PHILOSOPHICAL
TRANSACTIONS
OF
THE ROYAL
SOCIETY B

Phil. Trans. R. Soc. B (2012) **367**, 374–384
doi:10.1098/rstb.2011.0188

Research

No evidence for biased co-transmission of speciation islands in *Anopheles gambiae*

Matthew W. Hahn[1,2,*], Bradley J. White[3], Christopher D. Muir[1] and Nora J. Besansky[3,*]

[1]*Department of Biology, and* [2]*School of Informatics and Computing,*
Indiana University, Bloomington, IN 47405, USA
[3]*Eck Institute for Global Health, Department of Biological Sciences,*
University of Notre Dame, Notre Dame, IN 46556, USA

Genome-scale scans have revealed highly heterogeneous levels of divergence between closely related taxa in many systems. Generally, a small number of regions show high differentiation, with the rest of the genome showing no or only low levels of divergence. These patterns have been interpreted as evidence for ongoing speciation-with-gene-flow, with introgression homogenizing the whole genome except loci involved in reproductive isolation. However, as the number of selected loci increases, the probability of introgression at unselected loci decreases unless there is a transmission ratio distortion causing an over-representation of specific combinations of alleles. Here we examine the transmission of three 'speciation islands' that contain fixed differences between the M and S forms of the mosquito, *Anopheles gambiae*. We made reciprocal crosses between M and S parents and genotyped over 2000 F_2 individuals, developing a hierarchical likelihood model to identify specific genotypes that are under- or over-represented among the recombinant offspring. Though our overall results did not match the expected number of F_2 genotypes, we found no biased co-transmission among M or S alleles in the three islands. Our likelihood model did identify transmission ratio distortion at two of the three islands, but this distortion was small (approx. 3%) and in opposite directions for the two islands. We discuss how our results impinge on hypotheses of current gene flow between M and S and ongoing speciation-with-gene-flow in this system.

Keywords: *Anopheles gambiae*; speciation; centromeric drive

1. INTRODUCTION

Many closely related taxa show heterogeneous levels of divergence across the genome (reviewed in [1]). Some regions show little genetic differentiation, while others—usually only a small fraction of the genome—show high levels of divergence and may even contain fixed differences distinguishing the taxa ('genomic islands of speciation' [2]). This heterogeneity may result from two alternative models that differ mainly in the role played by gene flow. In the first model, loci conferring higher fitness in different environments are the first to diverge, with ongoing gene flow homogenizing the majority of loci not directly involved in isolation [3,4]. In this model, one expects to find the loci responsible for reproductive isolation in the regions of highest divergence, with levels of differentiation declining at neighbouring loci as recombination breaks up associations between linked sites. In the second model, which is simply an extreme alternative along a spectrum of intermediate possibilities, reproductive isolation is instantaneous and complete, with no ongoing gene flow, possibly due to geographical isolation. In this model, heterogeneity among loci in their levels of differentiation is due to stochastic variation in coalescent times [5], variable mutation rates [6] or heterogeneous natural selection. The targets of natural selection may be directly involved in reproductive isolation between the taxa, completely orthogonal to the isolating barriers between them, or some mixture of the two. Importantly, in the second model the regions of highest differentiation do not necessarily indicate the location of genes underlying reproductive isolation ('incidental islands' [7,8]).

Direct evidence for ongoing gene flow and the countervailing effects of natural selection can be most easily recognized in studies of hybrid zones between already diverged lineages [9–12]. In many cases, F_1 and backcross individuals are phenotypically distinguishable, and evidence that the species are coming together after a substantial period of allopatry (i.e. secondary contact) provides strong support for the inference that shared alleles are due to

*Authors for correspondence (mwh@indiana.edu;nbesansk@nd.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2011.0188 or via http://rstb.royalsocietypublishing.org.

One contribution of 13 to a Theme Issue 'Patterns and processes of genomic divergence during speciation'.

introgression and not ancestral variation. Genome-wide studies of hybrid zones in mice [13,14], rabbits [15] and butterflies [16–19] demonstrate a large amount of heterogeneity in the ability of individual loci to introgress between species, with very strong selection against introgression at the genes presumably responsible for reproductive isolation. Because multiple loci can be resistant to introgression across hybrid zones—in both directions—patterns of differential hybridization will generate significant linkage disequilibrium among these loci [14]. This non-random association of alleles, even between unlinked loci, is taken as further evidence for the strong barriers that exist at specific genes.

