Inferring selection in the Anopheles gambiae species complex: an example from immune-related serine protease inhibitors

Malaria Journal

BioMed Central

Research

Open Access

Inferring selection in the *Anopheles gambiae* species complex: an example from immune-related serine protease inhibitors

Darren J Obbard*, John J Welch and Tom J Little

Address: Institute of Evolutionary Biology, University of Edinburgh, Kings Buildings, West Mains Road, Edinburgh, UK

Email: Darren J Obbard* - darren.obbard@ed.ac.uk; John J Welch - j.j.welch@ed.ac.uk; Tom J Little - tom.little@ed.ac.uk

* Corresponding author

Published: 4 June 2009

Malaria Journal 2009, **8**:117 doi:10.1186/1475-2875-8-117

Received: 19 February 2009

Accepted: 4 June 2009

This article is available from: http://www.malariajournal.com/content/8/1/117

Abstract

Background: Mosquitoes of the *Anopheles gambiae* species complex are the primary vectors of human malaria in sub-Saharan Africa. Many host genes have been shown to affect *Plasmodium* development in the mosquito, and so are expected to engage in an evolutionary arms race with the pathogen. However, there is little conclusive evidence that any of these mosquito genes evolve rapidly, or show other signatures of adaptive evolution.

Methods: Three serine protease inhibitors have previously been identified as candidate immune system genes mediating mosquito-Plasmodium interaction, and serine protease inhibitors have been identified as hot-spots of adaptive evolution in other taxa. Population-genetic tests for selection, including a recent multi-gene extension of the McDonald-Kreitman test, were applied to 16 serine protease inhibitors and 16 other genes sampled from the *An. gambiae* species complex in both East and West Africa.

Results: Serine protease inhibitors were found to show a marginally significant trend towards higher levels of amino acid diversity than other genes, and display extensive genetic structuring associated with the 2La chromosomal inversion. However, although serpins are candidate targets for strong parasite-mediated selection, no evidence was found for rapid adaptive evolution in these genes.

Conclusion: It is well known that phylogenetic and population history in the *An. gambiae* complex can present special problems for the application of standard population-genetic tests for selection, and this may explain the failure of this study to detect selection acting on serine protease inhibitors. The pitfalls of uncritically applying these tests in this species complex are highlighted, and the future prospects for detecting selection acting on the *An. gambiae* genome are discussed.

Background

By vectoring *Plasmodium* parasites, *Anopheles* mosquitoes are a central component of the Malaria crisis. Consequently, there has been a substantial effort to identify the genes involved in the mosquito immune response against *Plasmodium*, including studies to identify genes associated with variation in vector competence [1-4]. It has been widely hypothesized that these immune response genes may be subject to strong parasite-mediated selection, such as that which occurs in a coevolutionary 'arms-race' [5,6]. Such arms-races involve strong reciprocally-antagonistic selection, leading to the frequent and rapid fixation of

new alleles. This reduces within-species diversity, while driving between-species protein divergence, and leaves a genomic signature of past selection that can be identified through DNA sequence analysis [7,8]. Thus, DNA sequence analysis and the tools of population genetics can augment understanding of immune gene function in host-parasite interaction by identifying genes that are the target of parasite adaptation, and even distinguish between forms of parasite-mediated selection [5,6,9].

Population genetic methods have previously shed light on the nature and intensity of selection in both mammalian and *Drosophila* immune systems. For example, *Drosophila* studies have suggested that pathogens which manipulate signal transduction pathways or the antiviral RNAi pathway have been a major selective force [10,11]. In *Anopheles* mosquitoes, the potential for immune-related genes to determine vector competence provides a clear incentive to elucidate the selective forces that drive evolution. Serine protease inhibitors (serpins, or SRPNs) are prime candidates for such parasite-mediated selection in *Anopheles* mosquitoes. Serpins comprise a large and rapidly evolving super-family of proteins (reviewed in [12,13]) with key roles in the immune systems of vertebrates [14] and invertebrates [15]. In particular, *Drosophila* serpins, such as Nec and SRPN27A, modulate two of the most important defense pathways: the Toll-pathway [16,17], and the melanization cascade [18,19], and many are up-regulated on septic injury (Spn28D, SRPN27A, Spn5, CG6687 and Spn4, see [20]). Moreover, some *Drosophila* serpins display very high rates of amino acid substitution, and/or other signatures of adaptive evolution, e.g. [21-23].

Three *Anopheles* serpins have been experimentally associated with *Plasmodium*-interaction phenotypes (see Table 1). In *Anopheles gambiae* and *Anopheles stephensi* SRPN10 is expressed in the mosquito midgut and in haemocytes [24], and during *Plasmodium berghei* (a rodent parasite) invasion of the midgut epithelium SRPN10 moves from the nucleus to the cytoplasm, and its expression is strongly induced [25]. SRPN6 is also expressed in infected midgut epithelial cells and in haemocytes, and again its expression is strongly induced by *P. berghei* invasion in both *An. gambiae* and *An. stephensi*. The expression of SRPN6 is also

Table 1: Locus Details and location

NAME	Identifier	Putative function	Genomic location
SRPN1	AGAP006909	Inhibitory Serine Protease inhibitor	2L:39892128-39893864
SRPN2	AGAP006911	Plasmodium-related Inhibitory Serine Protease inhibitor	2L:39897002-39899744
SRPN3	AGAP006910	Inhibitory Serine Protease inhibitor	2L:39895229-39896338
SRPN4C	AGAP009670	Inhibitory Serine Protease inhibitor	3R:38145527-38154288
SRPN5	AGAP009221	Inhibitory Serine Protease inhibitor	3R:28858000-28859778
SRPN6	AGAP009212	Plasmodium-related Inhibitory Serine Protease inhibitor	3R:28811997-28818217
SRPN7	AGAP007693	Inhibitory Serine Protease inhibitor	2L:49090665-49091915
SRPN8	AGAP003194	Inhibitory Serine Protease inhibitor	2R:33744972-33746720
SRPN9	AGAP003139	Inhibitory Serine Protease inhibitor	2R:33148444-33154607
SRPN10	AGAP005246	Plasmodium-related Inhibitory Serine Protease inhibitor	2L:12996143-13001508
SRPN11	AGAP001377	Non-inhibitory Serine Protease inhibitor	2R:4017728-4019706
SRPN12	AGAP001375	Non-inhibitory Serine Protease inhibitor	2R:4010431-4012512
SRPN14	AGAP007692	Non-inhibitory Serine Protease inhibitor	2L:49084812-49086463
SRPN16	AGAP009213	Inhibitory Serine Protease inhibitor	3R:28824548-28826209
SRPN17	AGAP001376	Inhibitory Serine Protease inhibitor	2R:4015617-4016537
SRPN18	AGAP007691	Non-inhibitory Serine Protease inhibitor	2L:49086842-49088278
Control1	AGAP006906	Adenosine deaminase-related growth factor	2L:39852471-39854636
Control2	AGAP006904	Matrix metalloproteinase	2L:39831595-39836700
Control3	AGAP006918	Putative NADH:ubiquinone dehydrogenase	2L:39995907-39997095
Control4	AGAP009673	glutaminyl-peptide cyclotransferase	3R:38248845-38249780
Control5	ENSANGG8091	(retrotransposon)	3R:28965847-28968579
Control6	AGAP009207	Mitogen-activated protein kinase ERK	3R:28697030-28708787
Control7	AGAP007712	Putative RHO guanyl-nucleotide exchange factor	2L:49181235-49190516
Control8	AGAP003205	Similar to Drosophila CG8468	2R:33825401-33827998
Control9	AGAP003143	Similar to Drosophila CG9904	2R:33211906-33213476
Control10	AGAP005247	no annotation	2L:13062962-13067750
Control11	AGAP001384	cAMP-dependent protein kinase, beta-catalytic subunit	2R:4098545-4103634
Control12	AGAP001371	Similar to Drosophila CG18643	2R:3885127-3885956
Control14	AGAP007713	Similar to human solute carrier family 39	2L:49196817-49198177
Control16	AGAP900209	DNA-directed RNA polymerase II subunit J	3R:28746535-28747425
Control17	AGAP001388	Similar to human mab-3-related transcription factor 3	2R:4120810-4122586
Control18	AGAP007717	Similar to Drosophila CAP CG18408-PE	2L:49212258-49224889

induced by the human parasite *Plasmodium falciparum* [26]. RNAi knockdown of SRPN6 in *An. stephensi* resulted in a significant increase in the number of developing *P. berghei* oocysts, and although knockdown had no effect on oocyst numbers in susceptible strains of *An. gambiae*, in a resistant strain, the number of melanized *P. berghei* ookinetes was significantly increased [26]. More recently it has also been shown that SRPN6 is induced in the salivary glands of *An. gambiae* in response to *P. berghei* sporozoite invasion, and knock-down of SRPN6 by RNAi significantly increases the number of sprozoites reaching the salivary glands [27]. Finally, knockdown of SRPN2 in *P. berghei*-susceptible *An. gambiae* has a broadly opposite effect, resulting in a 97% decrease in oocyst formation through increased lysis and melanization, following midgut invasion [28].

Here, population-genetic approaches are used to search for evidence of natural selection acting on 16 serpin genes in the *An. gambiae* species complex, including those implicated in immune function. First, by comparing serpins to other nearby genes, patterns of genetic diversity within and between populations of *An. gambiae*, *Anopheles arabiensis*, and *Anopheles melas* are used to identify loci that deviate strongly from neutral predictions. Second, a recent extension of the McDonald-Kreitman test is used to test for evidence of adaptive substitution between species [29,30]. The data are then discussed in terms of on-going population processes in the *An. gambiae* complex, many of which have important implications for the robust inference of selection.

Serpins were found to have slightly higher levels of amino acid diversity than other genes, consistent with either reduced constraint, or potentially with balancing selection. In common with previous analyses. [31], considerable structuring of genetic diversity in the SRPN1-2-3 cluster was found in association with the 2La chromosomal inversion. However, although serpins are good *a priori* candidates as targets for strong 'arms-race' selection, as with similar studies on other *Anopheles gambiae* immune-related genes (e.g. [6,32-34]), the tests for adaptive evolution presented here are largely inconclusive. The results show how standard population-genetic tests for selection may be difficult to apply in the *An. gambiae* species complex; this is due to for both demographic and phylogenetic factors that are already widely known, and further supported by the present data.

