that of *An. merus* confirming the ancestral state of the 2Ro inversion (Figure 5). The *An. stephensi* sequences, which correspond to the 2Ro breakpoints, had sizes more similar to those in *An. merus* than in *An. gambiae*, and they did not display any TEs or repetitive elements, further supporting the 2Ro ancestral state. However, we found TEs in sequences corresponding to one of the 2Ro breakpoints in *Ae. aegypti*. Incidentally, the areas between the homologous breakpoint-flanking genes were 12,055 bp in *Culex* and 31,352 bp in *Aedes*, and this probably reflects the repeat-rich nature of the *Culicinae* genomes. The demonstrated conservation of gene orders between *Anophelinae* and *Culicinae* species is remarkable given the ~145–200 million years of divergence time between these two lineages [33].

Approximate genomic positions of the $2R+^{p}$ breakpoints were determined between AGAP001983 and AGAP001984 and between AGAP003327 and AGAP003328 by physical mapping of *An. merus* chromosomes (Figure 2). Using these genes as probes, we obtained a positive Phage 3B of *An. merus* that was mapped to the proximal 2Rp breakpoint in *An. merus* (Figure S6). Sequencing and molecular analyses of Phage 3B revealed the presence of

AGAP001983 and AGAP013533 in this clone indicating that the actual distal breakpoint is located between AGAP013533 and AGAP001984 in *An. gambiae*. However, the available Phage 3B sequence ended at gene AGAP013533 and, thus, did not encompass the actual breakpoint sequence in *An. merus*. We performed the comparative analysis of gene orders at the 2Rp breakpoints in three outgroup species, *An. stephensi*, *C. quinquefasciatus*, and *Ae. aegypti*. The results demonstrated the common organization of the distal $2R+^{p}$ breakpoint in *An. gambiae* and outgroup species, indicating that this arrangement is ancestral (Figure 6). Interestingly, a gene similar to AGAP013533 was absent, but genes similar to AGAP001983 and AGAP001984 were present in supercontig 3.153 of *C. quinquefasciatus.* Genes similar to AGAP003327 and AGAP003328 were found in different scaffolds and supercontigs of the outgroup species. This pattern was expected because AGAP003327 and AGAP003328 were mapped to neighboring but different subdivisions on the *An. stephensi* chromosome map (Figure 2). Therefore, it is possible that an additional inversion separated these two genes in the *An. stephensi* lineage. The highly fragmented nature of the *C. quinquefasciatus* and *Ae. aegypti* genome assemblies could also explain the observed pattern. No TEs were found in the breakpoint regions of *An. stephensi* and *C. quinquefasciatus*. However, multiple TEs were found in the intergenic regions of *An. gambiae* and *Ae. aegypti* (Figure 6).

Using sequencing and cytogenetic approaches, the common 2La arrangement was previously found in *An. gambiae, An. merus*, and *An. arabiensis* [4,20], as well as in several outgroup species, including *An. subpictus* [18], *An. nili*, and *An. stephensi* [22]. Here, we used sequences available for breakpoints of the 2La inversion [20] to execute BLAST searches against genomes of more distantly related outgroup species *C. quinquefasciatus* and *Ae. aegypti.* BLAST results of genes adjacent to the 2La proximal breakpoint, AGAP007068 and AGAP005778, identified

orthologs CPIJ004936 and CPIJ004938 in the *Culex* genome as well as orthologs AAEL001778 and AAEL001757 in the *Aedes* genome. These genes were found within supercontig 3.77 in *C. quinquefasciatus* and within supercontig 1.42 in *Ae. aegypti*. Similarly, BLAST results of genes neighboring with the 2La proximal breakpoint, AGAP007069 and AGAP005780, identified homologous genes CPIJ005693 and CPIJ005692 in the *Culex* genome (supercontig 3.99) as well as AAEL011139 and AAEL011140 in the *Aedes* genome (supercontig 1.543). The obtained data confirmed the identical gene arrangement in distant outgroup species and the ancestry of the 2La inversion.

Chromosomal phylogeny of the *An. gambiae* complex

Physical chromosome mapping and bioinformatic analyses identified the 2Ro and $2R+^{p}$ arrangements in several outgroup species indicating that these arrangements are ancestral (Figure 5 and Figure 6). Because these two inversions overlap, only certain evolutionary trajectories and inversion combinations are possible (Figure 2). Specifically, the 2Rop–$2Ro+^{p}$–$2R+^{op}$ order of inversion events is possible, while the 2Rop–$2R+^{o}$p–$2R+^{op}$ evolutionary sequence is not possible, regardless of the direction. Identification of 2Ro and $2R+^{p}$ as the ancestral arrangements agrees well with this argument. We have also examined three different scenarios in reconstructing chromosomal phylogeny based on the established ancestry of 2Ro, $2R+^{p}$, and 2La and on the alternative hypothetical ancestries of X chromosomal arrangements (X+, Xag, or Xbcd) using the Multiple Genome Rearrangements (MGR) program [34]. Three different X chromosome arrangements (X+, Xag, and Xbcd) in an outgroup species were examined (Figure S7). The MGR program calculated the phylogenetic distances among species related to the ancestry of the X chromosome arrangement. Three hypothetical trees were obtained and used for interpretation of phylogenetic relationship and inversion reuse in the complex. Of the three scenarios, only the

phylogeny based on the ancestry of 2Ro, $2R+^{p}$, 2La, and Xag had all inversions originating only once in the evolution of the *An. gambiae* complex. The other scenarios (with X+ and Xbcd being ancestral) had multiple origins of one of the inversions implying that they are less parsimonious (Figure S7). Because Xag uniquely characterize the *An. gambiae–An. merus* clade, these two species have the least chromosomal differences from the ancestral species of the complex as compared with other members (Figure 7). The ancestry of Xag can be tested by mapping of the X chromosome genome sequences from several species of the *An. gambiae* complex, which soon will be available [10]. Importantly, the new phylogeny is in complete agreement with the previous discoveries of 2La being the ancestral arrangement [18,20]. Moreover, this is the first phylogeny based on knowledge about the status of a species-specific inversion (2Ro of *An. merus*). Therefore, the future data on the ancestry of the X chromosome arrangement are expected to support the new phylogeny.

Hypothetical evolutionary history of the *An. gambiae* complex

Speciation in the *An. gambiae* complex has been accompanied by fixation of chromosomal inversions, except for speciation within the *An. quadriannulatus* lineage [7,35]. Therefore, the chromosomal phylogeny likely reflects the species' evolutionary history. For a long time, the *An. quadriannulatus* lineage had been traditionally considered ancestral [4,7,16,17] (Figure 8A). This evolutionary history was reconstructed from an unrooted phylogeny without any knowledge about chromosomal arrangements in outgroup species. Later, the *An. arabiensis* lineage had been assumed basal because it has the fixed ancestral 2La inversion and based on knowledge about biogeography and ecology of *An. arabiensis* [18] (Figure 8B). In these two scenarios, saltwater species *An. merus* and *An. melas* had been assumed the most recently originated members in the complex. However, the ancestry and the unique origin of the 2La inversion [20] imply that *An.*

arabiensis, *An. gambiae*, or *An. merus* could be the closest to the ancestral species. The new chromosomal phylogeny led us to the substantial revision of the evolutionary history of the *An. gambiae* complex (Figure 8C). Accordingly, the ancestral species with 2Ro, 2R+^{p}, and 2La arrangements might have arisen in East Africa where *An. merus* and *An. gambiae* are present in sympatry. The ancestral species may have been polymorphic for the 2Rp and 2R+^{o} inversions and one lineage or population gave rise to *An. merus* with the 2Rp inversion while the other gave rise to the sister species *An. gambiae* containing the 2R+^{o} inversion. Otherwise one would have to postulate that *An. gambiae* and *An. merus* arose from independent ancestors. At some point in evolutionary history, *An. gambiae* acquired polymorphic 2La/+ inversion and entered forested regions in central Africa. Later, *An. gambiae* acquired multiple polymorphic inversions on 2R, which allowed this species to spread to the arid areas of West Africa [4]. A hypothetical karyotype might have originated from the *An. gambiae* chromosomal arrangements by acquiring X+^{ag} inversions. This karyotype in turn gave rise to the *An. arabiensis* chromosomes by generating the Xbcd inversions and fixing 2La and to the *An. quadriannulatus* karyotype by fixing the 2L+^{a} arrangement. The 3La inversion in *An. bwambae* originated from the *An. quadriannulatus* karyotype, followed by the origin of the 2Rm inversion in *An. melas*.

The two major malaria vectors *An. arabiensis* and *An. gambiae* are sympatric species in most of their distribution range, allowing for introgressive hybridization between them. Available data support the hypothesis of introgression of the 2La arrangement from *An. arabiensis* into *An. gambiae* [9,36,37]. According to the new chromosomal phylogeny, introgression of 2La has been happening from the more derived karyotype of *An. arabiensis* to the more ancestral karyotype of *An. gambiae*. Therefore, the 2La arrangement in isolated *An. gambiae* populations must retain alleles that are more distantly related to alleles of the 2La arrangement in *An. arabiensis*. This

hypothesis can be tested by the genomic analysis of *An. gambiae* island populations that do not have a history of hybridization with *An. arabiensis.* Because the 2La inversion in *An. gambiae* mainland populations has been associated with a tolerance to aridity and slightly reduced susceptibility to *Plasmodium falciparum* [4,38,39], the expected differences between the "original" and "introgressed" 2La arrangements could impact our understanding of a role of the inversion polymorphism in mosquito adaptation and malaria transmission.

Repeated origin of vectorial capacity and ecological adaptations

The results of this study indicate that *An. merus* is closely related to an ancestral species from which the *An. gambiae* complex arose. *Anopheles merus* is a minor vector of human malaria in African mainland. A role of *An. merus* in malaria transmission in Madagascar has also been documented [40]. Based on the unique origin of fixed inversions and X-linked sequences, *An. merus* and *An. gambiae* are considered sister taxa [9,10]. Therefore, according to the new chromosomal phylogeny, these two species possess the most "primitive" karyotypes in the complex. Our data suggest that the major malaria vector in Africa *An. gambiae* could be more closely related to the ancestral species than was previously assumed. Unexpectedly, we found that the karyotype of nonvectors *An. quadriannulatus* A and B was derived from the karyotype of *An. gambiae* (Figure 7 and Figure 8). *Anopheles quadriannulatus* is not involved in malaria transmission in nature due to its strong preference for feeding on animals [7]. *Anopheles melas* has the most recently formed karyotype and is a malaria vector in West Africa [41,42].

The new chromosomal phylogeny strongly suggests that vectorial capacity evolved repeatedly in the *An. gambiae* complex. Increased anthropophily could not have evolved in *An. gambiae* and *An. arabiensis* before humans originated and evolved to high enough densities. Therefore, the ability to effectively transmit human malaria must be a relatively recent trait in the complex. If

An. quadriannulatus were the ancestral species, as it was assumed earlier [4,7], then vectorial capacity could have originated only once when all other members split from the *An. quadriannulatus* lineage (Figure 8A). However, if the *An. gambiae–An. merus* clade is ancestral, as we demonstrated here, then vectorial capacity must have arisen independently in different lineages after the species were diversified. The available data cannot clearly delineate between the loss of vectorial capacity in *An. quadriannultus* and its subsequent reappearance in *An. bwambae* and *An. melas* with a possible alternative that vectorial capacity in present day *An. quadriannulatus* was only lost after *An. bwambae* and *An. melas* split from the *An. quadriannulatus* lineage. Depending on when the phenotypic change occurred (before or after *An. bwambae*/*An. melas* split from the *An. quadriannulatus* lineage) different scenarios are possible. However, even if a zoophilic behavior was acquired by *An. quadriannulatus* after the split from *An. bwambae* and *An. melas*, one still has to assume repeated origin of vectorial capacity. In this case, it originated independently in *An. gambiae*, *An. merus*, *An. arabiensis*, and the lineage that led to *An. quadriannulatus*/*An. bwambae*/*An. melas*. This alteration of the phylogeny of the *An. gambiae* species complex will likely have direct impact on studies aimed at understanding the genetic basis of traits important to vectorial capacity.

The chromosomal phylogeny also supports the idea of multiple origins of similar ecological adaptations in the complex. An early cytogenetic and ecological study postulated the repeated evolution of saltwater tolerance in the complex [4]. *Anopheles melas* and *An. merus* breed in saltwater pools in western and eastern Africa, respectively. Our finding revealed that the physiological adaptation to breeding in saltwater originated first in *An. merus* and then independently in *An. melas*.

Conclusion

Because of the high degrees of genetic similarities among sibling species, attempts to use molecular markers to reconstruct phylogenetic trees often fail [10]. Our study provides the methodology for rooting chromosomal phylogenies of sibling species complexes, which are common among disease vectors, including blackflies, sandflies, and mosquitoes [1-3]. The robustness of this methodology is supported by the agreement between the two alternative approaches to breakpoint mapping (cytogenetics and sequencing) and by the consensus among the three inversions in the phylogenic analysis (2Ro, 2Rp, and 2La). As we demonstrated, inversion breakpoints can be physically mapped on polytene chromosomes by FISH and identified within genomes by mate-pair and clone sequencing. Importantly, the increasing availability of sequenced and assembled genomes provides an opportunity for identification of gene orders in multiple outgroup species for rooting chromosomal phylogenies.

The high genetic similarity among the species of the *An. gambiae* complex suggests their recent evolution [10,18]. The identified chromosomal relationships among the species demonstrate rapid gains and losses of traits related to vectorial capacity and ecological adaptations. This study reinforces the previous observations that vectors often do not cluster phylogenetically with nonvectors [1,10]. The genome sequences for several members of the *An. gambiae* complex are soon to be released [10], and the new chromosomal phylogeny will provide the basis for proposing hypotheses about the evolution of epidemiologically important phenotypes. An intriguing question is whether or not evolution of independently originated traits, such as anthropophily and salt tolerance, is determined by changes of the same genomic loci in different species. In addition, the revised phylogeny will affect the interpretation of results from population genetics studies such as shared genetic variation and the detection of signatures of

selection. Specifically, variations shared with *An. merus* but not with *An. quadriannulatus* would be interpreted now as ancestral. Knowledge about how evolutionary changes related to ecological and behavioral adaptation and how susceptibility to a pathogen in arthropod vectors had happened in the past may inform us about the likelihood that similar changes will occur in the future.

Materials and methods

Mosquito strains and chromosome preparation

The OPHASNI strain of *An. merus*, the Indian wild-type laboratory strain of *An. stephensi*, and the SUA2La strain of *An. gambiae* were used for chromosome preparation. To obtain the polytene chromosomes, ovaries were dissected from half-gravid females and kept in Carnoy's fixative solution (3 ethanol: 1 glacial acetic acid) in room temperature overnight. Follicles of ovaries were separated in 50% propionic acid and were squashed under a cover slip. Slides with good chromosomal preparations were dipped in liquid nitrogen. Then cover slips were removed, and slides were dehydrated in a series of 50%, 70%, 90%, and 100% ethanol.

FISH

Multiple *An. gambiae* DNA probes derived from the cytological breakpoints of *An. gambiae* were physically mapped to the chromosomes of *An. merus* and *An. stephensi*. DNA probes obtained from PCR products were labeled by the Random Primers DNA Labeling System (Invitrogen Corporation, Carlsband, CA), and phage clones were labeled by the Nick Translation Kit (Amersham, Bioscience, Little Chalfont Buckinghamshire, UK). DNA probes were hybridized to chromosome slides overnight at 39°C. Then chromosomes were washed with 1X SSC at 39°C and room temperature. Chromosomes were stained with 1 mM YOYO-1 iodide (491/509) solution in DMSO (Invitrogen Corporation, Carlsbad, CA, USA) and were mounted in

DABCO (Invitrogen Corporation, Carlsbad, CA, USA). Images were taken by a laser scanning microscope and by the fluorescent microscope. Location of the signals was determined by using a standard photomap of *An. stephensi* [43] and *An. gambiae* [44].

Genome sequencing

Mate-paired whole genome sequencing was done on genomic DNA isolated from five adult males and females of *An. merus*. Genomic DNA of *An. merus* was isolated using the Blood and Cell Culture DNA Mini Kit (Qiagen Science, Germantown, MD, USA). Three libraries of 2 kb, 3 kb, and 5 kb were obtained. These libraries were used for 36 bp paired-end sequencing utilizing the Illumina Genome Analyzer IIx at Ambry Genetics Corporation (Aliso Viejo, CA, USA). The 16X coverage genome assembly for *An. stephensi* was obtained by sequencing genomic DNA isolated from Indian wild-type laboratory strain. The sequencing was done using Illumina and 454 platforms at the Core Laboratory Facility of the Virginia Bioinformatics Institute, Virginia Tech.

Phage and BAC library screening

Screening the *An. merus* Lambda DASH II phage library with genes adjacent to standard $2R+^{o}$ and $2R+^{p}$ was performed. To prepare probes for screening phage and BAC libraries, genomic DNA of *An. gambiae* was prepared using the Qiagen DNeasy Blood and Tissue Kit (Qiagen Science, Germantown, MD, USA). Primers were designed for genes adjacent to breakpoints using the Primer3 program [45]. PCR conditions were the following: 95°C for 4 min; 35 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 30 s; and 72°C for 5 min. All PCR products were purified from the agarose gel using GENECLEAN III kit (MP Biomedicals, Solon, OH, USA). DNA probes were labeled based on random primer reaction with DIG-11-dUTP from DIG DNA

Labeling Kit (Roche, Indianapolis, IN, USA). *Anopheles merus* Lambda DASH II phage library and *An. stephensi* BAC library (Amplicon Express, Pullman, WA, USA) were screened. Library screening was performed using the following kits and reagents (Roche Applied Science, Indianapolis, IN) according to protocols supplied by the manufacturer: Nylon Membranes for Colony and Plaque Hybridization, DIG easy Hyb, DIG Wash and Block Buffer Set, Anti-Dioxigenin-AP, and CDP Star ready to use. Positive phages were isolated with Qiagen Lambda midi Kit (Qiagen Science, Germantown, MD, USA), and positive BAC clones were isolated using the Qiagen Large Construct Kit (Qiagen Science, Germantown, MD, USA).

Clone sequencing

Primers 1760RCL (5'AGCAACAGGGACGATTTGTT3') and 2933RCL (5'CTCGCTTTGGTTTGTGCTTT3') were designed based on AGAP001760 and AGAP002933 sequences, and they were used to obtain the distal 2Ro breakpoint from Phage 7D DNA. The PCR conditions with Platinum *PfX* DNA polymerase (Invitrogen, Carlsbad, CA, USA) were: 94°C for 2 min; 35 cycles of 94°C for 15 s, 55°C for 30 s, and 68°C for 2 min; and 68°C for 10 min. Sanger sequencing of Phage 7D was performed using an ABI machine at the Core Laboratory Facility of the Virginia Bioinformatics Institute, Virginia Tech. Other positive phage and BAC clones were completely sequenced by the paired-end approach using an Illumina platform. Libraries of phages and BAC clones were made using Multiplex Sample Preparation Oligonucleotide Kit and Paired End DNA Sample Prep Kit (Illumina, Inc., San Diego, CA). Paired-end sequencing was performed on the Illumina Genome Analyzer IIx using 36 bp paired-end processing at Ambry Genetics Corporation (Aliso Viejo, CA, USA).

Bioinformatics analysis

Phage clone of *An. merus*, BAC clone of *An. stephensi*, and genome sequences *of An. merus, An. stephensi*, *An. gambiae, C. quinquefasciatus*, and *Ae. aegypti* were analyzed with BLASTN, TBLASTX, and BLAST2 using the laboratory server and the Geneious 5.1.5 software (www.geneious.com), a bioinformatics desktop software package produced by Biomatters Ltd. (www.biomatters.com). Identification of the accurate breakpoint was performed by aligning the *An. merus* sequences and *An. gambiae* AgamP3, AgamM1, and AgamS1 genome assemblies available at VectorBase [27]. The DNA transposons and retroelements were analyzed by using the RepeatMasker program [46] and by comparing to Repbase [47] and TEfam (http://tefam.biochem.vt.edu/tefam/) databases. To characterize novel TEs in the breakpoint, each candidate sequence was used as a query to identify repetitive copies in the genome using BLASTN searches. These copies, plus 1000 bp flanking sequences, were aligned using CLUSTAL 2.1 to define the 5' and 3' boundaries. Using this approach, a novel MITE was discovered in the *An. gambiae* breakpoint. According to the TEfam naming convention, this MITE was named Aga_m3bp_Ele1 because it was associated with a 3 bp target site duplication.

Accession numbers

All sequence data have been deposited at the National Center for Biotechnology Information short read archive (www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) as study no. SRP009814 of submission no. SRA047623 and to the GenBank database (http://www.ncbi.nlm.nih.gov/Genbank/) as accession nos.: JQ042681-JQ042688.

Acknowledgements

We thank Nora Besansky for providing the *An. merus* Lambda DASH II phage library and for fruitful discussions, Marco Pombi for useful comments, Maria Sharakhova for help with

chromosome mapping, Melissa Wade for editing the text, and Fan Yang for assistance with BAC clone isolation. Comments provided by two anonymous reviewers helped to improve the manuscript. The *An. gambiae* ND-TAM BAC library, the OPHASNI strain of *An. merus*, and SUA2La strain of *An. gambiae* were obtained from the Malaria Research and Reference Reagent Resource Center (MR4).

Author contributions

Conceived and designed the experiments: IVS. Performed the experiments: MK, AX, ZT, IVS. Analyzed the data: MK, AX, ZT, IVS. Wrote the paper: MK, ZT, IVS.

Figures

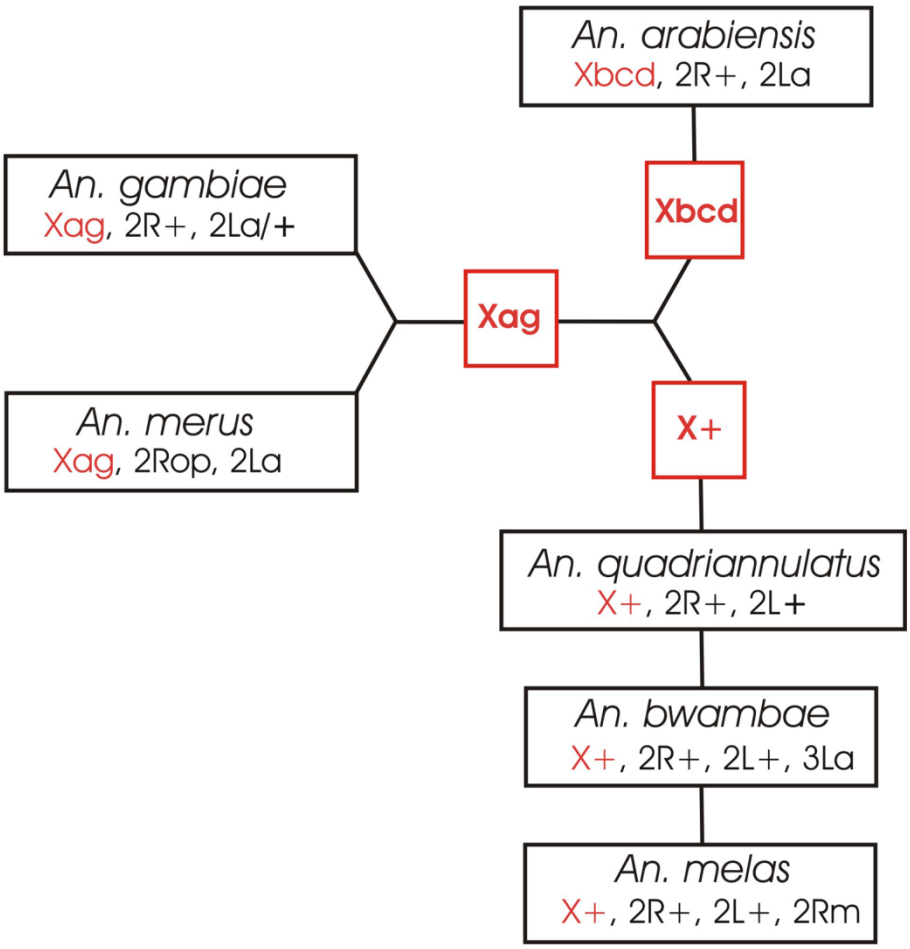

Figure 1. The three species clades identified based on the X chromosome fixed inversions in the *An. gambiae* complex.
The X chromosome arrangements are shown in red.

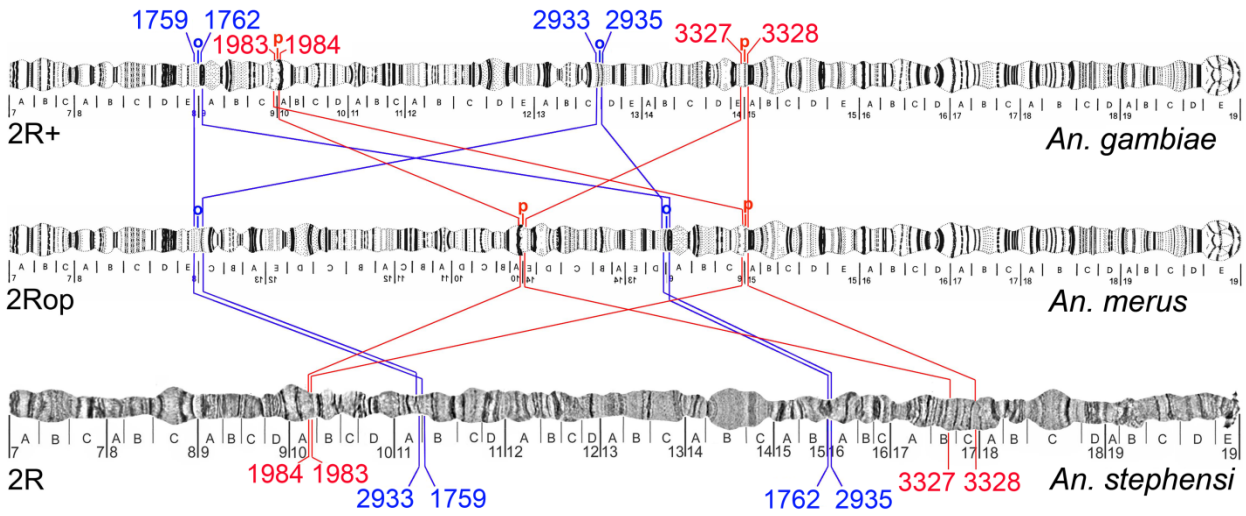

Figure 2. Gene orders in the polytene chromosomes at 2Ro/2R+o and 2Rp/2R+p breakpoints.
Genes of ingroup species *An. merus*, *An. gambiae*, and outgroup species *An. stephensi* are shown on polytene chromosomes. Genes AGAP001759, AGAP001762, AGAP002933, and AGAP002933 of 2Ro/2R+^{o} (in blue), and genes AGAP001983, AGAP001984, AGAP003327, and AGAP003328 of 2Rp/2R+^{p} (in red) are indicated by their last four digits.

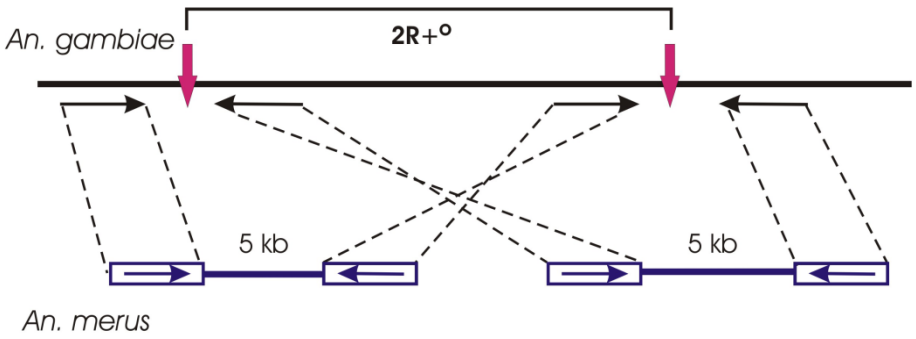

Figure 3. A scheme showing the utility of mate-paired sequencing for identifying inversion breakpoints.
The BLASTN search of *An. merus* mate-paired sequencing reads (horizontal arrows) detects the 2R+^{o} inversion breakpoints (vertical arrows) in the *An. gambiae* AgamP3 genome assembly.

