Published in final edited form as:
*Infect Genet Evol*. 2007 March ; 7(2): 285–292.

# Polymorphism at the *defensin* gene in the *Anopheles gambiae* complex: testing different selection hypotheses

**Frédéric Simard**[1,2,*], **Monica Licht**[3], **Nora J. Besansky**[4], and **Tovi Lehmann**[3,5]

***1*** *Organisation de Coordination pour la lutte Contre les Endémies en Afrique Centrale (OCEAC), BP 288, Yaoundé, Cameroon*

***2*** *Institut de Recherche pour le Développement (IRD) - UR016, BP 1857, Yaoundé, Cameroon*

***3*** *Division of Parasitic Diseases, Centers for Diseases Control and Prevention, Entomology Branch, MS F-22, 4770 Buford Highway, Chamblee, GA 30341, USA*

***4*** *Center for Tropical Disease Research and Training, Dept of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA*

***5*** *Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), 12735 Twinbrook Parkway, Room 2W13A, Rockville MD 20852, USA*

## Abstract

Genetic variation in *defensin*, a gene encoding a major effector molecule of insects immune response was analyzed within and between populations of three members of the *Anopheles gambiae* complex. The species selected included the two anthropophilic species, *An. gambiae* and *An. arabiensis* and the most zoophilic species of the complex, *An. quadriannulatus*. The first species was represented by four populations spanning its extreme genetic and geographical ranges, whereas each of the other two species was represented by a single population. We found (i) reduced overall polymorphism in the mature peptide region and in the total coding region, together with specific reductions in rare and moderately frequent mutations (sites) in the coding region compared with non coding regions, (ii) markedly reduced rate of nonsynonymous diversity compared with synonymous variation in the mature peptide and virtually identical mature peptide across the three species, and (iii) increased divergence between species in the mature peptide together with reduced differentiation between populations of *An. gambiae* in the same DNA region. These patterns suggest a strong purifying selection on the mature peptide and probably the whole coding region. Because *An. quadriannulatus* is not exposed to human pathogens, identical mature peptide and similar pattern of polymorphism across species implies that human pathogens played no role as selective agents on this peptide.

### Keywords

Anopheles gambiae; Africa; malaria; vector; arthropod; immunity; Defensin; evolution; polymorphism; selection

*Corresponding author's address: Frederic Simard, OCEAC, P.O. Box 288, Yaounde, Cameroon, Phone: +237 223 2232; Fax: +237 220 1854; E-mail: simard@ird.fr

## INTRODUCTION

The completion of the sequencing of the genome of *An. gambiae* (Holt *et al.* 2002), together with the successful germ line transformation of this mosquito (Grossman *et al.*, 2001), and the identification of key molecules and genes affecting susceptibility of mosquitoes to *Plasmodium* under laboratory conditions (Barillas-Mury *et al.*, 1996;Blandin *et al.*, 2004;Osta *et al.*, 2004a,b) provide strong support for malaria control via the introduction and spread of refractory genes into vector populations (Collins *et al.*, 2000). Immune-response molecules are particularly promising as determinants of vector susceptibility and have been the focus of a number of recent studies (Christophides *et al.*, 2002;Dimopoulos, 2003;Osta *et al.*, 2004a,b;Meister *et al.*, 2005). Some evidence suggests that *Anopheles* susceptibility to *Plasmodium* depends on the specific genotype of the vector and the parasite (Tahar *et al.*, 2002;Lambrechts *et al.*, 2005). Such finely tuned host-pathogen relationships are expected to mark their signature on the molecular make-up of the genes involved. However, molecular evolution of genes encoding immune response molecules of arthropod diseases vectors has received little attention, despite providing unique and valuable insights into the susceptibility to pathogens in other host species (Schlenke & Begun, 2003). Here, we describe and analyze polymorphism in the *defensin* gene within and between populations of the *An. gambiae* complex to address the following questions. Can selection be detected on this gene? If so, what mode of selection? And finally, do human pathogens mediate selection on *defensin*?

Defensin, a member of the cysteine-rich immune peptides, is a primary effector molecule produced by mosquitoes in response to infection with various pathogens (Richman *et al.*, 1996). Defensin is synthesized mainly in the fat body of both larvae and adults and secreted into the haemolymph. It is expressed constitutively at low rates in adults and larvae, but following an infection challenge expression increases dramatically (Richman *et al.*, 1997;Dimopoulos *et al.*, 1998;Eggleston *et al.*, 2000). Sporozoites of *Plasmodium gallinaceum* (and oocysts to a lower extent) are killed by defensin, but the relevance of this *in-vitro* study to natural defense needs to be determined (Shahabuddin *et al.*, 1998). Silencing of defensin in *An. gambiae* demonstrated that it is required for antimicrobial defense against Gram-positive bacteria (Blandin *et al.*, 2002). Defensin is encoded by a single copy gene located at division 41 on the third chromosome of *An. gambiae* (Vizioli *et al.*, 2001). It is comprised of two exons separated by a short intron. The 102 amino acids (aa) pre-pro-defensin includes a 25 aa signal peptide and a 77 aa segment that is cleaved to produce a 40 aa mature defensin. The signal and pro-peptide sequences of *An. gambiae* share little similarity with those from other insects, but the mature peptide is conserved and all insect defensins contain six cysteine residues. The promoter region is rich with sequence motifs similar to transcription regulatory elements of insect and mammalian immune response genes. These include binding sites for nuclear factor kappa B, GATA factors, nuclear factor interleukin 6 and interferon consensus elements among others (Eggleston *et al.*, 2000). Induction of *defensin* transcription can be mediated by Gambif1 (Barillas-Mury *et al.*, 1996), a member of the Rel protein group that mediates transcription regulation of immune response of Drosophila and other insects.