Methods

Samples

Anopheles gambiae individuals were collected from West Africa ('BK': Burkina Faso, Koubri village, 12°11'54 N; 1°23'43W) and East Africa ('KY': Kenya, Mbita, Suba District). *Anopheles arabiensis* individuals were also collected from West Africa ('BK', in the same collections as *An. gambiae*, above) and East Africa ('TZ', Tanzania, Ifakara). All *An. gambiae* and *An. arabiensis* used in this study were provided by H. M. Ferguson (University of Glasgow, UK). *Anopheles melas* individuals were collected from Coastal Ghana (Ghana, Essiama, 4°57.4 N; 2°24.1W) by N. Tuno (Institute of Tropical Medicine, Nagasaki University, Japan). *Anopheles merus* and *Anopheles quadriannulatus* species *A* (hereafter *An. quadriannulatus*) were both obtained from laboratory colonies, maintained by the Medical Research Council of South Africa (provided by R. Maharaj; MRC, Durban, South Africa) and the University of Wageningen (Strain 'Sangqua', Zimbabwe, provided by W. Takken), respectively.

The species identity of all *gambiae* complex members was verified by diagnostic PCR [35], and the *M* and *S* molecular forms of *An. gambiae* were distinguished by PCR-RFLP [e.g. [36]]. As expected from their known geographic distributions [37], all KY *An. gambiae* individuals were *S* form, and all but two of the BK *An. gambiae* sample were *M*-form. All *An. gambiae* individuals were also surveyed for 2La/+ chromosomal inversion status, using the PCR assay of White et al [38], derived from the sequenced breakpoints [39]. As reported previously, in addition to the expected diagnostic 207 bp and 492 bp fragment lengths, these primers were found to amplify fragments of lengths *ca*. 687 bp, 672 bp, 760 bp and 1020 bp in some individuals [32]. Direct sequencing of these fragments from a subset of individuals suggest they are insertion/deletion derivatives of expected assay products [32,40], and within the polymorphic KY population the 2La/2L+^{a} amplification fragments were in Hardy-Weinberg equilibrium (51 individuals, $\chi^2 = 0.44$, 1df., $p = 0.51$), allowing us to tentatively assign 2La/2L+^{a} inversion status to all individuals [32].

Loci

Thirty-two loci were selected for sequencing and analysis, including 16 of the 18 serpins currently identified in the *An. gambiae* genome (M. Kanost and K. Michel, pers. comm.), and 16 other protein-coding loci chosen to match the genomic position of the serpins without regard to function. SRPN19 (a non-inhibitory serpin with 1:1:1 orthologs in *Drosophila melanogaster*, *Aedes aegypti* and *An. gambiae*) and SRPN13, which does not appear in the current *An. gambiae* assembly, were not sequenced. The total sequenced length was ~19 Kbp of coding sequence per individual (i.e. approximately 600 bp from each locus; range 240 bp–800 bp). Not all loci were sequenced from the same individuals within populations, and not all loci were amplified from *An. melas*. A full summary of gene names, locations, and classification is presented in Table 1.

The 'control' loci should represent an unbiased sample of *Anopheles* genes, to which serpins can be compared. Because these control genes are position-matched, each lying ~90 Kbp (range 40–125 Kbp) from a 'partner' serpin, they should control for the effects of large-scale position-based variation in recombination and mutation rates. Note that improvements to the *An. gambiae* annotation have subsequently identified control locus 5 (previously annotated as ENSANGG000008091) as deriving from a retrotransposon (AgamP3.4, July, 2007).

PCR and sequencing

Genomic DNA was extracted from single mosquitoes using DNeasy kits (QIAgen). PCR primers were designed from the published *An. gambiae* genome sequence [41], and the final primer sequences selected after trouble-shooting (sequences are given in Additional file 1). Only one PCR amplicon was used per locus, thus sequences do not represent entire genes. Following PCR, unincorporated primers and dNTPs were removed using exonuclease I (New England BioLabs) and shrimp alkaline phosphatase (Amersham). PCR products were sequenced in both directions using BigDye™ reagents (v3.1, Applied BioSystems) and an ABI capillary sequencer. In some amplicons, indel polymorphism required the use of additional sequencing primers. The sequence chromatograms were assembled using SeqManII (DNAstar Inc., Madison USA) then inspected by eye to confirm the validity of all differences within and between species and all heterozygous base-calls. The heterozygous sequence from each diploid individual was decomposed into two pseudohaplotypes for analysis using PHASE [42,43]. However, the presented analyses should be highly robust to any errors in phase assignment, as only explicitly tree-based results, such as Hudson's nearest neighbour statistic (Snn), are affected by allelic phase. All unphased sequences have been submitted to GenBank as population sets, using ambiguity codes to indicate heterozygous sites. Sequence accession numbers span the range GQ146469–GQ148534.

Divergence, diversity and differentiation

The number of synonymous and non-synonymous polymorphisms and substitutions, and the average pairwise genetic diversity (π) were calculated using DnaSP [version 4.50.3, ref [44]]. Diversity was calculated separately for synonymous (π_s) and non-synonymous (π_a) sites, and used a Jukes-Cantor correction for multiple substitutions, as implemented in DnaSP. Departures from the allele frequency spectrum expected under the standard neutral model were quantified using Tajima's *D* statistic [45], also calculated using DnaSP. Tajima's *D* (which measures departures from the expected allele frequency distribution under a standard neutral model) was calculated using synonymous sites only, and was calculated separately for both populations of *An. gambiae* and *An. arabiensis* (but not for *An. melas*, where sample sizes were too small to give meaningful results). The significance of departures from the expected allele frequency distribution were tested using 10,000 rounds of coalescent simulation (as implemented in DnaSP) conditional on the number of segregating sites and conservatively assuming no recombination within loci.

Genetic differentiation between populations was quantified in DnaSP using Hudson's K_{ST} statistic [equations 7 to 9 in reference 46] which is calculated from the average number of pairwise differences between sequences taken within populations and between all populations, and is identical to Nei's γ_{ST} [47] except for the population-size weighting scheme [see [46]]. Significant departures from zero population differentiation were inferred by permuting sequences between populations to create a null distribution of K_{ST} values. All non-parametric statistical tests on diversity and differentiation were performed using the R statistical language (R Development Core Team, 2008 http://www.R-project.org). Although non-parametric tests (Spearman's rank correlation, paired Wilcoxon tests) are presented below, except where noted explicitly, parametric equivalents (Pearson's correlation, paired t-tests) gave qualitatively identical results.

The proportion of adaptive substitutions

If it is assumed that synonymous mutations are effectively neutral, and that the fixation or loss of selected amino-acid variants is so rapid that the vast majority of non-synonymous polymorphisms are also effectively neutral, then the relative numbers of polymorphisms (*P*, within species) and fixed differences (*D*, between species) at synonymous and non-synonymous sites can be used to identify the action of selection [see [7] for an introduction]. This forms the basis of the McDonald-Kreitman test [MK test, [29]], which seeks to detect a departure from independence in a simple 2 × 2 contingency table of polymorphisms (P_N and P_S) and fixed differences (D_N and D_S). For a single gene, the departure from neutrality can easily be quantified by summary statistic such as the neutrality index (N.I.= $(P_N/P_S)/(D_N/D_S)$ [48]), or by the estimated proportion of adaptive substitutions ($\alpha = 1 - (D_S P_N)/(D_N P_S) = 1$ - N.I., [49]). This approach can be extended to multiple genes using D_S, P_N, D_N and P_S averaged across genes [49], or using a more sophisticated maximum-likelihood estimator of α, such as that of Bierne and Eyre-Walker [50].

Here, an extension of the maximum-likelihood method of Welch [30] was used. This method is very closely related to that of Bierne and Eyre-Walker [50], but additionally allows for the possibility that some apparent fixed differences may actually be polymorphisms that only appear

fixed due to small sample size [51]. The method was extended to include polymorphism values from two species simultaneously e.g. [29,51]. Using this approach, models were fitted in which expected neutral divergence, $\lambda = \mu t$, took a single value at all loci; expected neutral diversity, $\theta = 4N_e\mu$, was also shared between all loci, but free to vary between species; and selective constraint, *f*, was free to vary between loci [see [30] for other details of the model]. Three nested models were fitted, in which (1) α was constrained to zero at all loci, i.e. no adaptive evolution, (2) a single α was shared by all genes, and (3) α was free to differ between serpins and 'control' genes. In this way, it was possible to test both for evidence of adaptive evolution, and whether serpins have a different rate of adaptive evolution to other genes. Model fit was tested using both likelihood ratio tests and Akaike weighting (derived from the Akaike information criterion) [52]. Confidence intervals on α were obtained by adjusting α away from its maximum likelihood value, and allowing the other parameters to take their maximum likelihood value, conditional on that α, until log likelihood decreased by 2 units [50]. C code to fit these models is available on request from the authors, or from [53].

Results

Synonymous site diversity

Across all 32 loci, average pairwise genetic diversity at synonymous sites (π_s) was highest in *An. gambiae* (π_s = 2.82%, 95% bootstrap interval 2.25–3.41%), lowest in *An. melas* (π_s = 0.86%, 0.53–1.19%), and intermediate in *An. arabiensis* (π_s = 1.97%, [1.50, 2.46]; Figure 1a). The difference in π_s between *An. gambiae* and *An. arabiensis* was highly significant (Paired Wilcox test using 32 loci, V = 444, p = 0.0004), and π_s correlated strongly between these species (Spearman's rho = 0.72, S = 1504, $p < 6 \times 10^{-6}$, Additional file 2). Diversity in *An. gambiae* and *An. arabiensis* did not correlate with diversity in *An. melas* ($p > 0.5$ in both cases, 26 loci). For a full summary of synonymous diversity, and all other summary stats that follow, see Additional file 3.

For An. gambiae, genetic diversity was slightly higher in West Africa (πs = 2.74%; BK M-form) than in East Africa (πs = 2.17%; KY S-form), and although the effect was small, it was statistically significant (Paired Wilcox V = 380, p = 0.03, Figure 1a). Diversity was also highly correlated between East- and West-African populations (Spearman's rho = 0.72, S = 1520, p = 3 × 10-6). Although only two S-form individuals were sampled from West Africa, they displayed higher diversity than the either of the other two An. gambiae populations (BK S-form; πs = 2.98%). For An. arabiensis, diversity correlated even more strongly between East and West Africa (rho = 0.87, S = 729, p < 2 × 10-10) and did not differ significantly between the populations (πs = 1.96% in BK vs. πs = 1.79% in TZ, Paired Wilcoxon test V = 316, p = 0.19).