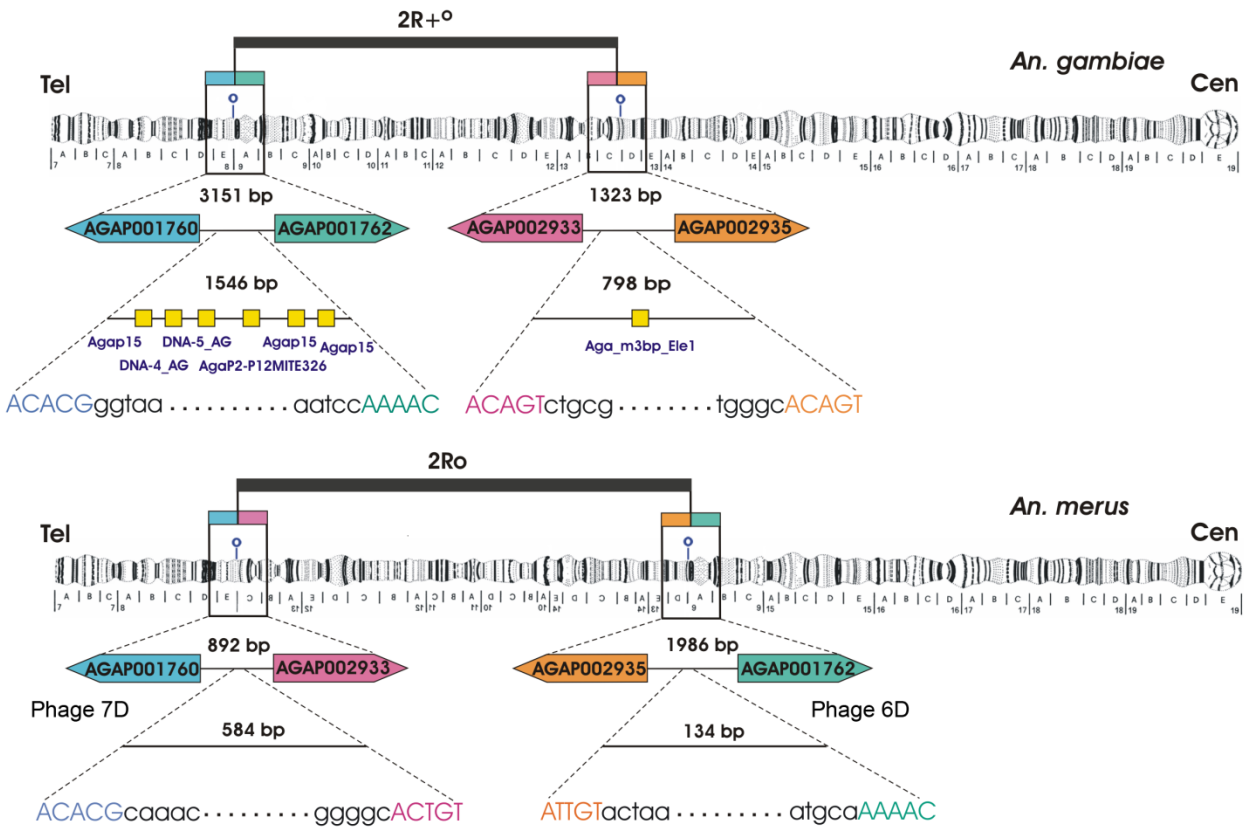

Figure 4. Structure of the 2R+o and 2Ro inversion breakpoint sequences in *An. gambiae* and *An. merus.*

Distal and proximal breakpoints are shown on polytene chromosomes and in the *An. gambiae* genome assembly. Breakpoint sequences are shown with small letters, and their sizes are indicated in base pairs. Genes at the breakpoints are shown in their 5'-3' orientation with boxes of similar colors. Distances between the genes are shown above the intergenic regions. Homologous sequences are represented by identically colored capital letters. Yellow boxes show assemblies of degenerate TEs in *An. gambiae*. The sizes of genes and intergenic regions are not drawn to scale. Cen, centromere. Tel, telomere.

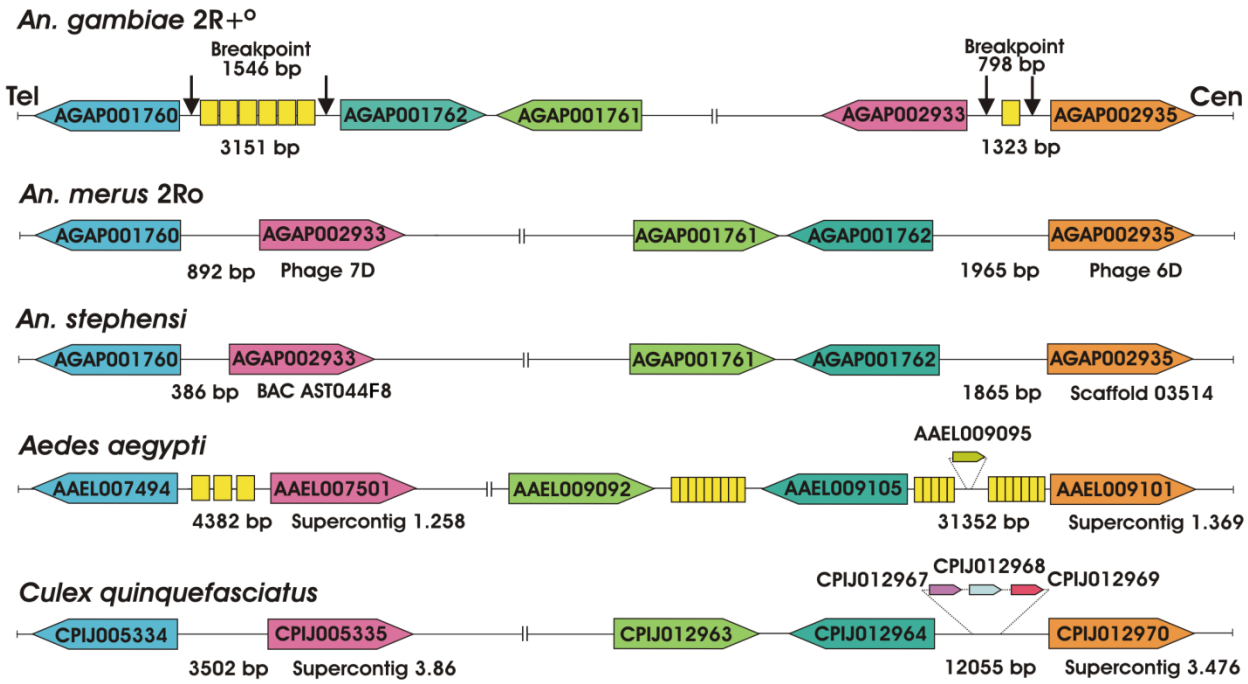

Figure 5. Gene orders in assembled sequences of the 2R+^{o} and 2Ro breakpoints.
Genes of *An. gambiae* and *An. merus* as well as three outgroup species *An. stephensi*, *Ae. aegypti*, and *C. quinquefasciatus* are shown. Breakpoint regions in *An. gambiae* are represented by vertical black arrows with their sizes in base pairs. Homologous genes are show in their 5'-3' orientation with boxes of similar colors. Distances between genes are shown in base pairs and are not depicted proportionally. The correct orientation of genes with respect to the centromere (Cen) and telomere (Tel) is shown only for *An. gambiae*. Additional genes at the breakpoints of *Ae. aegypti* and *C. quinquefasciatus* are shown in a smaller scale. Yellow boxes show assemblies of degenerate TEs. In the *An. gambiae* breakpoints, TEs are shown in the following order from left to right: AgaP15, DNA-4_AG, DNA-5_AG, AgaP2-P12MITE326, AgaP15, AgaP15 (distal breakpoint), and Aga_m3bp_Ele1 (proximal breakpoint). The sizes of genes and intergenic regions are not drawn to scale.

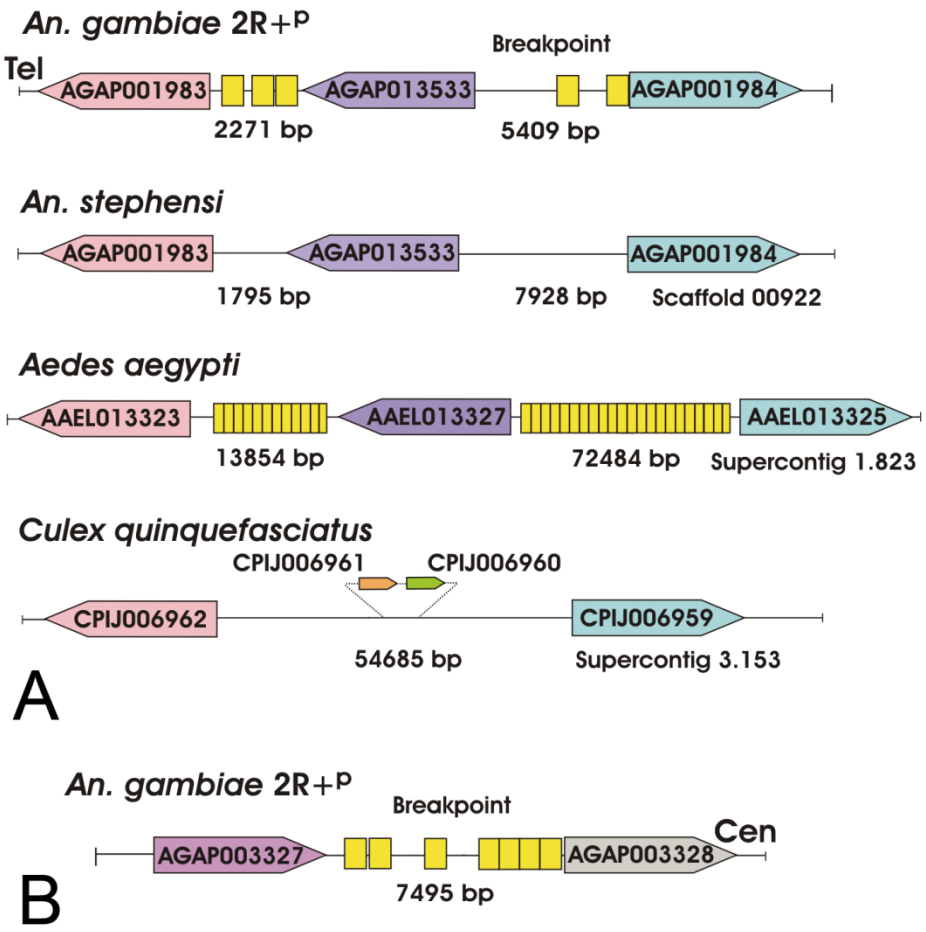

Figure 6. Gene order in assembled sequences of the 2R+^{p} breakpoints.

Genes of *An. gambiae* as well as three outgroup species *An. stephensi*, *Ae. aegypti*, and *C. quinquefasciatus* are shown. (A) The distal 2R+^{p} breakpoint region. Distances between genes are indicated in base pairs, and they are not depicted proportionally. Homologous genes are shown in their 5'-3' orientation with boxes of similar colors. Additional genes at the breakpoint of *C. quinquefasciatus* are shown in a smaller scale. Yellow boxes show assemblies of degenerate TEs. In *An. gambiae*, TEs are shown in the following order from left to right: SINEX-1_AG, P4_AG, SINEX-1_AG, RTE-1_AG, and SINEX-1_AG. (B) The proximal 2R+^{p} breakpoint region. The *An. stephensi*, *Ae. aegypti*, and *C. quinquefasciatus* genes homologous to genes from the proximal 2R+^{p} breakpoint of *An. gambiae* are found in different scaffolds and supercontigs and, therefore, are not shown. In *An. gambiae*, TEs are shown in the following order from left to right: AARA8_AG, CR1-8_AG, Copia-6_AG-LTR, Clu-47_AG, Clu-47_AG, SINEX-1_AG, and Clu-47_AG. The sizes of genes and intergenic regions are not drawn to scale. The correct orientation of genes with respect to the centromere (Cen) and telomere (Tel) is shown only for *An. gambiae*.

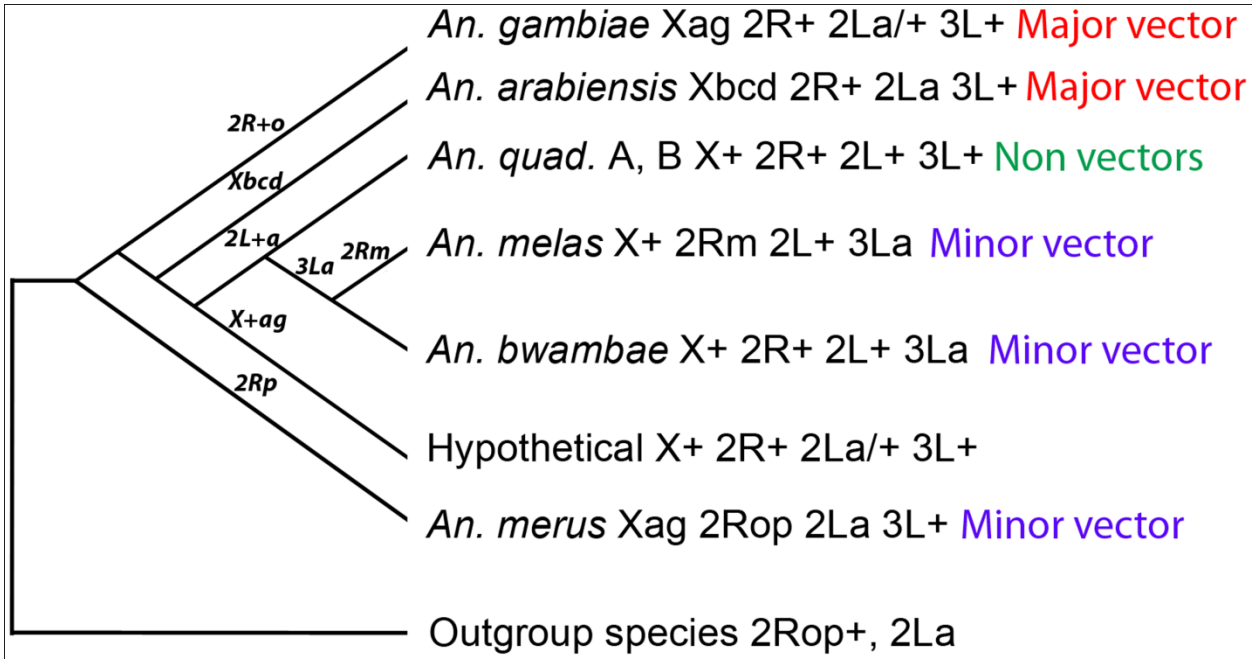

Figure 7. A rooted chromosomal phylogeny of the *An. gambiae* complex.
The phylogeny is based on the ancestry of the 2Ro, $2R+^{p}$, and 2La arrangements found in outgroup species. The vector status for each species is indicated. Inversion fixation events are shown above the branches.

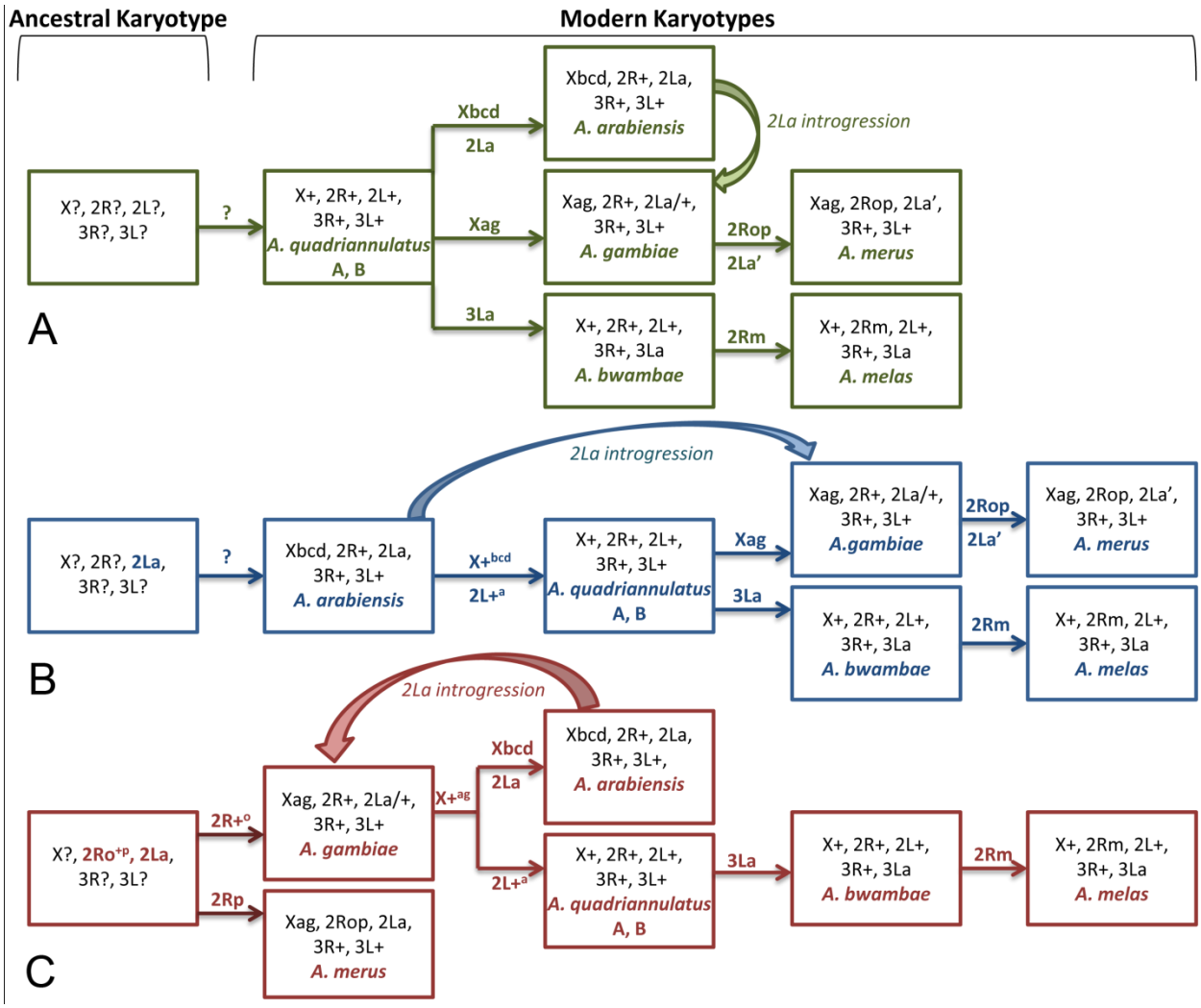

Figure 8. Alternative scenarios of karyotypic evolution in the *An. gambiae* complex.
(A) A chromosomal phylogeny based on the ancestral state of the "standard" karyotype of *An. quadriannulatus* [4,7]. (B) A karyotypic evolution based on the ancestral position of the *An. arabiensis* karyotype inferred from the finding of the fixed 2La inversion in outgroup species [18]. Scenarios A and B assume an independent origin of the 2La' inversion in *An. merus*. (C) A chromosomal phylogeny based on the established ancestry of the shared inversion 2La [20] and arrangements 2Ro and $2R^{+p}$ (this study). The introgression of 2La from *An. arabienstis* to *An. gambiae* is shown in all three scenarios. Inversion fixation events are shown above and below arrows.

Supplementary

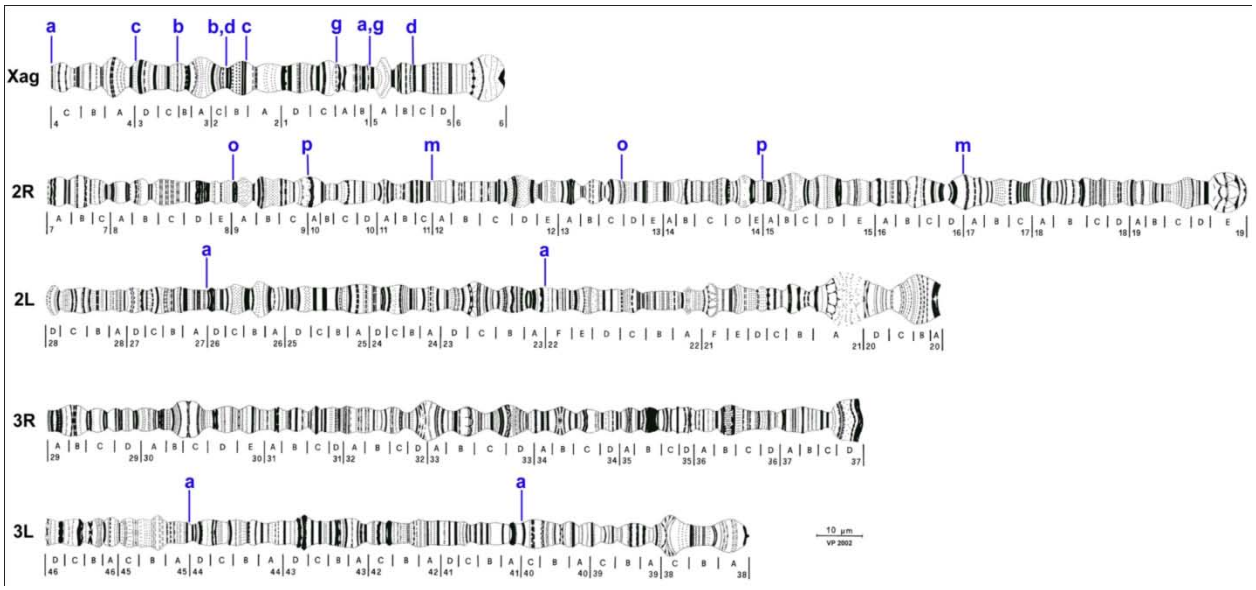

Figure S1. The 10 fixed paracentric inversions in sibling species of the *An. gambiae* complex.
The positions of breakpoints are shown in blue with small letters above the chromosomes.

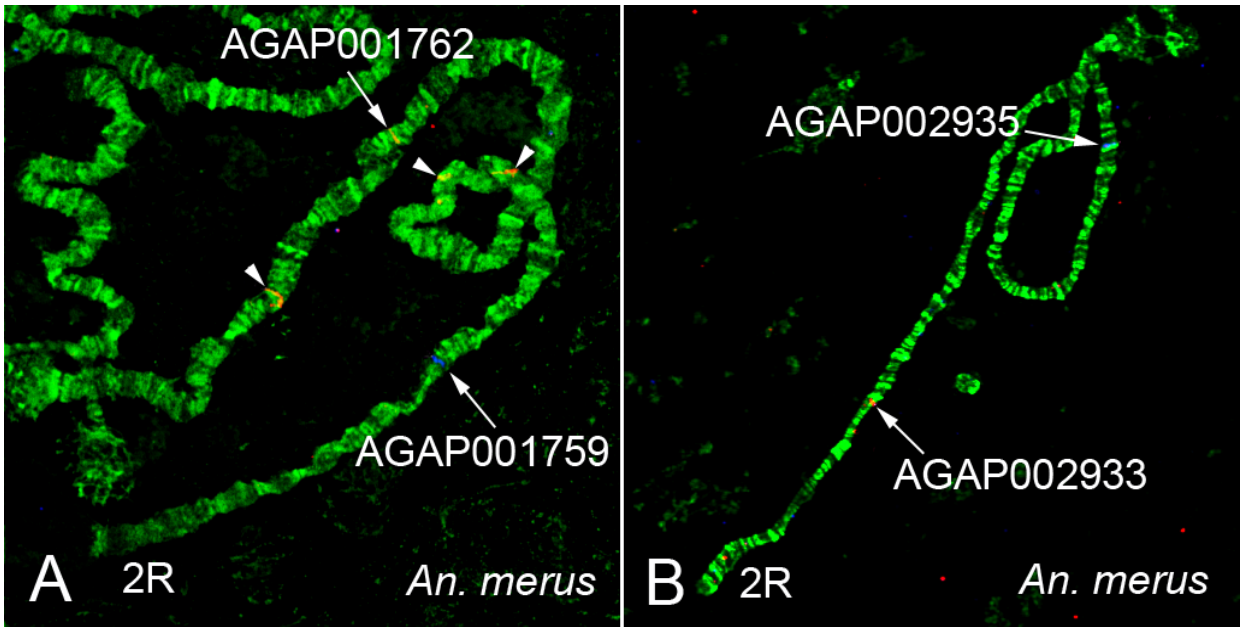

Figure S2. Physical mapping of genes at the 2Ro inversion breakpoints on polytene chromosomes of *An. merus*.

A) FISH of AGAP001759 (blue signal) and AGAP001762 (red signal) to subdivisions 8E and 9A, which are located at the distal and proximal breakpoints, respectively. B) Localization of AGAP002933 (red signal) in the distal breakpoint (13C) and AGAP002935 (blue signal) in the proximal breakpoint (13D). Arrows point at the hybridization signals. Arrowheads show additional signals from AGAP001762. Chromosomes are counterstained with the fluorophore YOYO-1.

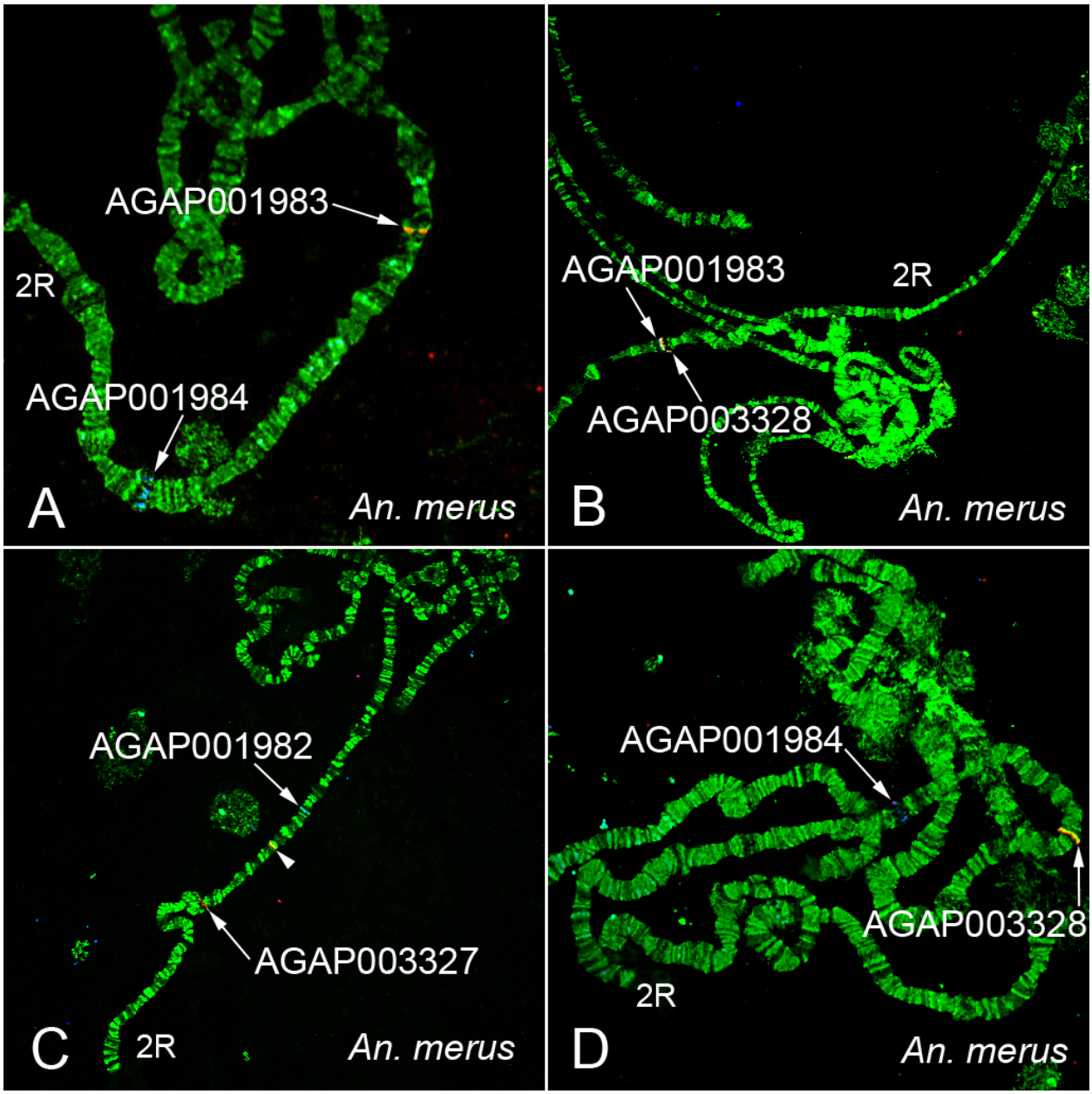

Figure S3. Physical mapping of genes at the 2Rp inversion breakpoints on polytene chromosomes of *An. merus*.

A) FISH of AGAP001983 (red signal) and AGAP001984 (blue signal) to subdivisions 9C and 10A, which are located at the proximal and distal breakpoints, respectively. B) Localization of AGAP001983 (blue signal) and AGAP003328 (red signal) in the neighboring subdivisions 9C and 15A of the proximal breakpoint. C) FISH of AGAP003327 (red signal) with the distal breakpoint (10A) and of AGAP001982, the neighboring gene of AGAP001983, (blue signal) with the proximal breakpoint (9C). D) Mapping of AGAP001984 (blue signal) to the distal breakpoint (14E) and of AGAP003328 (red signal) to the proximal breakpoint (15A). Arrows point at the hybridization signals. Arrowhead shows an additional signal from AGAP003327.

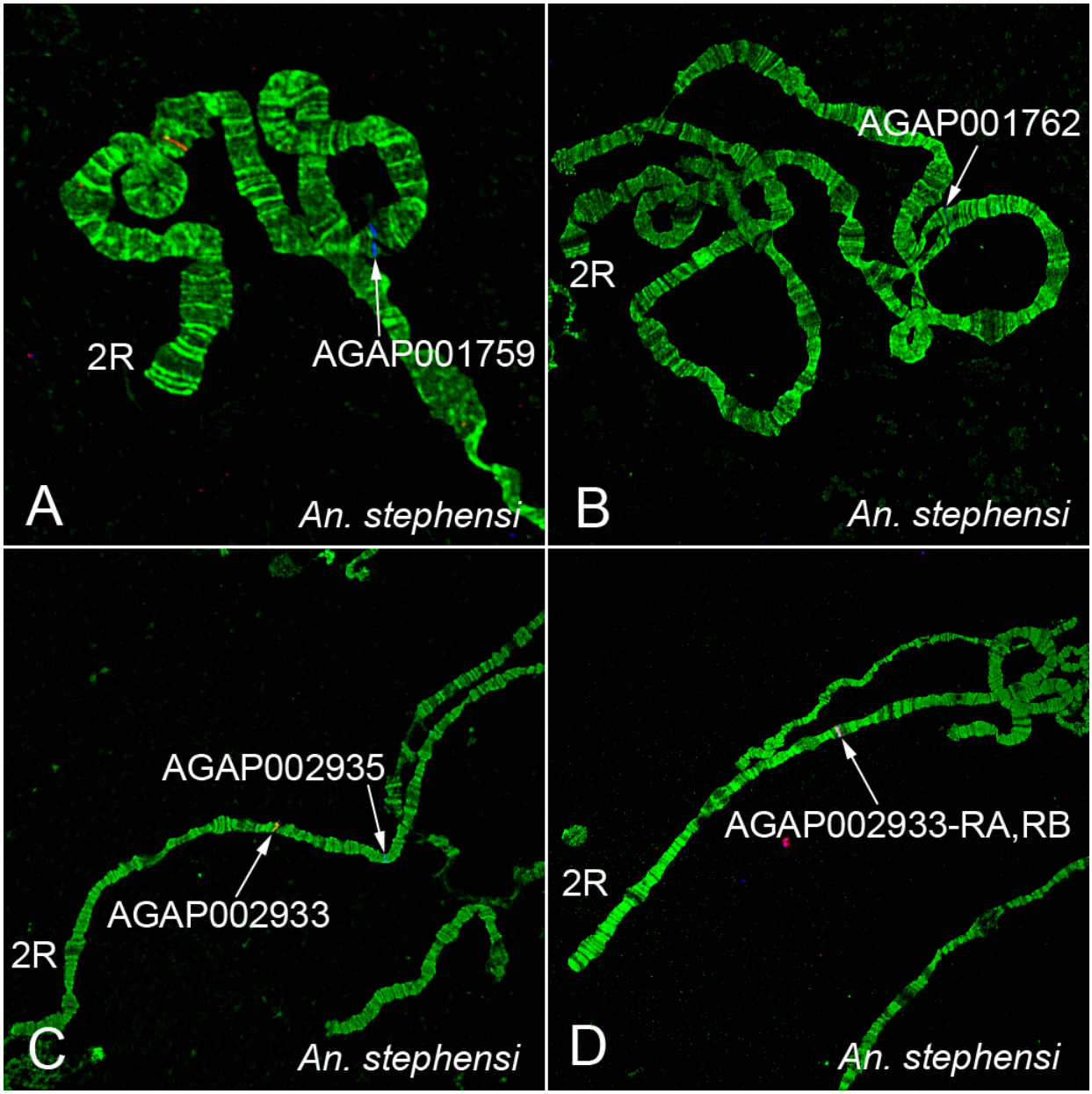

Figure S4. Physical mapping of genes from the 2Ro inversion breakpoints on polytene chromosomes of *An. stephensi.*

A) FISH of AGAP001759 (blue signal) to subdivision 11AB. B) Localization of AGAP001762 (blue signal) in subdivision 15B-16A. C) FISH of AGAP002933 (red signal) with subdivision 11AB and of AGAP002935 (blue signal) in subdivision 15B-16A. D) Colocalization of probes derived from transcripts AGAP002933-RA (red signal) and AGAP002933-RB (blue signal) in subdivision 11AB. Arrows point at the hybridization signals.

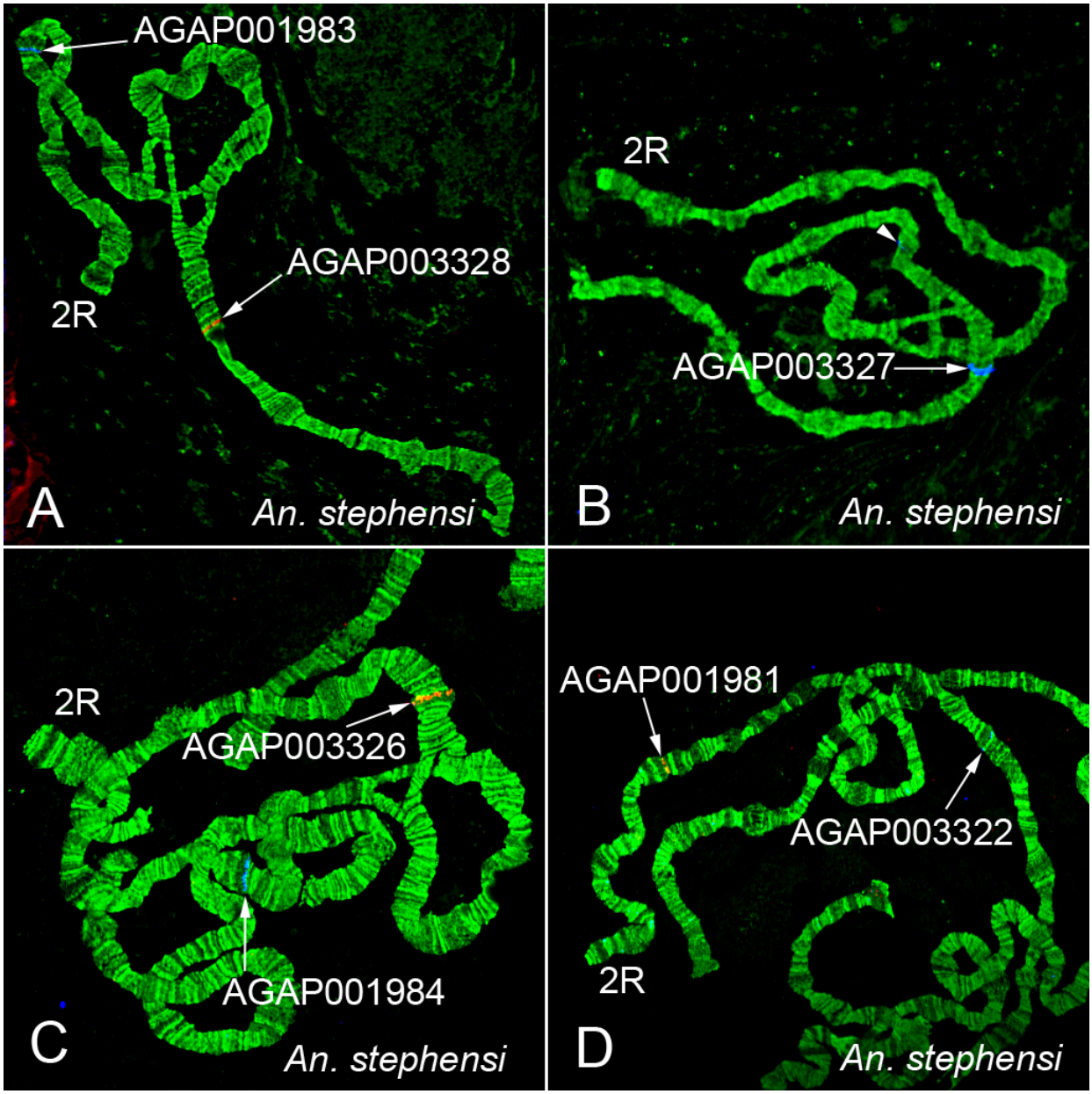

Figure S5. Physical mapping of genes from the 2Rp inversion breakpoints on polytene chromosomes of *An. stephensi.*

A) FISH of AGAP001983 (blue signal) and AGAP003328 (red signal) to subdivisions 10A and 17C, respectively. B) Localization of AGAP003327 (blue signal) in subdivision 17B. C) FISH of AGAP001984 (blue signal) to subdivision 10A and of AGAP003326, the neighboring gene of AGAP003327, (red signal) to subdivision 17B. D) Mapping of AGAP001981, a gene located in the vicinity of AGAP001983, (red signal) in subdivision 10A and of AGAP003322, a gene located in the vicinity of AGAP003327, (blue signal) in subdivision 17B. Arrows point at the hybridization signals. Arrowhead shows an additional minor signal from AGAP003327.

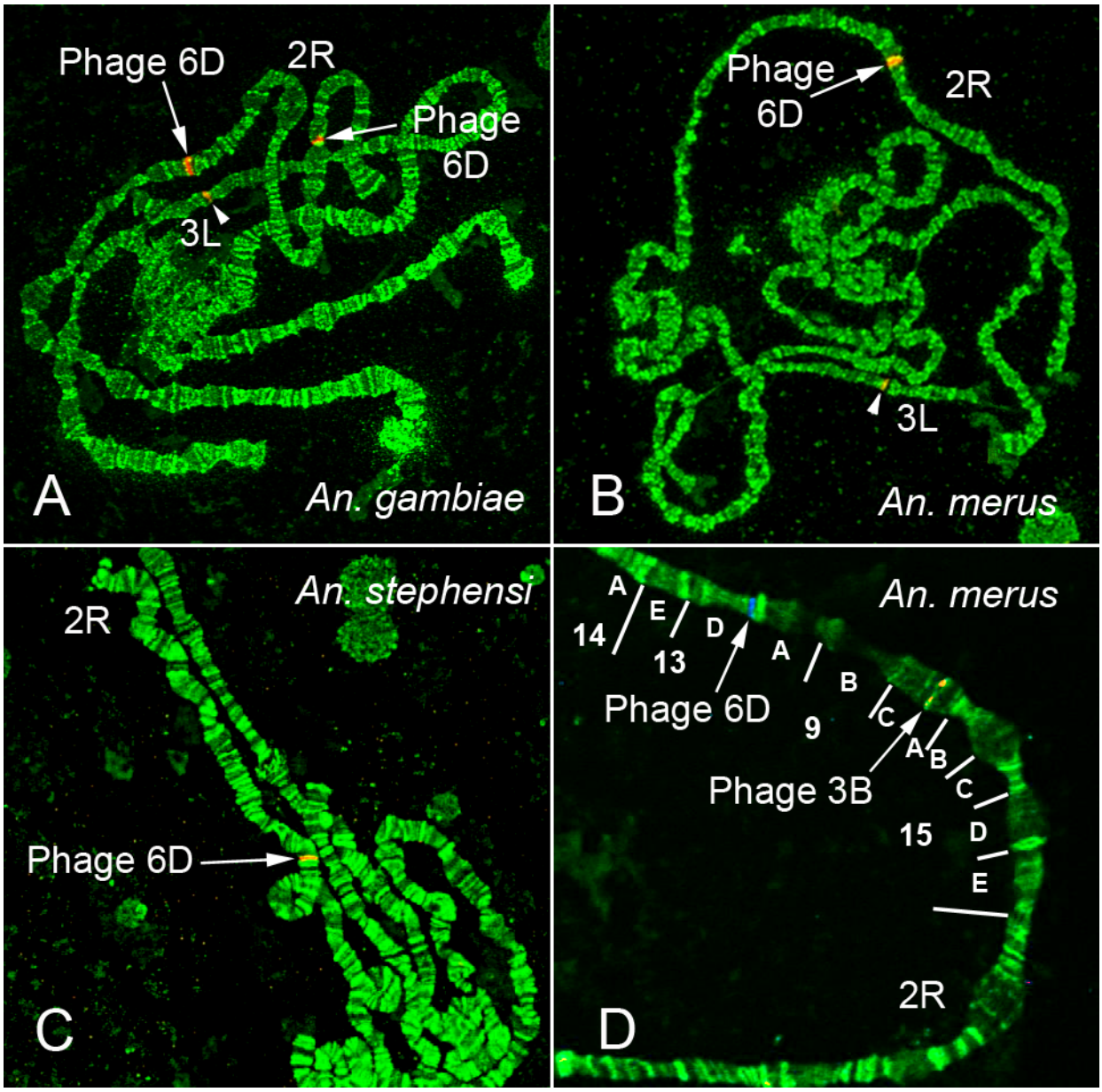

Figure S6. Chromosome mapping of positive phage from the *An. merus* Lambda DASH II phage library.

A) FISH of Phage 6D to both proximal (13D) and distal (9A) 2R+^{o} breakpoints on the 2R arm of *An. gambiae* (red signals). B) Hybridization of Phage 6D to the proximal 2Ro breakpoint (9A/13D) in *An. merus*. C) FISH of Phage 6D to the unique locus 15B-16A on polytene chromosomes of outgroup species *An. stephensi*. D) Detailed mapping of Phage 6D to the proximal 2Ro breakpoint in the region 9A/13D and Phage 3B to the proximal 2Rp breakpoint in the region 9C on a highly polytenyzed chromosome 2R of *An. merus*. Arrowheads show an additional signal on 3L in *An. gambiae* (A) and *An. merus* (B).

Figure S7. Unrooted trees of karyotype evolution in the *An. gambiae* complex recovered by the MGR program.

Each tree includes an outgroup species with different X chromosome arrangements: (A) X+, (B) Xbcd, and (C) Xag indicated with a blue font. The number of rearrangements that occurred on each edge is shown. The names of fixed inversions are shown in parentheses. A7 – A11 are putative intermediate karyotypes. The second origin of 2Ro is highlighted with yellow in (A) and (B).

References

1. Krzywinski J, Besansky NJ (2003) Molecular systematics of *Anopheles*: from subgenera to subpopulations. Annu Rev Entomol 48: 111-139.
2. Adler PH, Cheke RA, Post RJ (2010) Evolution, epidemiology, and population genetics of black flies (Diptera: Simuliidae). Infect Genet Evol 10: 846-865.
3. Yin H, Norris DE, Lanzaro GC (2000) Sibling species in the *Lutzomyia longipalpis* complex differ in levels of mRNA expression for the salivary peptide, maxadilan. Insect Mol Biol 9: 309-314.
4. Coluzzi M, Sabatini A, Petrarca V, Di Deco MA (1979) Chromosomal differentiation and adaptation to human environments in the *Anopheles gambiae* complex. Trans R Soc Trop Med Hyg 73: 483-497.
5. Takken W, Eling W, Hooghof J, Dekker T, Hunt R, et al. (1999) Susceptibility of *Anopheles quadriannulatus* Theobald (Diptera: Culicidae) to *Plasmodium falciparum*. Trans R Soc Trop Med Hyg 93: 578-580.
6. Habtewold T, Povelones M, Blagborough AM, Christophides GK (2008) Transmission blocking immunity in the malaria non-vector mosquito *Anopheles quadriannulatus* species A. PLoS Pathog 4: e1000070.
7. Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V (2002) A polytene chromosome analysis of the *Anopheles gambiae* species complex. Science 298: 1415-1418.
8. Besansky NJ, Powell JR, Caccone A, Hamm DM, Scott JA, et al. (1994) Molecular phylogeny of the *Anopheles gambiae* complex suggests genetic introgression between principal malaria vectors. Proc Natl Acad Sci U S A 91: 6885-6888.
9. Besansky NJ, Krzywinski J, Lehmann T, Simard F, Kern M, et al. (2003) Semipermeable species boundaries between *Anopheles gambiae* and *Anopheles arabiensis*: evidence from multilocus DNA sequence variation. Proc Natl Acad Sci U S A 100: 10818-10823.
10. White BJ, Collins FH, Besansky NJ (2011) Evolution of *Anopheles gambiae* in relation to humans and malaria. Annual Review of Ecology, Evolution, and Systematics, Vol 42 42: 111-132.
11. Hittinger CT, Johnston M, Tossberg JT, Rokas A (2010) Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. Proc Natl Acad Sci U S A 107: 1476-1481.
12. Bhutkar A, Gelbart WM, Smith TF (2007) Inferring genome-scale rearrangement phylogeny and ancestral gene order: a *Drosophila* case study. Genome Biol 8: R236.
13. della Torre A, Merzagora L, Powell JR, Coluzzi M (1997) Selective introgression of paracentric inversions between two sibling species of the *Anopheles gambiae* complex. Genetics 146: 239-244.
14. Gonzalez J, Casals F, Ruiz A (2007) Testing chromosomal phylogenies and inversion breakpoint reuse in *Drosophila*. Genetics 175: 167-177.
15. O'Grady PM, Baker RH, Durando CM, Etges WJ, DeSalle R (2001) Polytene chromosomes as indicators of phylogeny in several species groups of *Drosophila*. BMC Evol Biol 1: 6.
16. Coluzzi M, Sabatini A (1969) Cytogenetic observations on the salt water species, *Anopheles merus* and *Anopheles melas*, of the gambiae complex. Parassitilogia 11: 177-187.
17. Coluzzi M, Sabatini A (1968) Cytogenetic observations on species C of the *Anopheles gambiae* complex. Parassitilogia 10: 156-164.
18. Ayala FJ, Coluzzi M (2005) Chromosome speciation: humans, *Drosophila*, and mosquitoes. Proc Natl Acad Sci U S A 102 Suppl 1: 6535-6542.
19. Caccone A, Min GS, Powell JR (1998) Multiple origins of cytologically identical chromosome inversions in the *Anopheles gambiae* complex. Genetics 150: 807-814.
20. Sharakhov IV, White BJ, Sharakhova MV, Kayondo J, Lobo NF, et al. (2006) Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in the *Anopheles gambiae* complex. Proc Natl Acad Sci U S A 103: 6258-6262.
21. Xia A, Sharakhova MV, Sharakhov IV (2008) Reconstructing ancestral autosomal arrangements in the *Anopheles gambiae* complex. J Comput Biol 15: 965-980.

22. Sharakhova MV, Antonio-Nkondjio C, Xia A, Ndo C, Awono-Ambene P, et al. (2011) Cytogenetic map for *Anopheles nili*: Application for population genetics and comparative physical mapping. Infect Genet Evol 11: 746-754.
23. Caccone A, Garcia BA, Powell JR (1996) Evolution of the mitochondrial DNA control region in the *Anopheles gambiae* complex. Insect Mol Biol 5: 51-59.
24. Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, et al. (2010) Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. Science 330: 86-88.
25. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, et al. (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 316: 1718-1723.
26. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298: 129-149.
27. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, et al. (2009) VectorBase: a data resource for invertebrate vector genomics. Nucleic Acids Res 37: D583-587.
28. Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12: 363-376.
29. 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.
30. Lawniczak MK, Emrich SJ, Holloway AK, Regier AP, Olson M, et al. (2010) Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. Science 330: 512-514.
31. Megy K, Emrich SJ, Lawson D, Campbell D, Dialynas E, et al. (2012) VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. Nucleic Acids Res 40: D729-734.
32. Mathiopoulos KD, della Torre A, Predazzi V, Petrarca V, Coluzzi M (1998) Cloning of inversion breakpoints in the *Anopheles gambiae* complex traces a transposable element at the inversion junction. Proc Natl Acad Sci U S A 95: 12444-12449.
33. Krzywinski J, Grushko OG, Besansky NJ (2006) Analysis of the complete mitochondrial DNA from *Anopheles funestus*: an improved dipteran mitochondrial genome annotation and a temporal dimension of mosquito evolution. Mol Phylogenet Evol 39: 417-423.
34. Bourque G, Pevzner PA (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Res 12: 26-36.
35. Hunt RH, Coetzee M, Fettene M (1998) The *Anopheles gambiae* complex: a new species from Ethiopia. Trans R Soc Trop Med Hyg 92: 231-235.
36. Neafsey DE, Lawniczak MK, Park DJ, Redmond SN, Coulibaly MB, et al. (2010) SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. Science 330: 514-517.
37. White BJ, Cheng C, Sangare D, Lobo NF, Collins FH, et al. (2009) The population genomics of trans-specific inversion polymorphisms in *Anopheles gambiae*. Genetics 183: 275-288.
38. Gray EM, Rocca KA, Costantini C, Besansky NJ (2009) Inversion 2La is associated with enhanced desiccation resistance in *Anopheles gambiae*. Malar J 8: 215.
39. Petrarca V, Beier JC (1992) Intraspecific chromosomal polymorphism in the *Anopheles gambiae* complex as a factor affecting malaria transmission in the Kisumu area of Kenya. Am J Trop Med Hyg 46: 229-237.
40. Pock Tsy JM, Duchemin JB, Marrama L, Rabarison P, Le Goff G, et al. (2003) Distribution of the species of the *Anopheles gambiae* complex and first evidence of Anopheles merus as a malaria vector in Madagascar. Malar J 2: 33.
41. Ridl FC, Bass C, Torrez M, Govender D, Ramdeen V, et al. (2008) A pre-intervention study of malaria vector abundance in Rio Muni, Equatorial Guinea: their role in malaria transmission and the incidence of insecticide resistance alleles. Malar J 7: 194.
42. Mourou JR, Coffinet T, Jarjaval F, Pradines B, Amalvict R, et al. (2010) Malaria transmission and insecticide resistance of *Anopheles gambiae* in Libreville and Port-Gentil, Gabon. Malar J 9: 321.

43. Sharakhova MV, Xia A, McAlister SI, Sharakhov IV (2006) A standard cytogenetic photomap for the mosquito *Anopheles stephensi* (Diptera: Culicidae): application for physical mapping. J Med Entomol 43: 861-866.
44. George P, Sharakhova MV, Sharakhov IV (2010) High-resolution cytogenetic map for the African malaria vector *Anopheles gambiae*. Insect Mol Biol 19: 675-682.
45. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.
46. Smit AFA, Hubley R, Green P (2004) RepeatMasker, version Open-3.0. Available: http://repeatmasker.org/cgi-bin/WEBRepeatMasker. Accessed 29 August 2012.
47. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462-467.

CHAPTER 3 Molecular Phylogenies of Inversion Breakpoints Support Evolutionary History of the Fixed Gene Arrangements in the *Anopheles gambiae* Complex

Abstract

Attempts to construct the molecular phylogeny of *Anopheles gambiae* complex led to conflicting results. A possible reason for these conflicts is chromosomal location of a genetic marker with respect to breakpoints of fixed inversions because genes at the breakpoints are less subject to gene flow. To test this hypothesis we obtained sequences for 12 genetic markers adjacent to breakpoints of three inversions with known evolutionary history: 2Ro, 2Rp and 2La. Distal and proximal 2Ro and 2Rp breakpoint genes of *An. gambiae* complex were amplified by PCR and were sequenced. The 2La breakpoint genes and all sequences for outgroup species *An. stephensi*, *An. nili*, *Aedes aegypti* and *Culex quinquefasciatus* were obtained from GenBank. The sequences were aligned and analyzed in MEGA 5.05 and Geneious 5.1.5 programs. Phylogenetic analyses were performed and the evolutionary histories of distal and proximal breakpoint genes were constructed. The molecular phylogeny obtained from breakpoint genes is in agreement with the chromosomal phylogeny of the *An. gambiae* complex. The molecular phylogeny of 2Ro breakpoint genes confirms the ancestry of *An. merus* 2Ro arrangement and placed *An. merus* in a separate clade closer to outgroup species. In contrast, the molecular phylogeny of 2Rp breakpoint genes placed *An. gambiae* PEST strain with standard ancestral $2R+^{p}$ arrangement in a separate clade closer to the outgroup species. The molecular phylogeny of 2La breakpoint genes confirmed that *An. merus* with inverted 2La arrangement is ancestral. In addition, we confirmed the ancestry of $2R+^{p}$ arrangement with outgroup species, *An. nili*. We conclude that the knowledge about the position of the genetic markers with respect to chromosomal inversion is crucial for constructing phylogenetic trees.

Introduction

Gene order data can be used to reconstruct phylogeny in different organisms. In spite of availability of molecular markers and sequencing the genome of *An. gambiae* [1], reconstructing phylogeny based on molecular markers is complicated because of high degree of genetic similarity, caused by ancestral polymorphism and introgression. Molecular phylogeny in *An. gambiae* complex was performed based on mitochondrial DNA, rDNA and other gene markers [2,3,4]. However, phylogenies inferred from individual genes and DNA markers led to conflicting results mostly due to high levels of sequence similarity and gene flow among members of the complex. For example, phylogeny based on molecular data are incongruent with phylogeny obtained from inversion data [3,5]. Molecular phylogeny based on the nuclear rDNA intergenic spacer sequence, which are X chromosome linked but are outside of the breakpoints, and mitochondrial DNA sequence (mtDNA), places *An. gambiae* and *An. arabiensis* as sister taxa and suggest introgression among them [3]. However, this phylogeny contradicts with the phylogeny based on the fixed inversion, which places *An. gambiae* and *An. merus* as sister taxa [3,6,7,8]. In another study an anchored restriction mapping technique was used to study different regions in genome, and phylogenetic analysis based on the analysis of mtDNA, a rDNA, and an anonymous single-copy nuclear DNA locus (scnDNA), placed *An. gambiae*, *An. arabiensis*, *An. melas*, and *An. merus* as sister taxa [5], however the analysis provided a mixed information of nuclear and mtDNA and based on RFLP, and not the direct sequences. Phylogeny based on sequences inside the Xag inversion in *An. gambiae* complex indicated that *An. gambiae* and *An. merus* are sister taxa [9]. Moreover, analysis of AT-rich region of mtDNA of six species in *An. gambiae* complex as well as closely related African species, *An. christyi*, placed *An. gambiae* as a sister taxa to *An. arabiensis* [2].

In a study on the pattern of mitochondrial variation between *An. gambiae* and *An. arabiensis*, an extensive gene flow was observed [10]. Phylogeny result based on sequencing of four DNA regions between or very close to breakpoints of 2La inversion in the *An. gambiae* complex placed *An. gambiae* and *An. merus* as sister taxa [11]. *An. arabiensis* and *An. gambiae* are sister clades in the phylogenetic tree inferred from analysis of satellite DNA (stDNA) of the Y chromosome [12]. A population genetic approach was carried out to understand the evolution of genes for *Plasmodium* resistant or susceptibility in *An. gambiae*. However, no evidence of strong selection was observed [13]. However, in a study of molecular evolution of four antimalaria immune genes in the *An. gambiae* complex, introgression and the sharing of ancestral polymorphisms was verified in all four studied genes [14].

In general most of the molecular phylogenies agree on the sister taxa relationship between *An. gambiae* and *An. arabiensis*. In a recent phylogenetic tree based on sequences obtained from ribosomal second internal transcribed spacer (ITS2) and the mitochondrial CO1 loci, novel vectors of malaria in western highland of Kenya were detected. According to this phylogeny, *An. merus* and *An. gambiae* are more closely related to each other compared to *An. gambiae* and *An. arabiensis* [15].

Inversions have an important role in speciation because they can suppress recombination in heterozygotes and can be a postzygotic barrier by reducing the heterozygotes fecundity [16]. They can link and protect co- adapted alleles and protect them from recombination. For example species specific loci are mapped inside the inversions and inverted regions are suppressed from recombination [17,18,19]. Inversions can also be spread in natural populations because they protect favorable alleles from recombination, as a result inversions are involved in some important ecological adaptations, for example an increase in the ability to tolerate aridity in

Anopheles species [7]. Genes at inversion breakpoints are less subject to gene flow and as a result they are good candidates to construct molecular phylogeny.

There have been phylogenetic studies based on gene arrangements in *Drosophila pseudoobscura* [20,21,22]. Genes that were analyzed in these studies were mostly selected from the central regions of inverted arrangements. Therefore, the phylogenetic relationship might not be accurate because of the high frequency of genetic exchange in the central region of inverted arrangements, compared to breakpoint regions which have the lowest amount of genetic flow [23,24]. Later, a phylogenetic tree based on polymorphic gene arrangement on the third chromosome of *Drosophila pseudoobscura* was constructed [25]. Genes at the inversion breakpoints have been sequenced and used for phylogenetic analysis, because breakpoint genes are expected to have lower levels of genetic introgression compared to genes away from breakpoint regions [25].

In *An. gambiae* complex, no phylogenetic tree was constructed based on the genes adjacent to the breakpoints . In this study we aim to determine the molecular phylogeny of genes located at breakpoints of fixed 2Ro, 2Rp and 2La inversions, and our results revealed that the molecular phylogeny based on breakpoint genes are in agreement with chromosomal phylogeny. Moreover, we intend to elucidate how position of the genes in respect to chromosomal rearrangement, specifically inversions, would affect the accuracy of the molecular phylogeny data. The ancestry of $2R+^{p}$ standard arrangement is also tested and confirmed by analyzing the breakpoint region in outgroup species, *An. nili*.

Materials and methods

Mosquito strains and DNA extraction

Genomic DNA of *An. gambiae* SUA strain, *An. arabiensis* DONDOLA strain, *An.merus* OPHASNI strain and *An. quadriannualus* SKUQUA strain, were extracted using DNeasy blood and Tissue Kit (Qiagen, Maryland, USA). Eight genes adjacent to breakpoint of 2Ro and 2Rp in four members of *An. gambiae* complex were amplified.

PCR primer design

Primers were designed based on the conserved 2Ro and 2Rp exon regions using the Primer3 program [26]. These primers were used to amplify ~500 bp fragments of eight breakpoint genes in members of *An. gambiae* complex (Table 1).

PCR amplification

Breakpoint genes of 2Ro, AGAP001760, AGAP001762, AGAP002933, AGAP002935 and 2Rp, AGAP013533, AGAP001984, AGAP003327 and AGAP003328 were amplified using high-fidelity Platinum® *PfX* DNA polymerase (Invitrogen, Carlsbad, CA, USA). Each PCR reaction was a total of 50 ul in volume containing 10X *PfX* amplification buffer, 50mM $MgSO_4$, 10mM dNTP mix, forward and reverse primers, distilled water, isolated template DNA and *Pfx* DNA polymerase. The PCR conditions were: 94°C for 2 min; 35 cycles of 94°C for 30 s, 55°C for 30 s, and 68°C for 1 min; and 68°C for 10 min. After the PCR, about 5 ul of PCR products were run on 1% gel to verify the PCR fragment.

DNA purification and sequencing

DNA samples obtained by PCR were run on 1.5% gel electrophoresis. PCR fragments of 2Ro breakpoints AGAP001760, AGAP001762, AGAP002933, AGAP002935, and 2Rp breakpoints AGAP013533, AGAP001984, AGAP003327, AGAP003328 were cut from gel and purified using Qiaquick Gel Extraction Kit (Qiagen Science, MD, USA). The concentrations of purified

DNA samples were measured by Nanodrop 200c spectrophotometer. (Thermo Scientific, DE, USA). 10 ul of PCR product with concentration of 10 ng/ul were used for sequencing (100ng of purified DNA for each sample). Primers were supplied at 5 pmol/ul, and 5 ul of primers were used per reaction. Sanger sequencing was performed using an ABI machine at core laboratory facility at Virginia Bioinformatics Institute (VBI, Blacksburg, VA). Samples were sequenced in both forward and reverse directions. The sequences obtained for analysis in this study were submitted to the GenBank database.

BLAST searches

Breakpoint sequences of different forms within *An. gambiae* including *An. gambiae* M, *An. gambiae* S and *An. gambiae* PEST were obtained online from VectorBase [27]. The 2La breakpoint genes and all sequences for outgroup species *An. stephensi, An. nili, Aedes aegypti* and *Culex quinquefasciatus* were obtained from GenBank and VectorBase. Homologous sequences for additional outgroup species, *An. nili* were obtained from a low coverage assembly of *An. nili* genome using Basic Local Alignment Search Tool (BLAST) in Geneious Pro 5.1.5 software. Homologous sequences of *An. stephensi* were acquired from genome assembly.

DNA sequences alignment and phylogenetic analysis

Consensus sequences from forward and reverse coverage reads were made for each sequenced samples using Geneious Pro 5.1.5 software (www.geneious.com). Consensus nucleotide sequences and all sequences obtained from VectorBase, in addition to outgroup species sequences, were aligned and analyzed in Geneious Pro 5.1.5. Quality of sequence reads were examined manually and all the sequences were imported to Molecular Evolutionary Genetics Analysis (MEGA) 5.05 [28]. Accurate multiple alignments are critical in order to obtain the

correct phylogeny tree. Alignments were performed by adding the most closely related species, followed by adding the outgroup species. All sequences were selected and aligned using the ClustalW alignment option in the MEGA 5.05 program.

Phylogenetic trees for each breakpoint were constructed separately. Each phylogenetic tree was constructed by the neighbor joining method and from 1000 bootstrap replicates. In order to obtain a cumulative tree, sequence alignments of all four genes were transferred into MEGA software, end of sequences were trimmed and a concatenated tree was created by the neighbor joining method**.** Confidence values for each clade were generated by 1000 bootstrap replicates. Bootstrap values are shown on branches of phylogeny trees as percentages.

Results and discussion

To reconstruct the molecular phylogeny, sequencing data from 2Ro, 2Rp breakpoints of *An. gambiae, An. arabiensis*, *An. merus* and *An. quadriannulatus* were amplified (Figure 1). Sequence analyses include ~500 bp of breakpoint genes (Table 1). Additional sequences of different forms within *An. gambiae*, *An. gambiae* S, *An. gambiae* M and *An. gambiae* PEST strain, as well as homologous sequences of *Aedes aegypti* and *Culex quinquefasciatus*, were obtained from VectorBase [27]. Homologous sequences of breakpoint genes from *An. stephensi* and *An. nili* were obtained from genome assembly. Nucleotide sequences for 2La breakpoint genes were collected from GenBank. Nucleotide alignments and phylogenetic trees were generated for each of the breakpoint genes separately.

Ancestral gene order at the 2R+^{p} proximal breakpoint

In our previous study, we performed the comparative analysis of gene orders at the 2R+^{p} breakpoints in three outgroup species, *An. stephensi*, *Cx. quinquefasciatus*, and *Ae. aegypti*. The

results demonstrated the common organization of the distal $2R+^{p}$ breakpoint in *An. gambiae* and the outgroup species, indicating that this arrangement is ancestral [29]. However, the organization of the proximal $2R+^{p}$ breakpoint in outgroup species could not be established at that time because genes similar to AGAP003327 and AGAP003328 were found in different scaffolds and supercontigs of the outgroup species. An additional inversion in *An. stephensi* and highly fragmented genome assemblies in *Cx. quinquefasciatus* and *Ae. aegypti* could explain the observed pattern. Here, we analyzed gene order at the proximal $2R+^{p}$ breakpoints in a new outgroup species, *Anopheles nili*. We found that the order of AGAP003327 and AGAP003328 is preserved within the same supercontig of *An. nili* (Figure 2). Therefore, both distal and proximal $2R+^{p}$ breakpoint arrangements can be found in outgroup species. This finding also indicates that the inversion, which separates AGAP003327 and AGAP003328, is not common to all Anopheline species but was likely occurred in the *An. stephensi* lineage.

Molecular phylogeny of 2Ro breakpoint genes supports the ancestral status of the 2Ro arrangement

In the *An. gambiae* complex, *An. merus* carries the fixed 2Ro inversion which is ancestral [30]. In three phylogenetic trees based on AGAP001760, AGAP001762, and AGAP002933 genes, *An. merus* with the ancestral 2Ro arrangement is clustered separately from the rest of the *An. gambiae* complex (Figure 3, A, B and C). According to this phylogeny, *An. stephensi* is a closer outgroup species to *An. gambiae* complex, followed by *An. nili*, compared to *Ae. aegypti* and *Cx. quinquefasciatus*.

The phylogenetic tree in AGAP002935 (Figure 3, D) is not in agreement with rest of the 2Ro breakpoint genes and places *An. merus* as a more derived species and *An. gambiae*-M closer to outgroup species, *An. stephensi*. This could be due to nature of phylogeny tree construction

based on individual genes. Although most of the phylogeny trees according to breakpoint genes are in agreement with each other, phylogeny based on one breakpoint gene behaves differently, which could be due to high level of sequence similarities within the *An. gambiae* complex.

It has been shown that concatenation of trees provides a better support [31]**.** Phylogenetic trees obtained from concatenated sequences of the 2Ro inversion breakpoints are in agreement with the trees obtained from individual genes of AGAP001760, AGAP001762, AGAP002933, and in addition it provides a robust support for ancestral status of 2Ro arrangement (Figure 4). Based on the concatenated 2Ro phylogeny tree, *An. merus* is clustered separately from the rest of the *An. gambiae* complex, and is placed closer to outgroup species *An. stephensi*, which also has the inverted 2Ro arrangement (Figure 4). Other members of the *An. gambiae* complex with $2R+^{o}$ standard arrangement cluster together, and are considered more derived, confirming the ancestry of the 2Ro arrangement**.**

Molecular phylogeny of 2Rp breakpoint genes supports the ancestral status of the $2R+^{p}$ arrangement

Figure 5 shows the separate phylogenetic trees based on 2Rp breakpoints. According to phylogenetic tree of two breakpoint genes, AGAP013533 and AGAP003328 (Figure 5, A and D), *An. gambiae*-M and *An. gambiae*-S with $2R+^{p}$ standard arrangements are clustered separately and closer to outgroup species *An. stephensi.* Moreover, in both phylogenies *An. merus* with the inverted 2Rp arrangement is found as a more derived species. In the phylogeny of the AGAP003327 gene, out of four available species, *An. gambaie* SUA, *An. quadriannulatus, An. arabiensis*, and *An. merus*, only the breakpoint fragment from *An. quadriannulatus* gDNA was amplified by PCR, therefore the rest of species were not available to obtain a phylogeny tree for the AGAP003327 gene. However, sequences from all other species which were accessible on

Vectorbase have been used for creating the tree. Accordingly, *An. gambiae* M form is clustered separately from the rest of the forms and *An. quadriannulatus,* and *An. stephensi* is also a closer outgroup species compared to *An. nili*. Similar to one of the genes in the 2Ro breakpoint which was contradictory to the rest of the phylogeny trees, the phylogeny based on AGAP001984 in the 2Rp inversion, placed *An. merus* as a more ancestral species.

Figure 6 shows the concatenated phylogenetic tree obtained from four 2Rp breakpoint genes. Based on the previous findings, the standard $2R+^{p}$ arrangement is ancestral and *An. merus* with an inverted 2Rp has a more derived arrangement [29]. Phylogeny based on combined 2Rp breakpoint genes places *An. gambiae*-PEST strain in a separate clade. *An. gambiae*-PEST contains the ancestral standard $2R+^{p}$ arrangement and is placed closer to outgroup species *An. stephensi* which also has a standard $2R+^{p}$ arrangement. On the other hand, *An. merus* with an inverted 2Rp arrangement is considered a more derived species in this phylogeny, which is in agreement with previous findings [29]. *Ae. aegypti* and *Cx. quinquefasciatus* are placed as a more distance outgroup species.

Molecular phylogeny of 2La breakpoint genes supports the ancestral status of the 2La arrangement

Figure 7, shows the separate phylogeny tree based on 2La breakpoint genes, AGAP005778, AGAP005779, AGAP007068, and AGAP007069 (Figure 7, A, B, C, D). Phylogeny trees according to three breakpoint genes, AGAP005778, AGAP007068, AGAP007069, places *An. merus* with 2La inverted arrangement closer to outgroup species. However, phylogeny based on AGAP005779 gene was in contradiction to other breakpoint genes, and considers *An. gambiae* PEST strain a closer species to outgroup species. Homologous sequences of outgroup species *Cx. quinquefasciatus,* and *Ae. aegypti* were obtained. According to phylogenetic tree of

AGAP005778, AGAP007068, and AGAP007069 genes, *Cx. quinquefasciatus,* and *Ae. aegypti* are clearly considered a more distant outgroup species. However, in AGAP005779 gene, homologous sequences obtained from these two outgroups were too short (about 138-150bp) and did not yield to a correct alignment, as a result we did not include them in the phylogeny tree.

A phylogeny based on sequences located inside 2La inversions indicate that the *An. gambiae* 2La/a arrangement clusters together with the *An. arabiensis* 2La/a arrangement. However, the *An. gambiae* 2L+/+ arrangement clusters together with the *An. merus* 2La/a arrangement [11]. According to the chromosomal phylogeny, *An. merus—An. gambiae* clade is ancestral [29], therefore, introgression of 2La had been happening from the more derived *An. arabiensis* to the more ancestral *An. gambiae*. A more recent study also supports the hypothesis of the introgression from *An. arabiensis* to *An. gambiae* [32]. According to our 2La breakpoint gene molecular phylogeny, *An. merus* has a more ancestral 2La arrangement, followed by *An. gambie* PEST with $2L+^{a}$ standard arrangement. 2La in *An. gambiae* SUA is considered more derived compared to other members of *An. gambiae* complex, and it agrees with the hypothesis of 2La introgression from the more derived *An. arabiensis* to the more ancestral *An. gambiae* species.

Molecular phylogeny based on concatenated 2La breakpoint genes is presented in Figure 8. Based on this phylogeny tree, *An. merus* is a more ancestral species, which posses the inverted 2La inversion. *An. gambiae* PEST strain with the standard $2L+^{a}$ arrangement is then branched out followed by *An. arabiensis* with 2La arrangement, and *An. gambiae* SUA. Molecular phylogeny of concatenated 2La breakpoint genes confirms the ancestry of 2La arrangement.

Conclusion

In this study, for the first time, we have used breakpoint genes of fixed inversions to construct a molecular phylogeny in the *An. gambiae* complex. Several outgroup species have been included in this study. We have chosen genes adjacent to three fixed inversions, 2Ro and 2Rp inversions that are specific to *An. merus* and 2La inversion which is found in *An. gambiae*, *An. merus* and An. *arabiensis*. Moreover, we have analyzed gene order at the proximal $2R+^{p}$ breakpoints in a new outgroup species, *Anopheles nili*, and confirmed the ancestry of $2R+^{p}$ breakpoint arrangements.

Molecular phylogeny data are usually in contradictory with inversion phylogeny [3,5]**.** However, our molecular phylogeny data analysis reveals that molecular phylogeny of breakpoint genes are in agreement with chromosomal phylogeny. Since frequency of genetic exchange at the breakpoints are very low [23], genetic markers selected from these regions would be good candidates for constructing molecular phylogeny and inferring the evolutionary history among species**.** We are suggesting that the position of the genes used for molecular phylogeny is very crucial in respect to structural rearrangement in the genome such as inversions. Overall, genes that are used in this study are in agreement with chromosomal phylogeny and it proves that breakpoint genes are more intact and less subject to gene flow. We conclude that chromosomal position of genetic markers with respect to inversion breakpoints must be considered, as they may reflect the chromosomal phylogeny.

Figures

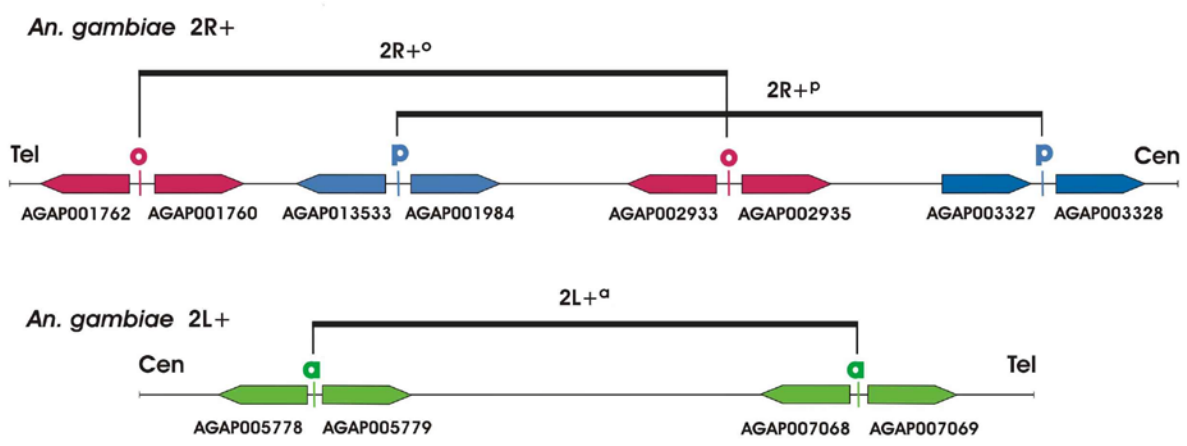

Figure 1. Schematic representation of genetic markers with respect to breakpoints of fixed inversions in *An. gambiae*.

Genes of each inversion are shown with different colors in their 5'-3' orientation. The orientation of genes with respect to the centromere (Cen) and telomere (Tel) is shown for the *An. gambiae* AgamP3 genome assembly.

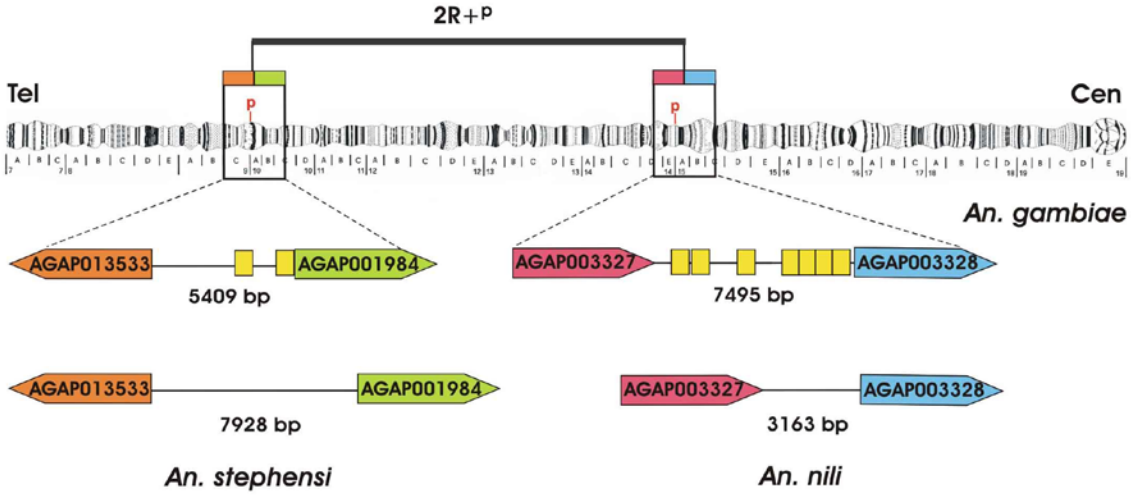

Figure 2. Gene order in assembled sequences of the 2R+^{p} breakpoints.

Genes of *An. gambiae* as well as two outgroup species *An. stephensi* and *An. nili* are shown. Breakpoint regions in *An. gambiae* are represented by vertical black arrows with their sizes in base pairs. Distances between genes are indicated in base pairs, and they are not depicted proportionally. Homologous genes are shown in their 5'-3' orientation with boxes of similar colors. Yellow boxes show assemblies of degenerate TEs. In the *An. gambiae* distal breakpoint, TEs are shown in the following order from left to right: RTE-1_AG, and SINEX-1_AG. In the *An. gambiae* proximal breakpoint, TEs are shown in the following order from left to right: AARA8_AG, CR1-8_AG, Copia-6_AG-LTR, Clu-47_AG, Clu-47_AG, SINEX-1_AG, and Clu-47_AG. The sizes of genes and intergenic regions are not drawn to scale. The correct orientation of genes with respect to the centromere (Cen) and telomere (Tel) is shown only for *An. gambiae*.

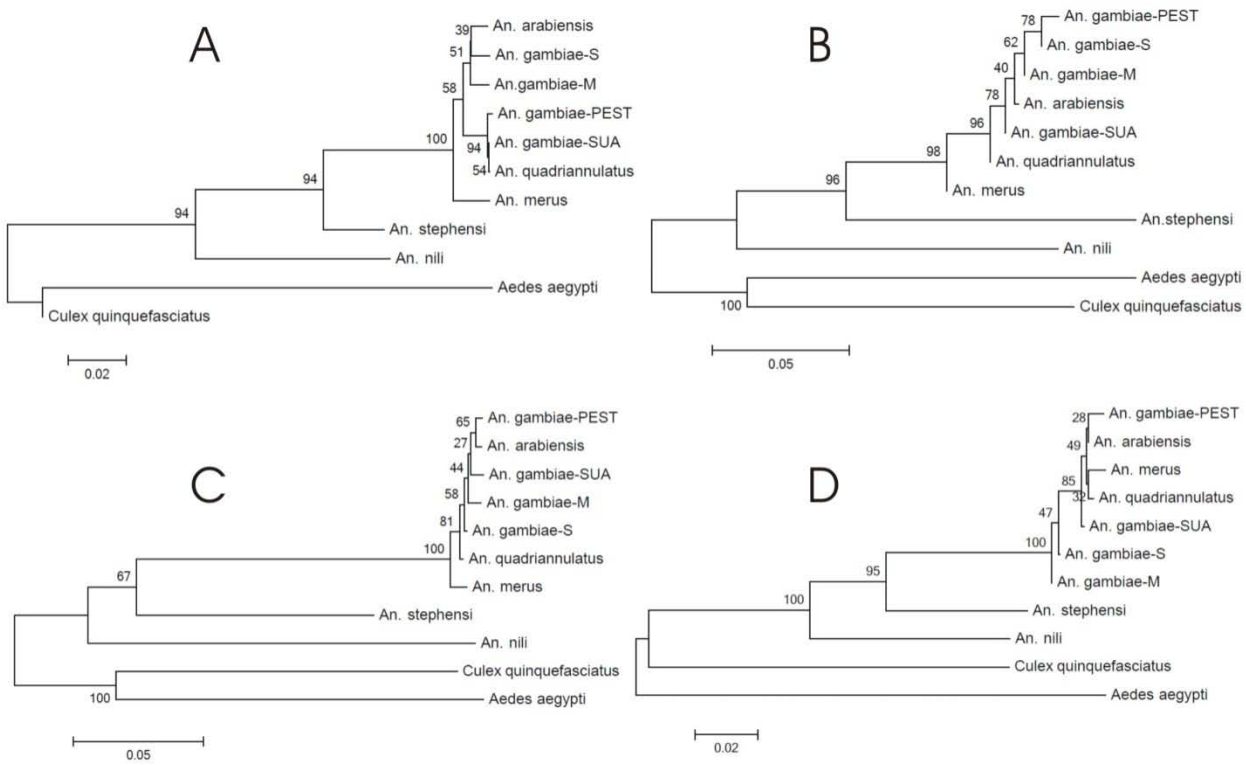

Figure 3. Phylogenetic trees of 2Ro breakpoint genes inferred from 1000 replicate bootstrap consensus tree based on neighbor-joining statistical method.

Numbers on branches represent bootstrap values. Scale bars show a number of amino acid substitutions per site. (A) AGAP001760 (B)AGAP001762 (C) AGAP002933 (D) AGAP002935.

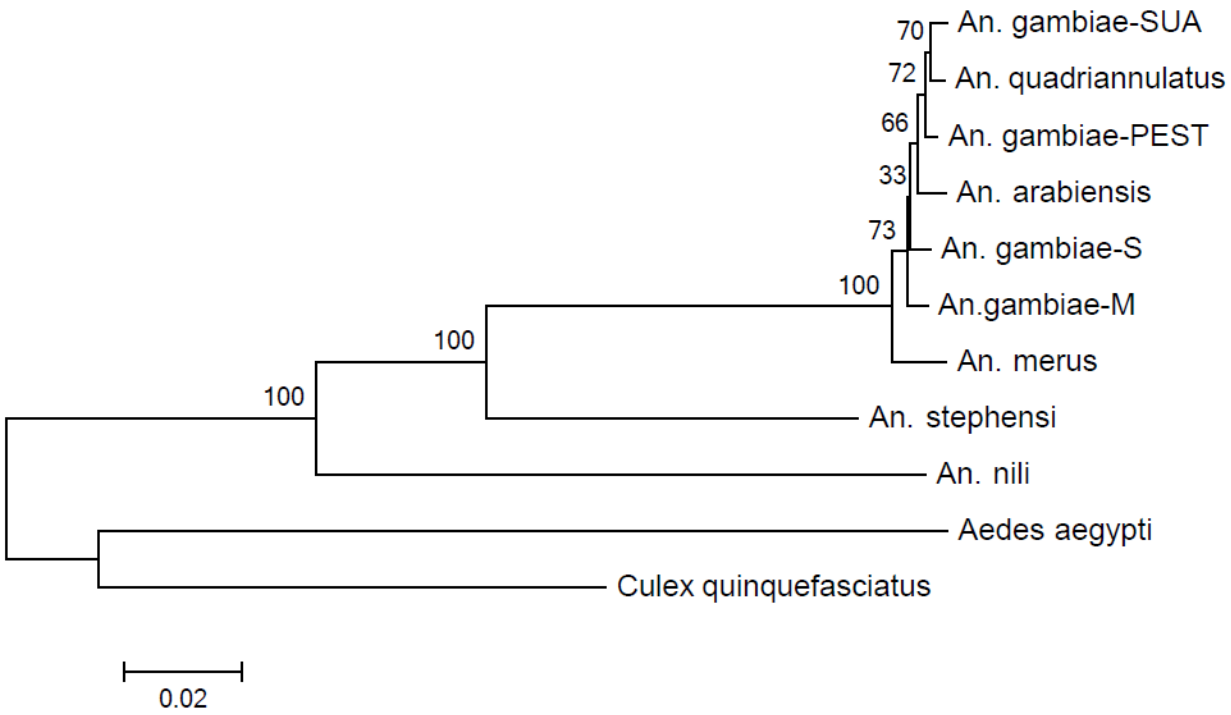

Figure 4. Concatenated phylogenetic tree of 2Ro breakpoint genes, AGAP001760, AGAP001762, AGAP002933 and AGAP002935 in seven members and forms of *An. gambiae* complex and homologous sequences in four outgroup species.

Numbers on branches show bootstrap values. Bootstrap consensus tree is inferred from 1000 replicates based on neighbor- joining statistical method. Scale bar corresponds to 0.02 amino acid substitutions per site.

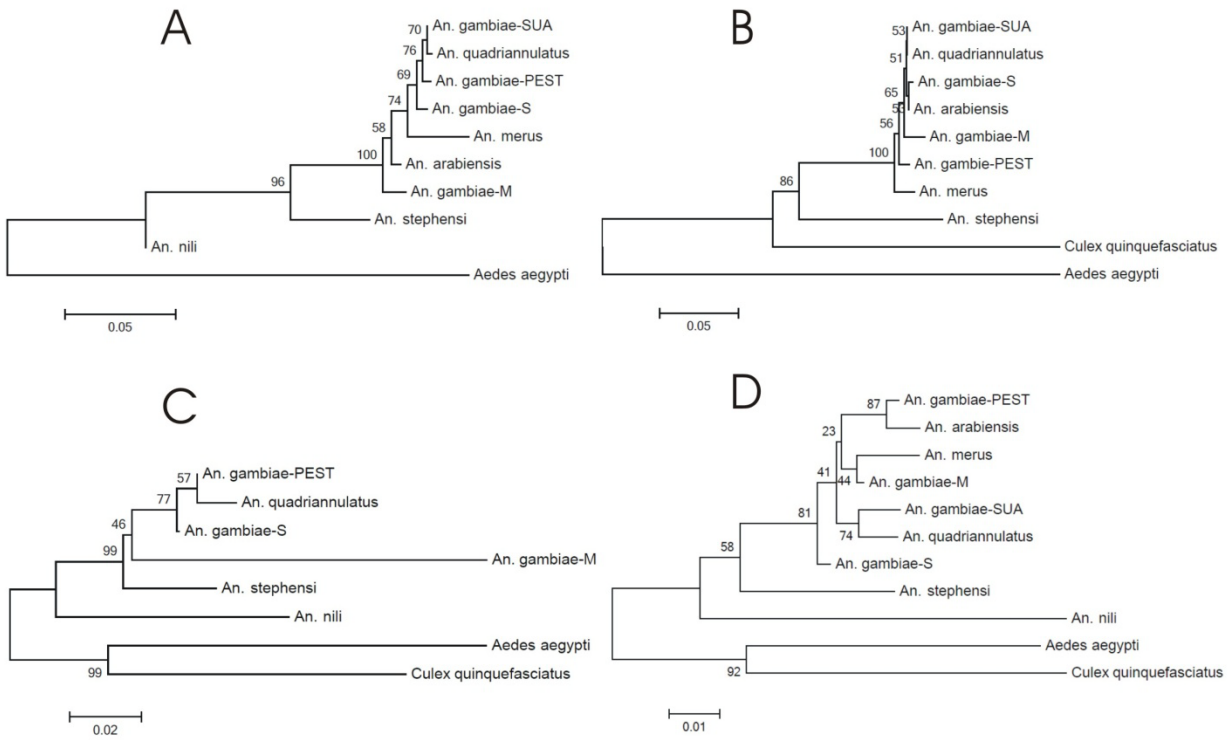

Figure 5. Phylogenetic trees of 2Rp breakpoint genes inferred from 1000 replicate bootstrap consensus tree based on neighbor-joining statistical method.

Numbers on branches represent bootstrap values. Scale bars show a number of amino acid substitutions per site. (A) AGAP013533 (B) AGAP001984 (C) AGAP003327 (D) AGAP003328.

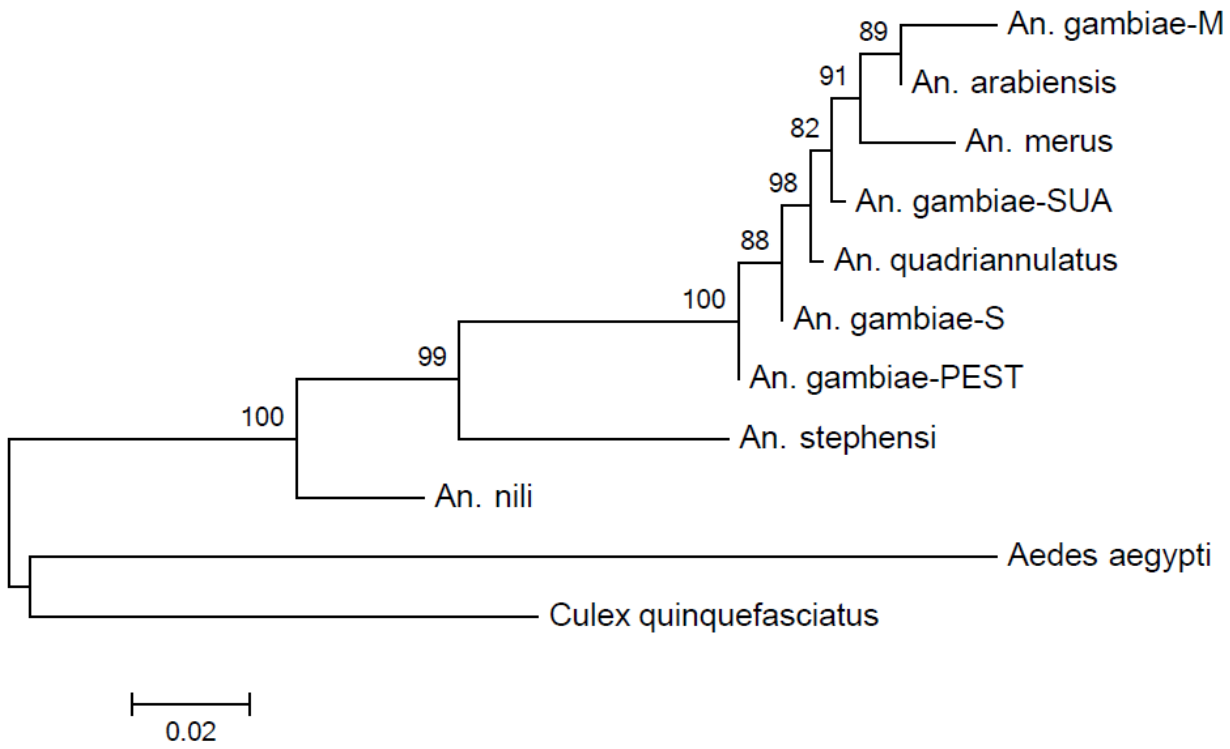

Figure 6. Concatenated phylogenetic tree of 2Rp breakpoint genes, AGAP013533, AGAP001984, AGAP003327 and AGAP003328 in seven members and forms of *An. gambiae* complex and homologous sequences in four outgroup species.

Numbers on branches show bootstrap values. Bootstrap consensus tree is inferred from 1000 replicates based on neighbor- joining statistical method. Scale bar corresponds to 0.02 amino acid substitutions per site.

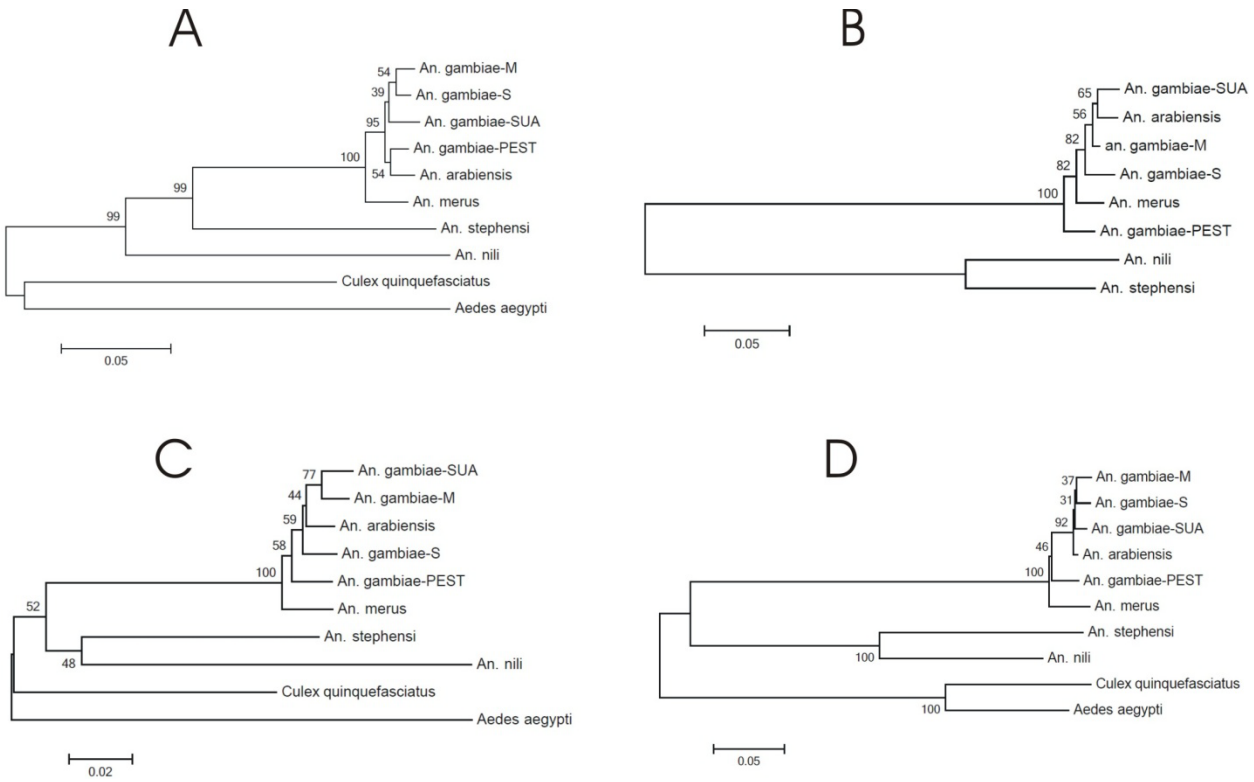

Figure 7. Phylogenetic trees of 2La breakpoint genes inferred from 1000 replicate bootstrap consensus tree based on neighbor-joining statistical method.

Numbers on branches represent bootstrap values. Scale bars show a number of amino acid substitutions per site. (A) AGAP005778 (B) AGAP005779 (C) AGAP007068 (D) AGAP007069.

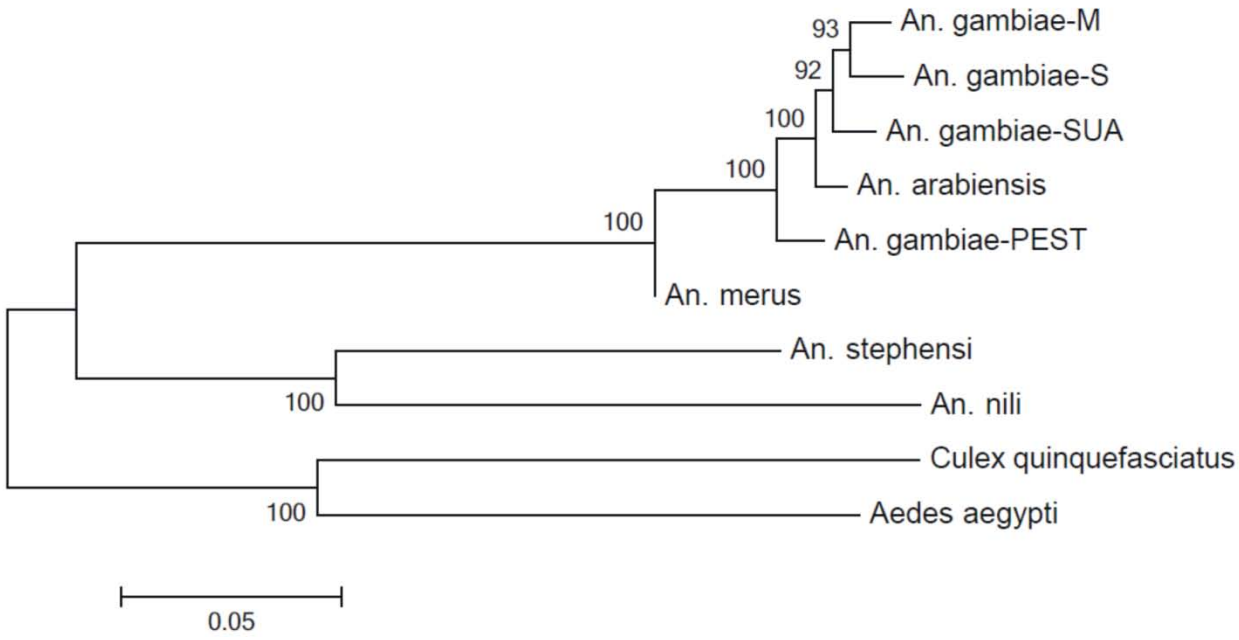

Figure 8. Concatnated phylogenetic tree of 2La breakpoint genes, AGAP005778, AGAP005779, AGAP007068 and AGAP007069 in six members and forms of *An. gambiae* complex and homologous sequences in two outgroup species.

Numbers on branches show bootstrap values. Bootstrap consensus tree is inferred from 1000 replicates based on neighbor- joining statistical method. Scale bar corresponds to 0.05 amino acid substitutions per site.

Tables

Table 1. Primer sequences used for PCR amplification of genetic markers for the 2Ro and 2Rp inversion breakpoints

Primer name	Primer sequence	Coordinate intervals	Length (bp)	Cytological position
AGAP001760 F	ATCAAGCCGAACAAGGAGAA	9484040…9484021	503	2R: 8E
AGAP001760 R	ACCATCCCCATCGTTCATTA	9483518…9483537		
AGAP001762 F	TCGACTTCTTCCGCAACTTC	9491455…9491474	499	2R: 9A
AGAP001762 R	GCGAGTTGGAAAGGTTGTGT	9491973…9491954		
AGAP002933 F	TGTGTGAGCAATCGTTCCAT	29836515…29836496	483	2R: 13C
AGAP002933 R	GTACAGGTCCAGCTCGGTGT	29836013…29836032		
AGAP002935 F	CTGCAGCTGTGCATGAAGAC	29840115…29840134	485	2R: 13D
AGAP002935 R	CCACTGTTTTTCGGACACCT	29840619…29840600		
AGAP013533 F	CGACGAACCTGTTCCTGAAG	13138673…13138654	503	2R: 9C
AGAP013533 R	CTTTAGGCTCGCCTTTGGAG	13138151…13138170		
AGAP001984 F	CTGTCCAACGTGTCCGAGTA	13154289…13154308	517	2R: 10A
AGAP001984 R	CACCTGATCCAGTGGGAAGT	13154825…13154806		
AGAP003327 F	CTGCTCTCCCTTGCCTGTAG	36017908…36017927	478	2R: 14E
AGAP003327 R	TGTACAACCTGATATGTATCCCTGT	36018410…36018386		
AGAP003328 F	TGGACTATGACATCCCGAAG	36027936…36027955	498	2R: 15A
AGAP003328 R	GCAACTGTTTCACCCTTTCG	36028453…36028434		

References

1. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science 298: 129-149.
2. Caccone A, Garcia B, Powell J (1996) Evolution of the mitochondrial DNA control region in the Anopheles gambiae complex. Insect Molecular Biology 5: 51-59.
3. Besansky NJ, Powell JR, Caccone A, Hamm DM, Scott JA, et al. (1994) Molecular phylogeny of the Anopheles gambiae complex suggests genetic introgression between principal malaria vectors. Proceedings of the National Academy of Sciences 91: 6885.
4. Besansky N, Krzywinski J, Lehmann T, Simard F, Kern M, et al. (2003) Semipermeable species boundaries between Anopheles gambiae and Anopheles arabiensis: evidence from multilocus DNA sequence variation. Proceedings of the National Academy of Sciences of the United States of America 100: 10818.
5. Mathiopoulos KD, Powell JD, McCutchan TF (1995) An anchored restriction-mapping approach applied to the genetic analysis of the Anopheles gambiae malaria vector complex 1. Molecular Biology and Evolution 12: 103-112.
6. Pape T (1992) Cladistic analyses of mosquito chromosome data in Anopheles subgenus Cellia (Diptera: Culicidae). Mosquito Systematics 24: 1-11.
7. Coluzzi M, Sabatini A, Petrarca V, Di Deco M (1979) Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. Transactions of the Royal Society of Tropical Medicine and Hygiene 73: 483-497.
8. 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.
9. Garcia BA, Caccone A, Mathiopoulos KD, Powell JR (1996) Inversion monophyly in African anopheline malaria vectors. Genetics 143: 1313.
10. Besansky NJ, Lehmann T, Fahey GT, Fontenille D, Braack LEO, et al. (1997) Patterns of mitochondrial variation within and between African malaria vectors, Anopheles gambiae and An. arabiensis, suggest extensive gene flow. Genetics 147: 1817-1828.
11. Caccone A, Min GS, Powell JR (1998) Multiple origins of cytologically identical chromosome inversions in the Anopheles gambiae complex. Genetics 150: 807.
12. Krzywinski J, Sangaré D, Besansky NJ (2005) Satellite DNA from the Y chromosome of the malaria vector Anopheles gambiae. Genetics 169: 185-196.
13. Obbard D, LINTON YM, Jiggins F, Yan G, Little T (2007) Population genetics of Plasmodium resistance genes in Anopheles gambiae: no evidence for strong selection. Molecular Ecology 16: 3497-3510.
14. Parmakelis A, Slotman M, Marshall J, Awono-Ambene P, Antonio-Nkondjio C, et al. (2008) The molecular evolution of four anti-malarial immune genes in the Anopheles gambiae species complex. BMC Evolutionary Biology 8: 79.
15. Stevenson J, Laurent BS, Lobo NF, Cooke MK, Kahindi SC, et al. (2012) Novel Vectors of Malaria Parasite in the Western Highlands of Kenya. Emerging Infectious Diseases 18: 1547.
16. Kirkpatrick M (2010) How and why chromosome inversions evolve. PLoS biology 8: e1000501.
17. Noor MAF, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and the reproductive isolation of species. Proceedings of the National Academy of Sciences 98: 12084.
18. Noor MAF, Grams KL, Bertucci LA, Almendarez Y, Reiland J, et al. (2001) The genetics of reproductive isolation and the potential for gene exchange between Drosophila pseudoobscura and D. persimilis via backcross hybrid males. Evolution 55: 512-521.
19. Rieseberg LH (2001) Chromosomal rearrangements and speciation. Trends in Ecology & Evolution 16: 351-358.
20. Bartolomé C, Charlesworth B (2006) Rates and patterns of chromosomal evolution in Drosophila pseudoobscura and D. miranda. Genetics 173: 779-791.

21. Popadić A, Anderson WW (1994) The history of a genetic system. Proceedings of the National Academy of Sciences 91: 6819-6823.
22. Aquadro C, Weaver A, Schaeffer S, Anderson W (1991) Molecular evolution of inversions in Drosophila pseudoobscura: the amylase gene region. Proceedings of the National Academy of Sciences 88: 305-309.
23. Navarro A, Betrán E, Barbadilla A, Ruiz A (1997) Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. Genetics 146: 695-709.
24. Schaeffer SW, Goetting-Minesky MP, Kovacevic M, Peoples JR, Graybill JL, et al. (2003) Evolutionary genomics of inversions in Drosophila pseudoobscura: evidence for epistasis. Proceedings of the National Academy of Sciences 100: 8319-8324.
25. Wallace AG, Detweiler D, Schaeffer SW (2011) Evolutionary History of the Third Chromosome Gene Arrangements of Drosophila pseudoobscura Inferred from Inversion Breakpoints. Molecular Biology and Evolution 28: 2219-2229.
26. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.
27. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, et al. (2009) VectorBase: a data resource for invertebrate vector genomics. Nucleic Acids Research 37: D583-D587.
28. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28: 2731-2739.
29. Kamali M, Xia A, Tu Z, Sharakhov IV (2012) A New Chromosomal Phylogeny Supports the Repeated Origin of Vectorial Capacity in Malaria Mosquitoes of the Anopheles gambiae Complex. PLoS Pathog 8(10): e1002960.
30. Kamali M, Sharakhova MV, Baricheva E, Karagodin D, Tu Z, et al. (2011) An Integrated Chromosome Map of Microsatellite Markers and Inversion Breakpoints for an Asian Malaria Mosquito, Anopheles stephensi. Journal of Heredity 102: 719-726.
31. Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798-804.
32. Neafsey D, Lawniczak M, Park D, Redmond S, Coulibaly M, et al. (2010) SNP genotyping defines complex gene-flow boundaries among african malaria vector mosquitoes. Science 330: 514.

CHAPTER 4 Multigene Phylogeny of African Malaria Vectors Places *Anopheles nili* in the Basal Clade

Abstract

Mosquito species of subgenus *Cellia*, *Anopheles gambiae*, *An. arabiensis*, *An. funestus*, and *An. nili*, are the major vectors of malaria in sub-Saharan African savannas. *Anopheles gambiae* and *An. arabiensis* are closely related species of the *An. gambiae* complex, which belongs to Pyretophorus series. *Anopheles funestus* and *An. nili* belong to Myzomia and Neomyzomiya series, respectively. Although previous phylogenetic studies using ribosomal and mitochondrial DNA sequences demonstrated paraphyletic relationships among the major African malaria vectors, the basal clade has not been unambiguously determined. This knowledge could be useful for understanding of association between evolutionary genomic changes and selective pressures from a malaria parasite on the immune systems of vectors. In this study, we reconstructed a molecular phylogeny using 49 genic sequences for *An. gambiae*, *An. funestus*, *An. nili*, the Asian malaria mosquito *An. stephensi*, and outgroup species *Culex quinquefasciatus* and *Aedes aegypti*. To this end, we identified orthologous sequences distributed approximately every 5 Mb in the *An. gambiae* reference genome and in the genomes or transcriptomes of the other species. The sequences were aligned using ClustalW and the phylogenetic trees were obtained using neighbor-joining method in the MEGA 5.05 program. Most of the trees from individual genes were consistent with each other, placing *An. nili* as the basal clade among the studied malaria mosquito species. The African *An. funestus* and Asian *An. stephensi* lineages have been diversified most recently. These results were supported by high bootstrap values in concatenated phylogenetic trees generated separately for each chromosomal arm. In conclusion, multigene genome-wide molecular phylogenic analysis is a useful approach for discerning historic

relationships among malaria vectors.

Introduction

Malaria in tropical, humid savannah environments of Africa is quite stable with entomological inoculation rates (EIR: number of infective bites per person per year) varying between 50 and 300 [1]. Four major vector species are responsible for malaria transmission in these areas: *An. gambiae*, *An. arabiensis*, *An. funestus*, and *An. nili* [1]. These species, together with *An. moucheti*, a major vector in the equatorial forest of Central Africa, are responsible for >95% of the total malaria transmission on the African continent [2]. *Anopheles gambiae*, *An. arabiensis*, and *An. funestus* breed in temporal or permanent freshwater pools. *Anopheles gambiae* is found mostly in humid savannas, *An. arabiensis* occupies arid savannas and steppes [3], while *An. funestus* has a continent-wide distribution [4]. *Anopheles nili* has as wide distribution as *An. gambiae*, *An. arabiensis*, and *An. funestus*, spreading across most of West, Central, and East Africa, mainly in humid savannah and degraded rainforest areas [5,6]. However, unlike other major vectors, *An. nili* breeds in slow-moving streams and large lotic rivers exposed to light and containing vegetation or debris [1,6]. A study of the ecological niche profile of major malaria vectors in Cameroon demonstrated that the habitats of *An. gambiae*, *An. arabiensis*, and *An. funestus* have more overlap with each other than with the habitat of *An. nili* [5]. This difference results in a much more unusual geographic distribution of *An. nili*, revealing its crucial role in malaria transmission in degraded forests in Cameroon [6].

The genus *Anopheles* is divided in to six subgenera including *Anopheles*, *Cellia*, *Kerteszia*, *Lophopodomyia*, *Nyssorhynchus* and *Stethomyia* [7]. All major malaria vectors in Sub-saharan Africa belong to subgenus *Cellia*, which consist of six series: Cellia, Neocellia, Myzomyia, Pyretophorus, Paramyzomyia, and Neomyzomyia. Previous analyses of the rDNA and combined

rDNA plus mtDNA data have supported the monophyly of the Pyretophorus, Myzomyia, Neocellia and Neomyzomyia series [8]. *Anopheles gambiae* and *An. arabiensis* are members of the *An. gambiae* complex, which belongs to series Pyretophorus. *Anopheles gambiae* consists of two molecular forms; the S form is widely distributed and the M form is restricted to West and Central Africa [9,10]. *Anopheles funestus* belongs to the Funestus subgroup, which is classified under series Myzomyia and is divided in to five subgroups: Aconitus, Culicifacies, Funestus, Minimus, and Rivulorum [7]. *Anopheles moucheti* also belongs to series Myzomyia. *Anopheles nili* is a member of the *An. nili* group that belongs to series Neomyzomiya [7]. Four species had been identified within the *An. nili* group based on the ribosomal DNA second internal transcribed spacer (rDNA ITS2) and D3 28S region: *An. nili s. s., An. somalicus, An. carnevalei,* and *An. ovengensis* [11]. A recent study on the genetic structure of species of the *An. nili* group using a combination of microsatellites, rDNA and mitochondrial DNA markers, demonstrated high genetic divergence among new cryptic members of the *An. nili* group [12]. A comparative cytogenetic analysis of polytene chromosomes revealed significant differences in banding pattern and structure of heterochromatin between *An. nili* and *An. ovengensis*, suggesting either a rapid rate of chromosome evolution or that the two species belong to different taxonomic groups [13]. The Asian malaria mosquito *An. stephensi*, which is often used for phylogenetic comparisons with African mosquito species [8,14-17], belongs to series Neocellia of subgenus *Cellia* [7].

Knowledge about the phylogenetic relationships among the major African malaria vectors could be useful for understanding of association between evolutionary genomic changes and selective pressures from *Plasmodium falciparum* on the immune systems of vectors. However, it is still unknown if a particular lineage has been transmitting *P. falciparum* for a long time or has evolved only recently. The diverse taxonomic positions of the major malaria vectors suggest that

vectorial capacity evolved independently in these species. Each species group or complex, to which the major vectors belong, also includes nonvectors [7]. Indeed, a phylogeny reconstruction using mtDNA and rDNA has demonstrated a paraphyletic relationship among *An. arabiensis*, *An. gambiae*, *An. funestus*, *An. moucheti*, and *An. nili* [14]. However, the basal lineage among the major African malaria vectors has not been unambiguously identified. For example, a Bayesian phylogenetic analysis using the combined rDNA (18S and 28S) sequences places the *An. gambiae* complex in the basal branch, *An. stephensi* in the intermediate branch, while *An. funestus* and *An. nili* have diversified more recently. However, bootstrap consensus tree created with the mtDNA sequences has been unable to infer the relationships due to a very low phylogenetic signal [14]. In contrast, other phylogenetic trees based on combined rDNA and mtDNA sequences have placed *An. dirus* and *An. farauti* (species from the same series as *An. nili*) in a basal clade, *An. gambiae* in an intermediate clade, and *An. funestus* and *An. stephensi* in a more recently diversified clades [8].

An alternative approach to inferring the phylogenetic relationships among species is a multigene phylogenetic analysis, which has been performed in many organisms. For example, 78 protein-coding genes have been successfully used to reconstruct a multigene phylogenetic tree of Choanozoa (unicellular protozoan phylum) and their evolutionary relationship to animals and fungi [18]. In another study, phylogenetic tree based on 22 gene segments in Mustelidae, carnivorous mammals, has been created [19]. Importantly, a multigene phylogeny based on concatenated sequences provides more resolution and support [18-20]. According to a phylogenetic study of 106 orthologous genes from eight yeast species, the sufficient number of concatenated genes that are required to achieve the mean bootstrap value of 70% is three; and a minimum of 20 genes is required to recover >95% bootstrap values for each branch of the

species tree [21]. Another study demonstrated an efficient phylogenetic approach by sampling and assembling transcriptomes of 10 mosquito species into data matrices containing hundreds of thousands of orthologous nucleotides from hundreds of genes [15]. However, that study did not include major African malaria vectors except 3 species from the *An. gambiae* complex.

In our study, we investigated the phylogenetic relationships among African malaria mosquito species as well as an Asian mosquito, *An. stephensi*. We have selected 49 genes from the *An. gambiae* genome [22,23] distributed throughout 5 chromosomal arms. We identified orthologous sequences in the genomes of *An. nili* [24], *An. stephensi* [17,25], *Culex quinquefasciatus* [26], and *Aedes aegypti* [27], and in the transcriptome of *An. funestus* [28]. Phylogenetic trees were generated using the neighbor-joining method from all individual genes and genes concatenated according to chromosomal arms. Results from different chromosomal arms were consistent, placing *An. nili* in the most basal branch as compared with the other African Anopheline species. The *An. gambiae* lineages emerged more recently, while African *An. funestus* and Asian *An. stephensi* were the most closely related and most recently diversified lineages.

Material and methods

Genome assemblies for *An. nili* and *An. stephensi*

The genome assembly for *An. nili* was obtained by sequencing of genomic DNA isolated from two larvae collected in Dinderesso (11°14’N; 4°23’W), Burkina Faso [24]. The assembly consisted of 51,048 contigs with a total length of 98,320,874 bp. The average contig length was 1,926 bp and the maximum contig length was 30,512 bp. Repetitive sequences were a problem for the assembly, only 34,956,095 reads (58.57%) aligned to the contigs with 95% base matching. The 16X coverage genome assembly for *An. stephensi* was obtained by sequencing genomic DNA isolated from the Indian wild-type laboratory strain [17,25]. The assembly

consisted of 33,024 contigs and 6,150 scaffolds with a total length of 158 megabase pairs (Mb).

Genome-wide selection of genetic markers

In order to select genes distributed throughout the genome, the AgamP3 genome assembly of the *An. gambiae* PEST strain (https://www.vectorbase.org/organisms/anopheles-gambiae/pest/AgamP3) [23] was divided into 5 Mb segments. Genes were randomly selected within the 5 Mb segments of the *An. gambiae* genome with approximate exon lengths between 364 and 1400 bp. Selected exons were transferred to the Geneious 5.1.5 software (www.geneious.com) and used for finding homologous sequences in the *An. nili* genome assembly [24] using Basic Local Alignment Search Tool (BLAST). If no BLAST hit were present, another exon or gene was selected. Appropriate exons from the *An. gambiae* PEST genome that had significant E-values in BLAST against the *An. nili* genome were the candidate genes for further BLAST analysis. These exons were used for BLAST against the *An. gambiae* M5 (M form, Mali strain) (https://www.vectorbase.org/organisms/anopheles-gambiae/mali-nih-m-form/m5) and G4 (S form, Pimperena strain) (https://www.vectorbase.org/organisms/anopheles-gambiae/pimperena-s-form/g4) genome assemblies [29], the *An. stephensi* AsteI1 (https://www.vectorbase.org/organisms/anopheles-stephensi/indian/AsteI1) [17,25], the *Culex quinquefasciatus* (https://www.vectorbase.org/organisms/culex-quinquefasciatus/johannesburg-jhb/cpipj1) [26], and *Aedes aegypti* (https://www.vectorbase.org/organisms/aedes-aegypti/liverpool-lvp/aaegl1) [27] genome assemblies, as well as *An. funestus* transcriptome sequences (http://funcgen.vectorbase.org/annotated-transcriptome/Crawford_et_al_Anopheles_funestus) [28].

Orthology detection

Two genes are orthologs if they diverged after a speciation event and, therefore, they are related by common ancestry [30]. To find orthologous genes, the Reciprocal Best Hits (RBH) method was used [31]. In this method, orthologous pairs should have the best reciprocal BLAST hits. In order to detect the RBH, *An. nili*, *An. stephensi*, *An. funestus*, *Culex* and *Aedes* sequences were used for BLAST against *An. gambiae* PEST strain in VectorBase. Orthology was confirmed if reciprocal BLAST finds the originally selected sequences in *An. gambiae* PEST genome as the best hits.

Gene alignment and phylogenetic analysis

Orthologous sequences with significant E-values in the BLAST search were transferred to Molecular Evolutionary Genetics Analysis (MEGA 5.05) program [32]. The sequences were aligned using ClustalW alignment option in the MEGA program. Alignments were performed by adding the most closely related species followed by outgroup species. Phylogenetic trees for each gene were constructed using the neighbor-joining statistical method [33]. Confidence values for each clade were generated by 1000 bootstrap replicates.

Results and discussion

Genome-wide approach to multigene phylogeny

It is important that genetic markers are distributed as uniformly as possible across the genome rather than cluster in a particular genomic region. To resolve the molecular phylogeny of African malaria mosquito species, we have selected 49 genes as molecular markers. These genes were distributed throughout the genome in all five chromosomal arms of the *An. gambiae* cytogenetic map [34] (Figure 1). Based on the data obtained from VectorBase [35], the length of each chromosomal arm (in base pairs) is the following: X, 24393108; 2R, 61545105; 2L, 49364325;

3R, 53200684 and 3L, 41963435 (Table 1). According to the length of each arm, 3, 19, 13, 8 and 6 genes were selected from X (Table S1), 2R (Table S2), 2L (Table S3), 3R (Table S4) and 3L (Table S5), respectively. Our sequencing data consisted of ≥364 bp-long gene fragments resulting in a total alignment of 41,124 bp based on the *An. gambiae* AgamP3 genome assembly. Orthologous sequences of selected genes from 8 genome assemblies representing 6 species were obtained. Confidence values for each clade was generated by 1000 bootstrap replicates and the reliability of phylogenetic trees was assessed based on 70% as the cut-off value [36].

Phylogenetic relationships among African malaria vectors

Trees obtained from individual genes agreed with each other, with few exceptions. The phylogenies based on three X chromosome genes placed *An. nili* in a separate cluster from the other *Anopheles* vectors. However, only in one of the trees, the *An. nili* branch was supported by a high (85%) bootstrap value (Figure 2). However, all of the trees had a high (100%) bootstrap value for a separation of the *An. gambiae* complex from the *An. funestus-An. stephensi* clade (Figure 2). For 2R arm, 13 out of 19 phylogenetic trees had bootstrap values ranging from 76% to100% in supporting the basal position of the *An. nili* clade. Similarly, 12 of 19 gene trees had a bootstrap value of 100% in supporting the intermediate position of the *An. gambiae* complex (Figure 3). For 2L arm, 9 out of 13 phylogenetic trees had high bootstrap values ranging from 83% to 100% for *An. nili* being in a separate more basal clade. Also, 9 out of 13 trees for 2L had a high bootstrap value of 100% for the separation of the *An. gambiae* complex from the *An. funestus-An. stephensi* clade (Figure 4). Eight phylogenetic trees were constructed for 3R arm, and 6 trees had high bootstrap values ranging from 78% to 99% placing *An. nili* in a separate basal lineage. Similarly, 5 of 8 phylogenetic trees had a high bootstrap value of 100% supporting the intermediate position of the *An. gambiae* complex in relation to the *An. nili* and the *An.*

funestus-An. stephensi clades (Figure 5). Finally, 6 phylogenetic trees were constructed based on the genes located on the 3L chromosomal arm, 5 of which had bootstrap values ranging rom 77% to 100% supporting the basal position of *An. nili* compared with other African mosquitoes. Moreover, 4 phylogenetic trees had high bootstrap values (86% and 100%) for separation of the *An. gambiae* complex from the *An. funestus-An. stephensi* clade (Figure 6).

Previous studies have shown that analysis of entire data set of concatenated genes is useful for resolving species phylogenetic tree [21]. We created trees using sequences for all genes concatenated according to five chromosomal arms (Figure 7). In all trees, *Aedes* and *Culex* were clustered separately as outgroup species. Trees of concatenated genes provided a robust support with 92% (X chromosome) and 100% (autosomes) bootstrap values, showing that *An. nili* is clustered separately from the rest of African *Anopheles* species. On the other hand, *An. funestus* is clustered together with *An. stephensi.* Our results indicate that *An. nili* belongs to the most basal clade among the African malaria mosquitoes. African *An. funestus* and Asian *An. stephensi* are the most recently evolved sister taxa. *Anopheles gambiae* branch has the intermediate position in relation to the *An. nili* and the *An. funestus-An. stephensi* clades.

The only other phylogenetic study that included *An. gambiae*, *An. funestus*, *An. nili*, and *An. stephensi* has used mtDNA and rDNA [14]. A parsimony bootstrap consensus tree based on the mtDNA sequence had a very low phylogenetic signal and, therefore, could not resolve the phylogeny. A Bayesian phylogenetic analysis using rDNA sequences placed the *An. gambiae* complex in the basal clade, *An. stephensi* in the intermediate clade, while *An. funestus* and *An. nili* have diversified most recently [14]. This is in contradiction with our concatenated trees, which show that *An. nili* is in the basal clade, while *An. funestus* and *An. stephensi* have split most recently (Figure 7). Interestingly, another phylogenetic study using combined rDNA and

mtDNA sequences has demonstrated a basal position of *An. dirus* and *An. farauti*, intermediate position of *An. gambiae*, and more recent diversification of the *An. funestus* and *An. stephensi* lineages [8]. Asian mosquitoes *An. dirus* and *An. farauti* together with an African mosquito *An. nili* belong to series Neomyzomiya, suggesting that this series is the most basal taxon followed by the splitting of series Pyretophorus (*An. gambiae, An. arabiensis*) and the most recent diversification of series Myzomyia (*An. funestus, An. moucheti*) and series Neocellia (*An. stephensi*) [7].

Hypothetic evolutionary history of African vectors

Our phylogenetic analysis of multiple genes showed that the *An. nili* lineage split early from the other African *Anopheles* species. Larvae of *An. nili* breed in rivers and fast flowing streams and adults are distributed in forested areas and humid savannas. Recent studies have demonstrated the high genetic and chromosomal divergence within the *An. nili* group in equatorial Africa [12,13]. The high genetic divergence within the *An. nili* group and the basal phylogenetic position of *An. nili* in our study suggest that the *An. nili* lineage originated in the equatorial forest before other lineages that led to major African malaria vectors. According to our phylogenic tree, African *An. funestus* and Asian *An. stephensi* are the most recently diversified taxa while *An. gambiae* branch has an intermediate position. The members of both the *An. gambiae* complex and the *An. funestus* group are not restricted to the equatorial Africa but distributed across the continent. A recent study of chromosomal phylogeny has demonstrated that the ancestral species of the *An. gambiae* complex had the 2Ro, $2R+^{p}$, and 2La chromosomal arrangements [17]. The ‘‘inverted’’ 2Ro arrangement uniquely characterizes an east African mosquito *An. merus* as the basal species in the complex. Therefore, the *An. gambiae* complex is likely evolved from the East African ancestor.

Interestingly, each of the major African malaria vectors clusters together with Asian malaria mosquito species. For example, *An. gambiae* and *An. arabiensis* were sister taxa with *An. subpictus* and *An. sundaicus* in a combined phylogenetic analysis of mtDNA and rDNA [14]. Moreover, the fixed 2La chromosomal inversion typical to *An. gambiae*, *An. arabiensis*, and *An. merus* was also found in two species from the Middle Eastern *An. subpictus* complex [37]. A study based on morphological characteristics as well as rDNA and mtDNA sequences considered Afrotropical Funestus and Afro-Oriental Minimus group as sister taxa [14,38]. Finally, a Bayesian phylogenetic analysis using rDNA sequences indicated that *An. nili* is a sister taxon with Asian *An. dirus* and *An. farauti* [14]. Overall, these data suggest multiple migrations between Africa and Asia in different time points during evolution of subgenus *Cellia*.

Conclusion

Comparative genomic analyses of epidemiologically important traits will be more informative if performed within a phylogenetic framework. Inferring basal and more recently diversified lineages in African malaria mosquitoes could be helpful for establishing the association between evolutionary genomic changes and the origin and loss of human blood choice, ecological and behavioral adaptations, and association with human habitats. A recent reconstruction of chromosomal phylogeny in the *An. gambiae* complex has found more ancestral chromosomal arrangements in mosquito species that are vectors of human malaria, while the more derived arrangements in both nonvectors and vectors suggesting a repeated origin of vectorial capacity during the recent evolution of these African mosquitoes [17]. Spreading over large distances of the phylogenetic tree, African malaria mosquitoes of subgenus *Cellia* represent a unique system for studying evolution of vectorial capacity. The multigene phylogenetic analysis is an important step toward elucidating the historic relationship among African malaria vectors. Our study

concluded that *An. nili* belongs to the most basal group probably originated in the equatorial forest. Other African malaria vectors had their different and independent histories of acquiring traits related to vectorial capacity. We found good agreement between molecular phylogenies constructed by using multiple unlinked genes indicating that next-generation sequencing data are highly valuable for inferring phylogenetic relationships among malaria mosquitoes.

Figures

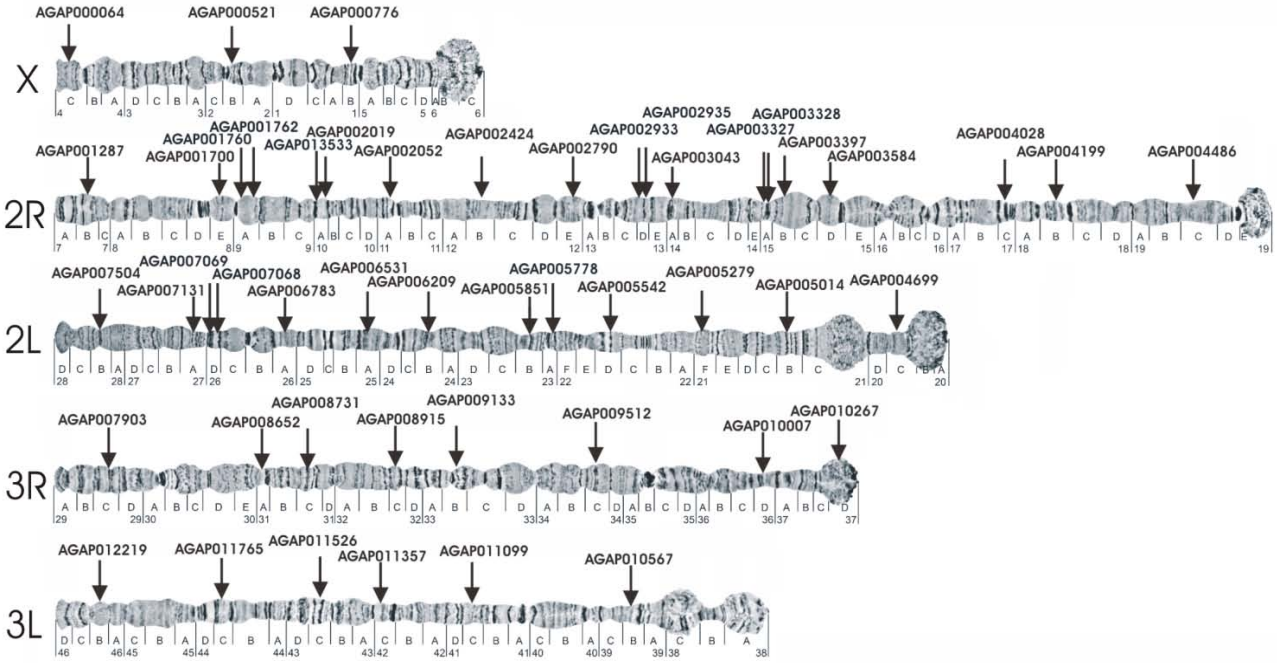

Figure 1. Distribution of genic phylogenetic markers in five chromosomal arms of *An. gambiae*.

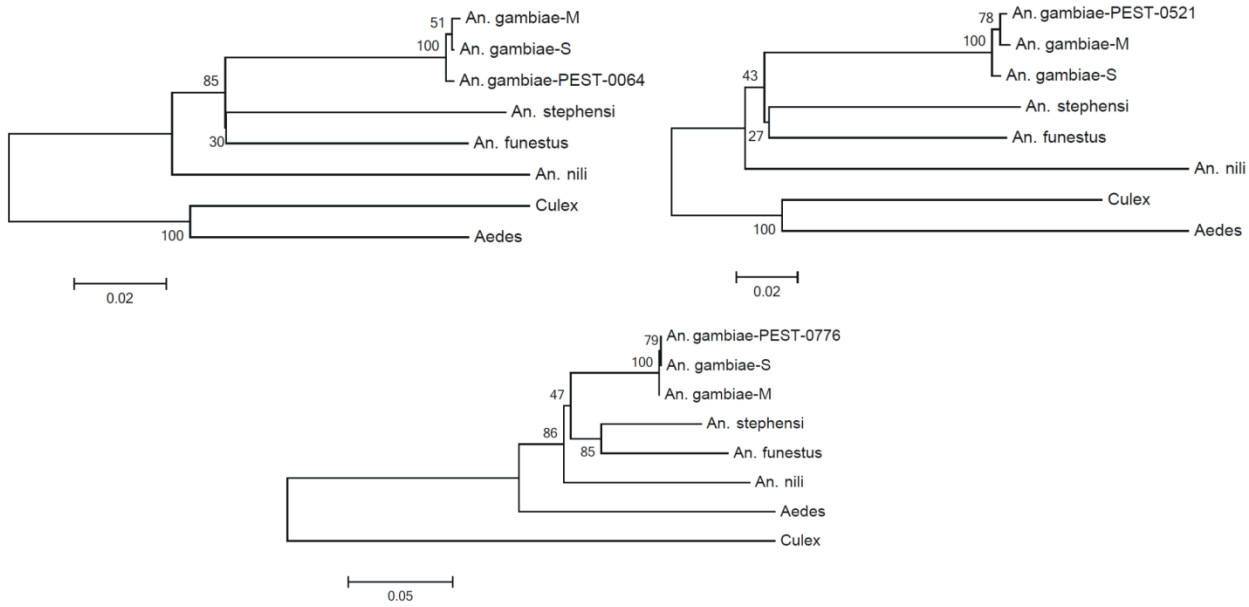

Figure 2. Molecular phylogeny of X chromosome genes.

Last four digits of corresponding gene are stated next to *An. gambiae*-PEST. Bootstrap values are shown on branches of phylogenetic trees as percentages.

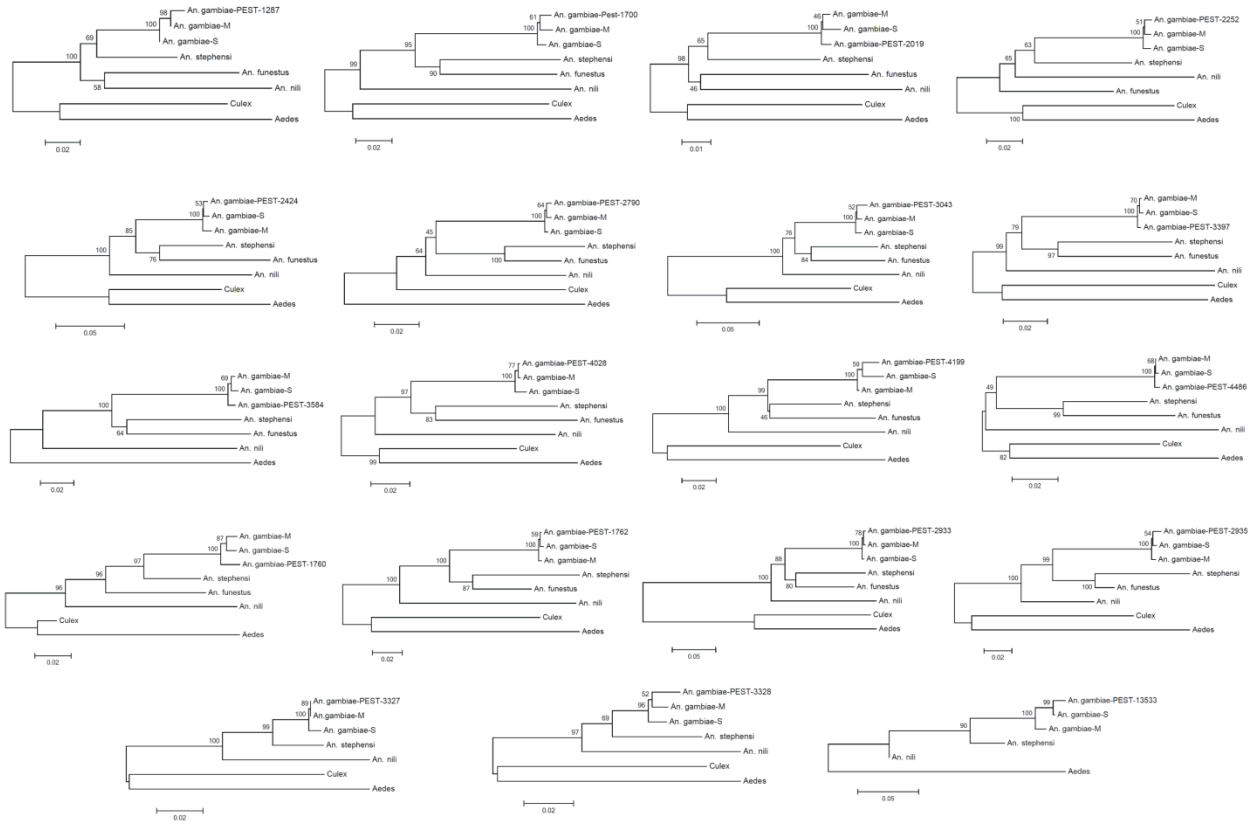

Figure 3. Molecular phylogeny of 2R chromosome genes.
Last four digits of corresponding gene are stated next to *An. gambiae*-PEST. Bootstrap values are shown on branches of phylogenetic trees as percentages.

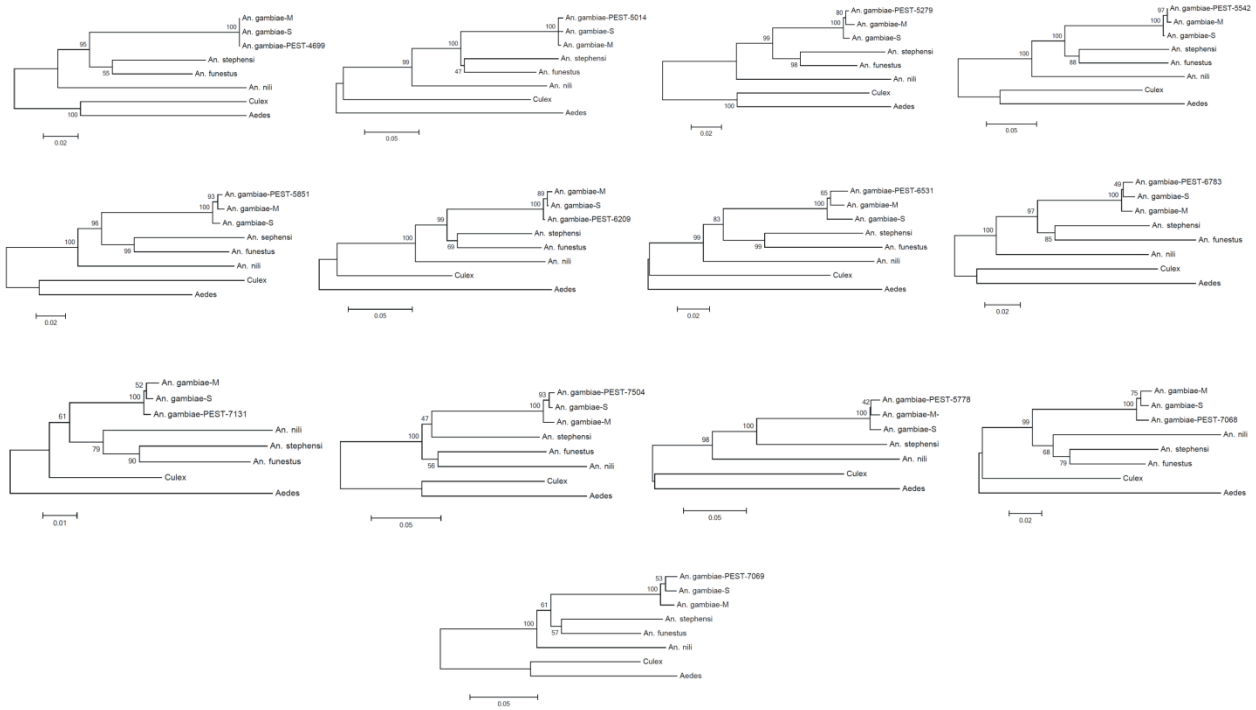

Figure 4. Molecular phylogeny of 2L chromosome genes.

Last four digits of corresponding gene are stated next to *An. gambiae*-PEST. Bootstrap values are shown on branches of phylogenetic trees as percentages.

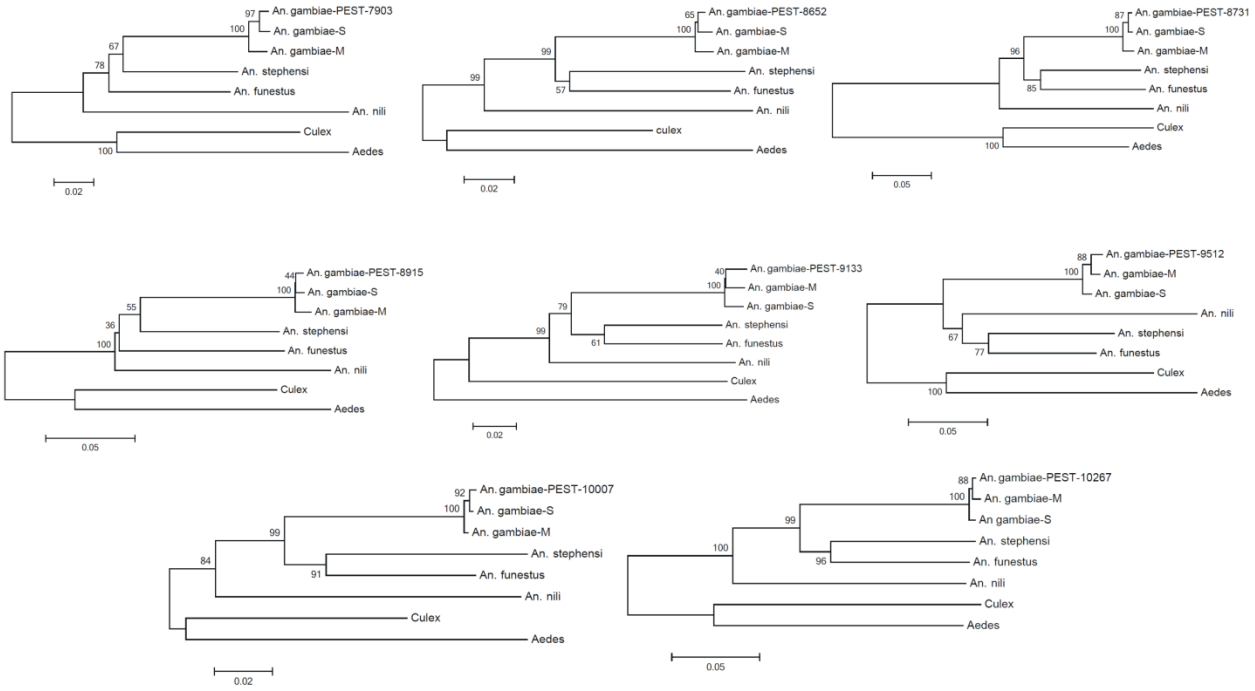

Figure 5. Molecular phylogeny of 3R chromosome genes.
Last four digits of corresponding gene are stated next to *An. gambiae*-PEST. Bootstrap values are shown on branches of phylogenetic trees as percentages.

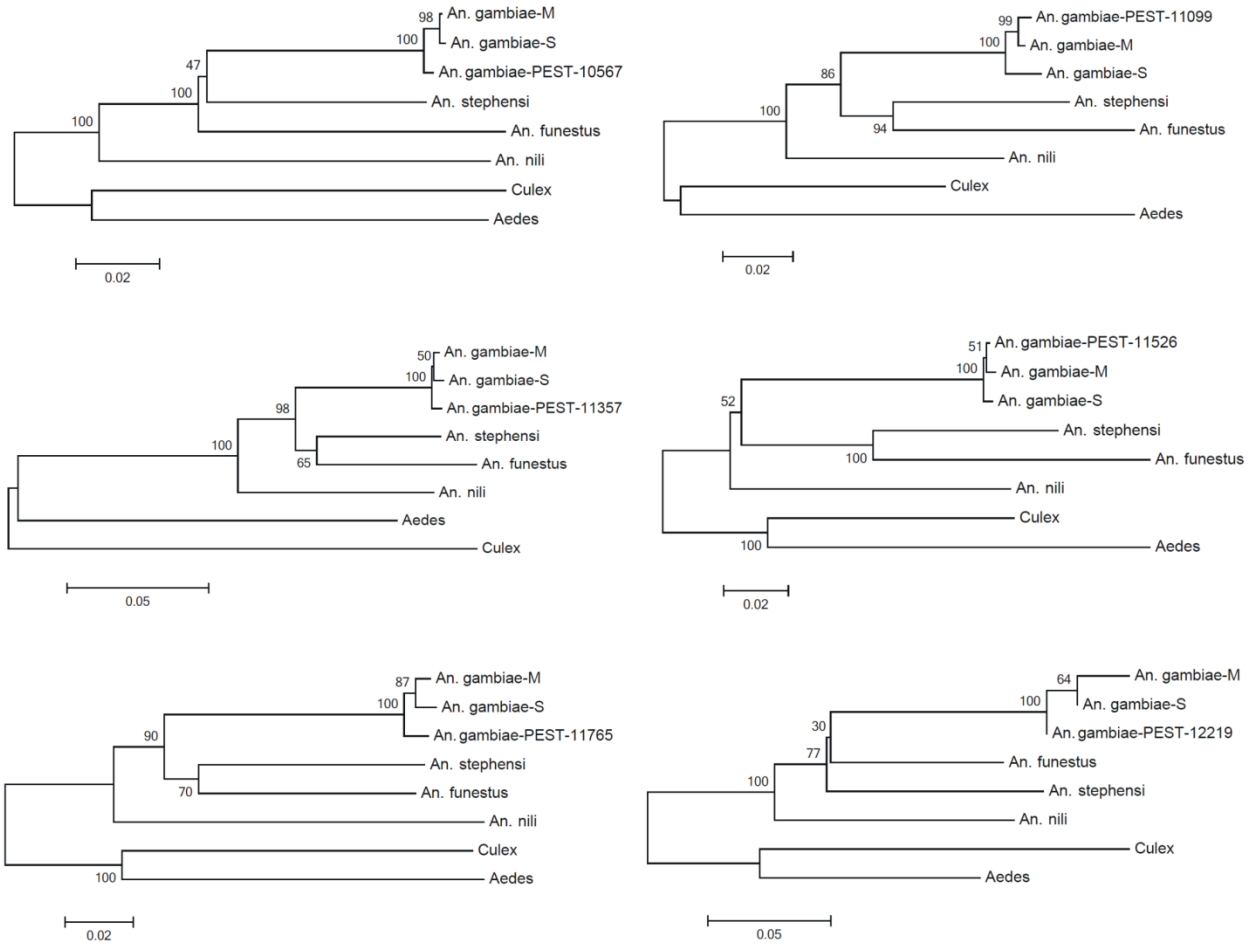

Figure 6. Molecular phylogeny of 3L chromosome genes.

Last four digits of corresponding gene are stated next to *An. gambiae*-PEST. Bootstrap values are shown on branches of phylogenetic trees as percentages.

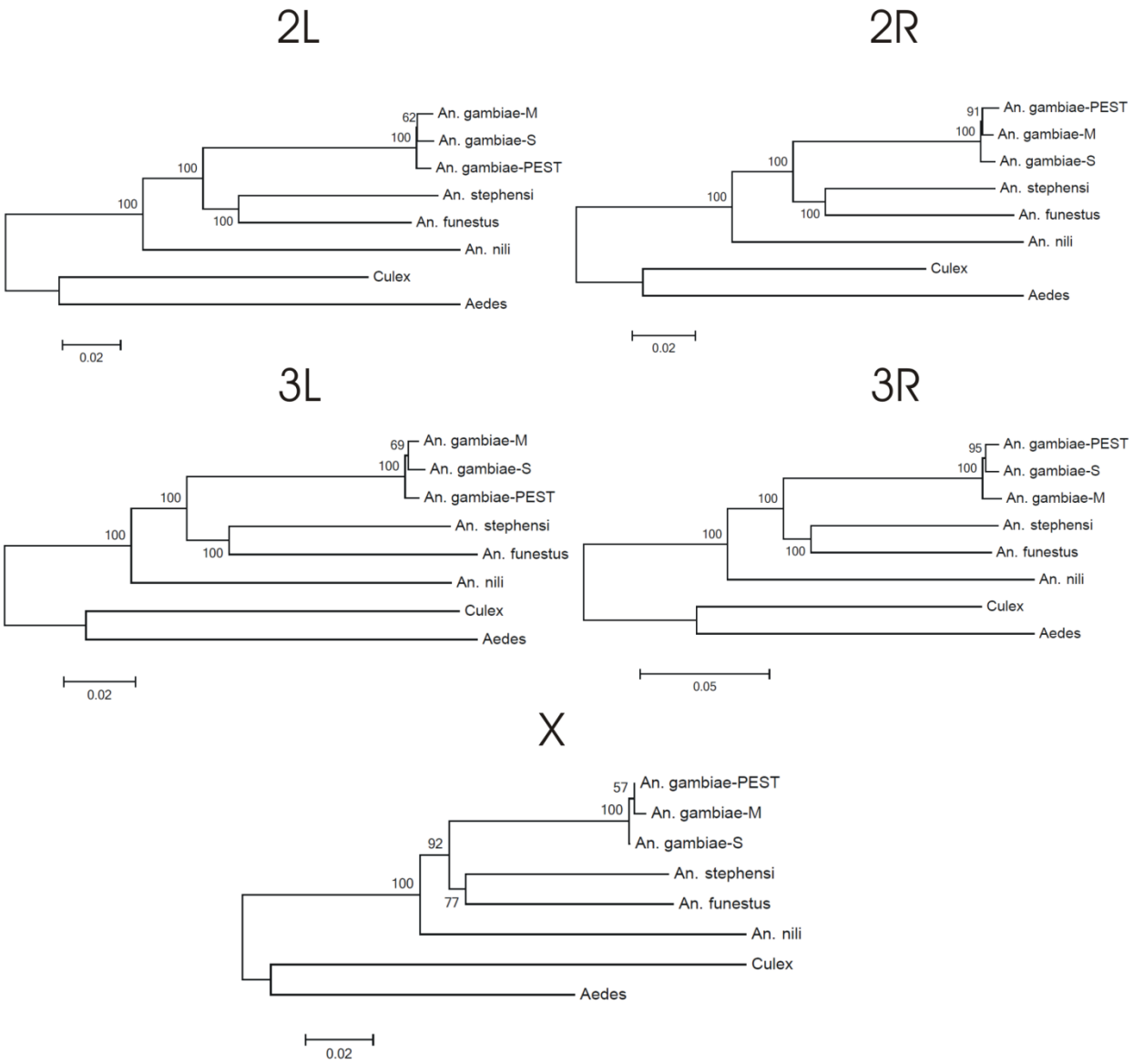

Figure 7. Phylogenetic trees build from concatenated sequences located in 5 chromosomal arms. Bootstrap values are shown on branches of phylogenetic trees as percentages.

Tables

Table 1. Genome-wide distribution of genes used in the phylogenetic study.

Chromosome arm	Length (Mb)	Number of Genes	Genes per 5 Mb
X	24.4	3	0.6
2R	61.5	19	1.5
2L	49.4	13	1.3
3R	53.2	8	0.8
3L	42.0	6	0.7
Total	**230.5**	**49**	**1.0**

Supplementary

Table S1. Selected genes from X chromosome and length of orthologous sequences in 7 species.

X chromosome	*An. gambiae*-PEST	*An. gambiae*-M	*An. gambiae*-S	*An. stephensi*	*An. nili*	*An. funestus*	*Aedes*	*Culex*
AGAP000064	774	508	774	772	680	769	597	597
AGAP000521	1183	512	655	1193	661	617	417	420
AGAP000776	640	593	640	631	619	637	486	252

Table S2. Selected genes from 2R chromosome and length of orthologous sequences in 7 species.

2R chromosome	*An. gambiae*-PEST	*An. gambiae*-M	*An. gambiae*-S	*An. stephensi*	*An. nili*	*An. funestus*	*Aedes*	*Culex*
AGAP001287	967	505	505	971	959	493	501	501
AGAP001700	983	983	983	969	962	482	408	408
AGAP002019	1039	1039	1039	818	815	808	624	624
AGAP002252	823	823	823	821	821	440	648	648
AGAP002424	717	717	711	717	712	380	564	564
AGAP002790	776	481	481	481	481	475	417	417
AGAP003043	912	912	913	648	675	633	321	321
AGAP003397	859	859	859	859	859	798	858	858
AGAP003584	811	811	811	811	811	580	807	339
AGAP004028	541	541	541	539	408	487	540	540
AGAP004199	643	409	643	644	643	639	642	642
AGAP004486	776	773	773	757	758	695	507	507
AGAP001760	821	817	821	504	503	502	771	502
AGAP001762	1118	707	1118	1072	742	335	941	1100
AGAP002933	1086	1086	1086	969	895	908	210	207
AGAP002935	827	826	826	802	486	406	798	1022
AGAP013533	618	445	446	235	149	-	680	-
AGAP003327	507	507	507	503	505	505	513	508
AGAP003328	570	569	569	245	243	-	132	132

Table S3. Selected genes from 2L chromosome and length of orthologous sequences in 7 species.

2L chromosome	*An. gambiae*-PEST	*An. gambiae*-M	*An. gambiae*-S	*An. stephensi*	*An. nili*	*An. funestus*	*Aedes*	*Culex*
AGAP004699	640	640	640	640	639	633	627	582
AGAP005014	1400	1385	1385	1187	1134	510	381	390
AGAP005279	531	531	531	499	496	484	489	489
AGAP005542	959	746	959	962	941	894	675	750
AGAP005851	831	831	831	829	770	770	828	570
AGAP006209	971	717	717	726	705	706	690	399
AGAP006531	903	901	900	879	884	457	885	873
AGAP006783	1017	1017	1016	1016	1016	492	1017	1017
AGAP007131	730	730	730	730	730	730	453	456
AGAP007504	850	850	850	843	851	450	849	849
AGAP005778	1216	1216	1216	1208	981	-	552	546
AGAP007068	712	712	712	718	712	539	555	555
AGAP007069	364	364	364	364	364	293	360	360

Table S4. Selected genes from 3R chromosome and length of orthologous sequences in 7 species.

3R Chromosome	*An. gambiae*-PEST	*An. gambiae*-M	*An. gambiae*-S	*An. stephensi*	*An. nili*	*An. funestus*	*Aedes*	*Culex*
AGAP007903	921	927	927	906	767	650	402	402
AGAP008652	597	595	596	571	561	575	453	453
AGAP008731	800	800	800	799	524	542	573	549
AGAP008915	795	795	795	801	419	599	411	408
AGAP009133	880	880	880	769	610	612	483	483
AGAP009512	549	549	549	541	530	353	207	393
AGAP010007	912	912	912	912	719	821	498	498
AGAP010267	943	946	943	798	801	796	438	438

Table S5. Selected genes from 3L chromosome and length of orthologous sequences in 7 species.

3L chromosome	*An. gambiae*-PEST	*An. gambiae*-M	*An. gambiae*-S	*An. stephensi*	*An. nili*	*An. funestus*	*Aedes*	*Culex*
AGAP010567	1173	1173	1173	1171	1173	938	1170	1023
AGAP011099	759	759	759	758	580	497	759	759
AGAP011357	743	743	743	741	742	741	381	396
AGAP011526	828	828	828	801	803	401	744	744
AGAP011765	1069	1069	1069	1069	1073	658	843	1062
AGAP012219	1040	607	1040	1035	720	1023	489	489

References

1. Fontenille D, Simard F (2004) Unravelling complexities in human malaria transmission dynamics in Africa through a comprehensive knowledge of vector populations. Comp Immunol Microbiol Infect Dis 27: 357-375.
2. Mouchet J, Carnevale P, Coosemans M, Julvez J, Manguin S, et al. (2004) Biodiversité du paludisme dans le monde Editions. . John Libbey Eurotext Montrouge, France.
3. Coluzzi M, Sabatini A, Petrarca V, Di Deco M (1979) Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. Transactions of the Royal Society of Tropical Medicine and Hygiene 73: 483-497.
4. Hay SI, Guerra CA, Gething PW, Patil AP, Tatem AJ, et al. (2009) A world malaria map: Plasmodium falciparum endemicity in 2007. PLoS Medicine 6: e1000048.
5. Ayala D, Costantini C, Ose K, Kamdem GC, Antonio-Nkondjio C, et al. (2009) Habitat suitability and ecological niche profile of major malaria vectors in Cameroon. Malar J 8: 307.
6. Antonio-Nkondjio C, Ndo C, Costantini C, Awono-Ambene P, Fontenille D, et al. (2009) Distribution and larval habitat characterization of Anopheles moucheti, Anopheles nili, and other malaria vectors in river networks of southern Cameroon. Acta Trop 112: 270-276.
7. Harbach R (2004) The classification of genus Anopheles (Diptera: Culicidae): a working hypothesis of phylogenetic relationships. Bulletin of Entomological Research 94: 537-554.
8. Sallum MAM, Schultz TR, Foster PG, Aronstein K, Wirtz RA, et al. (2002) Phylogeny of Anophelinae (Diptera : Culicidae) based on nuclear ribosomal and mitochondrial DNA sequences. Systematic Entomology 27: 361-382.
9. Favia G, Della Torre A, Bagayoko M, Lanfrancotti A, Sagnon NF, et al. (2003) Molecular identification of sympatric chromosomal forms of Anopheles gambiae and further evidence of their reproductive isolation. Insect molecular biology 6: 377-383.
10. della Torre A, Tu Z, Petrarca V (2005) On the distribution and genetic differentiation of Anopheles gambiae ss molecular forms. Insect biochemistry and molecular biology 35: 755-769.
11. Kengne P, Awono-Ambene P, Nkondjio CA, Simard F, Fontenille D (2003) Molecular identification of the Anopheles nili group of African malaria vectors. Medical and veterinary entomology 17: 67-74.
12. Ndo C, Simard F, Kengne P, Awono-Ambene P, Morlais I, et al. (2013) Cryptic Genetic Diversity within the Anopheles nili group of Malaria Vectors in the Equatorial Forest Area of Cameroon (Central Africa). PLOS ONE 8: e58862.
13. Sharakhova MV, Peery A, Antonio-Nkondjio C, Xia A, Ndo C, et al. (2013) Cytogenetic analysis of Anopheles ovengensis revealed high structural divergence of chromosomes in the Anopheles nili group. Infection, Genetics and Evolution.
14. Marshall JC, Powell JR, Caccone A (2005) Short report: Phylogenetic relationships of the anthropophilic Plasmodium falciparum malaria vectors in Africa. Am J Trop Med Hyg 73: 749-752.

15. Hittinger CT, Johnston M, Tossberg JT, Rokas A (2010) Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. Proc Natl Acad Sci U S A 107: 1476-1481.
16. Sharakhova MV, Antonio-Nkondjio C, Xia A, Ndo C, Awono-Ambene P, et al. (2011) Cytogenetic map for Anopheles nili: Application for population genetics and comparative physical mapping. Infect Genet Evol 11: 746-754.
17. Kamali M, Xia A, Tu Z, Sharakhov IV (2012) A New Chromosomal Phylogeny Supports the Repeated Origin of Vectorial Capacity in Malaria Mosquitoes of the Anopheles gambiae Complex. PLoS Pathog 8(10): e1002960.
18. Shalchian-Tabrizi K, Minge MA, Espelund M, Orr R, Ruden T, et al. (2008) Multigene phylogeny of choanozoa and the origin of animals. PLOS ONE 3: e2098.
19. Koepfli K-P, Deere K, Slater G, Begg C, Begg K, et al. (2008) Multigene phylogeny of the Mustelidae: resolving relationships, tempo and biogeographic history of a mammalian adaptive radiation. BMC biology 6: 10.
20. Gao F, Katz LA, Song W (2013) Multigene-based analyses on evolutionary phylogeny of two controversial ciliate orders: Pleuronematida and Loxocephalida (Protista, Ciliophora, Oligohymenophorea). Molecular Phylogenetics and Evolution.
21. Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798-804.
22. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science 298: 129-149.
23. Sharakhova MV, Hammond MP, Lobo NF, Krzywinski J, Unger MF, et al. (2007) Update of the Anopheles gambiae PEST genome assembly. Genome Biol 8: R5.
24. Peery A, Sharakhova MV, Antonio-Nkondjio C, Ndo C, Weill M, et al. (2011) Improving the population genetics toolbox for the study of the African malaria vector Anopheles nili: microsatellite mapping to chromosomes. Parasites and Vectors in review.
25. Kamali M, Sharakhova M, Baricheva E, Karagodin D, Tu Z, et al. (2011) An integrated chromosome map of microsatellite markers and inversion breakpoints for an Asian malaria mosquito, Anopheles stephensi. Journal of Heredity doi:10.1093/jhered/esr072.
26. Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, et al. (2010) Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. Science 330: 86-88.
27. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, et al. (2007) Genome sequence of Aedes aegypti, a major arbovirus vector. Science 316: 1718-1723.
28. Crawford JE, Guelbeogo WM, Sanou A, Traore A, Vernick KD, et al. (2010) De novo transcriptome sequencing in Anopheles funestus using Illumina RNA-seq technology. PLoS ONE 5: e14202.
29. Lawniczak MK, Emrich SJ, Holloway AK, Regier AP, Olson M, et al. (2010) Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences. Science 330: 512-514.
30. Fitch WM (2000) Homology: a personal view on some of the problems. Trends in genetics 16: 227-231.
31. Moreno-Hagelsieb G, Latimer K (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 24: 319-324.

32. Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings in bioinformatics 9: 299-306.
33. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution 4: 406-425.
34. George P, Sharakhova MV, Sharakhov IV (2010) High-resolution cytogenetic map for the African malaria vector Anopheles gambiae. Insect Mol Biol 19: 675-682.
35. Megy K, Emrich SJ, Lawson D, Campbell D, Dialynas E, et al. (2012) VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. Nucleic Acids Res 40: D729-734.
36. Maes P, Matthijnssens J, Rahman M, Van Ranst M (2009) RotaC: a web-based tool for the complete genome classification of group A rotaviruses. BMC microbiology 9: 238.
37. Ayala FJ, Coluzzi M (2005) Chromosome speciation: humans, Drosophila, and mosquitoes. Proc Natl Acad Sci U S A 102 Suppl 1: 6535-6542.
38. Garros C, Harbach RE, Manguin S (2005) Morphological assessment and molecular phylogenetics of the Funestus and Minimus Groups of Anopheles (Cellia). Journal of medical entomology 42: 522-536.

CHAPTER 5 An Integrated Chromosome Map of Microsatellite Markers and Inversion Breakpoints for an Asian Malaria Mosquito, *Anopheles stephensi*

The following chapter is published in the Journal of Heredity. As an author I retain the right to include this article in dissertation.

Kamali M, Sharakhova MV, Baricheva E, Karagodin D, Tu Z, et al. (2011) An Integrated Chromosome Map of Microsatellite Markers and Inversion Breakpoints for an Asian Malaria Mosquito, *Anopheles stephensi*. Journal of Heredity 102: 719-726.

An Integrated Chromosome Map of Microsatellite Markers and Inversion Breakpoints for an Asian Malaria Mosquito, *Anopheles stephensi*

Maryam Kamali[1], Maria V. Sharakhova[1], Elina Baricheva[2], Dmitrii Karagodin[2], Zhijian Tu[3], Igor V. Sharakhov[1]

[1]Department of Entomology, Virginia Polytechnic Institute and State University, Blacksburg, VA, United States

[2]Institute of Cytology and Genetics, Russian Academy of Sciences, Novosibirsk, Russia

[3]Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, VA, United States

Author's emails:

MK: mkamali@vt.edu

MVS: msharakh@vt.edu

EB: barich@bionet.nsc.ru

DK: gtc2000trit@gmail.com

ZT: jaketu@vt.edu

IVS: igor@vt.edu

Running title: A Microsatellite Map of *Anopheles stephensi*

Abstract

Anopheles stephensi is one of the major vectors of malaria in the Middle East and Indo-Pakistan subcontinent. Understanding the population genetic structure of malaria mosquitoes is important for developing adequate and successful vector control strategies. Commonly used markers for inferring anopheline taxonomic and population status include microsatellites and chromosomal inversions. Knowledge about chromosomal locations of microsatellite markers with respect to polymorphic inversions could be useful for better understanding a genetic structure of natural populations. However, fragments with microsatellites used in population genetic studies are usually too short for successful labeling and hybridization with chromosomes. We designed new primers for amplification of microsatellite loci identified in the *An. stephensi* genome sequenced with next-generation technologies. Twelve microsatellites were mapped to polytene chromosomes from ovarian nurse cells of *An. stephensi* using fluorescent *in situ* hybridization. All microsatellites hybridized to unique locations on autosomes, and seven of them localized to the largest arm 2R. Ten microsatellites were mapped inside the previously described polymorphic chromosomal inversions, including four loci located inside the widespread inversion 2Rb. We analyzed microsatellite-based population genetic data available for *An. stephensi* in light of our mapping results. This study demonstrates that the chromosomal position of microsatellites may affect estimates of population genetic parameters and highlights the importance of developing physical maps for nonmodel organisms.

Keywords: genome sequence, polymorphic inversion, malaria vector, polytene chromosome, population structure.

Introduction

Anopheles stephensi Liston (Diptera: Culicidae) is an important malaria vector in the Persian Gulf and South Iran (Manouchehri, et al., 1976; Vatandoost, et al., 2006), in urban areas of the Indian subcontinent (Hati, 1997; Pant, et al., 1981), as well as in rural areas of North Pakistan and East Afghanistan (Rowland, et al., 2002). This species is also an outstanding laboratory model system for malaria parasite transmission studies (Abraham, et al., 2004; Bass, et al., 2008; Baton and Ranford-Cartwright, 2007). Three morpho-ecological variants have been identified within *An. stephensi* populations—*type, intermediate,* and *mysorensis*—which can be identified by the number of ridges on the egg (Rao, et al., 1938; Subbarao, et al., 1987; Sweet and Rao, 1937). The *type* form is mainly an urban mosquito that breeds in temporary water pools, whereas the *mysorensis* and *intermediate* variants occupy predominantly rural areas (Nagpal, et al., 2003). Although *mysoriensis* is considered mostly zoophilic and is a poor vector of malaria in India (Subbarao, et al., 1987), it was found to be a major malaria vector in the southern part of Iran (Manouchehri, et al., 1976). In spite of the importance of bionomic differentiation for efficient malaria transmission, the genetic structure of the biological forms and populations of *An. stephensi* remains unclear, as tools are insufficient to precisely characterize the genetic variation. A study of PCR-based restriction fragment length polymorphism (PCR-RFLP) of 1512 bp of mitochondrial DNA cytochrome oxidase subunit I and II (COI-COII) and partial sequences of COI and COII genes has found extensive gene flow among the different forms of *An. stephensi* in Iran (Oshaghi, et al., 2006). Analysis of the ribosomal DNA internal transcribed spacer 2 (rDNA-ITS2) and random amplification of polymorphic DNA (RAPD) loci in different populations of *An. stephensi* has demonstrated very little genetic variation among different populations suggesting that *An. stephensi* in Iran is a single species with different ecological

forms in different zoogeographical zones (Djadid, et al., 2006). Sequencing of the rDNA-ITS2 and domain-3 (D3) of rDNA loci of the *An. stephensi type* and *mysorensis* in India did not find any intraspecies sequence variation (Alam, et al., 2008).

Microsatellites are informative markers for inferring population and taxonomic status of various organisms (Bruford and Wayne, 1993). These markers have high levels of polymorphism and tend to evolve neutrally. They have been successfully used to study gene flow in natural populations of *An. gambiae*, *An. funestus* and *An. nili* (Cohuet, et al., 2005; Lehmann, et al., 1996; Ndo, et al., 2010). A set of 16 microsatellite markers has been developed for *An. stephensi* (Verardi, et al., 2002). A study using seven of these microsatellite markers has revealed high levels of genetic diversity within populations but not among geographically isolated populations in Pakistan. Deviation from Hardy – Weinberg expectations has been observed for two microsatellite loci in 21 tests (Ali, et al., 2007). Another study of genetic variation at eight of the 16 microsatellite loci observed a significant differentiation and a low level of gene flow among three ecological variants of *An. stephensi* in India (Vipin and Gakhar, 2010). The study demonstrated that some microsatellites were in significant linkage disequilibrium, while other loci had a heterozygote deficit. Reduced recombination and selection can influence loci within polymorphic chromosomal inversions or near inversion breakpoints, resulting in estimates of gene flow that may depart significantly from those based on loci elsewhere in the genome (Lanzaro, et al., 1998; Tripet, et al., 2005). However, locations of the microsatellites on the chromosomes of *An. stephensi* were unknown, and interpretation of these differences among the loci was difficult.

Geographical and seasonal changes in chromosomal inversion frequencies have been interpreted as a signature of natural selection (Balanya, et al., 2006; Hoffmann and Rieseberg, 2008;

Hoffmann, et al., 2004). Therefore, polymorphic inversions are useful markers for studying ecological adaptations of various species, including malaria mosquitoes (Ayala, et al., 2011; Sharakhova, et al., 2011). An informative approach to determine the role of inversions in a population structure and species environmental adaptation is to study the genetic variation of molecular markers inside and outside of chromosomal rearrangements (Navarro and Barton, 2003). In Diptera, contrasting patterns of polymorphism at microsatellite loci and inversions have been revealed. For example, some microsatellite loci located within known inversions did not display clinal profiles even if inversions had clinal profiles (Ayala, et al., 2011; Cohuet, et al., 2005; Kennington, et al., 2003; Michel, et al., 2005; Onyabe and Conn, 2001). The number of genes under selection within inversions and the inversion age can be responsible for the observed patterns.

Polymorphic inversions have been employed to investigate the population structure and gene flow between ecological forms of *An. stephensi.* Breakpoints of at least 24 paracentric inversions have been mapped to polytene chromosomes of *An. stephensi* (Gayathri Devi and Shetty, 1992; Mahmood and Sakai, 1984). Seven inversions are located on 2R, four inversions on 2L, three inversions on 3R, and ten inversions on 3L. No polymorphic inversions have been identified on the X chromosome. The 2Rb inversion was found to be highly polymorphic and widespread in natural populations of *An. stephensi*. Population studies provide support for the variation in inversion polymorphism according to geographic distribution of *An. stephensi* especially with respect to the 2Rb inversion (Coluzzi, et al., 1973; Gayathri Devi and Shetty, 1992; Mahmood and Sakai, 1984). For example, a chromosomal study revealed striking differences in the kinds and frequencies of paracentric inversions between the urban population in Karachi city and rural populations in Lahore and Kasur districts in Pakistan (Mahmood and

Sakai, 1984). The *type* and *mysorensis* forms showed significant differences in frequencies of inversion 2Rb (Coluzzi, et al., 1973). However, another study did not find any correlation between the inversion and the number of ridges on the egg (Suguna, 1981).

In this study, we determined the locations of 12 microsatellite markers on a cytogenetic map of *An. stephensi.* All microsatellites hybridized to unique locations on autosomes both inside and outside polymorphic inversions. The chromosomal map of microsatellite markers and inversion breakpoints helps to understand better the genetic variation and differentiation in natural populations of *An. stephensi*.

Material and Methods

Mosquito strain and chromosome preparation

The Indian wild-type laboratory strain of *Anopheles stephensi* was used in this study. To obtain the polytene chromosomes, ovaries were taken from half-gravid females and placed in Carnoy's fixative solution (3 parts of ethanol: 1 part of glacial acetic acid by volume). Ovaries were kept at room temperature overnight before being stored at -20 °C. To obtain chromosomal slides, follicles of ovaries were separated in 50% propionic acid. Then a cover slip was used to squash the follicles. The quality of slides and the banding pattern of polytene chromosomes were analyzed using an Olympus CX-41 phase contrast microscope (Olympus America Inc., Melville, NY, USA). Slides then were dipped into liquid nitrogen, cover slips were removed, and slides were dehydrated in 50%, 70%, 90%, and 100% ethanol. Slides were air dried and used for further experiments.

Probe preparation

Three approaches were utilized for the microsatellite probe preparation. First, microsatellites were directly amplified from the genomic DNA using previously designed primers (Verardi, et al., 2002). Genomic DNA of *An. stephensi* was prepared using the Qiagen DNeasy Blood and Tissue Kit (Qiagen Science, Germantown, MD, USA). Approximately 80-200 bp-long fragments were amplified. Second, four microsatelites were ligated and cloned as a cluster of two or three repeats in the head-to-tail orientation in the plasmid pBluescript SK(+) (Agilent Technologies, Inc., Santa Clara, CA, USA). Amplicons prepared in the first approach were treated with T4 DNA polimerase (SibEnzyme Ltd., Novosibirsk, Russia) for blunting the ends. Resulting fragments were treated withT4 DNA ligase (SibEnzyme Ltd., Novosibirsk, Russia) at 14° C for 1-2 hours. The mixture was added to the dephosphorylated plasmid digested with restriction enzyme EcoRV (SibEnzyme Ltd., Novosibirsk, Russia). Ligation was performed overnight. A ligation mixture was used to transform the XL1-Blue *E. coli* strain (Agilent Technologies, Inc., Santa Clara, CA, USA). The blue-and-white screening revealed colonies containing insertions. The PCR screening revealed colonies containing a cluster of several repeats. Plasmids containing such clusters were sequenced. In this approach, 500-850 bp-long fragments were amplified from the plasmid DNA using standard T7 and T6 primers (Fermentas, Inc., Glen Burnie, MD, USA) and were labeled for *in situ* hybridization. Third, PCR products were obtained for 12 microsatellites using primers designed with the Primer3 program (Rozen and Skaletsky, 2000) based on sequences identified by BLASTN in the genome assembly of *An. stephensi* (accession numbers: HQ328840 - HQ328851). The genome assembly for *An. stephensi* was obtained using 16X coverage of 454 shotgun and pair-end sequences at the Core Laboratory Facility of the Virginia Bioinformatics Institute of Virginia Tech. The size of these fragments was from 493 to

584 bp. PCR conditions were as follows: 94°C for 5 min; 45 cycles of 94°C for 45 s and of 50°C for 45 s; 72 °C for 30 s; and 72 °C for 5 min. DNA was purified using the GE healthcare illustra GFX PCR DNA and Gel Band Purification Kit (GE Healthcare UK Ltd., Buckinghamshire, UK). Probes were labeled using Cy3-AP3-dUTP or Cy5-AP3-dUTP (GE Healthcare UK Ltd., Buckinghamshire, UK) fluorophores by a Random Primer DNA Labeling System (Invitrogen Corporation, Carlsbad, CA, USA).

Fluorescence in situ hybridization (FISH) and mapping

Labeled probes were hybridized at 42°C to *An. stephensi* polytene chromosome slides overnight. Then, slides were washed in 0.2 X SSC (Saline Sodium citrate, 0.03 M sodium chloride, and 0.03 M sodium citrate) at 42°C and room temperature. Chromosomes were stained using YOYO-1 (Invitrogen Corporation, Carlsbad, CA, USA), and slides were mounted in DABCO. A Zeiss LSM 510 Laser Scanning Microscope (Carl Zeiss MicroImaging, Inc., Thornwood, NY, USA) was used to detect fluorescent signals. Microscopic images were taken with a digital camera, and the locations of signals were determined using a standard cytogenetic photo map of *An. stephensi* (Sharakhova, et al., 2006).

Results

In the current study, we used three approaches to map the set of microsatellites to the polytene chromosomes from ovarian nurse cells of *An. stephensi*. In the first approach, microsatellites were amplified from genomic DNA using specific primers, which were developed before (Verardi, et al., 2002). All microsatellites were successfully amplified from the genomic DNA. However because of the small size of the products (between 82-200 bp), the probes failed to get labeled for successful *in situ* hybridization. In the second approach, the PCR products of four microsatellites (A1, A10, B1, C1) were cloned in a plasmid as a cluster of two or three repeats in

the head-to-tail orientation. In this approach, 500-850 bp-long fragments were amplified from the plasmid DNA using standard T7 and T6 primers and were labeled for *in situ* hybridization with chromosomes. Labeling and FISH of these microsatellite clones were successful. However, the probes hybridized either to multiple sites on chromosomes or to heterochromatic regions, probably because of the increased size of repetitive microsatellite motifs. In the third approach, we used recently obtained genomic sequence assembly of *An. stephensi* to design primers for 493-584 bp-long PCR products that contain microsatellites of interest (Table 1). We attempted to find sequences homologous to the previously described 16 microsatellite loci (Verardi, et al., 2002) in the *An. stephensi* genome by BLASTN. The BLASTN search did not yield either positive or unique hits for microsatellites E7*, A1, and D8T. Primers were designed but no PCR product was obtained for microsatellite H1. Therefore, PCR products were obtained for 12 out of 16 microsatellites using primers designed based on the *An. stephensi* genome sequences (accession numbers: HQ328840 - HQ328851). We successfully hybridized these probes to the polytene chromosomes from ovarian nurse cells of *An. stephensi* by FISH (Figure 1, Table 2).

The *An. stephensi* chromosomal complement in ovarian nurse cells consists of five chromosomal arms: X, 2R, 3R, 3L. All 12 microsatellites hybridized to unique locations on autosomes; no hybridization to the X chromosome was detected. We have mapped these microsatellites to the polytene chromosomes of *An. stephensi* (Figure 2). Seven of 12 microsatellites hybridized to the largest arm 2R; two microsatellites localized to each of the 2L and 3R arms, and only one microsatellite hybridized to the 3L arm. At least 24 paracentric inversions have been described for *An. stephensi* (Gayathri Devi and Shetty, 1992; Mahmood and Sakai, 1984). We placed breakpoints of these inversions on the most recently developed polytene chromosome map of *An. stephensi* (Sharakhova, et al., 2006) (Figure 2). Ten

microsatellites were mapped inside of the previously described polymorphic inversions. Four of them were found inside the large *2Rb* inversion, which is polymorphic in the Indian wild-type laboratory strain of *An. stephensi* (Figure 3). The locations of two microsatellites, D11 and C1, were cytogenetically indistinguishable; they hybridized to the same band in region 13C on the 2R arm.

Discussion

Among the three approaches to map microsatellite markers to chromosomes, the use of the genome sequence assembly of *An. stephensi* to amplify and hybridize microsatellites by FISH was successful (Figure 1). This success resulted because the genome sequence allowed us to design primers that would amplify 493-584 bp fragments, which are suitable for effective labeling by the random primer method. The developed microsatellite map (Figure 2) can greatly improve the understanding a population genetic structure of *An. stephensi*. Three ecological forms of *An. stephensi—type, intermediate,* and *mysorensis,* which differ in their habitat preferences and malaria transmission, have been described (Rao, et al., 1938; Subbarao, et al., 1987; Sweet and Rao, 1937). Population analysis using rDNA-ITS2 and mitochondrial DNA loci have demonstrated a low level of gene flow among the three variants (Djadid, et al., 2006; Oshaghi, et al., 2006). Genetic variation at seven microsatellite loci (F10, H2ii, E12, B1, A7, C1, and A10) has been studied in 153 individuals belonging to three populations in Pakistan (Ali, et al., 2007). The study has found significant deviation from Hardy – Weinberg equilibrium due to heterozygote deficit in two of the 21 tests (loci E12 and B1). Exact tests of linkage disequilibrium showed no significant departure from equilibrium between these or any other locus pairs in any population after Bonferroni correction (Ali, et al., 2007). Our study mapped

both microsatellite markers, E12 and B1, to the region near the telomere on the 2R arm (Figure 2) suggesting the presence of genes under selection in this chromosomal region.

Another population genetic study utilized eight microsatellite markers (F10, H2ii, E12, B1, G11, E7T, G1, and A10) to investigate the genetic isolation among the three morpho-ecological variants of *An. stephensi* in India (Vipin and Gakhar, 2010). These microsatellites had shown diverse patterns of genetic variation. Locus E7T was found to be highly differentiated between *mysorensis* and *type* forms (Fst = 0.890), between *intermediate* and *type* forms (Fst = 0.629), and between *mysorensis* and *intermediate* forms (Fst = 0.556). Moreover all three ecological forms had different nonoverlapping allele sizes for E7T locus (Vipin and Gakhar, 2010). These data suggest high genetic differentiation among the forms at the locus E7T. We mapped the microsatellite E7T inside the 2Rb inversion (Figures 2 and 3), which is the most widespread inversion in natural populations. It has been shown that the *type* and *mysorensis* forms significantly differ in frequencies of inversion 2Rb (Coluzzi, et al., 1973). Therefore, the presence of this microsatellite inside the 2Rb inversion confirms a possible role of the inversion in differentiating *An. stephensi* populations. It is possible that the inversion 2Rb acts as a barrier to gene flow among the forms. If this is the case, then E7T should be in linkage disequilibrium with the inversion. Other microsatellites located inside the 2Rb inversion (B2, D11, and C1) have not been included in the population genetics study (Vipin and Gakhar, 2010). Significant departure from Hardy – Weinberg equilibrium due to heterozygote deficits was found in *intermediate* and *mysorensis* forms across all loci, except for G1 and A10 in all three variants and F10 in *intermediate* (Vipin and Gakhar, 2010). Both G1 and A10 are located on 3R, and F10 is located on 2L. Microsatellite G11 has not been mapped in our study. Interestingly, the other four microsatellites with significant departure from Hardy – Weinberg equilibrium (H2ii, E12,

B1, E7T) were mapped on the 2R arm in our study (Figure 2). We have mapped microsatellites H2ii, B1, and E12 on the 2R arm outside of the 2Rb inversion; E12 and B1 are located close to the telomere. Linkage disequilibrium was found between H2ii and B1, E12 and B1 in *mysorensis*, and E12 and B1 in *intermediate* (Vipin and Gakhar, 2010). However, the Bonferroni correction for the linkage disequilibrium has not been performed.

