



**HAL**  
open science

## Field, genetic and modelling approaches show strong positive selection acting upon an insecticide resistance mutation in *Anopheles gambiae* s.s.

Amy Lynd, David Weetman, Susana Barbosa, Alexander Egyir Yawson, Sara Mitchell, Joao Pinto, Ian Hastings, Martin James Donnelly

### ► To cite this version:

Amy Lynd, David Weetman, Susana Barbosa, Alexander Egyir Yawson, Sara Mitchell, et al.. Field, genetic and modelling approaches show strong positive selection acting upon an insecticide resistance mutation in *Anopheles gambiae* s.s.. *Molecular Biology and Evolution*, 2010, 27 (5), pp.1117. 10.1093/molbev/msq002 . hal-00679169

**HAL Id: hal-00679169**

**<https://hal.science/hal-00679169>**

Submitted on 15 Mar 2012

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Field, genetic and modelling approaches show strong positive selection acting upon an insecticide resistance mutation in *Anopheles gambiae s.s.*

Amy Lynd<sup>a</sup>, David Weetman<sup>a</sup>, Susana Barbosa<sup>b</sup>, Alexander Egyir-Yawson<sup>c</sup>, Sara Mitchell<sup>a</sup>, Joao Pinto<sup>d</sup>, Ian Hastings<sup>b</sup> and Martin J Donnelly<sup>a,e</sup>

<sup>a</sup>Vector Group, Liverpool Tropical School of Medicine, Pembroke Place, Liverpool, L3 5QA, UK.

<sup>b</sup>Molecular and Biochemical Parasitology Group, Liverpool Tropical School of Medicine, Pembroke Place, Liverpool, L3 5QA, UK.

<sup>c</sup>Biotechnology and Nuclear Agriculture Research Institute, Ghana Atomic Energy Commission, Kwabenya, Accra, Ghana.

<sup>d</sup>Centro de Malária e outras Doenças Tropicais/IHMT/UNL, Rua da Junqueira 96, 1349-008 Lisbon, Portugal

<sup>e</sup>Corresponding author: Martin Donnelly, Vector Group, Liverpool Tropical School of Medicine, Pembroke Place, Liverpool, L3 5QA, UK. Telephone +441517053296, Fax +441517053369, E-mail [m.j.donnelly@liv.ac.uk](mailto:m.j.donnelly@liv.ac.uk)

**Submission:** Research article

**Keywords:** *Anopheles gambiae*, malaria, insecticide resistance, selection

**Running title:** Selection on an insecticide resistance locus

## **Abstract**

Alleles subject to strong, recent positive selection will be swept towards fixation together with contiguous sections of the genome. Whether the genomic signatures of such selection will be readily detectable in outbred wild populations is unclear. In this study we employ haplotype diversity analysis to examine evidence for selective sweeps around *knock down resistance (kdr)* mutations associated with resistance to DDT and pyrethroid insecticides in the mosquito *Anopheles gambiae*. Both *kdr* mutations have significantly lower haplotype diversity than the wild type (non-resistant) allele, with *kdr L1014F* showing the most pronounced footprint of selection. We complement these data with a time series of collections showing that the *L1014F* allele has increased in frequency from 0.05 to 0.54 in 5 years, consistent with a maximum likelihood-fitted selection coefficient of 0.16 and a dominance coefficient of 0.25. Our data show that strong, recent positive selective events, such as those caused by insecticide resistance, can be identified in wild insect populations.

## Introduction

The sequencing of the *Anopheles gambiae* genome has opened up the possibility for genome-wide SNP-based association mapping studies that have been successful in identifying positively selected loci in the human genome (Sabeti et al. 2002; Bersaglieri et al. 2004; Sabeti et al. 2007). The resolution of the association mapping approach is defined by the probability that recombination will have broken down the association between markers and a trait-associated functional polymorphism. Data from extensive re-sequencing of (primarily) detoxification genes in samples from wild populations of *An. gambiae* revealed a very high frequency of segregating sites (Wilding et al. 2009), consistent with high rates of recombination (Begun, Aquadro 1992; Begun et al. 2007) and/or a long history of outbreeding. In isofemale lab strains of *Drosophila spp.* it has been possible to observe selective sweeps around insecticide-resistance associated loci (Schlenke, Begun 2004; Aminetzach, Macpherson, Petrov 2005) but how long these signatures persist in wild populations is unknown. In this paper we use linkage disequilibrium-based haplotype diversity analysis (Sabeti et al. 2006) to investigate the pattern of molecular genetic variation associated with insecticide resistance mutations at the pyrethroid and DDT knock down resistance locus, *kdr*, in the African malaria mosquito *An. gambiae s.s.* Furthermore, as a corollary of this indirect genetic approach we demonstrate, using a series of temporal collections, a dramatic increase in *kdr* frequency in a population of *An. gambiae s.s.* over a period of approximately 72 generations. Data from these temporal collections are used to estimate the selection and dominance coefficients operating on *kdr* in the field to illustrate the potential levels of selection necessary to produce the patterns of linkage disequilibrium (LD) we observe.

Insecticide-treated bed nets are the principal method for preventing malaria in sub-Saharan Africa. Currently, pyrethroids are the only class of insecticides licensed for use on nets and there is concern that resistance will compromise control programmes. To date the most commonly recorded resistance mechanism is termed *knock down resistance*, and results from single base pair mutations in the voltage-gated sodium channel. The sodium channel gene, located within division 20C near the centromere of chromosome 2L, codes for a protein which is the target-site of pyrethroid insecticides. Two alternative single base pair mutations have been found in *An. gambiae* and these *kdr* mutations can cause target site insensitivity to pyrethroids as well as cross-resistance to DDT. The substitutions cause amino acid changes

at codon 1014 within the transmembrane structure of segment 6 in domain II of the voltage-gated sodium channel (numbering according to the housefly *para* sequence, Genbank X96668). The *L1014F* mutation, a leucine to phenylalanine change, was first observed in West Africa (Martinez-Torres et al. 1998) and the same substitution has been observed in a diverse array of insects (Davies et al. 2007a). A second substitution, *L1014S*, was observed more recently in East African *An. gambiae* (Ranson et al. 2000) and involves the adjacent base of the same codon, resulting in a leucine to serine change.