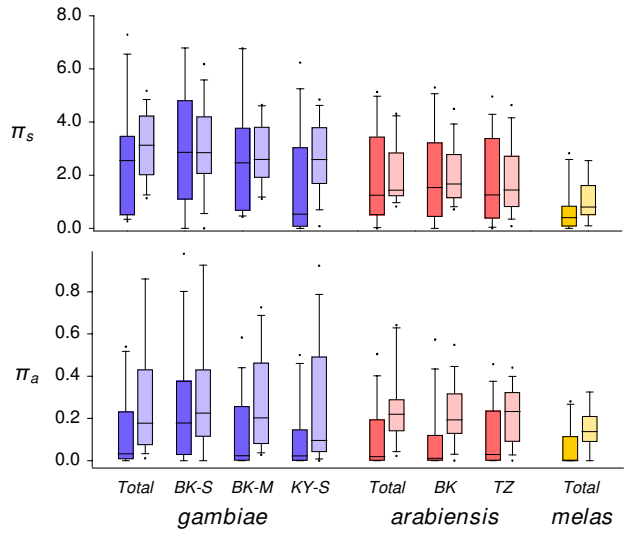

Figure 1

Genetic diversity at synonymous and non-synonymous sites. Genetic diversity (the percentage of sites that differ on average between haplotypes) at synonymous (π_s) and nonsynonymous (π_a) sites measured at 32 loci in populations of *An. gambiae*, *An. arabiensis* and *An. melas*. Diversity is shown separately for control loci (dark bars) and serpins (pale bars), and is shown for the species as a whole, and for each population separately. Note that although only two individuals (4 haplotypes) were sampled for S-form *An. gambiae* in population BK, ~19 Kbp of sequence will provide a good estimate of π if mating within the population is random. Diversity was significantly higher in *An. gambiae* than in *An. arabiensis*, and significantly lower in *An. melas*. For non-synonymous sites, serpins had significantly higher diversity than control loci, but this trend was non-significant at synonymous sites. See main text for details, and Additional File 1 for the raw data.

Synonymous site diversity did not differ significantly between serpins and other genes in either *An. gambiae* (3.10% vs. 2.53%, Paired Wilcoxon V = 48, p = 0.32, Figure 1a) or *An. arabiensis* (2.00% vs. 1.93%, Paired Wilcox V = 60, p = 0.71). Although position-matched, no correlation in diversity could be detected between serpins and their corresponding control genes ($p > 0.1$ in both cases), suggesting that the effect of genomic location on neutral diversity was relatively weak.

Non-synonymous site diversity

Non-synonymous diversity (π_a) was very similar between *An. gambiae* (π_a = 0.22%; 95% bootstrap interval 0.13–0.31%) and *An. arabiensis* (π_a = 0.18%; bootstrap interval: 0.12–0.24%, Figure 1b), and did not differ significantly between the species (Paired Wilcoxon test V = 285, p = 0.29). Although π_a did not correlate significantly with π_s in either *An. gambiae* or *An. arabiensis* (correlation coeffi-

cients were 0.17 and 0.07 respectively, $p > 0.3$ in both cases), there was a strong correlation in π_a between the two species (rho = 0.81, S = 1037, $p = 2 \times 10^{-8}$, Additional file 2). East and West African populations did not differ significantly in π_a for either *An. gambiae* or *An. arabiensis* ($p > 0.1$ in both cases), but π_a did correlate very highly between East and West African populations (rho = 0.80 $p = 6. \times 10^{-8}$, and rho = 0.88, $p = 2 \times 10^{-11}$, respectively). Interestingly, π_a was higher for serpins than for other genes in both *An. gambiae* and *An. arabiensis* (π_a = 0.30% vs. 0.13%, and π_a = 0.25% vs. 0.11%), although statistical significance was marginal for *An. gambiae* (Paired Wilcoxon V = 30 $p = 0.051$, and V = 24, $p = 0.024$ for *An. gambiae* and *An. arabiensis* respectively, Figure 1b). Despite fewer loci being sequenced, this effect could also be detected in *An. melas* (π_a = 0.15% vs. 0.06%, unpaired Wilcoxon test W = 25, $p = 0.034$, Figure 1b).

Allele frequency spectra

Tajima's *D* statistic for synonymous sites was negative in both populations of *An. arabiensis* (*D* = -0.29 and *D* = -0.38, averages across loci in TZ and BK respectively) and in *An. gambiae* BK (*D* = -0.71, *M*-form individuals only), and did not differ between serpins and other genes (Wilcoxon tests, $p > 0.5$ in all cases). Tajima's *D* did not differ significantly between the two populations of *An. arabiensis* (Paired Wilcoxon V = 193, $p = 0.42$), but did correlate between the populations (rho = 0.43 S = 2543.131, $p = 0.017$). In population BK, Tajima's *D* was correlated between *An. gambiae* and *An. arabiensis* (rho = 0.490 S = 2293, $p = 0.006$), but was significantly more negative in *An. gambiae* (Paired Wilcoxon V = 115, $p = 0.015$). Strikingly, Tajima's *D* was generally positive in *An. gambiae* population KY (mean across loci 0.77, 22 out of the 28 genes with non-zero diversity had $D > 0$, with overall 95% bootstrap interval [0.40, 1.15]). This was significantly higher than in BK (Paired Wilcoxon test V = 15, $p = 1 \times 10^{-6}$).

In *An. arabiensis*, five genes had individually significantly negative Tajima's D statistics ($p < 0.05$ in all cases, no correction for multiple tests): control loci 1 (BK and TZ) and 5 (BK), and serpins 10 (BK), 6 and 7 (TZ). In *An. gambiae* population BK (*M*-form only) 7 genes had significantly negative *D* values: serpins 7, 9 and 14, and control loci 5, 6, 10 and 11 ($p < 0.05$ in all cases).

Genetic differentiation between populations

In *An. arabiensis*, differentiation between East and West Africa was very low (K_{ST} = 0.08) and not significantly different from zero at 15 of the 32 loci examined. In *An. gambiae*, differentiation between East and West Africa was much higher (K_{ST} = 0.14; all loci except control locus 9 were individually significantly differentiated) and this difference between the species was statistically significant (paired Wilcoxon test V = 392, $p = 0.016$). Note that this differentiation reflects not only geographic separation, but also differentiation between *M* and *S* molecular forms of *An. gambiae*. In West Africa (BK) differentiation between *M*- and *S*-form was very low (K_{ST} = 0.016), and significantly lower than differentiation between *S*-form sampled from East Africa and *S*-form sampled from West Africa (K_{ST} = 0.073, paired Wilcoxon test V = 442, $p = 0.0005$). Only two S-form *An. gambiae* individuals were sampled from BK, making estimates of differentiation potentially poor and reducing the power of the test. However, assuming random mating, the estimates should not be biased by the small number of individuals sampled, and the large number of loci (32, ~19 Kbp of sequence) will reduce sampling error. For an overview of genetic differentiation see Figure 2.

The 2La chromosomal inversion

Anopheles gambiae population KY was highly polymorphic for the 2La/+ chromosomal inversion, allowing us to test for differentiation between the two inversion states. Dividing population KY into two groups on the basis of on inversion status (2La homozygotes *versus* $2L+^a$ homozygotes, heterozygotes excluded) identified very strong population structure associated with the inversion. At the six loci sampled from within the inversion (serpins 1–3 and control loci 1–3) mean differentiation between inversion-groups was K_{ST} = 0.25 (all six loci were signifi-

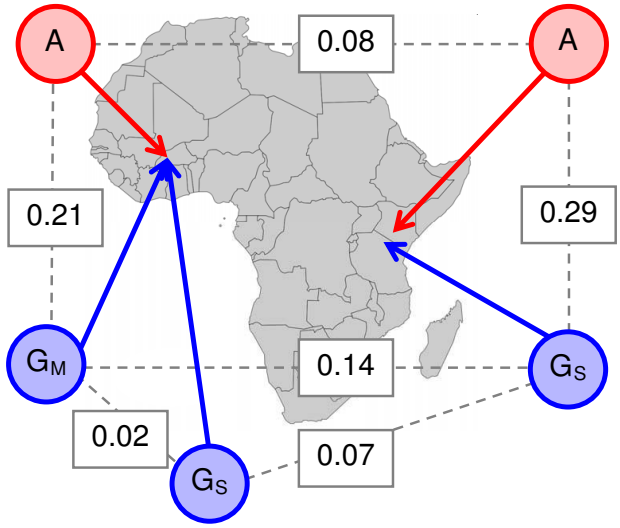

Figure 2
Genetic differentiation between populations. Arrows indicate approximate sampling locations within Africa, and letters identify species (A- *An. arabiensis*, G- *An. gambiae*, M-form and S-form). Dashed lines indicate pairs of populations for which genetic differentiation was calculated, and numbers are K_{ST} statistics, averaged across all 32 loci.

cantly differentiated: permutation $p < 0.01$ for each locus). Across the six other polymorphic loci sequenced from chromosome arm 2L (serpins 7, 10, 14 and 18, control loci 10 and 18), differentiation between these groupings was K_{ST} = 0.03, and was not significantly different from zero in any locus ($p > 0.14$ by permutation, in each locus).

Divergence and differentiation between species

Genetic divergence (substitutions per site) between *An. gambiae* and *An. arabiensis* was very low: K_S = 3.5% (averaged across loci), and when corrected for diversity (i.e. $K_s - \overline{\pi}_s$, page 220 in [54]), K_S = 1.2%. No synonymous fixed differences were identified between these species, and only three non-synonymous fixed differences (all in control locus 5, derived from a transposable element). Divergence from *An. melas* was higher for both *An. gambiae* (K_S = 6.4%, corrected K_S = 4.6%) and *An. arabiensis* (K_S = 6.3%, corrected K_S = 4.9%), as was divergence from *An. merus* (uncorrected K_S = 5.3% and K_S = 4.8%, respectively); divergence between these species and *An. quadriannulatus* was intermediate (uncorrected): K_S = 4.01% and K_S = 4.30% respectively. For an overview of interspecies divergence and an illustration of gene trees versus species trees, see Figure 3.