Reliably inferring that there is ongoing gene flow is much more difficult in nascent species that have recently arisen in sympatry or parapatry. In the mosquito malaria vector, *Anopheles gambiae sensu stricto* (hereafter, *A. gambiae*), two phenotypically indistinguishable species have recently formed in Africa. The two species—referred to as the S and M molecular forms based on the original diagnostic polymerase chain reaction (PCR) marker [20–22]—are found in a largely overlapping range in West and Central Africa, with only the presumed ancestral S form found in East Africa [23]. S-form mosquitoes only breed during the rainy season and have fast-developing larvae found in temporary pools and puddles [24]. M-form mosquitoes are reproductively active throughout the year, with more slowly developing larvae found in stable bodies of water such as rice fields that are closely associated with human activity [23,24]. The M form is therefore thought to have arisen since the emergence of semi-stable human settlements in Africa [25], though some data suggest an earlier split [26]. In the absence of predators the S-form larvae will out-compete M-form larvae [27], but in the presence of predators M-form larvae win (predators are more common in permanent bodies of water [28]). In addition to larval habitat preferences, the M and S mosquitoes show strong assortative mating. In most places where they are sympatric, mating swarms are composed almost exclusively of M- or S-form males with no mixing, even when swarms are located less than 200 feet apart [29]. In an extensive survey of mated females, Tripet *et al.* [30] found that only 1.2 per cent of individuals carried sperm of the ‘wrong’ form, indicating hybridization between forms (though not necessarily gene flow between forms, as the F_1 offspring of these individuals could be unfit). Given that females introduced to a swarm of the opposite form can still be inseminated [29], it appears that pre-mating signals [31] may be a very important barrier between the incipient species.

Levels of differentiation between M and S are highly variable across the *A. gambiae* genome, with some regions showing no differences (i.e. F_{ST} is close to 0) and some showing fixed differences [2,8,32]. The regions of highest differentiation are contained within three ‘speciation islands’ on chromosomes 2L, 3L and X [2,8]; a fourth region on chromosome 2R is not differentiated between M and S in every geographical area sampled and is therefore no longer considered a speciation island [33]. Fixed differences between the two forms are found in every island (the original diagnostic marker is in the X island), with lower levels of differentiation and no fixed differences in flanking regions tens to hundreds of kilobases away [2,8,33]. There is also near-complete association of alleles within the islands with each other: that is, there is strong linkage disequilibrium among the unlinked loci, with M genotypes at one locus found with M genotypes at the other loci, and similarly for S genotypes. Recent whole-genome studies have also found additional regions of increased differentiation [34,35], though they could not determine whether fixed differences existed among natural populations in these regions.

Because of highly similar allele frequencies across the majority of the genome (see also [36–38]) and biologically significant numbers of hybrid individuals found in nature (approx. 1%; [23,39–41]), it has largely been assumed that there is ongoing gene flow between M and S. Accordingly, the observed heterogeneity in divergence is thought to be due to selection against hybrid genotypes at loci contained within the islands, with recombination allowing the free introgression of flanking markers [2]. However, there are several alternative hypotheses that might explain these patterns. If there is actually only very little gene flow between M and S, then these regions need not be maintained in the face of introgression; instead, these incipient species may already be largely isolated and diverging independently [7]. Alternatively, if there are truly high levels of gene flow, in order to maintain the strong association between alleles at unlinked markers in the face of recombination and hybridization it must be the case that (i) a large fraction of offspring with recombinant genotypes do not survive, (ii) some form of transmission ratio distortion favours triply M ($X^M2L^M3L^M$) or triply S ($X^S2L^S3L^S$) gametes such that fully homozygous individuals at all three islands are more likely to be formed or (iii) some combination of these two processes (cf. [42]). One model linking transmission ratio distortion to incipient speciation is the centromeric drive hypothesis [43,44]. This model proposes that conflict between centromeric DNA repeats and centromere-binding proteins results in an arms race between the two structural units, driving rapid coevolution. If this arms race follows different trajectories in different populations, hybrid individuals may have lower fitness due to sub-optimal genetic interactions [43]. As all three speciation islands in *A. gambiae* are located next to a centromere, it is formally possible that some type of centromeric drive is responsible for keeping co-adapted combinations of M and S alleles at all three centromeres together.