There are two incipient species within the nominal taxon *Anopheles gambiae s.s.* which are characterized by mutations on the X-chromosome and are termed M and S form. The distribution of the *kdr* mutation is not uniform either within or between forms, although in general *kdr* alleles have been found at much higher frequencies in *Anopheles gambiae s.s.* S-form samples compared to M-form samples [reviewed in (Santolamazza et al. 2008)]. The reasons for the differences in distribution remain unclear since little is known about the origins of the *kdr* mutations and the selection pressures acting upon them in wild populations. In a sample from Benin, the *L1014F* was found in tight linkage disequilibrium with two upstream intronic polymorphisms in both M and S-form individuals. The two upstream polymorphisms associated with the *L1014F* variant were not found in wildtype M-form individuals but were common in wildtype S form individuals, suggestive of an introgression event from S-form to M-form populations (Weill et al. 2000). This linkage between *kdr* and the intronic polymorphisms was not seen in M form individuals from Bioko Island and was thought to indicate *de-novo* mutation (Reimer et al. 2008). More recently a study of S-form specimens from 15 countries suggested that the *L1014F* and *L1014S* mutations have both arisen independently on at least two separate occasions (Pinto et al. 2007).

Samples were obtained from three regions in sub-Saharan Africa; Kenya (East Africa) *Anopheles gambiae*, S molecular form, *kdr L1014S* allele present; Ghana (West Africa) both M and S molecular form, *kdr L1014F* allele present; Gabon (Central Africa) S molecular form, both the *L1014S* and *L1014F kdr* alleles present.

These population samples allow us to address a number of questions.

1. Available evidence suggests that the *L1014S* mutation has high penetrance for a DDT-resistant phenotype but lower penetrance for a pyrethroid-resistant phenotype than the *L1014F* mutation (Ranson et al. 2000). DDT was banned in Kenya in 1990 and we

can investigate the signature of positive selection associated with weaker selection or recombination and relaxed selection.

2. The populations from central Africa are some of the few locations where both the *L1014F* and *L1014S* alleles are observed sympatrically (Santolamazza et al. 2008). Indeed in an earlier study a significant, albeit marginal, *L1014F/L1014S* heterozygote excess was observed in samples from Libreville, Gabon (Pinto et al. 2006). By comparing patterns of linkage disequilibrium around the three alleles we investigate whether the unusually high frequency of the *L1014S* allele in these populations (63% (Pinto et al. 2006)) is a result of a recent selective sweep.
3. In many S-form populations in West Africa, including our collections from Ghana, the *L1014F* allele is close to fixation. In the absence of wild-type alleles we are unable to control for local variation in recombination rates (Sabeti et al. 2007) and it is therefore impossible to ascribe patterns of LD to a positive selection event. Recently developed approaches such as Cross Population Extended Haplotype Homozygosity have been developed to allow interpopulation comparisons in instances where alleles have proceeded to near fixation in some populations (Sabeti et al. 2007) but in our system resistance alleles may have multiple origins, presenting a confounding variable (Pinto et al. 2007). However, the presence of sympatric M-form populations in southern Ghana (Yawson et al. 2004; Yawson et al. 2007) allow us to both document the increase in frequency of the same *L1014F* haplotype, following an introgression event, over a period of 5 years; and to estimate the selection and dominance coefficients associated with the signatures of positive selection.

## Materials and methods

### *Sample sites, DNA extraction and species identification.*

Adult female *Anopheles gambiae s.s.* mosquitoes used in this study were obtained from aspirator and pyrethroid knock-down collections from the field in various geographic locations (table 1). DNA was extracted from single female *Anopheles gambiae* using either a modified Livak method or a phenol-chloroform method (Livak 1984; Ballinger-Crabtree, Black IV, Miller 1992). Species identification PCR, was carried out on *Anopheles gambiae s.l.* according to the protocol (Scott, Brogdon, Collins 1993). Reactions were then digested with CFO I restriction enzyme for 24 hours at 37°C in order to type *Anopheles gambiae s.s.* mosquitoes to M and S-form (Fanello, Santolamazza, della Torre 2002), and products visualised under UV light after electrophoresis on a 2% agarose TBE gel with Ethidium Bromide. *Kdr* genotypes were determined by allele specific PCR, Heated Oligonucleotide Ligation Assay (HOLA) (Lynd et al. 2005) or Taqman assay (Bass et al. 2007) depending upon year of collection.

### *Sodium channel SNP identification*

The voltage-gated sodium channel gene is nearly 74Kbp in length and is composed of 35 exons including two duplicate exons (Davies et al. 2007a). Ten regions of the sodium channel were amplified by PCR for direct sequencing. Where possible, primers were designed to bind within exons to produce amplicons that spanned an intron with a maximum size of 1.5Kbp. Exons (numbering as Davies et al. 2007a) 1-2, 3, 4, 7-9, 13-14, 15-17, 20c, 23-24, 28-30 and 32-33 were selected as targets for sequencing. Primer and amplification details are provided (supplementary table 1, Supplementary Material online). Sequencing for SNP detection was carried out on up to 12 individuals of known *kdr* genotype from Ghana, São Tomé, Gabon, Angola, Mozambique, Malawi, Kenya, from a susceptible laboratory strain (KISUMU) and from a permethrin tolerant resistant laboratory strain (RSP), both originating from Kenya. PCR products were cleaned using a Mini Elute PCR Purification kit (Qiagen) and then sequenced in both directions. Sequences were aligned using Bioedit software version 7.0.5.2 (Hall 1999) then manually annotated for polymorphisms and ambiguities.

In addition, seven M-form individuals from Accra, Ghana, homozygous for the *L1014F* allele were bidirectionally-sequenced across PCR amplicons 13-14, 15-17 and 21 to determine the associated haplotype of the *kdr* allele in this population.