Genetic differentiation (the proportion of total diversity attributable to between-species differences) between *An. gambiae* and *An. arabiensis* species was K_{ST} = 0.19 (95% bootstrap interval across loci [0.16, 0.24]). Differentiation between *An. gambiae* and *An. melas* was much higher (K_{ST} = 0.33 [0.27, 0.40]), as was differentiation between *An. arabiensis* and *An. melas* (K_{ST} = 0.50 [0.44, 0.57]). Differentiation between *An. gambiae* and *An. arabiensis* was lower for serpins than for control loci (K_{ST} = 0.14 vs. K_{ST} = 0.25, paired Wilcox test V = 113, p = 0.02). However, although the trend was in the same direction in East and West Africa, this effect was only significant in West Africa (V = 108, p = 0.04, as compared to V = 99, p = 0.12).

Inference of adaptive substitutions

Due to the lack of fixed differences between *An. gambiae*, *An. arabiensis* and *An. quadriannulatus*, McDonald-Kreitman based approaches were not applied to these data (see Discussion). However, it was possible to apply the method of Welch [30] using *An. gambiae*/*An. arabiensis*, and their divergence from *An. melas* and/or *An. merus*. According to the likelihood ratio test, neither analysis using *An. melas* provided any evidence supporting adaptive substitutions between the lineages (there was no significant improvement in model fit between fixing $\alpha = 0$ and allowing α to take its maximum likelihood value; $p >$ 0.19 in all cases: Table 2). This was also true for the test

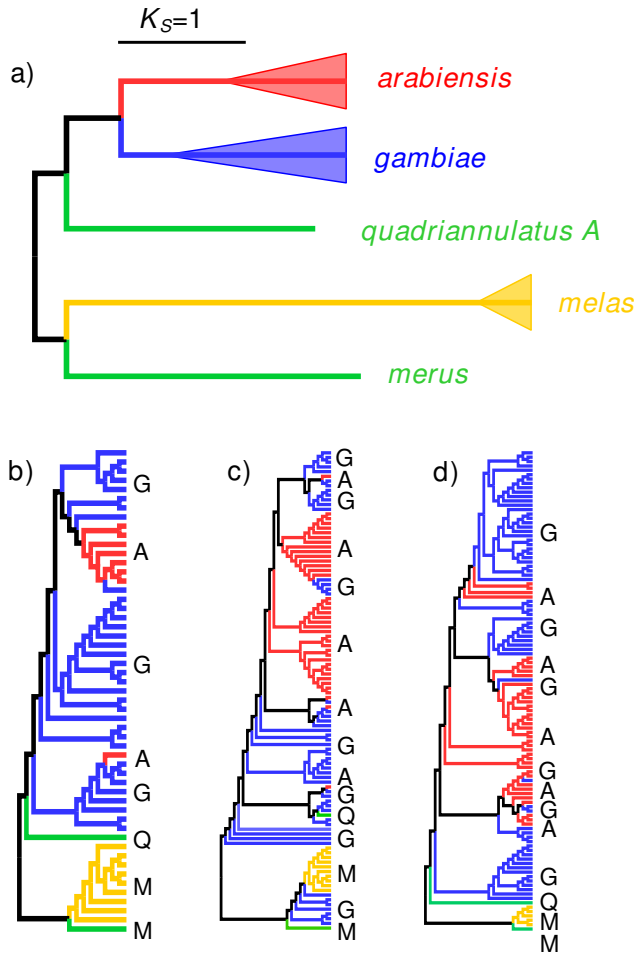

Figure 3
Genetic divergence within the *gambiae* species complex. (a) An un-rooted neighbour-joining tree, calculated from pairwise K_S between species averaged across loci. Branch lengths are to scale. The filled triangles illustrate the relative scale of diversity and divergence within the complex, such that the length of the triangle is half the divergence between haplotypes within species (i.e $\pi_s/2$) and net divergence (K_S-π_s) corresponds to branch-lengths that are not part of the triangle. (Note that population samples and thus π_s were not available for *An. quadriannulatus A* and *An. merus*). (b)-(d) Neighbour-joining cladograms (i.e. topology only, branch-lengths uninformative) showing the unique alleles sequenced from three loci. Note that in all cases *An. gambiae* and *An. arabiensis* alleles are intermixed. (b) to (d) are control locus 14, SRPN7 and SRPN11, selected to illustrate a wide range of K_{ST} values between *An. gambiae* and *An. arabiensis* (K_{ST} = 0.51, 0.10 and 0.09, respectively).

using *An. gambiae* and *An. merus* ($p > 0.2$: Table 2). For the test using *An. arabiensis* and *An. merus* there was some evidence that α was significantly greater than zero, (α = 0.34 [0.08, 0.53]; $2\Delta\ln L$ = 6.16, 1d.f., p = 0.013, significance lost if a correction is made for multiple tests), but no significant improvement in model fit was obtained by allow-

Table 2: Estimates of the proportion of adaptive substitutions

α model	Par	log(*L*)	2Δlog(*L*)	χ^2 *p*-value	AIC	Akaike weight	α_a	α_b
An. gambiae vs. An. melas								
α = 0	29	-336.49			730.98	**0.424**	[0]	[0]
α ~ (all loci)	30	-335.76	1.46	0.23	731.52	0.323	0.23	[0.23]
α ~ (control, serpin)	31	-335.66	0.19	0.91	733.33	0.131	0.18	0.25
α ~ (other, immune)	31	-335.73	0.06	0.97	733.46	0.123	0.24	0.07
An. arabiensis vs. An. melas								
α = 0	29	-329.36			716.73	**0.495**	[0]	[0]
α ~ (all loci)	30	-329.31	0.10	0.75	718.63	0.191	-0.05	[-0.05]
α ~ (control, serpin)	31	-328.07	2.48	0.29	718.15	0.243	-0.54	0.10
α ~ (other, immune)	31	-329.30	0.02	0.99	720.61	0.071	-0.03	-0.15
An. gambiae vs. An. merus								
α = 0	34	-333.36			734.72	**0.349**	[0]	[0]
α ~ (all loci)	35	-332.60	1.52	0.22	735.21	0.274	0.15	[0.15]
α ~ (control, serpin)	36	-331.59	2.02	0.36	735.18	0.277	0.34	0.03
α ~ (other, immune)	36	-332.60	0.00	1.00	737.21	0.101	0.15	0.17
An. arabiensis vs. An. merus								
α = 0	34	-311.04			690.07	0.053	[0]	[0]
α ~ (all loci)	35	-307.96	6.15	0.01	685.92	**0.422**	0.34	[0.34]
α ~ (control, serpin)	36	-307.22	1.48	0.48	686.44	0.325	0.47	0.25
α ~ (other, immune)	36	-307.71	0.50	0.78	687.42	0.200	0.36	0.10

α_a and α_b are estimates of the proportion of adaptive substitutions in each of the two classes of gene (control/serpin or non-immune/immune, respectively), Par is the number of parameters in the model. Where the value of α is constrained by the model it is marked in square brackets. Negative values arise from an 'excess' of non-synonymous polymorphism, and could represent sampling error, or mildly deleterious polymorphisms [7]. AIC is the Akaike Information Criterion. The Akaike weighting can be interpreted as the weight of evidence in favour of the corresponding model, given the relative support for all the available models [52].

ing α to differ between serpins and control loci, or between Plasmodium-related serpins and all other loci. This suggests that approximately 8–53% of amino-acid substitutions between *An. arabiensis* and *An. merus* were adaptive, but this value did not differ significantly between serpins and the control loci. Analysis of Akaike weights gives a qualitatively identical result: in each species-pair no model is strongly preferred, but in the *merus-arabiensis* comparison, α = 0 receives relatively little weight. Full results of the McDonald-Kreitman analysis are presented in table 2, and raw data are given in Additional file 4.

Discussion

Population history and speciation in Anopheles

The *An. gambiae* complex falls within the *Pyretophorus* series of the subgenus *Cellia*, and comprises a closely-related group of approximately eight species (*An. gambiae s.s., An. arabiensis, Anopheles bwambae, An. quadriannulatus A and B, An. merus, An. melas and Anopheles comorensis*) [55-57], plus at least one case of incipient speciation (*M* and *S* molecular forms of *An. gambiae* s.s. [37,58-60]). Because lineages within the complex differ in their importance as *Plasmodium* vectors [e.g., [61]], in their ecological preferences [62,63], and in their resistance to pesticides [e.g. [63]], there is considerable value in understanding both species relationships and how populations are structured. This may be of particular consequence if any attempt is ever made to genetically modify wild *Anopheles* populations to block or reduce *Plasmodium* transmission [64,65]. However, in addition to having important implications for vector control, as discussed below, understanding phylogeny and population history are also essential to the robust inference of selection.

Previous analyses suggest that *An. gambiae* and *An. arabiensis* are sister taxa, and the data presented here from five of the eight species are in strong agreement, placing *An. gambiae* and *An. arabiensis* as the most closely related species pair, with greater divergence to *An. merus* and *An. melas* (Figure 3a; note that the tree is unrooted). However, in common with previous studies [e.g. [33,66,67]], these data suggest extensive shared polymorphism (Figure 3) and very low differentiation between *An. gambiae* and *An. arabiensis* (K_{ST} = 0.19). The inter-species differentiation between *An. gambiae* and *An. arabiensis* is approximately the same as inter-population differentiation between African and European *D. melanogaster* (K_{ST} = 0.16 to 0.24, depending on population; pers. comm. P. R. Haddrill, data from [68]), and is lower than inter-population differ-

entiation in the predominantly selfing nematode *Caenorhabditis elegans* (K_{ST} = 0.38, pers. com. A. D. Cutter, data from [69]). Thus, although *An. gambiae* and *An. arabiensis* are largely reproductively isolated and significantly differentiated [70,71], these data confirm that either they share extensive ancestral polymorphism, or that there is considerable introgression between them [see also [67,72,73]]. This is further supported by the very high correlation in neutral diversity across genes, between these two species (Additional file 2).