We chose populations from the highly anthropophilic species, *An. gambiae*, the moderately anthropophilic species, *An. arabiensis*, and the highly zoophilic species, *An. quadriannulatus*, representing high, moderate and no exposure to human pathogens, respectively (Hadis *et al.*, 1997;Lemasson *et al.*, 1997;Mouchet *et al.*, 2004). The populations of *An. gambiae* and *An. arabiensis* are major vectors of *Plasmodium falciparum* with typical salivary glands infection rates of 3–9% (Hay *et al.*, 2000;Mouchet *et al.*, 2004 and references therein), but only the *An. gambiae* populations from Nigeria and eastern Kenya also transmit *Wucheraria bancrofti*, the causative agent of lymphatic filariasis (LF). The four *An. gambiae* populations include members of the M (Senegal) and S (West and East Kenya and Nigeria)

molecular forms and span the maximal genetic distance measured among *An. gambiae* populations across the continent (Lehmann *et al.*, 2003). Our study thus may help identify the potential and the limitation of such an approach for understanding the evolutionary forces that determine susceptibility to pathogens.

## MATERIALS AND METHODS

### Samples and collection methods

Mosquitoes were collected between 1994 and 1999 from Asembo Bay (1994, hereafter referred to as western Kenya), Jego (1996, hereafter referred to as eastern Kenya), Gwamlar in central Nigeria (1999), and Barkedji in Senegal (1995). Aliquots of *An. quadriannulatus* DNA were kindly provided by F. H. Collins, from specimens collected in 1986 in a rural area of southern Zimbabwe (see Collins *et al.*, 1988 for more details). Indoor-resting adult mosquitoes (mostly females) were collected by pyrethrum-spray or aspiration in eastern Kenya and Nigeria. In Senegal, blood-seeking mosquitoes were collected by human-baited night catches. In western Kenya, blood fed and blood seeking females were collected at dawn by aspiration from net traps hung over the beds of sleeping volunteers. At each site, mosquitoes were collected within one week from houses less than 5 km apart. Further details on sampling methods and sites of collection are given by Lehmann *et al.* 2003.

Anophelines were identified as members of the *An. gambiae* complex using morphological keys (Gillies & De Meillon, 1968;Gillies & Coetzee, 1987). Species identification was carried out using PCR (Scott *et al.*, 1993). The PCR-RFLP assay (Favia *et al.*, 1997) was used to determine *An. gambiae* molecular form. *An. gambiae* specimens from Kenya and Nigeria were all of the S form, while those from Senegal were of the M form. All *An. arabiensis* specimens included in this study were collected in Asembo Bay (western Kenya).

### DNA extraction and sequencing

Genomic DNA was extracted from whole mosquitoes as previously described (Lehmann *et al.*, 1996;2003) and suspended in 100 μl of TE. PCR reactions were carried out using 2 μl of template DNA (from an aliquot of whole-mosquito extracts diluted 1:20 in distilled water) in 50 μl reaction containing 5 units Taq polymerase (Boehringer Mannheim or Gibco BRL) in manufacturer's buffer, 1.5 mM $MgCl_2$, 200 μM each dNTP (PE Applied Biosystems) and 50 pmol each forward and reverse primers.

Primers were designed based on the complete *An. gambiae defensin* sequence (Eggleston *et al.*, 2000; GenBank Accession number: AF063402). A 1.4 Kb region (position 1524–2972 in the published sequence) encompassing the whole transcribed region of *defensin* as well as 5′ and 3′ non-transcribed flanking regions, was amplified with forward primer Df1524L (5′ GCG GGG TGA ATG TTA TCT CT 3′) and reverse primer Df2972R (5′ ACA ATA AAA GGA ACG CAA GC 3′). Cycling conditions for amplification included denaturation at 94°C for 5 minutes, followed by 35 cycles at 94°C for 30 seconds, 52°C for 30 seconds and 72°C for 1 minute, with a final extension step at 72°C for 5 minutes. PCR products were examined on a 1% agarose gel, and cloned using the pGem T-vector kit (Promega). Individual transformed colonies (white) were selected. The size of the DNA insert was screened by PCR using pUC/M13 forward and reverse primers. In most cases, a single appropriately sized insert was chosen at random, and sequenced in both directions after purification with the Wizard PCR Purification Kit (Promega). In addition to the previous forward and reverse primers, internal nested primers Df2217L (5′ CGG TGC CAA TCT CAA TAC CCT TT 3′) and Df2280R (5′ GAC AAC GGG AAA AAG GGA TG 3′) were used as sequencing primers. Cycle sequencing was performed using PE BigDye Terminator Ready Reaction Kit according to manufacturer's

recommendations (PE Applied Biosystems). Sequencing reaction products were analyzed on an ABI 377 automated sequencer (PE Applied Biosystems). Sequences were checked for accuracy on both strands using Sequence Navigator (PE Applied Biosystems). Multiple alignment was performed with the Pileup program of GCG (Genetics Computer Group, 1999) using default options, and was adjusted by eye. DNA sequences have been deposited in GenBank under accession numbers DQ211988–DQ212056. To avoid sampling bias, a single allele (haplotype sequence) was arbitrarily selected from each specimen for the analysis.

## PCR error

Because of multiple insertion/deletion (indels) in *defensin*, direct sequencing was not possible. Sequences were determined from 2–4 independent clones of the same allele, to identify errors resulting from mis-incorporation of nucleotides by Taq polymerase during the PCR amplification. We estimated PCR error rate to be 0.001 per bp in accordance with published records (Kwiatowski *et al.*, 1991). High variation between alleles, allowed distinguishing different alleles and different clones of the same allele. Although we used statistics that are less sensitive to the effect of PCR errors (e.g., nucleotide diversity instead of the number of segregating sites and theta, if derived based on the latter), the polymorphism reported here is biased upwards because of PCR errors. Nevertheless, our inference is unbiased because instead of relying on the absolute values of polymorphism, we compared polymorphism between different functional regions of the gene that have the same probability to include a PCR error once differences in sequence length were accommodated (below).

## Data analysis

Nucleotide diversity ($\pi$) was estimated using DnaSp 4.0 (Rozas & Rozas 1999). A more complete summary of polymorphism was obtained by the site frequency spectra (Tajima 1989;Braverman *et al.* 1995), which describes the frequency of sites that are invariant (f=0), singleton (f=1), and polymorphic (f=2, 3, … n/2), where f is the frequency of the rare nucleotide at this site/position and n is the number of sequences. These spectra distinguish between rare (e.g., singletons) and common mutations (sites where the rarest nucleotide was observed 4–7 times, which is the maximum possible frequency given 9–14 sequences per population). Most neutral mutations are lost or require a very long time to become common in a population and this time is expected to be shorter for positively selected mutations and longer for deleterious mutations. Hence, rare mutations represent a greater fraction of new and mildly deleterious mutations, whereas common ones represent a greater fraction of ancient and neutral mutations. Furthermore, the site frequency spectrum is especially suited to compare polymorphism in regions of the gene without bias due to PCR errors, because it accounts for sequence length variation. We compared and tested nucleotide diversity of synonymous and nonsynonymous sites using bootstrapping in MEGA 3.0 (Kumar *et al.*, 2004).

Differentiation between populations was assessed by sequence-based F statistics analogous to Wright F statistics (Wright, 1978), calculated according to Hudson *et al.* 1992 and tested for significance by permutation in DnaSp 4.0 (Rozas & Rozas 1999). Calculations not available in DnaSp and MEGA were carried out using programs written by TL in SAS (SAS Institute Inc., 1990).

Selection inference is primarily based on comparisons between different functional regions of the gene (defined below) to avoid confounding the effect of population demography and PCR errors that affect all regions of the gene equally. Similarly to a comparison of synonymous and non-synonymous mutations, this approach is conservative because polymorphism in shorter DNA fragments is subject to higher sampling variation, reducing the power to detect differences between regions. Physical linkage between adjacent regions may further reduce

the differences between them even if selection operated on only one region. The advantage of this approach, however, is that significant differences represent robust evidence for selection.

# RESULTS

## Genetic diversity

Within-population polymorphism in the *defensin* gene region was moderate to high (Table 1). The lowest polymorphism (in the whole gene) was observed in *An. arabiensis* (pi=0.015) and the highest in the western Kenyan population of *An. gambiae* (pi=0.028). Examination of intra-population variation using a sliding window revealed over ten fold difference across various segments of the gene, and considerable albeit lesser differences between species and populations (Fig. 1). In each collection, at least one site had three nucleotides segregating. Eight of 13 such sites were singletons that could have arisen as PCR errors. However, five of these were shared between populations or were observed at least twice within a collection, indicating that these were true triallelic SNPs. Although representing only 2–4% of the overall 305 segregating sites in the dataset, triallelic SNPs suggest some balanced and ancient polymorphism.

If selection has not shaped variation in *defensin*, a similar pattern of polymorphism is expected across its functional regions. These regions were defined *a-priori* and included: (i) mature peptide (120 bp), (ii) signal peptide and the clipped pro-peptide (186 bp), (iii) total transcribed but non coding regions including the intron (385 bp, excluding all gaps), and (iv) the flanking non-transcribed regions (600 bp, excluding all gaps and missing data). Nucleotide diversity in the coding region was lower than that of the non-coding region in all populations (Table 1, $P<0.016$, binomial test). Comparing the site frequency spectra between different functional regions provided a more comprehensive test of that variation. Frequency spectra were grouped into ‘rare alleles’ (singletons), ‘moderate alleles’ (sites where the rare nucleotide was observed two or three times), and ‘common alleles’ (sites where the rare nucleotide was observed four or more times). Invariant sites were included to accommodate total length variation between regions. Within population heterogeneity was detected in four out of six populations (Table 1). An overall within-population test was performed after finding no evidence for between-population heterogeneity (heterogeneity $X^2=56.5$, df=45, $P>0.1$). Marked deviations from homogeneity were concentrated in the coding region, showing deficiency of rare and moderate alleles in both the mature peptide and the signal/pro-peptide segments, but no significant deficiency at the high frequency sites. Additionally, an excess of rare alleles (singletons) and common alleles was observed in the transcribed-non-coding region (Table 1).

The heterogeneity in magnitude and pattern of polymorphism across functional regions and the reduced polymorphism in the coding region demonstrate selection on *defensin*. The expected signature of balancing (diversifying) selection would be increased polymorphism in the coding region while that of directional selection would involve reduced frequency of common alleles in that region; neither pattern was observed. The results, however, are consistent with purifying selection eliminating rare mutations from the coding regions.

## Synonymous vs nonsynonymous substitutions

In the coding region, within-population nucleotide diversity at synonymous sites was 4 fold higher than that of nonsynonymous sites and this difference was significant in five collections based on the two-sided test of neutral evolution (Table 2). This difference was most extreme in the mature peptide where no replacement mutations were observed in four populations and the overall difference in synonymous vs. nonsynonymous diversity (across populations) was 100 fold (Table 2). The short length and limited variation in this region resulted in low statistical

power, but the overall test (across collections) was significant (Table 2). At the protein level, except two singleton peptides, one single mature peptide was observed across the three species and six geographically distinct populations. Together with the high silent polymorphism across the gene and moderate synonymous diversity in the coding region, this extremely low diversity at the protein level in the mature peptide provides strong evidence that purifying selection is the main mode of selection operating on *defensin*.

### Species divergence and population differentiation

Divergence between species was highly significant (P<0.001) and Fst values exceeded 0.3 across all functional regions (Table 3). The mature peptide showed higher divergence between species than any other functional region (P<0.05, Ryan-Einot-Gabriel-Welsh Multiple Range Test following significant 'Region' effect in ANOVA, not shown). Moreover, it was the only region showing no differentiation among *An. gambiae* populations, whilst the magnitude of differentiation across the other functional regions were similar to each other (Table 3) and to the average of nine microsatellite loci (Fst=0.063, P<0.001; Lehmann *et al.* 2003). Heterogeneity in differentiation across functional regions is further evidence for selection. Increased inter-specific divergence and reduced differentiation among distant *An. gambiae* populations indicates strong purifying selection operating on the mature peptide in all species, hereby reducing within-species variation in this region and independently "fixing" few neutral mutations between species. This higher rate of fixation results in high fraction of between-species variation relatively to low fraction of within-species variation, yielding high Fst.

## DISCUSSION

Within- and between-population variation in *defensin* provides clear evidence that selection has shaped polymorphism in this gene. Purifying selection on the mature peptide explains (i) the overall reduced polymorphism in the mature peptide and the total coding region and the specific reductions in rare and moderately frequent alleles compared with non coding regions, (ii) the markedly reduced rate of nonsynonymous diversity compared with synonymous and the identity of the mature peptide across three species, and (iii) the increased divergence between species in the mature peptide together with the reduced differentiation between populations of *An. gambiae* in the same region. These patterns of variation were exhibited by all species and populations despite marked differences in their exposure to human pathogens due to their different preference to feed on human hosts (Hadis *et al.*, 1997;Lemasson *et al.*, 1997;Dekker&Takken, 1998;Mouchet *et al.*, 2004). High larval mortality (>95%) was measured under natural conditions and attributed to infections with nematodes and fungi (Jenkins, 1964;Service, 1973). These larval pathogens may be widespread throughout the species' range, explaining the identity of the mature defensin across the species and populations. However, it is likely that defensin is an effector molecule of broad target spectrum, so that it remains effective despite changes in pathogen composition affecting different populations.

Importantly, variation in susceptibility to pathogens (including human pathogens) among individual mosquitoes of the *An. gambiae* complex (Niare *et al.*, 2002;Lambrechts *et al.*, 2005) is not related to peptide variation in the mature defensin. The pattern of selection on defensin is inconsistent with positive and with balancing (diversifying) selection. Accordingly, the evolutionary dynamics between pathogens that mediated selection on defensin and mosquitoes in these populations are incompatible with the arms race dynamics. Consistent with purifying selection, the dynamics of the interactions may be governed by the cost of further increased anti-microbial effect on fitness as part of the functional constraints limiting variation of *defensin*, albeit not exclusively so.

**Acknowledgements**

The authors are grateful to Ananias Escalante, Jose Ribeiro, Randy Dejong, Franck Prugnolle, Adam Richman and Norio Kobayashi for useful discussions and comments. FS received financial support from the American Society for Microbiology postdoctoral fellowship program. This study was supported by UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases Grant A990476 to TL.

## References

Barillas-Mury C, Charlesworth A, Gross I, Richman A, Hoffmann JA, Kafatos FC. Immune factor Gambif1, a new rel family member from the human malaria vector, Anopheles gambiae. Embo J 1996;15:4691–4701. [PubMed: 8887560]

Blandin S, Moita LF, Kocher T, Wilm M, Kafatos FC, Levashina EA. Reverse genetics in the mosquito Anopheles gambiae: targeted disruption of the Defensin gene. EMBO Rep 2002;3:852–856. [PubMed: 12189180]

Blandin S, Shiao SH, Moita LF, Janse CJ, Waters AP, Kafatos FC, Levashina EA. Complement-like protein TEP1 is a determinant of vectorial capacity in the malaria vector Anopheles gambiae. Cell 2004;116:661–670. [PubMed: 15006349]

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 1995;140:783–796. [PubMed: 7498754]

Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S, Blass C, Brey PT, Collins FH, Danielli A, Dimopoulos G, Hetru C, Hoa NT, Hoffmann JA, Kanzok SM, Letunic I, Levashina EA, Loukeris TG, Lycett G, Meister S, Michel K, Moita LF, Muller HM, Osta MA, Paskewitz SM, Reichhart JM, Rzhetsky A, Troxler L, Vernick KD, Vlachou D, Volz J, von Mering C, Xu J, Zheng L, Bork P, Kafatos FC. Immunity-related genes and gene families in Anopheles gambiae. Science 2002;298:159–165. [PubMed: 12364793]

Collins FH, Petrarca V, Mpofu S, Brandling-Bennett AD, Were JB, Rasmussen MO, Finnerty V. Comparison of DNA probe and cytogenetic methods for identifying field collected *Anopheles gambiae* complex mosquitoes. Am J Trop Med Hyg 1988;39(6):545–550. [PubMed: 3207175]

Collins FH, Kamau L, Ranson HA, Vulule JM. Molecular entomology and prospects for malaria control. Bull Wld Hlth Organ 2000;78:1412–1423.

Dekker T, Takken W. Responses of Anopheles arabiensis Patton and Anopheles quadriannulatus Theobald to carbon dioxide, a man and a calf. Med Vet Entomol 1998;12:136–140. [PubMed: 9622366]

Dimopoulos G, Seeley D, Wolf A, Kafatos FC. Malaria infection of the mosquito Anopheles gambiae activates immune- responsive genes during critical transition stages of the parasite life cycle. EMBO J 1998;17:6115–6123. [PubMed: 9799221]

Dimopoulos G. Insect immunity and its implication in mosquito-malaria interactions. Cell Microbiol 2003;5(1):3–14. [PubMed: 12542466]

Eggleston P, Lu W, Zhao Y. Genomic organization and immune regulation of the defensin gene from the mosquito, Anopheles gambiae. Insect Mol Biol 2000;9:481–490. [PubMed: 11029666]

Favia G, della Torre A, Bagayoko M, Lanfrancotti A, Sagnon NF, Toure YT, Coluzzi M. Molecular identification of sympatric chromosomal forms of Anopheles gambiae and further evidence of their reproductive isolation. Insect Mol Biol 1997;6:377–383. [PubMed: 9359579]

Gillies MT, Coetzee M. A supplement to the Anophelinae of Africa south of the Sahara (Afrotropical Region). Johannesburg: South African Institute for Medical Research 1987;55:141p.

Gillies MT, De Meillon B. The Anophelinae of Africa south of the Sahara (Ethiopian zoogeographical region), 2nd ed. Johannesburg: South African Institute for Medical Research 1968;54:343p.

Grossman GL, Rafferty CS, Clayton JR, Stevens TK, Mukabayire O, Benedict MQ. Germline transformation of the malaria vector, Anopheles gambiae, with the piggyBac transposable element. Insect Mol Biol 2001;10:597–604. [PubMed: 11903629]

Hadis M, Lulu M, Makonnen Y, Asfaw T. Host choice by indoor-resting Anopheles arabiensis in Ethiopia. Trans R Soc Trop Med Hyg 1997;91(4):376–378. [PubMed: 9373624]

Hay SI, Rogers DJ, Toomer JF, Snow RW. Annual Plasmodium falciparum entomological inoculation rates (EIR) across Africa: literature survey, internet access and review. Trans R Soc Trop Med Hyg 2000;94:113–127. [PubMed: 10897348]

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O'Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, et al. The genome sequence of the malaria mosquito Anopheles gambiae. Science 2002;298:129–149. [PubMed: 12364791]

Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. Genetics 1992;132:583–589. [PubMed: 1427045]

Jenkins DW. Pathogens, parasites and predators of medically important arthropods. Annoted list and bibliography. Bull Wld Hlth Org 1964;30(suppl):5–150.

Kumar S, Tamura K, Nei M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 2004;5:150–163. [PubMed: 15260895]

Kwiatowski J, Skarecky D, Hernandez S, Pham D, Quijas F, Ayala FJ. High fidelity of the polymerase chain reaction. Mol Biol Evol 1991:884–887. [PubMed: 1775068]

Lambrechts L, Halbert J, Durand P, Gouagna LC, Koella JC. Host genotype by parasite genotype interactions underlying the resistance of anophelines mosquitoes to Plasmodium falciparum. Malaria Journal 2005;4:3–11. [PubMed: 15644136]

Lehmann T, Hawley WA, Kamau L, Fontenille D, Simard F, Collins FH. Genetic differentiation of Anopheles gambiae populations from East and West Africa: comparison of microsatellite and allozyme loci. Heredity 1996;77:192–200. [PubMed: 8760401]

Lehmann T, Licht M, Elissa N, Maega BT, Chimumbwa JM, Watsenga FT, Wondji CS, Simard F, Hawley WA. Population Structure of Anopheles gambiae in Africa. J Hered 2003;94:133–147. [PubMed: 12721225]

Lemasson JJ, Fontenille D, Lochouarn L, Dia I, Simard F, Ba K, Diop A, Diatta M, Molez JF. Comparison of behavior and vector efficiency of Anopheles gambiae and An. arabiensis (Diptera: Culicidae) in Barkedji, a sahelian area of Senegal. J Med Entomol 1997;34(4):396–403. [PubMed: 9220672]

Meister S, Kanzok SM, Zheng XL, Luna C, Li TR, Hoa NT, Clayton JR, White KP, Kafatos FC, Christophides GK, Zheng L. Immune signaling pathways regulating bacterial and malaria parasite infection of the mosquito Anopheles gambiae. Proc Natl Acad Sci USA 2005;102(32):11420–11425. [PubMed: 16076953]

Mouchet, J.; Carnevale, P.; Coosemans, M.; Julvez, J.; Manguin, S.; Richard-Lenoble, D.; Sircoulon, J. Biodiversité du paludisme dans le monde. John Libbey Eurotext; Paris: 2004. p. 428p

Niare O, Markianos K, Volz J, Oduol F, Toure A, Bagayoko M, Sangare D, Traore SF, Wang R, Blass C, Dolo G, Bouare M, Kafatos FC, Kruglyak L, Toure YT, Vernick KD. Genetic loci affecting resistance to human malaria parasites in a West African mosquito vector population. Science 2002;298:213–216. [PubMed: 12364806]

Osta MA, Christophides GK, Vlachou D, Kafatos FC. Innate immunity in the malaria vector Anopheles gambiae: comparative and functional genomics. J Exp Biol 2004a;207:2551–2563. [PubMed: 15201288]

Osta MA, Christophides GK, Kafatos FC. Effects of mosquito genes on Plasmodium development. Science 2004b;303:2030–2032. [PubMed: 15044804]

Richman AM, Bulet P, Hetru C, Barillas-Mury C, Hoffmann JA, Kafalos FC. Inducible immune factors of the vector mosquito Anopheles gambiae: biochemical purification of a defensin antibacterial

peptide and molecular cloning of preprodefensin cDNA. Insect Mol Biol 1996;5:203–210. [PubMed: 8799739]

Richman AM, Dimopoulos G, Seeley D, Kafatos FC. Plasmodium activates the innate immune response of Anopheles gambiae mosquitoes. EMBO J 1997;16:6114–6119. [PubMed: 9321391]

Rozas J, Rozas R. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 1999;15:174–175. [PubMed: 10089204]

SAS. SAS language: references, version 6. Cary, NC: SAS Institute; 1990.

Schlenke TA, Begun DJ. Natural selection drives Drosophila immune system evolution. Genetics 2003;164:1471–1480. [PubMed: 12930753]

Scott JA, Brogdon WG, Collin FH. Identification of single specimens of Anopheles gambiae complex by the polymerase chain reaction. Am J Trop Med Hyg 1993;49(4):520–529. [PubMed: 8214283]

Service MW. Mortalities of the larvae of the Anopheles gambiae Giles complex and detection of predators by the precipitin test. Bull Ent Res 1973;62:359–369.

Shahabuddin M, Fields I, Bulet P, Hoffmann JA, Miller LH. Plasmodium gallinaceum: differential killing of some mosquito stages of the parasite by insect defensin. Exp Parasitol 1998;89:103–112. [PubMed: 9603495]

Tahar R, Boudin C, Thiery I, Bourgouin C. Immune response of Anopheles gambiae to the early sporogonic stages of the human malaria parasite Plasmodium falciparum. EMBO J 2002;21(24): 6673–6680. [PubMed: 12485988]

Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989;123:585–595. [PubMed: 2513255]

Vizioli J, Richman AM, Uttenweiler-Joseph S, Blass C, Bulet P. The defensin peptide of the malaria vector mosquito Anopheles gambiae: antimicrobial activities and expression in adult mosquitoes. Insect Biochem Mol Biol 2001;31:241–248. [PubMed: 11167093]

Wright, S. Evolution and the genetics of populations: variability within and among natural populations. Chicago: University Press; 1978.

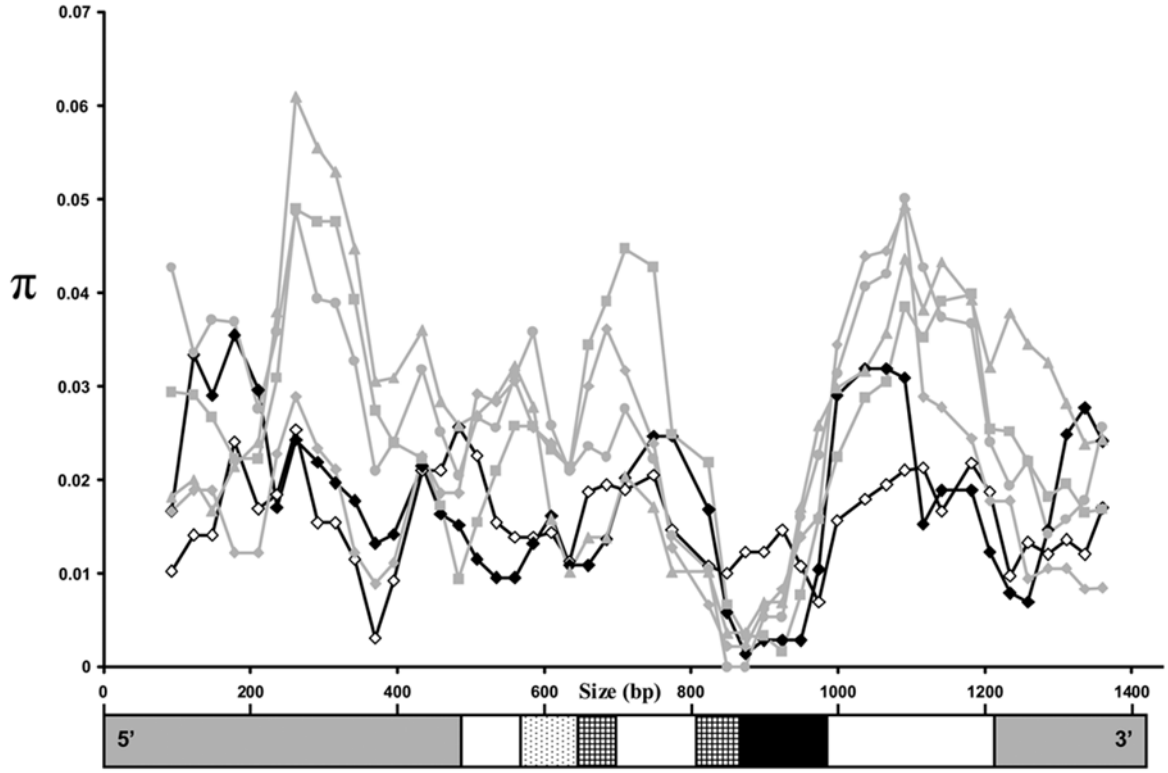


**Figure 1.**

Variation of nucleotide diversity (π) along the *defensin* gene sequence within species and populations of the *An. gambiae* complex.

Above:

Total nucleotide diversity variation within *An. arabiensis* (black line, closed diamonds), *An. quadriannulatus* (black line, open diamonds) and the *An. gambiae* populations from West-Kenya (grey line, closed circles), East-Kenya (grey line, closed triangles), Nigeria (grey line, closed squares) and Senegal (grey line, closed diamonds). Graphs were generated using DnaSp 4.0 (Rozas & Rozas, 1999) using a sliding window of 100bp and a step length of 25 bp.

Below:

Schematics of the *defensin* gene showing different functional regions: the 5′ and 3′ non-transcribed non-coding flanking regions (▭ NoTran), the transcribed but non coding region including the intron (□ TrNoCo), and both exons of the *defensin* gene encoding the signal peptide (▭ Sig), the clipped pro-peptide (▦ Pro) and the mature peptide (■ Mature).

## Table 1

Molecular polymorphism and site frequency spectra across functional regions of the *defensin* gene in populations of the *Anopheles gambiae* complex.

| Population | N | Region[a] | π | Allele frequency class (%)[b] | | | |
|---|---|---|---|---|---|---|---|
| | | | | Invariant f=0 | Rare alleles f=1 | Moderate alleles f=2–3 | Common alleles f=4–7 |
| *An. gambiae* | | | | | | | |
| W. Kenya | 10 (1284) | Mature | 0.0133 | 97.5 | -0^ | -0* | 2.5 |
| | | SigPro | 0.0176 | 96.2 | -0* | 2.2 | 1.6 |
| | | TrNoCo | 0.0316 | 92.5 | 1.4 | 4.4 | 1.7 |
| | | NoTran | 0.0313 | 90.8 | +4.1** | 4.1 | 1.1 |
| | | | | | $X^2$ = 25.2, df=9, P<0.004[c] | | |
| E. Kenya | 11 (1280) | Mature | 0.0142 | 96.7 | 0.8 | -0^ | 2.5 |
| | | SigPro | 0.0121 | 96.2 | 1.6 | 1.1 | 1.1 |
| | | TrNoCo | 0.0309 | 90.3 | 4.4 | 3.9 | 1.4 |
| | | NoTran | 0.0312 | 91.4 | 2.9 | 3.1 | 2.6 |
| | | | | | $X^2$ = 16.3, df=9, P<0.06[c] | | |
| Nigeria | 12 (1315) | Mature | 0.0078 | 96.7 | −1.7^ | 1.7 | 0 |
| | | SigPro | 0.0236 | 92.5 | 2.7 | 3.2 | 1.6 |
| | | TrNoCo | 0.0330 | 87.8 | +8.0* | 1.9 | 2.2 |
| | | NoTran | 0.0256 | 90.6 | 4.9 | 3.6 | 0.9 |
| | | | | | $X^2$ = 19.3, df=9, P<0.030[c] | | |
| Senegal | 9 (1272) | Mature | 0.0116 | 96.7 | 0.8 | 2.5 | 0 |
| | | SigPro | 0.0206 | 95.2 | 1.6 | 0.5 | +2.7^ |
| | | TrNoCo | 0.0303 | 92.0 | 3.7 | 2.0 | +2.3^ |
| | | NoTran | 0.0147 | 95.4 | 2.8 | 1.5 | −0.3^ |
| | | | | | $X^2$ = 18.9, df=9, P<0.022[c] | | |
| *An. arabiensis* | 13 (1301) | Mature | 0.0090 | 96.7 | 1.7 | 1.7 | 0 |
| | | SigPro | 0.0146 | 95.2 | 2.1 | 1.1 | 1.6 |
| | | TrNoCo | 0.0160 | 93.6 | 4 | 0.8 | 1.6 |
| | | NoTran | 0.0160 | 94.7 | 2.6 | 1.6 | 1.1 |
| | | | | | $X^2$ = 6.5, df=9, P<0.69[c] | | |
| *An. quadriannulatus* | 14 (1322) | Mature | 0.0024 | 98.3 | 1.7 | 0 | 0 |
| | | SigPro | 0.0059 | 97.9 | −1.1^ | 0.5 | 0.5 |
| | | TrNoCo | 0.0237 | 91.6 | 4.3 | 1.1 | +3.0* |
| | | NoTran | 0.0208 | 91.7 | 4.8 | +2.3^ | 1.2 |
| | | | | | $X^2$ = 22.7, df=9, P<0.012[c] | | |
| All pooled[d] | 69 (1248) | Mature | 0.0181 | 97.1 | −1.1*** | −1.0* | 0.8 |
| | | SigPro | 0.0199 | 95.5 | −1.5*** | −1.4^ | 1.5 |
| | | TrNoCo | 0.0393 | 91.3 | +4.3** | 2.3 | +2.0* |
| | | NoTran | 0.0319 | 92.4 | 3.7 | +2.7^ | 1.2 |
| | | Overall | 0.0308 | 93.0 | 3.3 | 2.2 | 1.5 |
| | | | | | $X^2$ = 52.4, df=9, P<0.0001[c] | | |