In this paper, we test for the under-representation of recombinant genotypes in a laboratory cross. To do this, we carried out reciprocal crosses between M and S mosquitoes to generate an F_2 population. By genotyping F_2 individuals at distinguishing markers in all three islands, we are able to test for deviations from expected proportions of all possible recombinant genotypes. Deviations from expected numbers of recombinant offspring would be expected under either strong early-viability effects or direct transmission ratio distortion interactions among alleles at all three islands. After controlling for one-locus and two-locus effects, we find no significant deviations from expectations for any three-locus combination of alleles. We conclude by discussing these results and their implications for speciation between M and S.

2. MATERIAL AND METHODS

(a) *Strains and crossing design*

Anopheles gambiae parental strains were Pimperena (S form) and Mali-NIH (M form) established in 2005 from Mali (www.mr4.org). Reciprocal crosses between M and S were performed en masse, and both types of F_1 hybrids were intercrossed separately to generate F_2 hybrids. All mosquito populations were maintained at the University of Notre Dame in the same insectary bay, under controlled conditions of 27°C, 80 per cent relative humidity, and a 12 L : 12 D hour light-dark cycle with 1 h sunrise and sunset light transitions. Larvae were reared in plastic trays (27 × 16 × 6.5 cm) at a density of approximately 100 per litre of deionized water, and fed a daily diet of a 2 : 1 mixture of finely ground tropical fish pellet: brewer's yeast. Pupae were transferred to 0.2 m^3 screened cages, where emerged adults were maintained with access to a 10 per cent solution of corn syrup.

(b) *Genotyping*

Daily upon emergence, F_2 hybrid adults were sexed, counted and held at −80°C until genotyping was performed. DNA was extracted from individual F_2 hybrids by heating a single leg in 50 μl of lysis buffer [45]. Restriction-fragment length polymorphism PCR assays for genotyping a single diagnostic single-nucleotide polymorphism (SNP) in each of the three speciation islands have been previously designed [8,46]. However, our goal was to streamline the protocol by eliminating the restriction-digest step. Conventional allele-specific PCR was not suitable, as Taq polymerase can extend over a single SNP difference between primer and target site, even when the mismatch is at the 3′-end of the primer. To overcome this obstacle, we adopted the artificial mismatch approach [47], in which a primer is designed with an intentional mismatch to the target site, three nucleotides from the 3′-end (see electronic supplementary material, figure S1). Contrary to the findings in Wilkins *et al.* [48], this approach was not successful for genotyping the X island despite repeated attempts. Accordingly, genotyping of this island followed Santolamazza *et al.* [49], except that three units of *Mse*1 enzyme were used to ensure complete digestion of products.

Genotyping of diagnostic SNPs between M and S in the 2L and 3L islands was performed with novel intentional-mismatch primers. First, we identified and aligned trace reads from the genome sequences of Mali-NIH (M) and Pimperena (S) [34] that mapped to exons in the 2L and 3L islands. After identifying exons with at least two nearby fixed SNP differences, we designed an M mismatch primer for one SNP, an S mismatch primer for the other SNP and a universal primer for each island (see electronic supplementary material, figure S1). SNPs were verified as fixed between colonies by genotyping at least 40 individuals per colony for both islands. Genotyping assays were performed individually for each island, using the primers and concentrations indicated in the electronic supplementary material, figure S1. Each 25 μl PCR reaction included 200 $\mu\mathrm{mol\,l}^{-1}$ each dNTP, 2.5 $\mathrm{mmol\,l}^{-1}$ MgCl_2, 20 $\mathrm{mmol\,l}^{-1}$ Tris-HCl (pH 8.4), 50 $\mathrm{mmol\,l}^{-1}$ KCl, 2.5 U Taq polymerase, and 1/5 of the DNA extracted from a single mosquito leg. Thermocycler conditions were 94°C for 2 min; 35 cycles of 94°C for 30 s, 58°C for 30 s and 72°C for 45 s; a final elongation at 72°C for 5 min; and a 4°C hold. The resulting products were analysed on 1.5 per cent agarose gels stained with ethidium bromide.

(c) *Statistical analysis*

In addition to comparing our results with the standard expectations under the assumptions of equal transmission of all alleles, Hardy–Weinberg equilibrium, and independent assortment (hereafter referred to as the null model) via a χ^2 goodness-of-fit test, we also wanted to determine which alleles or genotypic combinations were causing any observed deviations. We therefore employed a likelihood model to determine what factors could explain the data. Here, for simplicity, we describe the model in detail for the two-locus case; equations for the full three-locus model are given in the electronic supplementary material.