Given a particular divergence time, effective population size is the primary determinant of the amount of shared ancestral polymorphism between taxa, because drift (and therefore lineage-sorting) is faster in small populations [e.g. [74]]. Thus, although it is likely that *An. gambiae* and *An. arabiensis* share a more recent common ancestor that either does with *An. melas*, the lower differentiation between *An. gambiae* and *An. arabiensis* could also be explained (at least in part) by differences in effective population size within the complex. Since π_s is an estimator of $4N_e\mu$, differences in neutral diversity imply that *An. gambiae* and *An. arabiensis* have larger effective population sizes than *An. merus*, consistent both with their wider geographic range, and with potentially higher levels of shared ancestral polymorphism (see above: π_s~2.8%, 2.0%, and 0.9% for *An. gambiae*, *An. arabiensis* and *An. melas*, respectively). Diversity in *An. gambiae* and *An. arabiensis* is similar to that seen in African populations of *Drosophila simulans* and *D. melanogaster* (π_s = 3.2% and π_s = 1.7%, respectively) [e.g. [75]], and assuming the mutation rate (μ) is similar between mosquitoes and *Drosophila*, this also suggests a long-term effective population size for *An. gambiae* that is about 70% larger than *D. melanogaster* [76], i.e. well in excess of 1 million. This is broadly consistent with previous estimates from mitochondrial sequence, but is much larger than estimates based on microsatellites and allozyme variants [reviewed in [77]].

Incipient speciation between the *M* and *S* molecular forms of *An. gambiae* is a major focus of ongoing research [37,58-60], culminating in the recently completed sequencing of the *M* and *S* molecular-form genomes [78]. Although differentiation between the *M* and *S* form of *An. gambiae* in West Africa is extremely low (K_{ST} = 0.02; Figure 2), it is consistently non-zero at some loci, even where the lineages are sympatric (see e.g. [6]). This unambiguously identifies *M* and *S* form *An. gambiae* as being (at least partly) reproductively isolated [58], and it has been argued that different *M* and *S*-form niches may be distinguishable [37]. Moreover, differentiation is variable around the genome, being higher at so-called 'islands of speciation', potentially associated with adaptive differences [58,59]. However, other studies have shown that in some geographic regions microclimate is a better predictor of population divergence than is molecular form [79], and it is clear that the S-form of *An. gambiae* is not a single homogenous lineage [66,80]. Indeed, it is well-established that there is extensive genetic differentiation associated with the Great Rift Valley [80], and consistent with this, the data presented here not only identify considerable differentiation between East and West Africa (KY S-form *versus* BK M-form, K_{ST} = 0.12), but also show that West African *M*-form and *S*-form are less differentiated from each other than either is from East African S-form (Figure 2). This suggests either that the West African *M* and *S* lineages share a more recent common ancestor, or alternatively that they have experienced considerable recent gene flow [[37], e.g. [80], but see also [81]]. For *An. arabiensis*, differentiation across the width of the continent is only K_{ST} = 0.08, which is very similar to that between Eastern and Western *S*-form *An. gambiae* (K_{ST} = 0.07; Figure 2), and approximately twice that between *D. melanogaster* populations sampled across a similar geographic range (Gabon vs. Kenya or Zimbabwe, K_{ST} = 0.04) [P. R. Haddrill pers. comm., data from [68]].

Most populations showed a slight skew in the allele-frequency spectrum toward low-frequency variants (i.e., average Tajima's *D* was negative), consistent either with population growth, or with weak selection against some synonymous variants. In contrast, however, the data presented here also identify a strong skew toward intermediate frequency alleles in *An. gambiae* population KY (i.e., a positive Tajima's *D*, 0.77 averaged across loci, 95% bootstrap interval 0.40–1.15). One potential explanation for this is that population KY is admixed or contains cryptic population structure, for example as would be the case if two divergent lineages of *S*-form are coexisting there. Alternatively, a positive Tajima's *D* could also result from a recent decrease in effective population size.

The phylogenetic and phylogeographic complexity of the *An. gambiae* species group means that inferences drawn from single individuals should be treated with caution, as the low differentiation between species and high diversity within species means that any one individual is not necessarily typical or representative [e.g. [66]]. For example, the recently sequenced *M* and *S*-form genomes [78] were both obtained from mosquitoes sampled in Mali, but *S*-form divergence between East and West Africa is considerably greater than *M*-*S* divergence within West Africa (see Figure 2 above, and [66,80], cf. [81]). It is therefore not clear that any conclusions regarding *M* - *S* genome divergence will generalize to S-form individuals from the east African coast. The same issue arises with inter-species comparisons. For example, the divergence between two randomly sampled *An. gambiae* genomes is K_S~3%, and that between one randomly selected *An. gambiae* genome and one randomly selected *An. quadriannulatus* genome is

only K_S~4% (above, and Figure 3). This means that a single inbred strain of *An. quadriannulatus* is only marginally more informative about *An. quadriannulatus* than it is about *An. gambiae*, and without more extensive sampling it cannot reliably be used to identify genetic differences between the species [cf. [61]]. These issues will be of paramount importance in analysing the recently approved complete genomes of *An. arabiensis*, *An. quadriannulatus* and *An. merus* [82].

Evidence for adaptive evolution in Anopheles *serpins*

There is considerable evidence from other taxa that serpins are an evolutionarily dynamic gene family, with high turnover between lineages and occasional lineage-specific expansions (e.g. [23], see [13] for a review). For example, although most *Anopheles* and *Aedes* serpins have 1:1 orthologs, there are 29 serpins in *D. melanogaster* but only 18 in *An. gambiae*, and very few mosquito serpins have 1:1 orthologs in *Drosophila* [83]. Some serpins also show very high rates of adaptive evolution, such as *Drosophila* Spn28D [CG7219 in ref. [21]], and it has been suggested that, in general, rapid turnover and strong selection in serpins may be associated with serpin immune function, and could be driven by an evolutionary 'arms-race' [13]. In *An. gambiae*, three serpins are known to have immune-related function in response to *Plasmodium* infection (SRPN2, SRPN6 and SRPN10 in [24,26,28]).

Strong selection can affect patterns of genetic diversity, both between populations, and between chromosomal inversions. In *An. gambiae* population KY, considerable differentiation was identified between 2La and 2L+^{a} homozygotes around the SRPN1, 2 and 3 cluster (and control loci 1, 2 and 3). For these loci K_{ST} = 0.25 between inversion states, which is actually higher than the overall differentiation between *An. gambiae* and *An. arabiensis*, and twice as high as the differentiation between *M*-form and *S*-form *An. gambiae*. No such differentiation was seen for other loci on chromosome arm 2L, indicating that this is strongly associated with the inversion. Although chromosomal inversions are in general expected to suppress recombination, especially near breakpoints, this is unlikely to lead to extreme or long-term differentiation unless maintained by selection [e.g. [84,85]]. In particular, despite recombination being suppressed in heterozygous individuals, genetic exchange (including recombination and gene-conversion) within the region of the 2La/+ inversion is not zero [86], and this should allow such differentiation to break down rapidly if it is not selectively maintained. The finding of elevated differentiation around these loci agrees with previous analyses, which found the SRPN1-3 cluster to be close to the region of highest differentiation between 2La and 2L+^{a}. [31]. While this doesn't provide strong evidence that any of these serpins are being directly selected, it is interesting to note that SRPN2 is required for successful infection of *An. gambiae* by *P. berghei* [28], and that 2La inversion-status was identified with vector-competence in some early studies [87,88]. Thus it is possible that genetic structuring in serpins 1–3, introduced and maintained by the 2La inversion, may affect variation in vector competence, even if the underlying cause of 2La/+ differentiation is elsewhere.

In contrast, using a McDonald-Kreitman based approach to detecting adaptive substitutions [29,30], no strong or consistent evidence of selection could be detected, nor could differences in the rate of adaptive evolution between serpins and other genes (or between immune serpins and other genes; Table 2). Specifically, in three of the four species pairings that were analysed, the rate of adaptive evolution could not be distinguished from zero (Table 2). This may indicate that neither *Anopheles* immune-related serpins, nor *Anopheles* serpins as a family, are subject to selection for rapid change, and consequently that all selection acting on these genes is purifying. However, it may also be an artefactual result arising either from limitations of the McDonald-Kreitman framework, or from issues specific to the gambiae species complex (see next section). In particular, the McDonald-Kreitman approach assumes that non-synonymous substitutions can be divided into three classes: strongly advantageous mutations that fix rapidly, strongly deleterious mutations that are rapidly lost, and effectively neutral mutations that drift in frequency [50]. If there is also a large class of weakly deleterious mutations that remain polymorphic for an extended period, but are lost by selection in the long term, then this will reduce estimates of α [50].

In agreement with this, a trend toward higher amino-acid diversity was found in serpins, as compared to other genes, in all three species (*An. gambiae*, *An. arabiensis* and *An. melas*; Figure 1b). This could suggest that purifying selection on serpins is weak or intermittent (as compared to purifying selection on other genes), or that there is selection favouring diversity in serpins, such as balancing selection [89]. However, if the latter were the case, then one might also expect to find an increase in diversity at linked synonymous sites, and although slightly higher, synonymous diversity in serpins did not differ significantly from other genes (Figure 1a), suggesting that they neither experience long-term selection for increased polymorphism, nor undergo more frequent selective sweeps [e.g. [8]]. Moreover, in no population or species did a serpin display the highest or lowest neutral genetic diversity, and there was no clear pattern in the allele frequency spectrum (as measured by Tajima's *D* statistic) or inter-population differentiation (measured by K_{ST}) that supports the notion of strong selection acting on *Anopheles* serpins. These data therefore fail to identify any serpins as candi-

dates for recent strong selection within the gambiae complex.

Prospects and pitfalls in inferring adaptive evolution in Anopheles

A null result in tests for selection may also result from low power or model violations associated with the phylogenetic and population history in the *An. gambiae* complex. Motivated by an interest in identifying targets of pathogen-mediated selection, several studies have now attempted to identify adaptive evolution in immune-related genes from the *Anopheles gambiae* species complex [5,6,32-34,90-92]. However, in stark contrast to almost identical studies in taxa such as *Drosophila* [for example [10,11,21,93-95]], to date there is very little clear evidence supporting adaptive evolution in the *An. gambiae* complex [but see [90,92]].