N: Number of sequences in the population and, in parentheses, sequence length after removing all gaps; π: total nucleotide diversity (silent and non-synonymous).

[a] DNA functional regions including regions encoding the mature peptide (Mature), the signal peptide and the clipped pro-peptide (SigPro), the total transcribed but non coding regions including the intron (TrNoCo), and the 3′ and 5′ non-transcribed regions (NoTran). See Figure 1 for a sketch map of the total *defensin* gene region.

[b] Frequency spectra classes including invariant positions (f=0), low polymorphism represented by singleton positions (f=1), moderately polymorphic positions with the rare nucleotide observed two or three times (f=2–3), and highly polymorphic positions with the rare nucleotide observed four or more times (f=4–7). The relative distribution of each class is expressed as percentages. Excess and deficit of observed vs. expected frequency is marked by '+' and '−', respectively in cells with significant deviations based on 1 df $X^2$ test (^, *, **, ***, represent P<0.1, 0.05, 0.01, and 0.001, respectively).

[c] Within-population heterogeneity in the spectra was tested by exact tests using Monte Carlo simulations.

[d] Heterogeneity $X^2$ test across species and populations failed to reject that null hypothesis ($X^2$ = 56.5, df=45, P>0.1) before data were pooled.

**Table 2**

Nucleotide diversity ($\pi$) within population at non-synonymous and synonymous sites in the coding region of *defensin*.

| Population | N | Coding Region[a] (Total) 306 bp = 102 aa NonSyno[b] | P[c] | Syno[d] | Signal/Pro-peptide 186 bp = 62 aa NonSyno | P | Syno | Mature Peptide 120 bp = 40 aa NonSyno | P | Syno |
|---|---|---|---|---|---|---|---|---|---|---|
| *A. gambiae-pool* | 42 | 0.008 (9) | ** | 0.042 (18) | 0.012 (7) | ns | 0.041 (13) | 0.001 (2) | ns | 0.043 (5) |
| W. Kenya | 10 | 0.009 (4) | ns | 0.036 (6) | 0.015 (4) | ns | 0.025 (3) | 0 (0) | ns | 0.053 (3) |
| E. Kenya | 11 | 0.005 (4) | * | 0.035 (7) | 0.008 (3) | ns | 0.025 (4) | 0.002 (1) | ns | 0.050 (3) |
| Nigeria | 12 | 0.009 (6) | ** | 0.041 (12) | 0.014 (5) | * | 0.051 (9) | 0.002 (1) | ns | 0.026 (3) |
| Senegal | 9 | 0.006 (3) | ** | 0.049 (10) | 0.010 (3) | * | 0.051 (6) | 0 (0) | ns | 0.046 (4) |
| *An. arabiensis* | 13 | 0.006 (4) | * | 0.032 (9) | 0.009 (4) | ns | 0.030 (5) | 0 (0) | ns | 0.036 (4) |
| *An. quadriannulatus* | 14 | 0 (0) | * | 0.017 (6) | 0 (0) | ns | 0.022 (4) | 0 (0) | ns | 0.009 (2) |
| All pooled | 69 | 0.007 (11) | *** | 0.053 (28) | 0.012 (9) | * | 0.043 (17) | 0.0007 (2) | * | 0.070 (11) |

N: Number of sequences;

[a] Excluding the intron, no insertion/deletion in the coding region.

[b] Nucleotide diversity ($\pi$) at Non-Synonymous sites (amino acid replacements) and the number of such sites in parenthesis.

[c] P-value of the two-sided test of neutral evolution available in MEGA 3.0 software: ns, not significant (P>0.05),

* P<0.05,

** P<0.01,

*** P<0.001.

[d] Nucleotide diversity ($\pi$) at synonymous sites (silent mutations) and the number of such sites in parenthesis.

**Table 3**

Divergence between species and differentiation between populations of *An. gambiae* measured by Fst across functional regions of the *defensin* gene.

| | DNA functional region[a] | | | |
|---|---|---|---|---|
| **Mean Fst** | **Mature** | **SigPro** | **TrNoCo** | **NoTran** |
| across species | 0.77** | 0.35** | 0.51** | 0.44** |
| between *An. gambiae* populations | −0.003 | 0.095* | 0.077** | 0.084** |

Fst was estimated according to Hudson *et al.* 1992 implemented in Dnasp 4.0 (Rozas & Rozas, 1999).

*P<0.01;

**P<0.001.

[a]Functional regions of the *defensin* gene are as defined in Table 1 and Figure 1.