### *SNP screening*

SNPs discovered through resequencing were screened in the large scale SNP detection study using the SNPStart Primer Extension Kit on the Beckman CEQ™ 8000 Genetic Analysis System. Details of SNPs both included and excluded from the SNP screening are given in supplementary table 2 (Supplementary Material online). Multiplex PCR was carried out to amplify the regions of DNA containing SNPs of interest, including a region of exon 20 and the preceding intron to allow high throughput detection of the *kdr* mutation and three other well characterized SNPs (Weill et al. 2000; Diabate et al. 2004; Pinto et al. 2006) (primers and reaction conditions detailed in supplementary table 3, Supplementary Material online). Products were visualized on a 2% TBE agarose gel. Successfully multiplexed samples were prepared for subsequent SNP extension by ExoI/SAP enzymatic digestion. Interrogation primers were then designed for each individual SNP chosen for investigation according to the manufacturers recommendations (supplementary table 4, Supplementary Material online). Single base extension to the 3' end of the interrogation primer by a dye terminator molecule, corresponding to the nucleotide found at the SNP location, was carried out using a GenomeLab SNPStart Primer Extension Kit (Beckman Coulter, Amersham, UK). The SAP-digested product was then scored on the Beckman CEQ 8000 Genetic Analysis System.

### *Data analysis*

As reviewed exhaustively by (Sabeti et al. 2006) there are numerous statistical tests of positive selection which differ in their ability to detect selection events on different time scales. For the present SNP dataset it is not possible to use the suite of sequence-based tests that compare synonymous/ non-synonymous differences or detect an excess of rare alleles. We are therefore fortunate that on the timescales in which the emergence, and selection, of insecticide resistance is likely to occur, estimates of interpopulation divergence (*e.g.* based on F-statistics) and screens of LD around selected *vs* wildtype alleles are likely to be the two most powerful analytical approaches. With the sample sizes available in our study, single marker analyses based on F-statistic estimates would perform better as indicators of selection when markers can be typed at a more coarse scale, with consequently enhanced signal: noise ratio. However, with sample size constraints the signal would be difficult to localise. By contrast long range haplotype analyses such as extended haplotype homozygosity (EHH) (Sabeti et al. 2002) analysis performs very well at a fine physical scale in identifying narrow candidate regions (Sabeti et al. 2006).



EHH analysis was carried out to assess the patterns of LD associated with wild type and the two *kdr* alleles. EHH can be defined as the probability that two random chosen chromosomes carrying the core (eg. the wildtype or *kdr* allele) haplotype of interest are identical by descent. This approach first identifies core haplotypes surrounding the locus of interest and then examines the decay in LD from these core haplotypes to the surrounding loci. The resulting EHH can be used as evidence of recent positive selection at a locus in haplotypes that have high frequency and high EHH (Sabeti et al. 2002). EHH analysis requires haplotype information that cannot be empirically determined from the genotype data gathered by the methods used in this study. Therefore haplotypes were inferred using PHASE software version 2.1.1 using default parameters (Stephens, Smith, Donnelly 2001; Stephens, Scheet 2005). PHASE utilizes a Bayesian coalescent-based approach to determine phase and allows for varying rates of recombination at each SNP interval. The method is based on the idea that an unresolved haplotype is more likely to be the same or be similar to a previous haplotype. This approach was found to outperform other methods available for autosomal human data sets (Stephens, Smith, Donnelly 2001; Stephens, Scheet 2005). Data were analysed together rather than as separate subpopulations since (a) previous studies found this to be more accurate, and (b) haplotype determination methods of this nature are relatively insensitive to departures from Hardy-Weinberg equilibrium so are fairly robust to population sub-structuring. This approach is also more conservative than determining haplotypes for individual populations since the latter is liable to lead to an underestimation in differences in haplotype frequencies (Stephens, Scheet 2005). Phase reconstruction was executed ten times upon the total data set and differences in counts of best haplotypes were noted.

The estimated haplotypes obtained from PHASE were used as input for EHH analysis implemented by SWEEP version 2.1.1 (Sabeti et al. 2002). Core haplotypes were selected manually to include only the two adjacent *kdr*-causing loci. Significance of EHH values is usually assigned through comparison to an empirically generated null distribution from other regions of the genome. However given that we had already identified the causal mutations of interest we were able to make a comparison of patterns of LD around wild-type and resistant cores. The primary advantage of this approach is that it is not subject to the genome-wide variations in recombination rate which can affect the null distribution approach in species lacking detailed recombination maps. Significant differences in EHH values were determined in two ways. (1) Within country samples, at individual SNP positions with non-overlapping 95% confidence intervals. These confidence intervals were calculated at each SNP position

using a bootstrapping procedure, carried out in SAS Version 9 software. Re-sampling was carried out 1000 times. (2) Across all SNPs within and among country samples the diversity of the different *kdr* allele-bearing haplotypes were compared using sign tests, implemented by SPSS 14. Where exact sign test probabilities could not be calculated, a Monte Carlo procedure with 10000 permutations was performed. The sequential Bonferroni procedure was applied to determine statistical significance following correction for multiple testing (Holm 1979). Although our data - EHH values at each SNP position - are not independent it is this non-independence caused by LD that will cause departure from the null hypothesis of equality of median EHH values. Therefore the null hypothesis remains that there is no difference in median EHH between *kdr* and wild type alleles. Bifurcation plots were also created using the SWEEP software. In a bifurcation plot the core haplotype is represented as a black circle. Each SNP, moving out from the core both upstream and downstream of the *kdr* locus, is a potential site for a bifurcation which would result from the presence of two segregating alleles. Therefore the diagram provides a means of displaying the breakdown in LD at increasing distance from the core haplotypes. The radius of the circle at each node is proportional to the number of individuals with that haplotype.