One likely reason for the difference between *Anopheles* and *Drosophila* studies is the dearth of suitable outgroups for *An. gambiae* [e.g. [32]]. First, well-studied *Anopheles* species outside of the gambiae complex appear to be too distantly related to reliably infer the divergence between them. For example, the *An. stephensi* SRPN6 cDNA sequence [26] suggests that K_S between with *An. stephensi* and *An. gambiae* is ~1.33 (± 0.157) substitutions per site when estimated by maximum likelihood [96], or 0.92 (± 0.117) by the method of Li [97]. Second, low divergence within the complex means there is little power to infer the substitution rate between them [6,32,33]. For example, across all 8 serpins appearing in the extensive survey of Cohuet *et al*. [6], there are only two fixed amino acid differences between *An. arabiensis* and *An. gambiae*. Similarly, the present study found no fixed differences at all in the same genes, probably because wider geographic sampling provided greater power to distinguish between polymorphisms and fixed differences. The stochastic errors involved here mean that estimates of substitution rate are likely to be wildly variable, reducing the power to estimate the fraction of adaptive substitutions using a McDonald-Kreitman framework (Figure 4).

In principle, other species from within the gambiae complex might be informative outgroups for *An. gambiae* and *An. arabiensis*: although divergence is small (1 to 5%; Figure 3a), this may be sufficient in other taxa, such as the human-chimp comparison [e.g. [98]]. However, unlike the human-chimp case, diversity in *Anopheles* is very high compared to humans (π_s~3% in *An. gambiae*, ~2% in *An. arabiensis*, Figure 1, c.f. ~0.1% in humans, e.g. [99]), and this leads to two potential problems. First, a large proportion of apparent substitutions will in reality be polymorphisms [30,51], and although this effect is small enough to be negligible for pairs of species in which $K_S >> \pi_S$, it becomes a concern when comparing *An. gambiae* to *An. arabiensis*, for which $K_S < \pi_S$ (Figure 3). However, as here, this effect can be accounted for using models which include the sample size and diversity, and thereby infer the 'true' number of substitutions [30,51]. Second, and potentially more serious, is the opportunity for extensive shared polymorphism. The McDonald-Kreitman framework uses information from current diversity (i.e. π_s and π_a) to infer whether some proportion of historic substitutions (K_A) cannot be explained by purely neutral processes. This model implicitly assumes a time period when the two lineages were diverging from their common ancestor, during which selection (that is to be detected) was able to act. If there is extensive shared polymorphism between the species, for example if gene trees are rarely reciprocally monophyletic (as is the case with *An. gambiae* and *An. arabiensis*; see Figure 3b–d and e.g. [33]), then it is hard to see how the McDonald-Kreitman approach can ever be usefully applied. In other words, unlike the straight-forward cases where divergence is too high (the branch is too long) or divergence is too low (the branch is

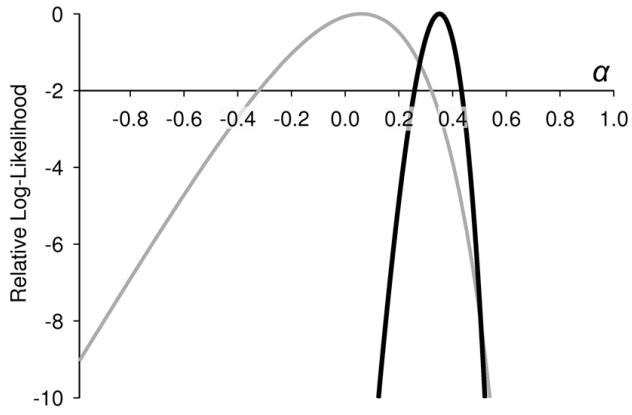

Figure 4

The power to estimate α using *An. gambiae* and *An. arabiensis*. The relative log-likelihood of α (the proportion of amino-acid substitutions that are adaptive) estimated using the modified McDonald-Kreitman approach [30]. The grey curve is calculated from all 102 genes for which both *An. arabiensis* and *An. gambiae* population samples were available in the dataset of Cohuet et al. [6]. The black curve shows an equivalent dataset of 102 genes from *Drosophila melanogaster* and *D. simulans*, with genes selected to be the same average length as those in the Cohuet dataset (D. J. Obbard, J. J. Welch and F. M. Jiggins, unpublished data). Despite both pairs of species having similar levels of diversity (π_s from 1.6% to 2.9%), for the *Anopheles* dataset the bounds (2 units of log Likelihood) stretch from -0.33 to 0.32 (and include zero) while for *Drosophila* the bounds only stretch from 0.26 to 0.44, and the maximum-likelihood estimate of α is 35%. The low precision in the second estimate reflects the very low power available due to the low divergence in *An. gambiae*-*An. arabiensis* comparisons

too short), for *An. gambiae* and *An. arabiensis*, in effect there is no branch at all (Figure 3b–d).

Unfortunately, for similar reasons, there are also serious concerns about the application of other selection-inference methods to the gambiae complex, such as the phylogenetic methods implemented in PAML [96] and HyPhy [100]. These methods use multiple sequences related by a gene (or species) tree to infer relative rates of synonymous and non-synonymous substitution, allowing variable rates at different sites or in different parts of the gene. Most phylogenetic methods assume there is no recombination within loci [but see OmegaMap, [101]], and simulation suggests false positives can reach >50% when $2N_e r > 0.01$ [102]. Because $2N_e r$ (i.e: 2 × effective population size × recombination rate per codon per generation) in *An. gambiae* is likely to be of the order 0.01 – 0.1 or higher – primarily due to the large effective population size – such phylogenetic approaches cannot be applied reliably to within-species *Anopheles* data.

Additionally, where the McDonald-Kreitman framework assumes that between-species K_A results from the joint action of selection and drift, and within-species π_a results only from drift, the phylogenetic approaches (as they are most commonly applied) assume either that all amino acid variants have the same cause (i.e. K_A and π_a do not provide independent information about selection and drift) or that most differences are fixed between species, (i.e. $K_A >> \pi_a$, such that π_a is negligible). Thus the phylogenetic and McDonald-Kreitman approaches constitute very different models that lend themselves to different datasets, and it is not clear that any single dataset can reasonably be analysed using both. Instead, to analyse a dataset that includes substantial within-species sampling using a phylogenetic approach, it may be more rational to fit a model which allows the relative rates of synonymous and non-synonymous substitution to differ between within- and between-species branches [e.g. [103]]. However, in the case of the *gambiae* complex, the low power associated with the low inter-species divergence will then be encountered again.

If McDonald-Kreitman tests have relatively low power within the gambiae complex, and phylogenetic methods cannot easily be applied to within-species data, how can selection be inferred from *Anopheles* population genetic data? One possibility is to use approaches based solely on within-population diversity to identify recent selective sweeps or regions of elevated polymorphism, and this will work for some loci [e.g. TEP1, [92]], though the possibility of introgression, and/or chromosomal inversions that might affect the distribution of diversity, should then be taken into account. Nevertheless, at present it seems the best options for outgroup-based analyses are *An. merus* and/or *An. melas*, and both the data presented here (Figure 3) and previous studies [e.g. [33,66]] suggest that their divergence from the *An. gambiae*/*An. arabiensis* clade should be sufficient in the case of genes evolving under very strong selection. However, the recently approved genome sequences from 13 more species of *Anopheles* mosquitoes [82] may hold the solution, and particularly the complete genome sequence of *Anopheles sundaicus* (also subgenus *Cellia*, *Pyretophorous* series).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DJO selected the loci, designed the primers, and performed all PCR, sequencing and analysis. JJW developed the likelihood-based methods for inferring the proportion of adaptive substitutions, and provided statistical support. TJL conceived the project, and all authors contributed to the final manuscript.

Additional material

Additional file 1

***PCR primers**. Locus names, identifiers, genomic locations and PCR primer sequences.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1475-2875-8-117-S1.doc]

Additional file 2

***Correlations in genetic diversity between species**. The correlation in genetic diversity for loci sampled from* Anopheles gambiae *and* Anopheles arabiensis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1475-2875-8-117-S2.doc]

Additional file 3

***Genetic diversity and differentiation**. Sheet 1: genetic diversity at synonymous and non-synonymous sites, and Tajima's D statistic for synonymous sites, for all loci and populations. Sheet 2: genetic differentiation (Fst, Kst and Snn) between populations for all loci.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1475-2875-8-117-S3.xls]

Additional file 4

***McDonald-Kreitman data**. Sample sizes, analysed gene lengths, polymorphisms and fixed differences for each locus, for each pair of species*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1475-2875-8-117-S4.xls]

Acknowledgements

We thank H. M. Ferguson and N. Tuno for providing field samples of *Anopheles* mosquitoes, R. Maharaj and W. Takken for providing preserved mos-

quitoes from laboratory culture, and G. Yan for ongoing logistical support with fieldwork. We thank Abraham Eappen and Marcelo Jacobs-Lorena for providing the cDNA sequence of *An. Stephensi* SRPN6, Asher Cutter and Penny Haddrill for unpublished K_{ST} statistics, and Mike Kanost and Kristin Michel for sharing pre-publication data on *Anopheles* serpin classification and function. We thank an anonymous reviewer for helpful comments on the manuscript, and Marcel Hommel for careful copyediting. This work was funded by Wellcome Trust Grant 073210 to TJL, and DJO also received Wellcome Trust funding in the form of 'Value In People' bridging salary and Wellcome Trust Research Career Development Fellowship 085064/Z/08/Z. JJW is funded by BBSRC grant DO17750 awarded to Andrew Rambaut.

References

1. Riehle M, Markianos K, Lambrechts L, Xia A, Sharakhov I, Koella J, Vernick K: **A major genetic locus controlling natural *Plasmodium falciparum* infection is shared by East and West African *Anopheles gambiae*.** *Malar J* 2007, **6:**87.
2. Dong Y, Aguilar R, Xi Z, Warr E, Mongin E, Dimopoulos G: ***Anopheles gambiae* immune responses to human and rodent *Plasmodium* parasite species.** *PLoS Pathog* 2006, **2:**e52.
3. Michel K, Kafatos FC: **Mosquito immunity against *Plasmodium*.** *Insect Bioch Mol Biol* 2005, **35:**677-689.
4. Osta MA, Christophides GK, Kafatos FC: **Effects of mosquito genes on *Plasmodium* development.** *Science* 2004, **303:**2030-2032.
5. Little TJ, Cobbe N: **The evolution of immune-related genes from disease carrying mosquitoes: diversity in a peptidoglycan- and a thioester-recognizing protein.** *Insect Mol Biol* 2005, **14:**599-605.
6. Cohuet A, Krishnakumar S, Simard F, Morlais I, Koutsos A, Fontenille D, Mindrinos M, Kafatos F: **SNP discovery and molecular evolution in *Anopheles gambiae*, with special emphasis on innate immune system.** *BMC Genomics* 2008, **9:**227.
7. Eyre-Walker A: **The genomic rate of adaptive evolution.** *Trends Ecol Evolution Trends Ecol Evol. 2006 Oct;21(10):569-75* 2006, **21(10):**569-575.
8. Nielsen R: **Molecular signatures of natural selection.** *Ann Rev Genetics* 2005, **39:**197-218.
9. Boëte C: ***Anopheles* mosquitoes: not just flying malaria vectors especially in the field.** *Trends Parasitol* 2009, **25:**53-55.
10. Obbard DJ, Jiggins FM, Halligan DL, Little TJ: **Natural selection drives extremely rapid evolution in antiviral RNAi genes.** *Current Biol* 2006, **16:**580-585.
11. Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG: **Dynamic evolution of the innate immune system in *Drosophila*.** *Nature Genetics* 2007, **39:**1461-1468.
12. Gettins PGW: **Serpin structure, mechanism, and function.** *Chem Rev* 2002, **102:**4751-4803.
13. Christeller JT: **Evolutionary mechanisms acting on proteinase inhibitor variability.** *FEBS J* 2005, **272:**5710-5722.
14. Mangan MSJ, Kaiserman D, Bird PI: **The role of serpins in vertebrate immunity.** *Tissue Antigens* 2008, **72:**1-10.
15. Jiravanichpaisal P, Lee BL, Söderhäll K: **Cell-mediated immunity in arthropods: Hematopoiesis, coagulation, melanization and opsonization.** *Immunobiology* 2006, **211:**213-236.
16. Levashina EA, Langley E, Green C, Gubb D, Ashburner M, Hoffmann JA, Reichhart JM: **Constitutive activation of toll-mediated antifungal defense in serpin-deficient *Drosophila*.** *Science* 1999, **285:**1917-1919.
17. Ligoxygakis P, Pelte N, Hoffmann JA, Reichhart JM: **Activation of *Drosophila* Toll during fungal infection by a blood serine protease.** *Science* 2002, **297:**114-116.
18. De Gregorio E, Han SJ, Lee WJ, Baek MJ, Osaki T, Kawabata SI, Lee BL, Iwanaga S, Lemaitre B, Brey PT: **An immune-responsive serpin regulates the melanization cascade in *Drosophila*.** *Developmental Cell* 2002, **3:**581-592.
19. Ligoxygakis P, Pelte N, Ji CY, Leclerc V, Duvic B, Belvin M, Jiang HB, Hoffmann JA, Reichhart JM: **A serpin mutant links Toll activation to melanization in the host defence of *Drosophila*.** *EMBO J* 2002, **21:**6330-6337.
20. De Gregorio E, Spellman PT, Rubin GM, Lemaitre B: **Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays.** *Proc Natl Acad Sci USA* 2001, **98:**12590-12595.
21. Jiggins FM, Kim KW: **A screen for immunity genes evolving under positive selection in *Drosophila*.** *J Evolut Biol* 2007, **20:**965-970.
22. Heger A, Ponting CP: **Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes.** *Genome Res* 2007. gr.6249707.
23. Borner S, Ragg H: **Functional diversification of a protease inhibitor gene in the genus *Drosophila* and its molecular basis.** *Gene* 2008, **415:**23-31.
24. Danielli A, Kafatos FC, Loukeris TG: **Cloning and characterization of four *Anopheles* gambiae serpin isoforms, differentially induced in the midgut by *Plasmodium berghei* invasion.** *J Biol Chem* 2003, **278:**4184-4193.
25. Danielli A, Barillas-Mury C, Kumar S, Kafatos FC, Loukeris TG: **Overexpression and altered nucleocytoplasmic distribution of *Anopheles* ovalbumin-like SRPN10 serpins in *Plasmodium*-infected midgut cells.** *Cell Microbiol* 2005, **7:**181-190.
26. Abraham EG, Pinto SB, Ghosh A, Vanlandingham DL, Budd A, Higgs S, Kafatos FC, Jacobs-Lorena M, Michel K: **An immune-responsive serpin, SRPN6, mediates mosquito defense against malaria parasites.** *Proc Natl Acad Sci USA* 2005, **102:**16327-16332.
27. Pinto SB, Kafatos FC, Michel K: **The parasite invasion marker SRPN6 reduces sporozoite numbers in salivary glands of *Anopheles gambiae*.** *Cell Microbiol* 2008, **10:**891-898.
28. Michel K, Budd A, Pinto S, Gibson TJ, Kafatos FC: ***Anopheles gambiae* SRPN2 facilitates midgut invasion by the malaria parasite *Plasmodium berghei*.** *EMBO Reports* 2005, **6:**891-897.
29. McDonald JH, Kreitman M: **Adaptive protein evolution at the adh locus in *Drosophila*.** *Nature* 1991, **351:**652-654.
30. Welch JJ: **Estimating the genomewide rate of adaptive protein evolution in *Drosophila*.** *Genetics* 2006, **173:**821-837.
31. White BJ, Hahn MW, Pombi M, Cassone BJ, Lobo NF, Simard F, Besansky NJ: **Localization of candidate regions maintaining a common polymorphic inversion (2La) in *Anopheles gambiae*.** *Plos Genetics* 2007, **3:**2404-2414.
32. Obbard DJ, Linton YM, Jiggins FM, Yan G, Little TJ: **Population genetics of *Plasmodium* resistance genes in *Anopheles gambiae*: no evidence for strong selection.** *Molecular Ecology* 2007, **16:**3497-3510.
33. Parmakelis A, Slotman M, Marshall J, Awono-Ambene P, Antonio-Nkondjio C, Simard F, Caccone A, Powell J: **The molecular evolution of four anti-malarial immune genes in the *Anopheles gambiae* species complex.** *BMC Evolutionary Biology* 2008, **8:**79.
34. Lehmann T, Hume JCC, Licht M, Burns CS, Wollenberg K, Simard F, Ribeiro JMC: **Molecular Evolution of Immune Genes in the Malaria Mosquito *Anopheles gambiae*.** *PLoS ONE* 2009, **4:**e4549.
35. Scott JA, Brogdon WG, Collins FH: **Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction.** *Am J Trop Med Hyg* 1993, **49:**520-529.
36. Favia G, della Torre A, Bagayoko M, Lanfrancotti A, Sagnon N, 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.
37. Lehmann T, Diabate A: **The molecular forms of *Anopheles gambiae*: A phenotypic perspective.** *Infection, Genetics and Evolution* 2008, **8:**737-746.
38. White BJ, Santolamazza F, Kamau L, Pombi M, Grushko O, Mouline K, Brengues C, Guelbeogo W, Coulibaly M, Kayondo JK, Sharakhov I, Simard F, Petrarca V, Della Torre A, Besansky NJ: **Molecular karyotyping of the 2La inversion in *Anopheles gambiae*.** *Am J Trop Med Hyg* 2007, **76:**334-339.
39. Sharakhov IV, White BJ, Sharakhova MV, Kayondo J, Lobo NF, Santolamazza F, della Torre A, Simard F, Collins FH, Besansky NJ: **Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2*La*) in the *Anopheles gambiae* complex.** *Proc Natl Acad Sci USA* 2006, **103:**6258-6262.
40. Ng'habi K, Meneses C, Cornel A, Slotman M, Knols B, Ferguson H, Lanzaro G: **Clarification of anomalies in the application of a 2La molecular karyotyping method for the malaria vector *Anopheles gambiae*.** *Parasites & Vectors* 2008, **1:**45.
41. 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 ZW, 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, Chatuverdi 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 ZP, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke ZX, 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 HG, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun JT, Thomasova D, Ton LQ, Topalis P, Tu ZJ, Unger MF, Walenz B, Wang AH, Wang J, Wang M, Wang XL, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang HY, Zhao Q, Zhao SY, Zhu SC, Zhimulev I, Coluzzi M, della Torre A, Roth CW, Louis C, Kalush F, Mural RJ, Myers EW, Adams MD, Smith HO, Broder S, Gardner MJ, Fraser CM, Birney E, Bork P, Brey PT, Venter JC, Weissenbach J, Kafatos FC, Collins FH, Hoffman SL: **The genome sequence of the malaria mosquito *Anopheles gambiae* .** *Science* 2002, **298:**129.