Differences from the expected values can arise at three different levels: single-locus deviations in the expected allele frequencies or genotypes; combinations of two-locus genotypes that deviate from the expected, taking into account all one-locus deviations; and combinations of three-locus genotypes that deviate from the expected, taking into account all one-locus and two-locus deviations. We employed a 'forward' selection process to estimate parameters describing the deviations from expected values: by a forward process, we mean that deviations of allele or genotype frequencies (e.g. too few M alleles at the 2L locus) will necessarily alter the frequency of all two-locus and three-locus genotypes containing that allele or genotype.

For the one-locus model, there are three parameters describing deviations from expectations. The parameter θ_i estimates the deviation from the expected 50 : 50 ratio of M and S alleles at each locus, i (X, 2L and 3L). We define positive values of θ_i as excesses of M alleles and negative values as deficiencies of M alleles. The parameter β_i estimates the deviation from expected values of heterozygotes or homozygotes for particular alleles at each locus, i. We define positive values of β as excesses of homozygotes, and negative values as deficiencies of homozygotes. Finally, the parameter α_i estimates the deviation from expected values of a particular genotype, i (MM or SS), where positive values are defined as an excess of that genotype. In other words, α_i measures the asymmetry in deviations from the expected proportions between the two homozygous genotypes. So the frequencies of the three possible genotypes at a single locus are

$$\left.\begin{aligned}
\hat{p}_{i\mathrm{MM}} &= \hat{p}_{i\mathrm{MM}}|\theta_i,\beta_i\left(1-\frac{\alpha_{i\mathrm{SS}}}{\hat{p}_{i\mathrm{MS}}|\theta_i,\beta_i+\hat{p}_{i\mathrm{MM}}|\theta_i,\beta_i}\right)+\alpha_{i\mathrm{MM}}\\
\hat{p}_{i\mathrm{MS}} &= \hat{p}_{i\mathrm{MS}}|\theta_i,\beta_i\left(1-\frac{\alpha_{i\mathrm{MM}}}{\hat{p}_{i\mathrm{MS}}|\theta_i,\beta_i+\hat{p}_{i\mathrm{SS}}|\theta_i,\beta_i}-\frac{\alpha_{i\mathrm{SS}}}{\hat{p}_{i\mathrm{MS}}|\theta_i,\beta_i+\hat{p}_{i\mathrm{MM}}|\theta_i,\beta_i}\right)\\
\text{and}\quad \hat{p}_{i\mathrm{SS}} &= \hat{p}_{i\mathrm{SS}}|\theta_i,\beta_i\left(1-\frac{\alpha_{i\mathrm{MM}}}{\hat{p}_{i\mathrm{MS}}|\theta_i,\beta_i+\hat{p}_{i\mathrm{SS}}|\theta_i,\beta_i}\right)+\alpha_{i\mathrm{SS}},
\end{aligned}\right\}\tag{2.1}$$

where

$$\hat{p}_{i\mathrm{MM}}|\theta_i,\beta_i=(0.5+\theta_i)^2+\frac{\beta_i}{2},$$

$$\hat{p}_{i\mathrm{MS}}|\theta_i,\beta_i=2(0.5+\theta_i)(0.5-\theta_i)-\beta_i,$$

$$\hat{p}_{i\mathrm{SS}}|\theta_i,\beta_i=(0.5-\theta_i)^2+\frac{\beta_i}{2}.$$

The denominators below α_i reflect the fact that the deficit (excess) of individuals in one genotype are deposited to (withdrawn from) the other genotypes in proportion to their frequency. Note that the equations will differ between males and females for the X locus because there are only two possible genotypes in males, the heterogametic sex.

Estimating the parameters in the one-locus model is straightforward. θ is simply the deviation of the observed allele frequencies from the expectation of 0.5. Since all excess M alleles must cause a deficit of S alleles, we can just calculate one θ for each locus:

$$\hat{\theta}_i=\frac{2N_{i\mathrm{MM}}+N_{i\mathrm{MS}}}{2N}-0.5,\tag{2.2}$$

where N_{iG} is the number of individuals of genotype G at locus i and N is the total number of individuals genotyped. