Published in final edited form as:

J Med Entomol. 2012 March ; 49(2): 307–315.

Single-Nucleotide Polymorphisms for High-Throughput Genotyping of *Anopheles arabiensis* in East and Southern Africa

YOOSOOK LEE^{1,2}, STEPHANIE N. SEIFERT¹, CHRISTEN M. FORNADEL³, DOUGLAS E. NORRIS³, and GREGORY C. LANZARO¹

¹ School of Veterinary Medicine, Department of Pathology, Micro-biology and Immunology, University of California–Davis, Davis, CA 95616

³ Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205

Abstract

Anopheles arabiensis Patton is one of the principal vectors of malaria in sub-Saharan Africa, occupying a wide variety of ecological zones. This species is increasingly responsible for malaria transmission in Africa and is becoming the dominant vector species in some localities. Despite its growing importance, little is known about genetic polymorphisms in this species. Multiple sequences of various gene fragments from *An. arabiensis* isolates from Cameroon were obtained from GenBank. In total, 20 gene fragments containing single-nucleotide polymorphisms (SNPs) at moderate density were selected for direct sequencing from field collected specimens from Tanzania and Zambia. We obtained 301 SNPs in total from the 20 gene fragments, 60 of which were suitable for Illumina GoldenGate SNP genotyping. A greater number of SNPs ($n = 185$) was suitable for analysis using Sequenom iPLEX, an alternative high-throughput genotyping technology using mass spectrometry. An SNP was present every 59 (± 44.5) bases on average. Overall, *An. arabiensis* from Tanzania and Zambia are genetically closer (mean $F_{ST} = 0.075$) than either is to populations in Cameroon ($F_{ST, TZ-CM} = 0.250$, $F_{ST, ZA-CM} = 0.372$). A fixed polymorphism between East/southern and Central Africa was identified on AGAP000574, a gene on the X chromosome. We have identified SNPs in natural populations of *An. arabiensis*. SNP densities in *An. arabiensis* were higher than *Anopheles gambiae* s.s., suggesting a greater challenge in the development of high-throughput SNP analysis for this species. The SNP markers provided in this study are suitable for a high-throughput genotyping analysis and can be used for population genetic studies and association mapping efforts.

Keywords

Anopheles arabiensis; single-nucleotide polymorphism; single-nucleotide polymorphism genotyping; Africa

Anopheles gambiae s.s. is often identified as the most important vector of malaria in Africa, and as such it has been the most extensively studied. It is the only malaria vector species whose genome sequence has been described in full (Holt et al. 2002, Lawniczak et al. 2010), and it has been the primary focus for genetic manipulation of refractoriness to malaria infection (Ito et al. 2002, Kim et al. 2004, Papathanos et al. 2009). Despite this attention, there is growing evidence that it is not this species, but its sister species *Anopheles arabiensis* Patton, that is increasingly responsible for malaria transmission in Africa. Recent reports indicate that in areas of high insecticide-treated net coverage, *An. arabiensis* may be replacing *An. gambiae* s.s. as the dominant malaria vector (Lindblade et al. 2006, Bayoh et al. 2010). Consequently, the ecology, vector competence, and population genetics of this somewhat neglected vector merit particular attention in preparation for future vector control scenarios.

Genes involved in the mosquito immune system have been intensively studied in efforts to understand interactions between the malaria parasite and mosquito vector and to manipulate vectors to reduce malaria transmission (Morlais et al. 2004; Lim et al. 2005; Riehle et al. 2006; Obbard et al. 2007; Cohuet et al. 2008; Parmakelis et al. 2008, 2010; Zou et al. 2008; Cirimotich et al. 2010; Horton et al. 2010). Genetic markers such as single-nucleotide polymorphisms (SNPs) located in or near immune genes can facilitate the identification of genes associated with resistance or refractoriness to *Plasmodium* infection and contribute to assessing the relative importance of ecological and genetic determinants of this phenotype.

Allelic variation and genetic differentiation among populations of *An. arabiensis* have been reported previously (Coluzzi et al. 1979, Nyanjom et al. 2003, Morlais et al. 2005). For example, there is a remarkable amount of chromosome inversion polymorphism within *An. arabiensis* (Coluzzi et al. 1979). Chromosome arrangements occur in *An. gambiae* s.s. and form the basis of the well known chromosomal forms (Coluzzi et al. 1985, Toure et al. 1998). Chromosomally defined *An. gambiae* s.s. subpopulations display associations with specific habitats, as demonstrated by their distribution along ecological clines (Coluzzi et al. 1979, Coluzzi et al. 1985, Bayoh et al. 2001). Similar inversion and ecological associations have been observed in *An. arabiensis* (Coluzzi et al. 1979). Population genetics studies based on microsatellite analyses have likewise revealed a high degree of genetic variability in *An. arabiensis*, both on the large (African continent, Simard et al. 1999) and small geographic scales (<25 km, Donnelly et al. 1999, Donnelly and Townson 2000, Onyabe and Conn 2001, Morlais et al. 2005, Temu and Yan 2005). These studies suggest the existence of genetically distinct subpopulations within *An. arabiensis*. It is well known that the presence of population structure can result in “spurious associations” between a phenotype and markers that are not linked to any causative loci (Lander and Schork 1994, Ewens and Spielman 1995, Pritchard and Rosenberg 1999, Kang et al. 2008). This becomes a problem when these subpopulations are not recognized so that a sample being used in an association mapping study consists of a mixture of individuals originating from two or more cryptically diverged sub-populations.

We conducted a preliminary study of nucleotide variation in field populations of *An. arabiensis* from Tanzania and Zambia at coding regions predicted from the *An. gambiae* genome as an attempt to develop SNP markers for high-throughput genotyping. These

markers provide the opportunity to perform population genetic studies. Here, the results of investigating SNPs in natural populations of *An. arabiensis* are presented. We report on methods and pitfalls associated with developing SNPs based on sequence data across broad geographic regions and limitations of two commonly used SNP detection technologies.

Materials and Methods

Sample Collection

In Tanzania, adult females of *An. arabiensis* were collected from Mkamba (−8.0333 S, 37.7666 E) and Lupiro (−8.3833 S, 36.6667 E) by using CDC light traps set in different randomly selected houses for three consecutive nights from January to May in 2007. In Macha, Zambia (−16.387 S, 26.792 E), adult females were collected using CDC light traps set in individual homes from January to March in 2008. Carcasses were stored in alcohol for subsequent DNA extraction. Genomic DNA was extracted using a DNeasy extraction kit (QIAGEN, Valencia, CA).

Species Diagnostic and Sequencing Analysis

The polymerase chain reaction (PCR) assay described by Scott et al. (1993) was used to identify members of the *An. gambiae* species complex. Eight samples of *An. arabiensis* were selected from Tanzania and up to 11 samples from Zambia for direct sequencing. We selected 20 genes of which isolate sequences from Cameroon were available on GenBank. These genes included CLIPA1, CLIPB8, CLIPB17, PPO9, REL1, SCRASP3, SCRB10, SRPN2, SRPN14, SRPN16, STAT1, TEP15, TOLL11, and seven other novel protein-encoding genes. Primers were designed for each gene fragment using Primer3 online tools (<http://frodo.wi.mit.edu/primer3/>). Primer sequences and related information are provided in Table 1.

PCR was carried out in 50- μ l reactions using 5Prime *Taq*DNA polymerase following the manufacturer's recommended thermal cycle protocols. PCR products were purified using a QIAquick PCR purification kit or QIAquick Multiwell purification kit (QIAGEN). The purified PCR amplicons were subjected to conventional post-PCR Sanger sequencing at the ^UC DNA Sequencing Facility (University of California, Davis, CA).

DnaSP version 5 (Librado and Rozas 2009) was used to identify haplotype sequences using Phase algorithm, to calculate nucleotide diversity, to perform Tajima's D statistics, and to estimate levels of gene flow.

SNP Genotyping Assay Design. A list of SNPs with up to 100-bp flanking sequence at each SNP position were sent to the technical support team at Illumina (San Diego, CA) to obtain design scores for each SNP. The design score that ranges from 0 to 1 is calculated using a proprietary algorithm developed by Illumina (http://www.illumina.com/technology/goldengate_genotyping_assay.ilmn) to give a metric of likelihood that an assay will succeed. A low-scoring SNP has higher chance of failing. We followed guidelines provided by Illumina to use a cutoff of 0.6 design score to select SNPs suitable for genotyping assay. The same list of SNPs, with the same flanking sequences, was entered into the Typer version 4.0 (Sequenom; <http://www.sequenom.com/home/products—services/genetic-analysis/>

[applications/snpgenotyping-with-iplex-gold/](#)) Assay Designer software to select SNPs suitable for iPLEX genotyping assay. We used the default preset for high-multiplexing setting, a setting that will search an assay suitable for genotyping up to 36 SNPs in a single reaction. Multiplex evaluation takes into account various aspects such as false priming potential, primer-dimer potential, amplicon length variation, PCR primer melting temperature variation, and self-priming potential. If an SNP has an overall confidence score exceeding the default threshold ($=0.4$), primers are designed and their sequences are provided in the Typer output.

Results

Sequence polymorphism in the 20 genes—three genes on the X chromosome, four genes on the right arm of chromosome 2 (2R), five genes on the left arm of chromosome 2 (2L), four genes on 3R, and four genes on 3L—were included in this study (Table 2). Isolate sequences from Tanzania and Zambia for these genes are available on GenBank, with accession numbers from JN011670 to JN011935. Seven are novel protein-encoding genes and the remaining 13 correspond to an exon and two to known protein coding genes. AGAP000574, SCRASP3, and CLIPB8 include an intron sequence between two exons.

We obtained 301 SNPs in total from 20 gene fragments, 60 of which were qualified for the GoldenGate SNP genotyping assay design (Illumina; Fan et al. 2006). Fisher exact tests were conducted on allele abundances between Cameroon and Tanzania and Zambia for each SNP. A list of SNPs, allele abundances in each country, and Fisher exact test *P* values is provided in Supp Table 1 (online only). After multiple comparison adjustment, SNPs showing significant divergence between the two countries are highlighted in red (Supp Table 1 [online only]). Loci at which no polymorphism was observed in each country is highlighted in yellow (Supp Table 1 [online only]).

A greater number of SNPs ($n = 185$) was qualified for the Sequenom iPLEX assay (Sequenom; Jurinke et al. 2001, Gabriel et al. 2009), an alternative high-throughput genotyping technology using mass spec-trometry. Out of 60 qualified SNPs, 11 were suitable for the GoldenGate assay only. SNPs that were unsuitable for GoldenGate but qualified for iPLEX assay were as many as 126. Forty-nine SNPs were qualified for both genotyping assay. Design Scores for each SNP for each genotyping platform is provided in Supp Table 2 (online only).

The number of segregating sites varies from three to 20 per gene fragment. The mean number of segregating sites was 11.9 ± 5.2 , and the median was 13. AGAP000193 contained the fewest SNPs (three in a 634-bp fragment). AGAP000574, AGAP001764, AGAP007050, AGAP008687, CLIPB8, CLIPB17, SRPN2, and TEP15 were among the most polymorphic regions (see SNP density in Table 2). We observed ≈ 12 segregating sites per gene fragment that varied in length from 355 to 634; this is equivalent to an SNP every 59 ± 44.5 bp on average. DNA polymorphism statistics for 20 genes is summarized in Table 2. Note that the numbers of synonymous, nonsynonymous, and noncoding silent mutations were based on the exons reported in the *An. gambiae* s.s. genome. The actual number of synonymous and

nonsynonymous mutations may change if *An. arabiensis* has different reading frames from *An. gambiae* s.s.

Genetic polymorphism data are missing for eight genes (AGAP001764, AGAP004978, AGAP006911, AGAP007050, AGAP009515, AGAP010423, AGAP010481, and AGAP011186) for the Zambia specimens, and data for the gene AGAP011791 are missing for Tanzania specimens. This is due to unsuccessful amplification of PCR product. We tried several variations of primer binding sites, but our attempts to develop a method to amplify these genes in the aforementioned samples was unfruitful. None of the Tajima's D values shown in Table 2 were statistically significant ($P > 0.05$ after multiple comparison adjustment), suggesting that these gene fragments are selectively neutral.

Of the 20 genes sequenced, two genes on the X chromosome (AGAP000193 and AGAP000574), two genes on chromosome 2R (AGAP001764 and SCRASP3), and one gene on chromosome 2L (PPO9) showed significant genetic divergence among populations of *An. arabiensis* between East/southern (Tanzania and Zambia) and Central (Cameroon) Africa.

In addition to SNPs, other sequence variations such as microsatellite DNA were observed. The *An. gambiae* s.s. genome sequence of SCRB10 has two repeats of TGT motif. A sample with three repeats was found in the heterozygous state in Zambia. In the intron sequence between exon 1 and exon 2 of the AGAP000574 gene, the *An. gambiae* genome sequence showed nine repeats of GC motif. *An. arabiensis* sequences typically had two repeats of GC, but a sample with eight repeats also was observed in Tanzania in the homozygous state. In the intron sequence between exon 8 and 9 of SCRASP3, the *An. gambiae* s.s. genome sequence has five repeats of AC; this is also the most common form in *An. arabiensis*. Zambian specimens included two samples with four repeats in the heterozygous state.

Exon 1 of the SRPN16 gene typically has one repeat of TGAGCTCG motif for most of the *An. arabiensis* specimens we analyzed. A sample from Tanzania was homozygous for two repeats. Exon 2 of AGAP006632 had a 22-bp nucleotide deletion (GAATACCACCGAGATCGAGAAG) in all *An. arabiensis* samples, in comparison with the *An. gambiae* s.s. genome sequence. These changes in coding regions may alter mRNA sequence and alter the phenotype of mosquitoes.

We compared sequence variation from Tanzania and Zambia with the publicly available sequences of *An. arabiensis* isolates from Cameroon. Up to 11 isolate sequences are available for each gene fragment, and the DNA polymorphism statistics for these isolate sequences are provided in Table 2. significant genetic differentiation within *An. arabiensis* was observed in five genes—AGAP000193, AGAP000574, AGAP001764, SCRASP3, and PPO9 ($P < 0.0016$; DnaSP gene flow test). Genetic distance was the smallest between Tanzania and Zambia (mean $F_{ST} = 0.069$). Mean F_{ST} between Cameroon and Tanzania was 0.447, and mean F_{ST} between Cameroon and Zambia was 0.568. In particular, AGAP000574 and AGAP008687 have fixed SNPs between East/southern Africa and Cameroon (Table 3). Little genetic differentiation was observed on AGAP007050, AGAP010481, SRPN2, REL1, STAT1, and TOLL11, with F_{ST} ranging from 0.025 to 0.111.

Discussion

Natural populations of *An. gambiae* s.l. have recently been found to have a much greater degree of polymorphism than previously reported. For example, the *An. gambiae* s.s. sequencing project reports that on average, an SNP occurs every 247 bp in the *An. gambiae* s.s. genome (data provided on Ensembl). Another study reports a SNP frequency of one every 125 coding base pairs in nuclear genome sequence obtained from a laboratory strain of *An. gambiae* (Morlais et al. 2004). The recently published genome sequence of M and S molecular form of *An. gambiae* s.s. indicates similar density of SNPs (≈ 1 SNP every 130 bp, Lawniczak et al. 2010) to the SNP density reported by Morlais et al. (2004). The average density of SNPs in *An. arabiensis* on the 20 genes screened in this study was one every 59 bp, far exceeding previous reports. *An. gambiae* s.s. isolate sequences for the corresponding 20 genes also had similarly high SNP density (an SNP every 39 bp on average, Cohuet et al. 2008). This level of variation poses a challenge to high-throughput genotyping options for large scale genome studies in these mosquito vector species because of difficulty in primer design.

We selected three to five genes from each chromosomal arms in an attempt to gather markers that cover the genome in a more or less unbiased manner. Tajima's D statistics indicates that all genes are selectively neutral, which suggests they are suitable as markers for population genetic studies. Some chromosome arms contained a higher density of informative SNPs than others as shown in Table 3. All three genes screened on the X chromosome showed significant divergence between East/southern Africa and Cameroon. Three of four genes on the right arm of chromosome two also showed significant geographic divergence. Genes on the other chromosomal arms (2L, 3R, and 3L) showed relatively low levels of divergence. A full list of mutations and corresponding *P* values are provided in Supp Table 1 (online only).

Our study highlights the fact that SNP variation tends to be very population specific, so that patterns of SNP variability reported from a population in Central Africa are very different from patterns of variation in other geographic regions. Some loci were polymorphic in East/southern Africa and monomorphic in Central Africa or vice versa (Supp Table 1 [online only]). This implies that separate SNP discovery efforts are required if a research project takes places in a different geographic area, in this case Cameroon versus Tanzania and Zambia.

Three fixed SNPs were identified from two gene fragments: AGAP000574 on the X chromosome and AGAP008687 on the right arm of chromosome 3. Fixed polymorphism on 3R, however, should be considered with caution because only two isolate sequences from Cameroon are available for comparison. According to *An. gambiae* genome coordinates, these are not located in proximity to the chromosome 3 centromere that has been shown to have a higher density of diverged regions among populations of *An. gambiae* (Stump et al. 2005, Turner et al. 2005, Slotman et al. 2006, Lawniczak et al. 2010, Neafsey et al. 2010). Sharakhov et al. (2002) have demonstrated that gene order shuffling can occur between *An. gambiae* and *Anopheles funestus* Giles. It is possible that fixed inversion polymorphism on the X chromosome between *An. gambiae* and *An. arabiensis* (Coluzzi et al. 2002) may put

AGAP000574 near the centromere of X in *An. arabiensis* rather than in the middle of the chromosome. Further genome sequencing and genome assembly effort for this species should resolve this question.

Among five microsatellites or multinucleotide scale mutations, three occurred in coding sequences. In particular, variation in AGAP006632 and AGAP009213 are expected to cause frame shifts in encoding amino acids. A study with a greater sample size may elucidate recombination, linkage disequilibrium, and selection pressure on the genetic variation we observed.

A greater number of SNPs were qualified for Sequenom iPLEX ($n = 185$) than Illumina GoldenGate assay ($n = 60$). Each method has pros and cons. The Illumina GoldenGate assay uses two fluorescence dyes to differentiate genotype of diallelic SNPs, whereas Sequenom iPLEX Assay uses mass differences of alleles due to the different molecular weight of each nucleotide. The GoldenGate assay can only genotype diallelic SNPs, whereas the iPLEX assay can assay triallelic SNPs. Although the GoldenGate assay will allow multiplex up to 3,072 SNPs in one well, the Sequenom iPLEX assay can only genotype a maximum of 36 SNPs per well. Thus, a set of SNPs that can be done in one reaction using GoldenGate assay may need a substantially greater amount of hands-on time for iPLEX assays. Because it is unlikely that one would genotype hundreds of SNPs from merely 20 gene fragments, the number of SNPs qualified for genotyping should not deter users from using the GoldenGate assay platform. The choice would more likely depend on which method would allow genotyping of the particular choice of SNPs.

In conclusion, our field isolate sequences combined with existing sequence data represent a good foundation for studying genetic polymorphism among *An. arabiensis* populations. SNP data from this study can facilitate the development of markers for population genetic studies of this medically important malaria vector. The significant levels of genetic differentiation we observed between East, southern, and Central Africa underscore this value.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Kija Ng'habi for the providing specimens from Tanzania. We thank Brandon Wilson for assistance in sequencing. This research was supported by funding National Institutes of Health grants R01AI085175 (to G.C.L.) and U19 AI089680 (to D.E.N.) and the Johns Hopkins Malaria Research Institute.

References Cited

- Bayoh MN, Thomas CJ, Lindsay SW. Mapping distributions of chromosomal forms of *Anopheles gambiae* in West Africa using climate data. *Med. Vet. Entomol.* 2001; 15:267–274. [PubMed: 11583443]
- Bayoh MN, Mathias DK, Odiere MR, Mutuku FM, Kamau L, Gimnig JE, Vulule JM, Hawley WA, Hamel MJ, Walker ED. *Anopheles gambiae*: historical population decline associated with regional distribution of insecticide-treated bed nets in western Nyanza Province, Kenya. *Malar. J.* 2010; 9:62. [PubMed: 20187956]

- Cirimotich CM, Dong Y, Garver LS, Sim S, Dimopoulos G. Mosquito immune defenses against *Plasmodium* infection. *Dev. Comp. Immunol.* 2010; 34:387–95. [PubMed: 20026176]
- Cohuet A, Krishnakumar S, Simard F, Morlais I, Koutsos A, Fontenille D, Mindrinos M, Kafatos FC. SNP discovery and molecular evolution in *Anopheles gambiae*, with special emphasis on innate immune system. *BMC Genomics.* 2008; 9:227. [PubMed: 18489733]
- Coluzzi M, Sabatini A, Petrarca V, Di Deco MA. Chromosomal differentiation and adaptation to human environments in the *Anopheles gambiae* complex. *Trans. R. Soc. Trop. Med. Hyg.* 1979; 73:483–497. [PubMed: 394408]
- Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V. A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science.* 2002; 298:1415–1418. [PubMed: 12364623]
- Coluzzi M, Petrarca V, Di Deco MA. Chromosomal inversion intergradation and incipient speciation in *Anopheles gambiae*. *Boll. Zool.* 1985; 52:45–63.
- Donnelly MJ, Townson H. Evidence for extensive genetic differentiation among populations of the malaria vector *Anopheles arabiensis* in Eastern Africa. *Insect Mol. Biol.* 2000; 9:357–367. [PubMed: 10971713]
- Donnelly MJ, Cuamba N, Charlwood JD, Collins FH, Townson H. Population structure in the malaria vector, *Anopheles arabiensis* Patton, in East Africa. *Heredity.* 1999; 83:408–417. [PubMed: 10583542]
- Ewens WJ, Spielman RS. The transmission/ disequilibrium test: history, subdivision, and admixture. *Am. J. Hum. Genet.* 1995; 57:455–464. [PubMed: 7668272]
- Fan JB, Chee MS, Gunderson KL. Highly parallel genomic assays. *Nat. Rev. Genet.* 2006; 7:632–644. [PubMed: 16847463]
- Gabriel S, Ziaugra L, Tabbaa D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Curr. Protoc. Hum. Genet.* 2009;12. Chapter 2 (Unit 2). [PubMed: 19170031]
- Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science.* 2002; 298:129–149. [PubMed: 12364791]
- Horton AA, Lee Y, Coulibaly CA, Rashbrook VK, Cornel AJ, Lanzaro GC, Luckhart S. Identification of three single nucleotide polymorphisms in *Anopheles gambiae* immune signaling genes that are associated with natural *Plasmodium falciparum* infection. *Malar. J.* 2010; 9:160. [PubMed: 20540770]
- Ito J, Ghosh A, Moreira LA, Wimmer EA, Jacobs-Lorena M. Transgenic anopheline mosquitoes impaired in transmission of a malaria parasite. *Nature.* 2002; 417:452–455. [PubMed: 12024215]
- Jurinke C, van den Boom D, Cantor CR, Koster H. Automated genotyping using the DNA MassArray technology. *Methods Mol. Biol.* 2001; 170:103–116. [PubMed: 11357675]
- Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. *Genetics.* 2008; 178:1709–1723. [PubMed: 18385116]
- Kim W, Koo H, Richman AM, Seeley D, Vizioli J, Klocko AD, O'Brochta DA. Ectopic expression of a cecropin transgene in the human malaria vector mosquito *Anopheles gambiae* (Diptera: Culicidae): effects on susceptibility to *Plasmodium*. *J. Med. Entomol.* 2004; 41:447–55. [PubMed: 15185949]
- Lander ES, Schork NJ. Genetic dissection of complex traits. *Science.* 1994; 265:2037–2048. [PubMed: 8091226]
- Lawniczak MK, Emrich SJ, Holloway AK, Regier AP, Olson M, White B, Redmond S, Fulton L, Appelbaum E, Godfrey J, et al. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science.* 2010; 330:512–514. [PubMed: 20966253]
- Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics.* 2009; 25:1451–1452. [PubMed: 19346325]
- Lim J, Gowda DC, Krishnegowda G, Luckhart S. Induction of nitric oxide synthase in *Anopheles stephensi* by *Plasmodium falciparum*: mechanism of signaling and the role of parasite glycosylphosphatidylinositols. *Infect. Immun.* 2005; 73:2778–2789. [PubMed: 15845481]

- Lindblade KA, Gimnig JE, Kamau L, Hawley WA, Odhiambo F, Olang G, Ter Kuile FO, Vulule JM, Slutsker L. Impact of sustained use of insecticide-treated bednets on malaria vector species distribution and culicine mosquitoes. *J. Med. Entomol.* 2006; 43:428–432. [PubMed: 16619629]
- Morlais I, Poncon N, Simard F, Cohuet A, Fontenille D. Intraspecific nucleotide variation in *Anopheles gambiae*: new insights into the biology of malaria vectors. *Am. J. Trop. Med. Hyg.* 2004; 71:795–802. [PubMed: 15642974]
- Morlais I, Girod R, Hunt R, Simard F, Fontenille D. Population structure of *Anopheles arabiensis* on La Reunion island, Indian Ocean. *Am. J. Trop. Med. Hyg.* 2005; 73:1077–1082. [PubMed: 16354815]
- Neafsey DE, Lawniczak MK, Park DJ, Redmond SN, Coulibaly MB, Traore SF, Sagnon N, Costantini C, Johnson C, Wiegand RC, et al. SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science.* 2010; 330:514–517. [PubMed: 20966254]
- Nyanjom SR, Chen H, Gebre-Michael T, Bekele E, Shililu J, Githure J, Beier JC, Yan G. Population genetic structure of *Anopheles arabiensis* mosquitoes in Ethiopia and Eritrea. *J. Hered.* 2003; 94:457–463. [PubMed: 14691312]
- Obbard DJ, Linton YM, Jiggins FM, Yan G, Little TJ. Population genetics of *Plasmodium* resistance genes in *Anopheles gambiae*: no evidence for strong selection. *Mol. Ecol.* 2007; 16:3497–3510. [PubMed: 17688548]
- Onyabe DY, Conn JE. Population genetic structure of the malaria mosquito *Anopheles arabiensis* across Nigeria suggests range expansion. *Mol. Ecol.* 2001; 10:2577–2591. [PubMed: 11883874]
- Papathanos PA, Windbichler N, Menichelli M, Burt A, Crisanti A. The vasa regulatory region mediates germline expression and maternal transmission of proteins in the malaria mosquito *Anopheles gambiae*: a versatile tool for genetic control strategies. *BMC Mol. Biol.* 2009; 10:65. [PubMed: 19573226]
- Parmakelis A, Slotman MA, Marshall JC, Awono-Ambene PH, Antonio-Nkondjio C, Simard F, Caccone A, Powell JR. The molecular evolution of four anti-malarial immune genes in the *Anopheles gambiae* species complex. *BMC Evol. Biol.* 2008; 8:79. [PubMed: 18325105]
- Parmakelis A, Moustaka M, Poulakakis N, Louis C, Slotman MA, Marshall JC, Awono-Ambene PH, Antonio-Nkondjio C, Simard F, Caccone A, et al. *Anopheles* immune genes and amino acid sites evolving under the effect of positive selection. *PLoS ONE.* 2010; 5:e8885. [PubMed: 20126662]
- Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* 1999; 65:220–228. [PubMed: 10364535]
- Riehle MM, Markianos K, Niare O, Xu J, Li J, Toure AM, Podiougou B, Oduol F, Diawara S, Diallo M, et al. Natural malaria infection in *Anopheles gambiae* is regulated by a single genomic control region. *Science.* 2006; 312:577–579. [PubMed: 16645095]
- Scott JA, Brogdon WG, Collins FH. Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. *Am. J. Trop. Med. Hyg.* 1993; 49:520–529. [PubMed: 8214283]
- Sharakhov IV, Serazin AC, Grushko OG, Dana A, Lobo N, Hillenmeyer ME, Westerman R, Romero-Severson J, Costantini C, Sagnon N, et al. Inversions and gene order shuffling in *Anopheles gambiae* and *A. funestus*. *Science.* 2002; 298:182–185. [PubMed: 12364797]
- Simard F, Fontenille D, Lehmann T, Girod R, Brutus L, Gopaul R, Dournon C, Collins FH. High amounts of genetic differentiation between populations of the malaria vector *Anopheles arabiensis* from West Africa and eastern outer islands. *Am. J. Trop. Med. Hyg.* 1999; 60:1000–1009. [PubMed: 10403334]
- Slotman MA, Reimer LJ, Thiemann T, Dolo G, Fondjo E, Lanzaro GC. Reduced recombination rate and genetic differentiation between the M and S forms of *Anopheles gambiae* s.s. *Genetics.* 2006; 174:2081–2093. [PubMed: 17057242]
- Stump AD, Fitzpatrick MC, Lobo NF, Traore S, Sagnon N, Costantini C, Collins FH, Besansky NJ. Centromere-proximal differentiation and speciation in *Anopheles gambiae*. *Proc. Natl. Acad. Sci. U.S.A.* 2005; 102:15930–15935. [PubMed: 16247019]
- Temu EA, Yan G. Microsatellite and mitochondrial genetic differentiation of *Anopheles arabiensis* (Diptera: Culicidae) from western Kenya, the Great Rift Valley, and coastal Kenya. *Am. J. Trop. Med. Hyg.* 2005; 73:726–733. [PubMed: 16222017]

- Toure YT, Petrarca V, Traore SF, Coulibaly A, Maiga HM, Sankare O, Sow M, Di Deco MA, Coluzzi M. The distribution and inversion polymorphism of chromosomally recognized taxa of the *Anopheles gambiae* complex in Mali, West Africa. *Parassitologia*. 1998; 40:477–511. [PubMed: 10645562]
- Turner TL, Hahn MW, Nuzhdin SV. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol*. 2005; 3:e285. [PubMed: 16076241]
- Zou Z, Shin SW, Alvarez KS, Bian G, Kokoza V, Raikhel AS. Mosquito RUNX4 in the immune regulation of PPO gene expression and its effect on avian malaria parasite infection. *Proc. Natl. Acad. Sci. U.S.A.* 2008; 105:18454–18459. [PubMed: 19011100]

Table 1

PCR primer sequence and related information

Chromosome	Ensemble Gene ID	Fragment ^a	Gene ^b	Annotation on Ensembl	Coding region ^c	μ d	Primer sequence ^e	GC%	T _m
X	AGAP000016	Exon 3	SCRB10	Class B scavenger receptor (CD36 domain)	1-576	576	cgacgtgctgagaagtta	50	59.27
X	AGAP000193	Exon 5	NOVEL		1-634	634	ctgaacacagcgaatagct	45	59.09
X	AGAP000574	Exon 1-2	NOVEL		1-87	382	ggcggcattattcaaaact	45	58.91
2R	AGAP001648	Exon 3	CLIPB17	Clip-Domain serine protease, family B	232-382		ttctgcccagatacagttct	50	59.69
2R	AGAP001764	Exon 2	NOVEL		1-498	498	ggcgggagaattcgacca	61	62.2
2R	AGAP001979	Exon 8-9	SCRASP3	Class A scavenger receptor (SRCR domain)	1-141;	356	ggcagttgtcgcgttgta	52	60.86
2R	AGAP003057	Exon 5-6	CLIPB8	Clip-Domain serine protease, family B	237-356	568	gtactcaaccgctggcact	63	61.4
2L	AGAP004978	Exon 1	PP09	Prophenoloxidase	1-581	581	gtactgacgctggcact	60	61.3
2L	AGAP006632	Exon 3	NOVEL		1-626	626	ggatgtagatgcccggatg	50	60.3
2L	AGAP006911	Exon 3	SRPN2	Serine protease inhibitor (serpin) likely cleavage at K/F	1-518	518	gactgctgtacgagaacga	55	60
2L	AGAP007050	Exon 3	NOVEL		1-632	632	agactgacgctggcact	50	59.58
2L	AGAP007692	Exon 1	SRPN14	Serine protease inhibitor (serpin) ho mo log unlikely to be inhibitory	1-422	422	gatcaagccagtggagctg	57.89	59.52
3R	AGAP008364	Exon 9	TEP15	Thioester-containing protein	1-505	505	gcaaccgattgcagaacc	55.56	60.2
3R	AGAP008687	Exon 4	NOVEL		1-513	513	cggacgcgtacatca	62.5	58.76
							ctcgcgcgtaccag	73.33	58.15
							caccgtatccagaacgata	50	59.52
							agfacatcgcgagctgttg	50	59.1
							agttgtatcctggccaca	52.63	60.17
							gggtactctgcccattctc	60	60.69
							ggagcagaacgtggcgt	61.11	59.83
							cggcactcagctgtatcac	55	59.42
							atggagcgggctatcag	61.11	61.14
							accacttgaatcacatctcg	50	58.72
							cctccctggcagcagatcat	57.89	58.84
							acttcgctccaccctgact	57.89	59.85

Chromosome	Ensemble Gene ID	Fragment ^a	Gene ^b	Annotation on Ensembl	Coding region ^c	Id	Primer sequence ^e	GC%	T _m
3R	AGAP009213	Exon 1	SRPN16	Serine protease inhibitor (serpin) likely cleavage at A/L	1-507	507	tttgcgagggggctc	55.56	59.92
3R	AGAP009515	Exon 9	REL1	TOLL pathway signaling NF-κB Relish-like transcription factor	1-556	556	catcttcgacacgagctcac	55	59.58
3L	AGAP010423	Exon 1	STAT1	TAK/ STAT pathway signaling signal transducer and activator of transcription 1	1-536	536	cgggccagacggctcacc	66.67	62.67
3L	AGAP010481	Exon 3	NOVEL		1-424	424	accggacgctgcatggt	55.56	60.12
3L	AGAP011186	Exon 1	TOLL11	Toll-like receptor	1-574	574	gcaacttcgctcgaggg	55.56	60.2
3L	AGAP011791	Exon 2	CLIPA1	Clip-Domain serine protease, family A protease homolog	1-602	602	atcacttgcgggttgatt	45	60.38
3L	AGAP011791	Exon 2	CLIPA1	Clip-Domain serine protease, family A protease homolog	1-602	602	cgatggcatcctgtattacg	50	59.02
3L	AGAP011791	Exon 2	CLIPA1	Clip-Domain serine protease, family A protease homolog	1-602	602	cgtaccacacggttccat	55.56	60.38
3L	AGAP011791	Exon 2	CLIPA1	Clip-Domain serine protease, family A protease homolog	1-602	602	cgaagtcgtcaacaacaaga	50	60.18
3L	AGAP011791	Exon 2	CLIPA1	Clip-Domain serine protease, family A protease homolog	1-602	602	gatacagcagcggcagatagt	55	60.54
3L	AGAP011791	Exon 2	CLIPA1	Clip-Domain serine protease, family A protease homolog	1-602	602	acggccagctgacctacac	66.67	62.43
3L	AGAP011791	Exon 2	CLIPA1	Clip-Domain serine protease, family A protease homolog	1-602	602	tggtaecgccacattcgttag	50	59.61

Genes successfully amplified in all three countries are in bold.

^aSequenced fragment of the corresponding gene.

^bEncoding gene (source: Ensembl).

^cCoding region (source: An. *gambiae* genome annotation from Ensembl).

^dExpected PCR product size.

^eF, forward; and R, reverse.

Table 2

Genetic polymorphism statistics

Chr ^a	Gene ^b	Pop ^c	N ^d	n _s ^e	π (%) ^f	D ^g	μ_s ^h	μ_{NS} ⁱ	μ_{NCS} ^j	T:Z ^k	T:C ^l	Z:C ^m	F _{ST, T:Z} ⁿ	F _{ST, T:C} ^o	F _{ST, Z:C} ^p	P ^q	SNP density ^r
X	16	TZ	16	9	0.473	0.01721	4	5	NA	0:8	0:5	0:5	-0.006	0.350	0.279	0.0202	1:64.3
		ZA	18	8	0.595	0.60858	4	6	NA								
		CM	16	9	0.497	-0.81685	3	2	NA								
X	193	TZ	16	3	0.446	0.06703	2	1	NA	0:4	0:1	0:0	0.050	0.362	0.400	0.0010	1:211.3
		ZA	16	1	0.079	1.30896	1	0	NA								
		CM	16	3	0.329	1.13586	1	2	0								
X	574	TZ	12	11	0.801	-0.66472	4	1	6	0:4	1:0	6:0	0.056	0.725	0.870	0.0000	1:34.7
		ZA	16	4	0.218	-0.96578	1	0	3								
		CM	12	3	0.328	1.13586	1	2	NA								
2R	1648	TZ	16	11	0.855	1.07508	8	3	NA	0:7	0:3	0:3	0.029	0.259	0.238	0.0124	1:31.1
		ZA	16	16	0.755	-0.86965	10	6	NA								
		CM	16	5	0.427	1.35198	2	3	NA								
2R	1764	TZ	16	15	0.790	-0.73336	14	1	NA		0:4			0.329		0.0020	1:33.1
		ZA	-														
		CM	10	6	0.644	2.10483	6	0	NA								
2R	1979	TZ	16	6	0.570	0.41042	4	0	2	0:5	0:2	0:2	0.101	0.266	0.434	0.0000	1:59.2
		ZA	22	5	0.399	0.37567	3	0	2								
		CM	16	3	0.192	-0.70788	3	0	NA								
2R	3057	TZ	16	19	0.876	-1.01168	10	0	NA	0:9	0:4	1:4	-0.003	0.358	0.422	0.0054	1:29.9
		ZA	16	9	0.634	0.72563	5	0	NA								
		CM	16	7	0.469	0.93053	7	0	NA								
2L	4978	TZ	16	11	0.613	0.29629	10	1	NA		0:3			0.554		0.0016	1:53.0
		ZA	-														
		CM	16	5	0.197	-0.7084	5	0	NA								
2L	6632	TZ	16	12	0.413	-1.31358	11	1	NA		0:5		0.287	0.157	0.497	0.0085	1:52.2
		ZA	6	0	NA	NA	NA	NA	NA								
		CM	16	6	0.221	-0.80427	6	0	NA								
2L	6911	TZ	16	11	0.648	0.04936	7	4	NA		0:7			0.086		0.0160	1:30.5

Chr ^a	Gene ^b	Pop ^c	N ^d	n ^e	π (%) ^f	D ^g	μ_s ^h	μ_{NS} ⁱ	μ_{NCS} ^j	T:Z ^k	T:C ^l	Z:C ^m	F _{ST, T:Z} ⁿ	F _{ST, T:C} ^o	F _{ST, Z:C} ^p	P ^q	SNP density ^r
2L	7050	ZA	-														
		CM	14	17	0.764	-1.07232	10	7	NA								
		TZ	16	20	1.005	0.21536	19	1	NA		0:10			0.058		0.0518	1:31.6
2L	7692	ZA	-														
		CM	14	14	0.746	-0.00216	13	1	NA								
		TZ	16	2	0.150	0.12996	2	0	NA	0:2	0:2	0:2	0.091	0.228	0.116	0.0164	1:140.7
3R	8364	ZA	22	3	0.295	1.33196	3	0	NA								
		CM	16	2	0.233	1.61632	2	0	NA								
		TZ	16	17	0.145	0.04014	16	2	NA	0:14	0:10	0:10	-0.033	0.217	0.268	0.0356	1:29.7
3R	8687	ZA	16	13	1.079	0.8055	13	1	NA								
		CM	16	11	0.990	1.06536	11	1	NA								
		TZ	16	17	1.256	1.0206	17	0	NA	0:8	2:0	3:0	0.177	0.426	0.672	0.0039	1:30.2
3R	9213	ZA	16	9	0.601	0.50243	8	1	NA								
		CM	4	4	0.422	-0.06501	1	3	NA								
		TZ	16	6	0.367	-0.41962	1	5	NA	0:4	0:3	0:3	0.077	0.175	0.119	0.0390	1:64.4
3R	9515	ZA	16	5	0.334	0.40425	1	4	NA								
		CM	16	8	0.501	-0.23104	5	3	NA								
		TZ	16	14	0.734	-0.12515	11	3	NA		0:6			0.044		0.0179	1:39.7
3L	10423	ZA	-														
		CM	16	6	0.337	0.12647	4	2	NA								
		TZ	18	10	0.839	1.4375	6	4	NA		0:7			0.025		0.0278	1:54.6
3L	10481	ZA	-														
		CM	16	9	0.636	0.89737	4	5	NA								
		TZ	16	4	0.169	-1.26856	4	0	NA		0:0			0.031		0.4060	1:84.8
3L	11186	ZA	-														
		CM	16	5	0.202	-2.08991	1	2	NA								
		TZ	16	5	0.781	-0.03219	14	1	NA		0:5			0.111		0.0379	1:38.3
3L	11791	ZA	-														
		CM	16	10	0.663	0.98345	8	2	NA			0:4					
		TZ	-												0.146	0.0265	1:60.2
		ZA	10	10	0.613	0.19314	8	2	NA								

Chr ^a	Gene ^b	Pop ^c	N ^d	n ^e	π (%) ^f	D ^g	μ_s ^h	μ_{NS} ⁱ	μ_{NCS} ^j	T:Z ^k	T:C ^l	Z:C ^m	F _{ST, T:Z} ⁿ	F _{ST, T:C} ^o	F _{ST, Z:C} ^p	P ^q	SNP density ^r
		CM	10	10	0.479	-0.16218	8	2	NA								

^aChromosome.

^bNumeric portion of Ensemble Gene ID starting with “AGAP.”

^cSource population. Specimens were grouped by country: TZ, Tanzania; ZA, Zambia; and CM, Cameroon.

^dNumber of haplotype sequences per group.

^eNumber of segregating sites.

^fNucleotide diversity.

^gTajima’s D.

^hNumber of synonymous mutations.

ⁱNumber of nonsynonymous mutations.

^jNumber of silent mutations in noncoding sequences.

^kFixed:shared polymorphism between Tanzania and Zambia groups.

^lFixed:shared polymorphism between Tanzania and Cameroon groups.

^mFixed:shared polymorphism between Zambia and Cameroon.

ⁿF_{ST} between Tanzania and Zambia.

^oF_{ST} between Tanzania and Cameroon.

^pF_{ST} between Zambia and Cameroon.

^qSignificance *P* value of a genetic differentiation estimates based on chi-square implemented in the DnaSP gene flow test.

^rSNP density of 1:x means a SNP present in every x number of bases.

Table 3

Significant divergence between East/southern Africa and Cameroon for selected SNPs

Chromosome	Gene	SNP location ^a	Allele	East/southern Africa	Cameroon	<i>p</i> ^b
X	AGAP000016	SNP at 453rd nucleotide	G/T	23/9	1/15	2.85E-05
X	AGAP000193	SNP at 460th nucleotide	T/A	10/22	16/0	4.36E-06
X	AGAP000574	SNP at 20th nucleotide	C/T	28/0	5/11	5.70E-07
X	AGAP000574	SNP at 139th nucleotide	A/C	7/21	16/0	6.37E-07
X	AGAP000574	SNP at 163th nucleotide	G/C	2/26	16/0	3.67E-10
X	AGAP000574	SNP at 168th nucleotide	G/A	0/28	16/0	2.40E-12
X	AGAP000574	SNP at 177th nucleotide	C/T	2/26	16/0	3.67E-10
X	AGAP000574	SNP at 184th nucleotide	C/T	2/26	16/0	3.67E-10
X	AGAP000574	SNP at 202th nucleotide	G/T	2/26	16/0	3.67E-10
2R	AGAP001764	SNP at 158th nucleotide	G/C	3/13	10/0	1.08E-04
2R	AGAP001979	SNP at 150th nucleotide	A/-	8/28	16/0	7.10E-08
2R	AGAP001979	SNP at 153th nucleotide	G/C-	8/28	16/0	7.10E-08
2R	AGAP003057	SNP at 276th nucleotide	A/C	11/21	16/0	6.44E-06
2R	AGAP003057	SNP at 299th nucleotide	G/T	4/28	16/0	2.15E-09
2L	AGAP004978	SNP at 545th nucleotide	C/A	2/14	14/2	4.88E-05
2L	AGAP004978	SNP at 554th nucleotide	C/T	2/14	14/2	4.88E-05
3R	AGAP008687	SNP at 421st nucleotide	T/C	0/32	4/0	1.70E-05
3R	AGAP008687	SNP at 487th nucleotide	T/G	0/32	4/0	1.70E-05

Fixed difference between East/southern Africa and Cameroon is in bold. A full list of mutations and corresponding *P* values is in Supp Table 1([online only]).

^a Location of a SNP in relation to the PCR amplicon.

^b Fisher exact test *P* values.