The genome sequence assembly of *An. stephensi* can also be used to discover and develop new microsatellite markers. Future studies of these microsatellites, together with inversion polymorphisms in the natural populations, will provide a better understanding of the population structure of *An. stephensi* and the effect of inversions on the behavior of microsatellites.

Funding

This work was supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health (grant 1R21AI081023 to I.V.S).

Acknowledgements

We thank Diego Ayala for helpful comments on the manuscript and Melissa Wade for editing the text. Insightful comments from two anonymous reviewers helped to improve the manuscript.

Figures

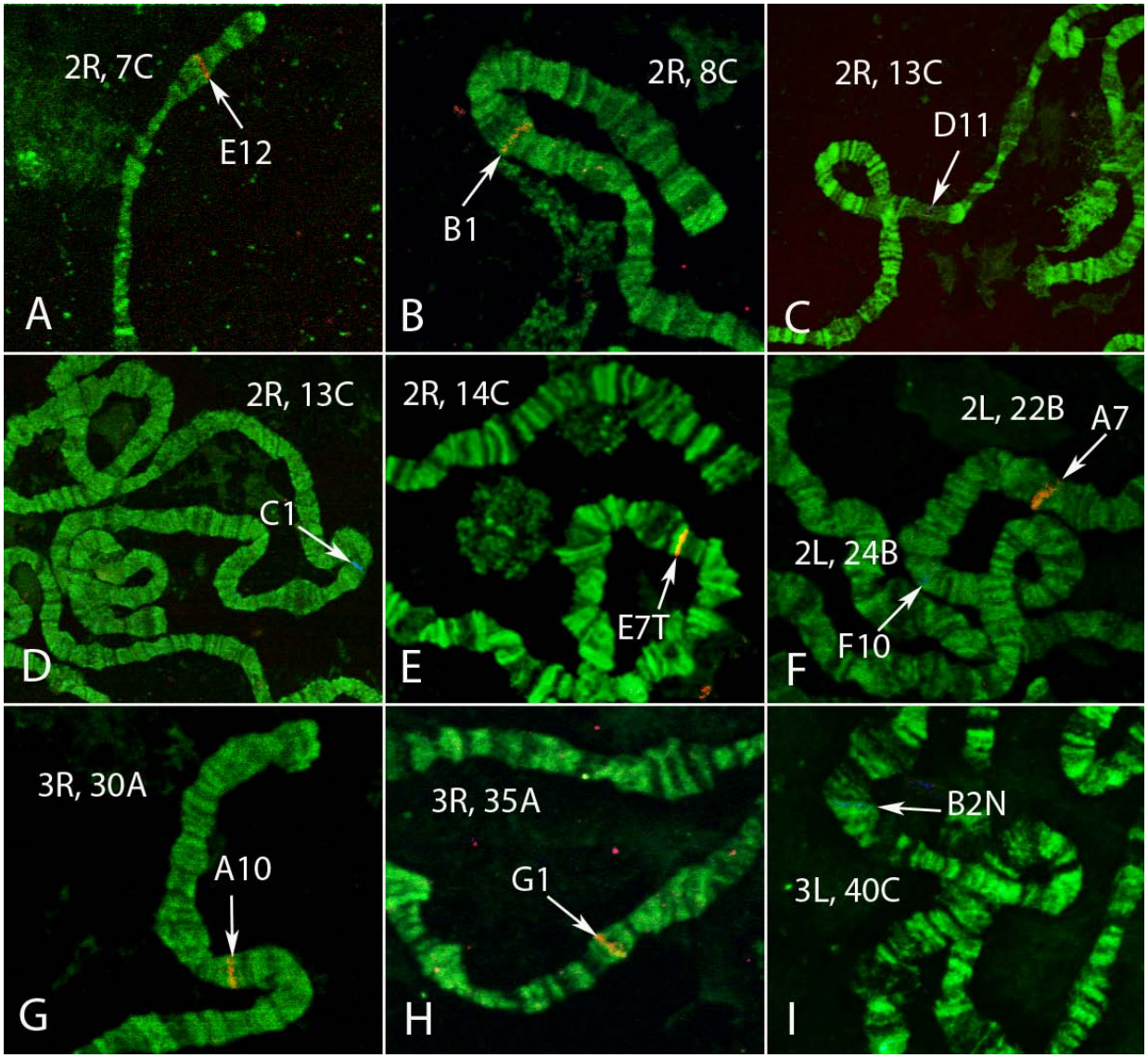

Figure 1. FISH performed on the chromosomes of *An. stephensi*.
Chromosomes counterstained with the fluorophore YOYO-1 and hybridized with fluorescently labeled probes Cy5 (blue) and Cy3 (red) are shown.

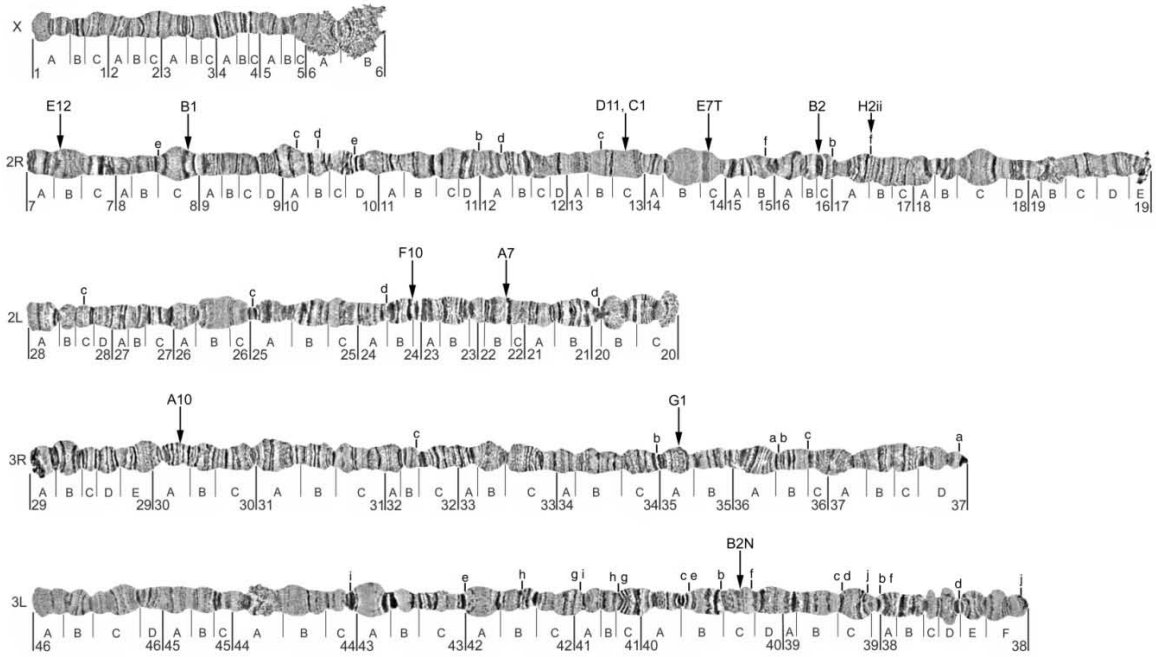

Figure 2. Physical map of 12 microsatellite markers on the *An. stephensi* polytene chromosomes.

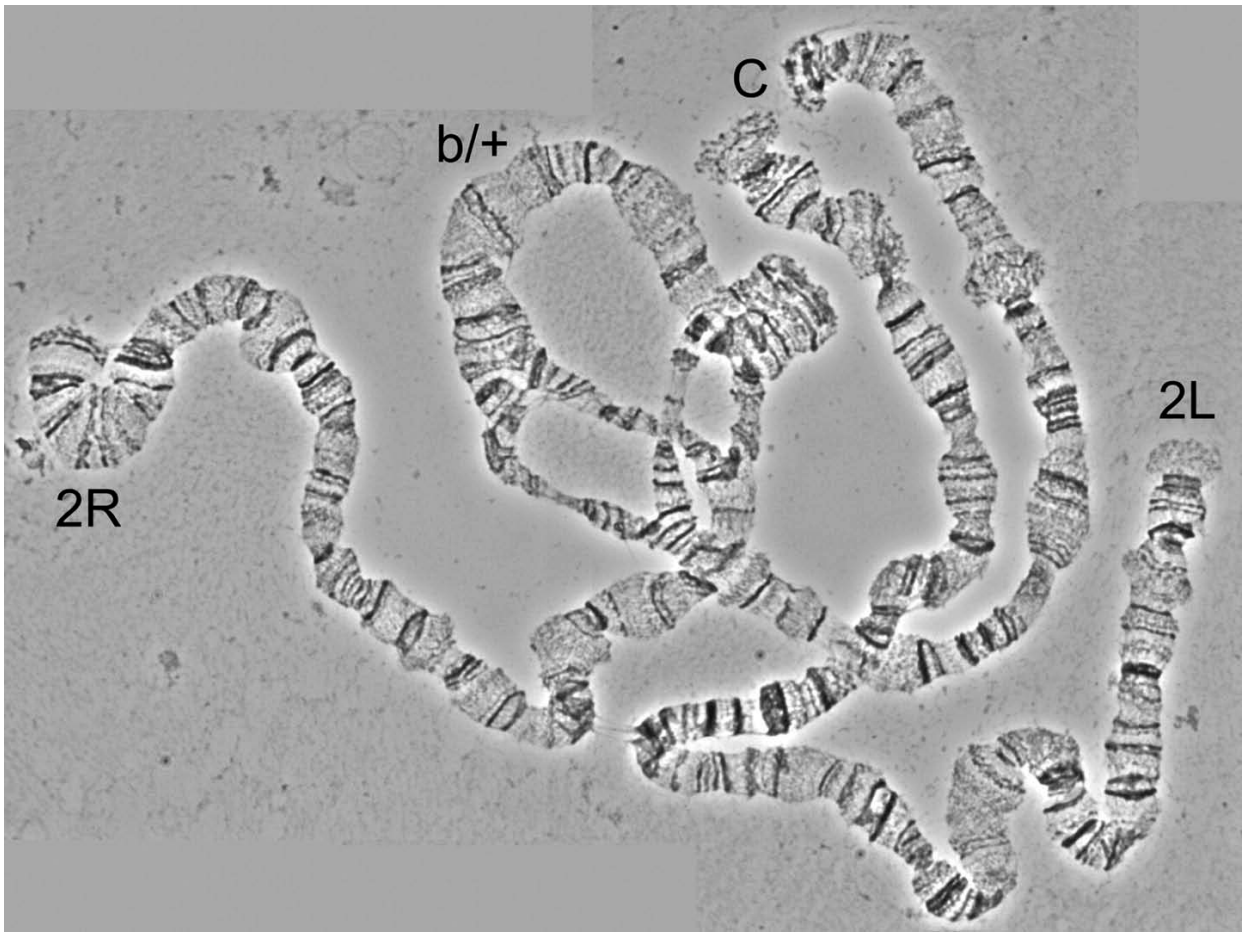

Figure 3. A photomicrograph of the polymorphic inversion 2Rb in a heterozygote state from the Indian wild-type laboratory colony of *An. stephensi.*

Tables

Table 1. Primers designed for the microsatellite loci using the *An. stephensi* genome sequence.

Locus	Forward primer	Reverse primer	Size of the PCR product (bp)
E12	GCGAGAGCGAGAGAGTGAGT	GGCGTTCAGTTCTGTGTGAA	503
B1	GCATGGGTATGAGCCAAGTT	ATGAGTGTCGTCGTCCGTTT	510
D11	ACCAGGGGTTCACAAATTCA	TGTACACAGGAGATAACGTGCAT	493
C1	CACAGGAGATAACGTGCATTT	ACGCTCACACCACAAAACC	505
E7T	ACGAAGGAGCTGTCCGAGTA	GTCGTGTGGGGATAGTTGCT	584
B2	CAGGAAAAGCGAGTGAAAGG	TGACAGCTGTGGAAGATTCG	531
H2ii	TCGATTCGAGGCATCTTTTC	CCTTCAATGTCCGTCACCTT	564
F10	ATCCCACTACTGCACCCACT	CATCGCATGCTGATTGTTCT	580
A7	CAGTTTTGCGCAGTAGTTGG	TTTCGCCTTTCATTCCTACG	549
A10	CACGCAAGTAGGCTTTGACA	TTGAAATCGCTTCACACGAC	513
G1	CAAGCGATTTTGGGGTAGAA	TACACCACCCACCCATAACC	514

Table 2. Locations of the *An. stephensi* microsatellite markers on polytene chromosomes.

	Microsatellite locus	Accessions for cloned sequences (with length in bp)	Accessions for assembled sequences (with length in bp)	Location; inside/outside inversions
1.	*E12*	AF418586 (229)	HQ328840 (794)	2R, B7; *outside*
2.	*B1*	AF418596 (150)	HQ328841 (847)	2R, C8; inside
3.	*D11*	AF418581 (82)	HQ328843 (630)	2R, 13C; inside
4.	*C1*	AF418583 (114)	HQ328844 (539)	2R, 13C; inside
5.	*E7T*	AF418593 (144)	HQ328842 (980)	2R, 14C; inside
6.	*B2*	AF418588 (210)	HQ328845 (739)	2R, 16C; inside
7.	*H2ii*	AF418595 (189)	HQ328846 (619)	2R, 17A; *outside*
8.	*F10*	AF418589 (175)	HQ328847 (700)	2L, 24B; inside
9.	*A7*	AF418592 (289)	HQ328848 (985)	2L, 22B; inside
10.	*A10*	AF418591 (186)	HQ328849 (666)	3R, 30A; *outside*
11.	*G1*	AF412812 (154)	HQ328850 (702)	3R, 35A; inside
12.	*B2N*	AF418590 (134)	HQ328851 (814)	3L, 40C; inside

References

Abraham EG, Islam S, Srinivasan P, Ghosh AK, Valenzuela JG, Ribeiro JM, Kafatos FC, Dimopoulos G, Jacobs-Lorena M. 2004. Analysis of the *Plasmodium* and *Anopheles* transcriptional repertoire during ookinete development and midgut invasion, J Biol Chem, 279:5573-5580.

Alam MT, Bora H, Das MK, Sharma YD. 2008. The type and mysorensis forms of the *Anopheles stephensi* (Diptera: Culicidae) in India exhibit identical ribosomal DNA ITS2 and domain-3 sequences, Parasitol Res, 103:75-80.

Ali N, Hume JC, Dadzie SK, Donnelly MJ. 2007. Molecular genetic studies of *Anopheles stephensi* in Pakistan, Med Vet Entomol, 21:265-269.

Ayala D, Fontaine MC, Cohuet A, Fontenille D, Vitalis R, Simard F. 2011. Chromosomal inversions, natural selection and adaptation in the malaria vector *Anopheles funestus*, Mol Biol Evol, 28:745-758.

Balanya J, Oller JM, Huey RB, Gilchrist GW, Serra L. 2006. Global genetic change tracks global climate warming in *Drosophila subobscura*, Science, 313:1773-1775.

Bass C, Nikou D, Blagborough AM, Vontas J, Sinden RE, Williamson MS, Field LM. 2008. PCR-based detection of *Plasmodium* in *Anopheles* mosquitoes: a comparison of a new high-throughput assay with existing methods, Malar J, 7:177.

Baton LA, Ranford-Cartwright LC. 2007. Morphological evidence for proliferative regeneration of the *Anopheles stephensi* midgut epithelium following *Plasmodium falciparum* ookinete invasion, J Invertebr Pathol, 96:244-254.

Bruford M, Wayne R. 1993. Microsatellites and their application to population genetic studies, Current Opinion in Genetics & Development, 3:939-943.

Cohuet A, Dia I, Simard F, Raymond M, Rousset F, Antonio-Nkondjio C, Awono-Ambene P, Wondji C, Fontenille D. 2005. Gene flow between chromosomal forms of the malaria vector *Anopheles funestus* in Cameroon, Central Africa, and its relevance in malaria fighting, Genetics, 169:301.

Coluzzi M, Di Deco M, Cancrini G. 1973. Chromosomal inversions in *Anopheles stephensi*, Parassitologia, 15:129-136.

Coluzzi M, Di Deco M, Cancrini G. 1973. Further observations on the egg length in *Anopheles stephensi* in relation to chromosomal polymorphism, Parassitologia, 15:213-215.

Djadid N, Gholizadeh S, Aghajari M, Zehi A, Raeisi A, Zakeri S. 2006. Genetic analysis of rDNA-ITS2 and RAPD loci in field populations of the malaria vector, *Anopheles stephensi* (Diptera: Culicidae): implications for the control program in Iran, Acta tropica, 97:65-74.

Gayathri Devi K, Shetty J. 1992. Chromosomal inversions in *Anopheles stephensi* Liston--a malaria mosquito., J Cytol Genet, 27:153-161.

Hati AK. 1997. Urban malaria vector biology, Indian J Med Res, 106:149-163.

Hoffmann AA, Rieseberg L. 2008. Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation. Annual Review of Ecology, Evolution, and Systematics, 39:21-42.

Hoffmann AA, Sgro CM, Weeks AR. 2004. Chromosomal inversion polymorphisms and adaptation, Trends Ecol Evol, 19:482-488.

Kennington WJ, Gockel J, Partridge L. 2003. Testing for asymmetrical gene flow in a *Drosophila melanogaster* body-size cline, Genetics, 165:667-673.

Lanzaro GC, Toure YT, Carnahan J, Zheng L, Dolo G, Traore S, Petrarca V, Vernick KD, Taylor CE. 1998. Complexities in the genetic structure of *Anopheles gambiae* populations in west Africa as revealed by microsatellite DNA analysis, Proc Natl Acad Sci U S A, 95:14260-14265.

Lehmann T, Hawley W, Kamau L, Fontenille D, Simard F, Collins F. 1996. Genetic differentiation of *Anopheles gambiae* populations from East and West Africa: comparison of microsatellite and allozyme loci, Heredity, 77:192-200.

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

Manouchehri A, Javadian E, Eshighy N, Motabar M. 1976. Ecology of *Anopheles stephensi* Liston in southern Iran, Trop Geogr Med, 28:228-232.

Michel AP, Guelbeogo WM, Grushko O, Schemerhorn BJ, Kern M, Willard MB, Sagnon N, Costantini C, Besansky NJ. 2005. Molecular differentiation between chromosomally defined incipient species of *Anopheles funestus*, Insect Mol Biol, 14:375-387.

Nagpal BN, Srivastava A, Kalra NL, Subbarao SK. 2003. Spiracular indices in *Anopheles stephensi*: a taxonomic tool to identify ecological variants, J Med Entomol, 40:747-749.

Navarro A, Barton NH. 2003. Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation, Evolution, 57:447-459.

Ndo C, Antonio-Nkondjio C, Cohuet A, Ayala D, Kengne P, Morlais I, Awono-Ambene PH, Couret D, Ngassam P, Fontenille D, Simard F. 2010. Population genetic structure of the malaria vector *Anopheles nili* in sub-Saharan Africa, Malar J, 9:161.

Onyabe DY, Conn JE. 2001. Genetic differentiation of the malaria vector *Anopheles gambiae* across Nigeria suggests that selection limits gene flow, Heredity, 87:647-658.

Oshaghi M, Yaaghoobi F, Abaie M. 2006. Pattern of mitochondrial DNA variation between and within *Anopheles stephensi* (Diptera: Culicidae) biological forms suggests extensive gene flow, Acta tropica, 99:226-233.

Pant CP, Rishikesh N, Bang YH, Smith A. 1981. Progress in malaria vector control, Bulletin of the World Health Organization, 59:325-333.

Rao BA, Sweet WC, Subbarao AM. 1938. Ova measurements of *A. stephensi type* and *A. stephensi* var. *mysorensis.*, J. Malar. Inst. India., 1:261-266.

Rowland M, Mohammed N, Rehman H, Hewitt S, Mendis C, Ahmad M, Kamal M, Wirtz R. 2002. Anopheline vectors and malaria transmission in eastern Afghanistan, Transactions of the Royal Society of Tropical Medicine and Hygiene, 96:620-626.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers, Methods Mol Biol, 132:365-386.

Sharakhova MV, Antonio-Nkondjio C, Xia A, Ndo C, Awono-Ambene P, Simard F, Sharakhov IV. 2011. Cytogenetic map for *Anopheles nili*: Application for population genetics and comparative physical mapping, Infect Genet Evol., 11:746-54.

Sharakhova MV, Xia A, McAlister SI, Sharakhov IV. 2006. A standard cytogenetic photomap for the mosquito *Anopheles stephensi* (Diptera: Culicidae): application for physical mapping, J Med Entomol, 43:861-866.

Subbarao S, Vasantha K, Adak T, Sharma V, Curtis C. 1987. Egg-float ridge number in *Anopheles stephensi*: ecological variation and genetic analysis, Medical and Veterinary Entomology, 1:265-271.

Suguna SG. 1981. Inversion(2)R1 in *Anopheles stephensi*, its distribution and relation to egg size, Indian J Med Res, 73 Suppl:124-128.

Sweet W, Rao B. 1937. Races of *Anopheles stephensi* Liston, 1901, Indian Medical Gazette, 72:665-674.

Tripet F, Dolo G, Lanzaro GC. 2005. Multilevel analyses of genetic differentiation in *Anopheles gambiae s.s.* reveal patterns of gene flow important for malaria-fighting mosquito projects, Genetics, 169:313-324.

Vatandoost H, Oshaghi MA, Abaie MR, Shahi M, Yaaghoobi F, Baghaii M, Hanafi-Bojd AA, Zamani G, Townson H. 2006. Bionomics of *Anopheles stephensi* Liston in the malarious area of Hormozgan province, southern Iran, 2002, Acta Trop, 97:196-203.

Verardi A, Donnelly M, Rowland M, Townson H. 2002. Isolation and characterization of microsatellite loci in the mosquito *Anopheles stephensi* Liston (Diptera: Culicidae), Molecular Ecology Notes, 2:488-490.

Vipin M, Gakhar S. 2010. Genetic differentiation between three ecological variants ('type','mysorensis' and 'intermediate') of malaria vector *Anopheles stephensi* (Diptera: Culicidae), Insect Science.

CHAPTER 6 Summary

The *Anopheles gambiae* complex consists of seven morphologically indistinguishable sibling species. Members of the complex have different geographical distribution, behaviors and vectorial capacity. However, the evolutionary history among the members of the complex is not fully understood. The goal for this dissertation research was to determine the evolutionary history among members of the *An. gambiae* complex and better understand the evolution of epidemiologically important traits by using cytogenetic and molecular approaches. Constructing a phylogeny based on the analysis of inversion breakpoints provided a methodology for rooting chromosomal phylogeny among sibling species. The molecular phylogeny of inversion breakpoints resulted in confirming the chromosomal phylogeny. The importance of genetic marker position with respect to chromosomal inversion in phylogeny construction is emphasized. In addition, we performed a genome-wide analysis to reconstruct a phylogenetic relationship of African malaria vectors and to reveal the basal position of *An. nili* among other major vectors of malaria in Arica. The final goal of this project included mapping microsatellites markers to polytene chromosomes of the Asian malaria vector *An. stephensi*. Our results provided new insights into understanding the evolutionary history of the *An. gambiae* complex as well as other major vectors of malaria in Africa and Asia.

Chapter 2

The phylogenetic relationship among the members of the *An. gambiae* complex is not resolved. This is due to a high degree of genetic similarity among sibling species. In this study we have analyzed the fixed 2Ro and 2Rp chromosomal inversions in *An. merus,* and the homologous sequences in several outgroup species to infer the ancestral–descendant relationships among members of *An. gambiae* complex. We have screened the *An. merus* phage, and *An. stephensi*

BAC library. Genes adjacent to breakpoints were identified by fluorescent *in situ* hybridization (FISH), phage/BAC clone sequencing, and whole-genome mate-paired sequencing. The gene structure of inversion breakpoints was also analyzed in several outgroup species including *An. stephensi*, *Aedes aegypti* and *Culex quinquefasciatus*. The same gene arrangement at the 2Ro inverted and $2R+^{P}$ standard arrangement was observed in outgroup species, confirming the ancestry of these arrangements. According to the ancestry of 2La, 2Ro and $2R+^{P}$, we have revised the chromosomal phylogeny and concluded that *An. gambiae–An. merus* clade is ancestral. *An. merus* gained the 2Rp inversion, and its sister taxa, *An. gambiae* acquired the $2R+^{P}$ inversion from the ancestral species. We conclude that the ability to transfer malaria has originated repeatedly in the *An. gambiae* complex. This knowledge can be used to detect genetic changes associated with human blood choice among members of the complex.

Chapter 3

Attempts to construct a molecular phylogeny among members of the *An. gambiae* complex often yielded contradictory results. We have hypothesized that chromosomal location of genetic markers with respect to breakpoints of fixed inversions is crucial in constructing a molecular phylogeny. Since genes at the breakpoints are less subject to gene flow, molecular phylogeny based on breakpoint genes should be consistent with the chromosomal phylogeny. We have amplified and sequenced breakpoint genes of 2Ro and 2Rp inversions in *An. gambiae* complex including *An. gambiae*-SUA, *An. arabiensis, An. merus* and *An. quadriannulatus.* Sequences from the *An. gambiae* M and *An. gambiae* S form, as well as outgroup species *Aedes aegypti* and *Culex quinquefasciatus*, were obtained from VectorBase. Sequences of *An. nili* were collected from the genome assembly. The 2La breakpoint genes were obtained from GenBank. Molecular phylogenies of all breakpoint genes, as well as concatenated trees. were constructed using the

neighbor-joining method in the MEGA 5.05 program. In addition, we have analyzed the gene order at the proximal 2R+^{p} breakpoints of *An. nili* and confirmed the ancestry of the 2R+^{p} arrangement in another outgroup species. According to phylogenetic trees obtained from the 2Ro and 2La breakpoint genes, *An. merus* with the inverted 2Ro and 2La arrangements is clustered separately from rest of the *An. gambiae* complex and is considered more ancestral. However, in the phylogenetic tree based on 2Rp breakpoint genes, *An. merus* with the 2Rp arrangement is a more derived species and *An. gambiae*-PEST is placed in a more ancestral clade with the 2R+^{p} arrangement. Our molecular phylogeny strongly supports the ancestral status of the 2Ro, 2R+^{p} and 2La arrangements. In conclusion, the position of the genes related to chromosomal inversions is important in constructing molecular phylogenies. We also conclude that due to a lack of gene flow in breakpoint regions, these genes are good candidates in constructing a molecular phylogeny and understanding the evolutionary history among species.

Chapter 4

In order to elucidate the phylogenetic relationship of *An. nili* among other African species, a genome wide phylogenetic analysis was performed. African species *An. gambiae*-PEST, *An. gambiae*-M, *An. gambiae*-S, and *An. funestus* were studied. We have also included an Asian Anopheles species, *An. stephensi*, and outgroup species *Aedes aegypti* and *Culex quinquefasciatus*. 49 genes were selected and these genetic markers were distributed on 5 chromosomal arms of *An. gambiae*-PEST. Phylogenetic trees were constructed based on individual genes by the neighbor-joining method in the MEGA 5.05 program. Concatenated phylogenetic trees of genes located at each chromosomal arm were obtained. Our phylogenetic analysis strongly supports the ancestral position of *An. nili* compared with other African species. Moreover, Asian *An. stephensi* clusters together with Affican *An. funestus* and both belong to the

most recently evolved lineages. On the other hand, the *An. gambiae* clade has an intermediate position relative to *An. nili* and the *An. funestus*/*An. stephensi* clade. We conclude that *An. nili*, which is mostly distributed in forested areas, belongs to the most basal lineage among African species.

Chapter 5

In population genetic studies it is important to know the location of microsatellite markers with respect to polymorphic inversions, because it will help us to better understand the genetic structure of the population. A microsatellite map of polytene chromosomes in *An. stephensi* was not available before our study. We have obtained polytene chromosomes of *An. stephensi* from the ovaries from half-gravid females. Primers were designed based on the *An. stephensi* genome sequence, microsatellite probes were obtained by PCR and were hybridized to polytene chromosomes using fluorescent *in situ* hybridization (FISH). We have developed a physical map of twelve microsatellite markers spread over all autosomal arms. Ten microsatellites were located inside inversions. Among seven microsatellites on the 2R arm, four were inside the 2Rb inversion. Most of the microsatellites, including H2II, B1, E12 with the most significant linkage disequilibrium in natural populations, hybridized to arm 2R which has the highly polymorphic inversion 2Rb. Our study shows that the position of microsatellites markers may affect estimates of population genetic parameters.