42. Stephens M, Donnelly P: **A comparison of Bayesian methods for haplotype reconstruction from population genotype data.** *Am J Hum Genet.* 2003, **73(5):**1162-1169.
43. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet.* 2001, **68(4):**978-989.
44. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R: **DnaSP, DNA polymorphism analyses by the coalescent and other methods.** *Bioinformatics* 2003, **19:**2496-2497.
45. Tajima F: **Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.** *Genetics* 1989, **123:**585-595.
46. Hudson RR, Boos DD, Kaplan NL: **A statistical test for detecting geographic subdivision.** *Mol Biol Evol* 1992, **9:**138-151.
47. Nei M: **Evolution of human races at the gene level.** In *Human genetics, part A: The unfolding genome* Edited by: Bonne-Tamir B, Cohen T, Goodman RM. New York.: Alan R. Liss; 1982.
48. Rand DM, Kann LM: **Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans.** *Mol Biol Evol* 1996, **13:**735-748.
49. Smith NGC, Eyre-Walker A: **Adaptive protein evolution in *Drosophila* .** *Nature* 2002, **415:**1022-1024.
50. Bierne N, Eyre-Walker A: **The Genomic Rate of Adaptive Amino Acid Substitution in *Drosophila* .** *Mol Biol Evol* 2004, **21:**1350-1360.
51. Sawyer SA, Hartl DL: **Population-Genetics Of Polymorphism And Divergence.** *Genetics* 1992, **132:**1161-1176.
52. Burnham KP, Anderson DR: *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach* 2nd edition. Springer; 2002.
53. **MKtest** [http://tree.bio.ed.ac.uk/software/mktest/]
54. Nei M: *Molecular Evolutionary Genetics* New York: Columbia University Press; 1987.
55. Anthony TG, Harbach RE, Kitching IJ: **Phylogeny of the Pyretophorus Series of *Anopheles* subgenus *Cellia* (Diptera: Culicidae).** *Systematic Entomol* 1999, **24:**193-205.
56. Harbach RE: **Review of the internal classification of the genus *Anopheles* (Diptera, Culicidae) – The foundation for comparative systematics and phylogenetic research.** *Bull Entomol Res* 1994, **84:**331-342.
57. Foley DH, Bryan JH, Yeates D, Saul A: **Evolution and Systematics of *Anopheles*: Insights from a Molecular Phylogeny of Australasian Mosquitoes.** *Mol Phylogenet Evol* 1998, **9:**262.
58. Turner TL, Hahn MW: **Locus- and Population-Specific Selection and Differentiation between Incipient Species of *Anopheles gambiae*.** *Mol Biol Evol* 2007, **24:**2132-2138.
59. Turner TL, Hahn MW, Nuzhdin SV: **Genomic islands of speciation in *Anopheles gambiae*.** *PLoS Biology* 2005, **3:**e285.
60. della Torre A, Costantini C, Besansky NJ, Caccone A, Petrarca V, Powell JR, Coluzzi M: **Speciation within *Anopheles gambiae* – the glass is half full.** *Science* 2002, **298:**115-117.
61. Habtewold T, Povelones M, Blagborough AM, Christophides GK: **Transmission blocking immunity in the malaria non-vector mosquito *Anopheles quadriannulatus* species A.** *PLoS Pathog* 2008, **4:**8.
62. Lindsay SW, Parson L, Thomas CJ: **Mapping the ranges and relative abundance of the two principal African malaria vectors, *Anopheles gambiae sensu stricto* and *An. arabiensis*, using climate data.** *Proc R Soc London Series B-Biol Sci* 1998, **265:**847-854.
63. Kerah-Hinzoumbe C, Peka M, Nwane P, Donan-Gouni I, Etang J, Same-Ekobo A, Simard F: **Insecticide resistance in *Anopheles gambiae* from south-western Chad, Central Africa.** *Malar J* 2008, **7:**192.
64. Tripet F, Dolo G, Lanzaro GC: **Multilevel analyses of genetic differentiation in *Anopheles gambiae* s.s. reveal patterns of gene flow important for malaria-fighting mosquito projects.** *Genetics* 2005, **169:**313-324.
65. Little TJ: **Immune system polymorphism: Implications for genetic engineering.** In *Genetically modified mosquitoes for malaria control* Edited by: Boete C. Georgetown, Texas: Landes Bioscience; 2006:36-59.
66. Wang-Sattler R, Blandin S, Ning Y, Blass C, Dolo G, Toure YT, Torre Ad, Lanzaro GC, Steinmetz LM, Kafatos FC, Zheng L: **Mosaic genome architecture of the *Anopheles gambiae* species complex.** *PLoS ONE* 2007, **2:**e1249.
67. Donnelly MJ, Pinto J, Girod R, Besansky NJ, Lehmann T: **Revisiting the role of introgression vs shared ancestral polymorphisms as key processes shaping genetic diversity in the recently separated sibling species of the *Anopheles gambiae* complex.** *Heredity* 2004, **92:**61-68.
68. Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P: **Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations.** *Genome Research* 2005, **15:**790-799.
69. Cutter AD: **Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*.** *Genetics* 2006, **172:**171-184.
70. Slotman MA, Della Torre A, Calzetta M, Powell JR: **Differential introgression of chromsomal regions between *Anopheles gambiae* and *An. arabiensis*.** *Am J Trop Med Hyg* 2005, **73:**326-335.
71. Slotman M, Della Torre A, Powell JR: **Female sterility in hybrids between *Anopheles gambiae* and *An. arabiensis*, and the causes of Haldane's rule.** *Evolution* 2005, **59:**1016-1026.
72. Besansky NJ, Krzywinski J, Lehmann T, Simard F, Kern M, Mukabayire O, Fontenille D, Toure Y, Sagnon NF: **Semipermeable species boundaries between *Anopheles gambiae* and *Anopheles arabiensis*: Evidence from multilocus DNA sequence variation.** *Proc Natl Acad Sci USA* 2003, **100:**10818-10823.
73. Black WC, Lanzaro GC: **Distribution of genetic variation among chromosomal forms of *Anopheles gambiae* s.s.: introgressive hybridization, adaptive inversions, or recent reproductive isolation?** *Insect Molecular Biology* 2001, **10:**3-7.
74. Rosenberg NA: **The Probability of Topological Concordance of Gene Trees and Species Trees.** *Theoretical Population Biology* 2002, **61:**225-247.
75. Andolfatto P: **Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*.** *Mol Biol Evol* 2001, **18:**279-290.
76. Andolfatto P, Przeworski M: **A Genome-wide departure from the standard neutral model in natural populations of *Drosophila*.** *Genetics* 2000, **156:**257-172.
77. Lehmann T, Hawley WA, Grebert H, Collins FH: **The effective population size of *Anopheles gambiae* in Kenya: implications for population structure.** *Mol Biol Evol* 1998, **15:**264-276.
78. Besansky NJ: **White Paper: Proposal for the Eight Genomes Cluster for Genus *Anopheles*.** NIH, National Human Genome Research Institute; 2005.
79. Yawson AE, Weetman D, Wilson MD, Donnelly MJ: **Ecological zones rather than molecular forms predict genetic differentiation in the malaria vector *Anopheles gambiae* s.s. in Ghana.** *Genetics* 2007, **175:**751-761.
80. 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(2):**133-147.
81. Esnault C, Boulesteix M, Duchemin JB, Koffi AA, Chandre F, Dabiré R, Robert V, Simard F, Tripet F, Donnelly MJ, Fontenille D, Biémont C: **High Genetic Differentiation between the M and S Molecular Forms of *Anopheles gambiae* in Africa.** *PLoS ONE* 2008, **3:**e1968.
82. Besansky NJ: **White Paper: Genome analysis of vectorial capacity in major Anopheles vectors of malaria parasites.** NIH, National Human Genome Research Institute; 2008.
83. Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, Barillas-Mury C, Bian G, Blandin S, Christensen BM, Dong Y, Jiang H, Kanost MR, Koutsos AC, Levashina EA, Li J, Ligox-

ygakis P, MacCallum RM, Mayhew GF, Mendes A, Michel K, Osta MA, Paskewitz S, Shin SW, Vlachou D, Wang L, Wei W, Zheng L, Zou Z, Severson DW, Raikhel AS, Kafatos FC, Dimopoulos G, Zdobnov EM, Christophides GK: **Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes.** *Science* 2007, **316:**1738-1743.
84. Andolfatto P, Depaulis F, Navarro A: **Inversion polymorphisms and nucleotide variability in *Drosophila*.** *Genetical Research* 2001, **77:**1-8.
85. Schaeffer SW, Anderson WW: **Mechanisms of genetic exchange within the chromosomal inversions of *Drosophila pseudoobscura*.** *Genetics* 2005, **171:**1729-1739.
86. Stump AD, Pombi M, Goeddel L, Ribeiro JMC, Wilder JA, Torre AD, Besansky NJ: **Genetic exchange in 2La inversion heterokaryotypes of *Anopheles gambiae*.** *Insect Mol Biol* 2007, **16:**703-709.
87. Vernick KD, Collins FH: **Association of a *Plasmodium* refractory phenotype with an esterase locus in *Anopheles gambiae*.** *Am J Trop Med Hyg* 1989, **40:**593-597.
88. Petrarca V, Beier JC: **Intraspecific chromosomal polymorphism in the *Anopheles gambiae* complex as a factor affecting malaria transmission in the Kisumu area of Kenya.** *Am J Trop Med Hyg* 1992, **46:**229-237.
89. Charlesworth D: **Balancing selection and its effects on sequences in nearby genome regions.** *PloS Genetics* 2006, **2(4):**e64.
90. Slotman MA, Parmakelis A, Marshall JC, Awono-Ambene PH, Antonio-Nkondjo C, Simard F, Caccone A, Powell JR: **Patterns of selection in anti-malarial immune genes in malaria vectors: evidence for adaptive evolution in LRIM1 in *Anopheles arabiensis*.** *PLoS ONE* 2007, **2:**e793.
91. Simard F, Licht M, Besansky NJ, Lehmann T: **Polymorphism at the defensin gene in the *Anopheles gambiae* complex: Testing different selection hypotheses.** *Infect Genet Evol.* 2007, **7(2):**285-292.
92. Obbard DJ, Callister D, Jiggins FM, Soares D, Yan G, Little TJ: **The evolution of TEP1, an exceptionally polymorphic immunity gene in *Anopheles gambiae*.** *BMC Evolutionary Biology* 2008, **8:**274.
93. Begun DJ, Whitley P: **Adaptive eolution of relish, a *Drosophila* NF-{kappa}B/I{kappa}B Protein.** *Genetics* 2000, **154:**1231-1238.
94. Schlenke TA, Begun DJ: **Natural selection drives *Drosophila* immune system evolution.** *Genetics* 2003, **164:**1471-1480.
95. Lazzaro BP: **Elevated polymorphism and divergence in the class C scavenger receptors of *Drosophila melanogaster* and *D. simulans*.** *Genetics* 2005, **169:**2023-2034.
96. Yang ZH: **PAML 4: Phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24:**1586-1591.
97. Li WH: **Unbiased estimation of the rates of synonymous and nonsynonymous substitution.** *J Mol Evol* 1993, **36:**96-99.
98. Ramensky VE, Nurtdinov RN, Neverov AD, Mironov AA, Gelfand MS: **Positive selection in alternatively spliced exons of human genes.** *Am J Human Genetics* 2008, **83:**94-98.
99. Li WH, Sadler LA: **Low nucleotide diversity in man.** *Genetics* 1991, **129:**513-523.
100. Kosakovsky Pond SL, Frost SDW, Muse SV: **HyPhy: hypothesis testing using phylogenies.** *Bioinformatics* 2005, **21:**676-679.
101. Wilson DJ, McVean G: **Estimating diversifying selection and functional constraint in the presence of recombination.** *Genetics* 2006, **172:**1411-1425.
102. Anisimova M, Nielsen R, Yang ZH: **Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites.** *Genetics* 2003, **164:**1229-1236.
103. Hasegawa M, Cao Y, Yang ZH: **Preponderance of slightly deleterious polymorphism in mitochondrial DNA: Nonsynonymous/synonymous rate ratio is much higher within species than between species.** *Mol Biol Evol* 1998, **15:**1499-1505.