#### *Calculation of selection and dominance coefficients*

The spread of the *L1014F* allele was modelled using the standard recursive population genetic formula:

$$p' = \frac{p^2(1+s) + p(1-p)(1+hs)}{W} \quad (1)$$

Where  $p$  is the frequency of the *L1014F* allele,  $p'$  is the frequency in the next generation,  $s$  is the selective coefficient of the resistance mutation,  $h$  is the dominance coefficient (1=complete dominance, 0=complete recessivity), and  $W$  is the normalising factor (Maynard-Smith 1998).

Tracking allele frequencies over time requires three input parameters: initial allele frequency at time zero;  $s$ ; and  $h$ . Estimates of all three unknown parameters were obtained by maximum likelihood assuming a binomial distribution of observed allele frequencies around the predicted frequency. The analysis was performed in R (<http://www.r-project.org>) using maximum likelihood functions and optimizing routines. The generation time was set at the standard of one generation per calendar month (Lehmann et al. 1998).

## Results

### *SNP discovery and screening.*

Ten genomic regions of a combined length of  $\approx 6.5$ Kb of DNA, spanning a region of  $\approx 73$ Kb of the voltage-gated sodium channel were amplified and sequenced in *Anopheles gambiae* s.s. individuals from seven countries across sub-Saharan Africa. A total of 62 potential SNPs were found, of which 14 were exonic (supplementary table 1, Supplementary Material online). Six intronic indels were observed, usually in poly-A or tandem AT repeats (table 1, Supplementary Material online). On average there was one SNP every 106 bp, which represents a low SNP frequency for *An. gambiae*, but similar to other genes in the same genomic locality (chromosome 2L division 20; (Wilding et al. 2009)). Thirty-two SNPs, including the two *kdr* mutations, were selected for screening in 258 individuals. In S-form individuals the SNP adjacent to the core in the upstream (centromeric) direction was excluded from further analysis as it was found to be monomorphic. Details of the populations and associated *kdr* genotypes are given in table 1. The genotypic data were resolved into haplotypes with ten runs of the analysis. In only one instance did the replicate runs resolve a novel estimated haplotype, which in a subsequent comparative analysis was found to exert no qualitative effect on the results. Therefore all analyses reported here are based upon the haplotypes resolved in the vast majority of the phasing runs.

Extended haplotype homozygosity (EHH) analysis was carried out to assess the patterns of LD associated with the wild type and the two *kdr* alleles. The intronic SNPs which have been used to identify the origin of the *kdr* mutations were the proximate SNPs in the centromeric direction (Weill et al. 2000; Pinto et al. 2007). Linkage disequilibrium decay was examined between these core haplotypes and the remaining 29 or 30 SNP loci (for S- or M-forms, respectively).

Only two core-alleles were present in the western Kenyan sample: wild-type and *L1014S*. In the downstream telomeric direction, EHH decays at a similar rate for both wildtype and *L1014S*, but there was a marked contrast between alleles in the centromeric direction, with entirely non-overlapping confidence limits from just a few kb away from the core (fig. 1 and 2a). In the Gabonese collection the difference between resistance associated alleles and wild type was even more marked with significantly lower EHH in the wildtype in both centromeric and telomeric directions less than 5kb from the core (fig. 2b). Indeed both of the

*L1014F* and *L1014S* resistance mutations showed little haplotype bifurcation in the Gabon samples over the length of the sodium channel (fig.1) suggesting a relatively recent origin for both these mutations accompanied by a strong selective sweep. The patterns of LD are most marked around the resistant *L1014F* haplotype in Ghanaian S-form samples in which the *L1014F kdr* allele was at very high frequency (fig. 1 and 2c), as would be expected given the near fixation of this allele in southern Ghana in the S molecular form (mean frequency = 0.96; 95% CI 0.95-0.97) (Yawson et al. 2004). The presence of only two wildtype haplotypes in the sample prevent any meaningful comparison of LD decay but it should be noted that there was complete LD over the entire 64kb length of the sodium channel in the centromeric direction. The wildtype allele, observed in the Ghanaian M-form populations (fig. 1 and 2c) showed marked LD, only in the telomeric direction, between exons 20 and 32, the opposite directional asymmetry to the *L1014F* mutation in Ghana S-form populations. Whilst simulation studies have shown that LD decay may be asymmetric even when rates of mutation and recombination are constant (Kim, Stephan 2002) it is possible that the LD observed in these samples may reflect the presence of one or more hitherto overlooked selectively advantageous mutants, although we cannot rule out recombination with unsampled haplotypes (supplementary table 5, Supplementary Material online). (Davies et al. 2007b) have summarised that there are a number of additional non-synonymous changes observed in a variety of taxa and detailed association mapping studies are presently underway to investigate this phenomenon. Comparing overall levels of EHH for the whole 72.6 kb regions typed, it is interesting to note that median EHH values are statistically indistinguishable for the same allele typed in different populations (table 2) and that a clear hierarchy of evidence for selective sweeps emerged. Median EHH levels were highest for the *L1014F* resistance mutation, followed by those for the *L1014S* mutations, with the lowest for the wildtype allele (table 2). The only exception to this pattern was within the Gabonese sample, the only one in which both resistance alleles were present, where median EHH was equal for the two resistance alleles. Nevertheless, despite the possibilities of different origins of the same allele, and local variation in recombination rates, EHH levels across the genomic region investigated suggest some degree of commonality in selection across populations for each allele, although the actual rate of change in LD with distance can be quite complex and dependent on direction from the core (fig. 1 and 2).

We investigated temporal change in the frequency of the *L1014F* allele and associated haplotype in sympatric populations of M-form individuals in a subset of the populations

previously described by (Yawson et al. 2004). Using the data reported in Yawson et al (2004) we estimated the *L1014F* allele frequency in M form populations from around Accra, southern Ghana ( $\approx 30\text{km}$  diameter collection area) during 2002 ( $\text{freq}_{L1014F}=0.03$ ; 95% CI 0.01-0.05). Additional screening in 2007 and 2008 from the same greater Accra regions revealed that within five years this frequency had reached  $\text{freq}_{L1014F}=0.54$  (95% CI 0.49-0.60; (fig. 3). The data from years 2007 and 2008 are reported here for the first time. Phasing of the SNP genotypes of two M-form individuals with a wild type/*L1014F* genotype showed that the *L1014F*-associated haplotype was identical to that found in the S-form. This was confirmed by sequences obtained from seven M-form individuals collected from Accra, Ghana in 2008, which were homozygous for the *L1014F* allele (supplementary table 6, Supplementary Material online). Therefore the *L1014F* allele, which has increased in frequency in M-form populations is the same that has been putatively swept towards fixation in sympatric S-form populations. Introgression of *kdr* alleles between forms has been documented previously (Weill et al. 2000) and is unsurprising given that in southern Ghana there is a low but temporally stable level of interform matings (Yawson et al. 2004; Yawson et al. 2007).

Using a maximum likelihood estimation procedure with random starting values for selection coefficient ( $s$ ), dominance ( $h$ ) and initial allele frequency ( $p_0$ ) the parameter estimates converged to  $s=0.163$  (sd= 0.052);  $h= 0.249$  (sd= 0.142) and initial frequency  $p_0 = 0.025$  (sd= 0.008) (fig. 3).

## Discussion

These data show that there is marked LD around *kdr* mutations, loci exhibiting high penetrance, and, for *L1014F* at least, subject to strong recent positive selection. Despite similar median EHH levels, there were differences in the patterns of LD associated with the *L1014S* mutation in Kenya and Gabon. In Kenyan samples the rate of dissipation of LD around the *L1014S* core was quite rapid suggesting that the mutation has not been subject to as recent or as strong a selective sweep as the same mutation in Gabon (or indeed as the *L1014F* mutation in Ghana). This is as predicted if the serine resistance allele was primarily selected by the use of DDT in the latter part of the 20<sup>th</sup> century rather than by the more recent use of pyrethroids in agriculture and insecticide control programmes. In *Culex* mosquitoes the equivalent *L1014S* mutation gives low levels of *knockdown* resistance to pyrethroids compared to the *L1014F* mutation but confers high levels of DDT resistance (Martinez-

Torres et al. 1998; Ranson et al. 2000). Stump *et al* (2004) investigated the change in allele frequency of the *L1014S* allele before and after the commencement of a large-scale ITN project in Asembo Bay Western Kenya, the site of our collections (Stump et al. 2004). The frequency of the *L1014S* allele in the region approximately ten years before bednet introduction was approximately 0.04 (95% CI 0.02-0.08). In 2002, fifteen years after this initial survey and five years after the introduction of nets the frequency of the *L1014S* allele had increased, non-significantly to only 0.075 (95% CI 0.05-0.12). This suggests that there is little selective advantage for this mutation in the present environment, although it should be noted that in a neighbouring district in Uganda a recent study reported that the *L1014S* mutation was at a frequency of 0.85 (95% CI 0.83-0.87) (Ramphul et al. 2009). An alternative explanation would be that in Uganda there is an epistatic interaction between *L1014S* and some, as yet unidentified locus, which may affect the selection, and indeed dominance coefficients, and thereby result in a higher *L1014S* frequency. The high frequency and marked LD associated with *L1014S* in Gabon may be a result of the co-occurrence in genotypes, though not haplotypes (supplementary table 5, Supplementary Material online), with *L1014F*. A recent study from Cameroon showed that whilst *L1014F/L1014S* heterozygotes were significantly less resistant to permethrin than *L1014F* homozygotes, *L1014F/L1014S* heterozygotes were significantly more resistant to all insecticides tested than *L1014F/L1014*-wild type heterozygotes (Reimer et al. 2008). Repetitive mutation at the *1014* locus could, at least in part, be responsible for the patterns of LD around the *kdr* locus in the Gabonese data. Indeed, there is evidence for repeated mutations of *kdr* alleles across the species range of *Anopheles gambiae* (Pinto et al. 2007). However, we argue that on the recent time-scales on which *kdr* has arisen and spread that it is more parsimonious to assume that recombination is the dominant influence on patterns of LD rather than postulating high rates of repetitive mutation.

Whilst *kdr* is the best-documented resistance mechanism in *An. gambiae* there are many other resistance-associated loci. Microarray and recombinant protein expression work has shown that resistant mosquitoes over express a small number of enzymes that catalyze insecticide degradation (Ortelli et al. 2003; Müller, Donnelly, Ranson 2007; Chiu et al. 2008; Müller et al. 2008b). Linkage disequilibrium-based screens could be a powerful way of identifying regions of the genome carrying the scars of recent selection that regulate such overexpression. However, whether association mapping approaches will effectively identify genes subject to much older and comparatively weaker selection is currently unclear. The

bounded estimate of the selection coefficient reported here is at the upper limit of estimates generated to date and of a similar magnitude to estimates generated for resistance alleles in the mosquito *Culex pipiens* (Labbe et al. 2009). In human populations mutations associated with resistance to malaria infection such as G6PD and sickle cell trait have coefficients of selection of 0.02-0.05 (Tishkoff, Williams 2002) and 0.05-0.18 (Li 1975) respectively. In the third actor in the malaria transmission cycle of *Plasmodium falciparum* a selection coefficient of 0.1 has been obtained for the locus *dhfr* that confers resistance to the chemotherapeutic agent, pyrimethamine (Nair et al. 2003).

Together with strong and recent positive selection, the major determinant of LD around selected loci will be the rate of recombination. Indications of dramatic variation in the recombination rate across the *An. gambiae* genome have already been reported (Pombi et al. 2006; Black IV et al. 2008) and it is possible that, being close to the centromere of chromosome 2L the sodium channel locus is in an area of reduced recombination. However, our Kenyan data are consistent with rates of recombination sufficient to reduce the region hitch-hiked with a selectively advantageous locus in a relatively short period of time. Indeed detection of the signatures of selection for loci with low selection coefficients will be more logistically challenging in *An. gambiae* than humans because of much lower background levels of LD (Weetman D, Wilding CS, Steen K, Donnelly MJ unpublished). We attempted to amplify microsatellites from around the sodium channel to fully define the extent of the swept region as has been done for drug resistance loci in *Plasmodium falciparum* (Wootton et al. 2002; Nair et al. 2003). However, the sodium channel is situated in a region with an abundance of repetitive sequences and it was not possible to identify unique locus-specific microsatellite primer pairs.

Given the apparently high selection pressure on the *L1014F* mutation, it is curious that there are no studies, with adequate sample size, that have observed either of the *kdr* alleles at fixation (Santolamazza et al. 2008). One explanation would be that of over-dominance; however, insecticide bioassays studies suggest this is unlikely to be the case (Chandre et al. 2000; Reimer et al. 2008) and our estimate of the dominance coefficient shows the *kdr* *L1014F* allele to be partially recessive. Therefore it is likely that there is some fitness cost to the *L1014F* allele, and that this could be attributable to heterogeneity in exposure to pyrethroids in the environment or a consequence of an Hill-Robertson effect where selection

at a *kdr* locus can interfere with the selection at nearby beneficial mutations (Hill, Robertson 1966).

The data presented herein show that it is possible to detect genomic signatures of strong positive selection in pest species with large effective population size and generally low levels of linkage disequilibrium. We suggest that such approaches are likely to be extremely powerful in many non-model taxa subject to similar selective events,

### **Supplementary Materials.**

Supplementary tables are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

### **Funding.**

This work was supported by the National Institutes for Health [U01 AI58542 to Dr Edward Walker, Michigan State University] the Innovative Vector Control Consortium and the Royal Society.

### **Acknowledgements.**

We thank Nelson Cuamba, Eveline Klinkenberg and Pie Muller for providing wild caught specimens and the associate editor for constructive comments.

### **Literature cited**

- Aminetzach, YT, JM Macpherson, DA Petrov. 2005. Pesticide resistance via transposiiton-mediated adaptive gene truncation in *Drosophila*. *Science* 309:764-767.
- Ballinger-Crabtree, ME, WC Black IV, BR Miller. 1992. Use of genetic polymorphisms detected by the random-amplified polymorphic DNA polymerase chain reaction (RAPD-PCR) for differentiation and identification of *Aedes aegypti* subspecies and populations. *American Journal of Tropical Medicine and Hygiene* 47:893-901.
- Bass, C, D Nikou, MJ Donnelly, MS Williamson, H Ranson, A Ball, J Vontas, LM Field. 2007. Detection of knockdown resistance (*kdr*) mutations in *Anopheles gambiae*: a comparison of two new high-throughput assays with existing methods. *Malaria Journal* 6:e111.
- Begun, D, C Aquadro. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356:519-520.
- Begun, DJ, AK Holloway, K Stevens, et al. 2007. Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology* 5:2534-2559.



- Bersaglieri, T, PC Sabeti, N Patterson, T Vanderploeg, SF Schaffner, JA Drake, M Rhodes, DE Reich, JNW Hirschhorn. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics* 74:1111-1120.
- Black IV, WC, N Gorrochetegui-Escalante, NP Randle, MJ Donnelly. 2008. The yin and yang of linkage disequilibrium: Mapping of genes and nucleotides conferring insecticide resistance in insect disease vectors. *Transgenesis and the Management of Vector-Borne Disease* 627:71-83.
- Chandre, F, F Darriet, S Duchon, L Finot, S Manguin, P Carnevale, P Guillet. 2000. Modifications of pyrethroid effects associated with *kdr* mutation in *Anopheles gambiae*. *Medical and Veterinary Entomology* 14:81-88.
- Chiu, T, Z Wen, S Rupasinghe, M Schuler. 2008. Comparative molecular modeling of *Anopheles gambiae* CYP6Z1, a mosquito P450 capable of metabolizing DDT. *Proceedings of the National Academy of Sciences of the United States of America* 105:8855-8860.
- Davies, TGE, LM Field, PNR Usherwood, MS Williamson. 2007a. A comparative study of voltage-gated sodium channels in the Insecta: implications for pyrethroid resistance in Anopheline and other Neopteran species. *Insect Molecular Biology* 16:361-375.
- Davies, TGE, LM Field, PNR Usherwood, MS Williamson. 2007b. DDT, pyrethrins, pyrethroids and insect sodium channels. *IUBMB Life* 59:151-162.
- Diabate, A, C Brengues, T Baldet, et al. 2004. The spread of the Leu-Phe *kdr* mutation through *Anopheles gambiae* complex in Burkina Faso: genetic introgression and *de novo* phenomena. *Tropical Medicine and International Health* 9:1267-1273.
- Fanello, C, F Santolamazza, AI della Torre. 2002. Simultaneous identification of species and molecular forms of the *Anopheles gambiae* complex by PCR-RFLP. *Medical and Veterinary Entomology* 16:461-464.
- Hall, TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98.
- Hill, WG, A Robertson. 1966. The effect of linkage on limits to artificial selection. *Genetical Research* 8:269-294.
- Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6:65-70.
- Kim, Y, W Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160:765-777.
- Labbe, P, N Sidos, M Raymond, T Lenormand. 2009. Resistance gene replacement in the mosquito *Culex pipiens*: fitness estimation from long-term cline series. *Genetics* 182:303-312.
- Lehmann, T, WA Hawley, H Grebert, FH Collins. 1998. The effective population size of *Anopheles gambiae* in Kenya: implications for population structure. *Molecular Biology and Evolution* 15:264-276.
- Li, W. 1975. The first arrival time and mean age of a deleterious mutant gene in a finite population. *American Journal of Human Genetics* 27:274-286.
- Livak, KJ. 1984. Organization and mapping of a sequence on the *Drosophila melanogaster* X and Y chromosomes that is transcribed during spermatogenesis. *Genetics* 107:611-634.
- Lynd, A, H Ranson, PJ McCall, NP Randle, WC Black, ED Walker, MJI Donnelly. 2005. A simplified high-throughput method for pyrethroid knock-down resistance (*kdr*) detection in *Anopheles gambiae*. *Malaria Journal* 4:e16.
- Martinez-Torres, D, F Chandre, MS Williamson, F Darriet, JB Berge, AL Devonshire, P Guillet, N Pasteur, D Pauron. 1998. Molecular characterization of pyrethroid

- knockdown resistance (*kdr*) in the major malaria vector *Anopheles gambiae* s.s. *Insect Molecular Biology* 7:179-184.
- Maynard-Smith, J. 1998. *Evolutionary Genetics*. Oxford: Oxford University Press.
- Müller, P, M Chouaibou, P Pignatelli, J Etang, ED Walker, MJ Donnelly, F Simard, HW Ranson. 2008a. Pyrethroid tolerance is associated with elevated expression of antioxidants and agricultural practice in *Anopheles arabiensis* sampled from an area of cotton fields in Northern Cameroon. *Molecular Ecology* 17:1145-1155.
- Müller, P, MJ Donnelly, H Ranson. 2007. Transcription profiling of a recently colonised pyrethroid resistant *Anopheles gambiae* strain from Ghana. *BMC Genomics* 8:e36.
- Müller, P, E Warr, BJ Stevenson, et al. 2008b. Field-caught permethrin-resistant *Anopheles gambiae* overexpress CYP6P3, a P450 that metabolises pyrethroids. *PLoS Genetics* 4:e1000286.
- Nair, S, JT Williams, A Brockman, et al. 2003. A selective sweep driven by pyrimethamine treatment in southeast asian malaria parasites. *Molecular Biology and Evolution* 20:1526-1536.
- Newcombe, RG. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 17:857-872.
- Ortelli, F, LC Rossiter, J Vontas, H Ranson, J Hemingway. 2003. Heterologous expression of four glutathione transferase genes genetically linked to a major insecticide-resistance locus from the malaria vector *Anopheles gambiae*. *Biochemical Journal* 373:957-963.
- Pinto, J, A Lynd, N Elissa, MJ Donnelly, C Costa, G Gentile, A Caccone, VEI Do Rosario. 2006. Co-occurrence of East and West African *kdr* mutations suggests high levels of resistance to pyrethroid insecticides in *Anopheles gambiae* from Libreville, Gabon. *Medical and Veterinary Entomology* 20:27-32.
- Pinto, J, A Lynd, JL Vicente, et al. 2007. Origins and distribution of knockdown resistance mutations in the afrotropical mosquito vector *Anopheles gambiae*. *PLoS ONE* 11:e1243.
- Pombi, M, AD Stump, A Della Torre, NJ Besansky. 2006. Variation in recombination rate across the X chromosome of *Anopheles gambiae*. *American Journal of Tropical Medicine and Hygiene* 75:901-903.
- Ramphul, U, T Boase, C Bass, LM Okedi, MJ Donnelly, P Müller. 2009. Insecticide resistance and its association with target-site mutations in natural populations of *Anopheles gambiae* from eastern Uganda. *Transactions of the Royal Society of Tropical Medicine and Hygiene* in press.
- Ranson, H, B Jensen, JM Vulule, X Wang, J Hemingway, FHI Collins. 2000. Identification of a point mutation in the voltage-gated sodium channel gene of Kenyan *Anopheles gambiae* associated with resistance to DDT and pyrethroids. *Insect Molecular Biology* 9:491-497.
- Reimer, L, E Fondjo, S Patchoke, et al. 2008. Relationship between *kdr* mutation and resistance to pyrethroid and DDT insecticides in natural populations of *Anopheles gambiae*. *Journal of Medical Entomology* 45:260-266.
- Sabeti, PC, DE Reich, JM Higgins, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832-837.
- Sabeti, PC, SF Schaffner, B Fry, J Lohmueller, P Varilly, O Shamovsky, A Palma, TS Mikkelsen, D Altshuler, ESW Lander. 2006. Positive natural selection in the human lineage. *Science* 312:1614-1620.
- Sabeti, PC, P Varilly, B Fry, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913-918.

- Santolamazza, F, M Calzetta, J Etang, et al. 2008. Distribution of knock-down resistance mutations in *Anopheles gambiae* molecular forms in west and west-central Africa Malaria Journal 7:e74.
- Schlenke, TA, DJ Begun. 2004. Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. Proceedings of the National Academy of Sciences 101:1626-1631.
- Scott, JA, WG Brogdon, FH Collins. 1993. Identification of single specimens of the *Anopheles gambiae* complex by the Polymerase Chain Reaction. American Journal of Tropical Medicine and Hygiene 49:520-529.
- Stephens, M, P Scheet. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. American Journal of Human Genetics 76:449-462.
- Stephens, M, NJ Smith, P Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics 68:978-989.
- Stump, AD, FK Atieli, JM Vulule, NJ Besansky. 2004. Dynamics of the pyrethroid knockdown resistance allele in western Kenyan populations of *Anopheles gambiae* in response to insecticide-treated bed net trials. American Journal of Tropical Medicine and Hygiene 70:591-596.
- Tishkoff, SA, SM Williams. 2002. Genetic analysis of African populations: human evolution and complex disease. Nature Reviews Genetics 3:611-621.
- Weill, M, F Chandre, C Brengues, S Manguin, M Akogbeto, N Pasteur, P Guillet, M Raymond. 2000. The *kdr* mutation occurs in the Mopti form of *Anopheles gambiae* s.s. through introgression. Insect Molecular Biology 9:451-455.
- Wilding, CS, D Weetman, K Steen, MJ Donnelly. 2009. High, clustered, nucleotide diversity in the genome of *Anopheles gambiae* revealed by SNP discovery through pooled-template sequencing: implications for high-throughput genotyping protocols. BMC Genomics 10:e320.
- Wootton, JC, XR Feng, MT Ferdig, RA Cooper, JB Mu, DI Baruch, AJ Magill, XZI Su. 2002. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. Nature 418:320-323.
- Yawson, AE, PJ McCall, MD Wilson, MJ Donnelly. 2004. Species abundance and insecticide resistance of *Anopheles gambiae* in selected areas of Ghana and Burkina Faso. Medical and Veterinary Entomology 18:372-377.
- Yawson, AE, D Weetman, MD Wilson, MJ Donnelly. 2007. Ecological zones rather than molecular forms predict genetic differentiation in the malaria vector *Anopheles gambiae* s.s. in Ghana. Genetics 175:751-761.

### **Table legends**

Table 1.

Origin and *kdr* genotype of specimens used in the study.

Table 2.

Comparison of median EHH levels between alleles at the *kdr* loci.

Population	Year Collected	Total N	Form	Number of each <i>kdr</i> genotype					
				<i>L1014S/</i>	<i>L1014F/</i>	<i>L1014F/</i>	<i>L1014S/wt</i>	<i>L1014F/wt</i>	Wt/wt
				<i>L1014S</i>	<i>L1014F</i>	<i>L1014S</i>			
Asembo Bay, Kenya, 00°10'S, 34°22'E	2005 <sup>1</sup>	48	S	11	-	-	17	-	20
Dienga, Gabon, 01°52'S, 12°40'E	'99-2000 <sup>2</sup>	30	S	-	-	-	4	2	24
Bakoumba, Gabon, 01°49'S, 13°01'E	'99-2000 <sup>2</sup>	42	S	-	5	8	5	7	17
Libreville, Gabon, 00°22'N, 09°26'E	'99-2000 <sup>2</sup>	73	S	34	8	31	-	-	-
Okyereko & Accra area, Ghana, 05° 24.9N 00° 36.6W, 05°38'N, 00°15'E	2002 <sup>3</sup>	35	S	-	33	-	-	2	-
Okyereko, Ghana, 05° 24.9N 00° 36.6W	2002 <sup>3</sup>	30	M	-	-	-	-	2	28

The population name and total numbers of each DNA sample utilized. Molecular form is indicated and the numbers of each *kdr* genotype shown. Additional information on the collection sites may be obtained from the publications where the specimens are originally described <sup>1</sup>(Müller et al. 2008a) <sup>2</sup>(Pinto et al. 2006) <sup>3</sup>(Yawson et al. 2004)

	<u>Kenya</u> <u>L1014S</u> (S form)	<u>Kenya</u> <u>Wildtype</u> (S form)	<u>Gabon</u> <u>L1014S</u> (S form)	<u>Gabon</u> <u>Wildtype</u> (S form)	<u>Gabon</u> <u>L1014F</u> (S form)	<u>Ghana</u> <u>Wildtype</u> (M form)
<b>Kenya Wildtype</b> (S form)	<u>0.0001</u>					
<b>Gabon L1014S</b> (S form)	0.26 NS	<b>0.0001</b>				
<b>Gabon_Wildtype</b> (S form)	<u>0.0001</u>	0.026NS	<u>0.0001</u>			
<b>Gabon L1014F</b> (S form)	<b>0.0001</b>	<b>0.0001</b>	1.00NS	<b>0.0001</b>		
<b>Ghana Wildtype</b> (M form)	<u>0.005</u>	1.00NS	<u>0.005</u>	0.86NS	<u>0.005</u>	
<b>Ghana L1014F</b> (S form)	<b>0.0005</b>	<b>0.0001</b>	<b>0.0001</b>	<b>0.0001</b>	0.04NS	<b>0.0003</b>

Probabilities from sign tests are shown. The values followed by NS were not significant after sequential Bonferroni corrections. Values which are underlined indicate that the EHH values were significantly higher for the sample given in the column heading; values which are in bold indicate that the EHH values were significantly higher for the sample given in the row heading;.

## Figure legends

FIG 1. Bifurcation plots showing patterns of recombination in the centromeric (5' towards the left) and telomeric (3' towards the right) directions. The core is marked by the dark circle and each of the 29 SNPs is represented by a node and a recombination event is represented by a bifurcation. The diameter of the circle at each SNP node is proportional to the numbers of individuals with the same long range haplotype at that position. No bifurcation plot is shown for the *L1014F* core in Ghanaian M-form populations as only a single haplotype was observed (see results)

FIG 2. Extended Haplotype Homozygosity analysis showing linkage disequilibrium decay with increasing distance from the core (marked as the origin on the X-axis). 95% confidence intervals estimated by bootstrapping (see methods). The X-axis is ordinal, negative numbers are in the centromeric direction, positive numbers in the telomeric direction. The scale bar at the top of the figure is 72.6 kb in length and shows the physical distance between the SNPs. Panel A Kenya data for *L1014S* and wildtype alleles; Panel B Gabon data for *L1014S*, *L1014F* and wildtype alleles and Panel C Ghana data for *L1014F* (S-form) and wildtype (M-form)

FIG 3. Observed and predicted changes in *L1014F* allele frequency in the *Anopheles gambiae* M- form populations from southern Ghana. Observed data obtained from surveys conducted in 2002, 2006 and 2007. First collection point (Generation 1) was June 2002. Data from 2002, first three data points, are taken from Yawson et al. 2004 all other data are novel. One generation per month is assumed following (Lehmann et al. 1998). 95% confidence intervals for each observed data point were calculated according to (Newcombe 1998). Expected data generated from simultaneous maximum-likelihood estimates of initial frequency and selection and dominance coefficients (see methods).



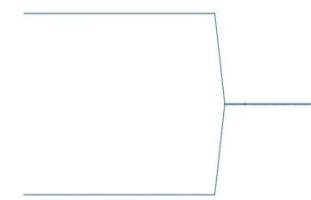
L1014S

L1014F

Wildtype

Ghana  
S-form

Not present  
in this  
population



Ghana  
M-form

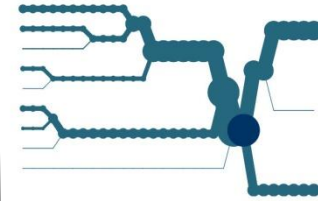
Not present  
in this  
population

Low frequency  
in this  
population



Kenya  
S-form

Not present  
in this  
population



Gabon  
S-form

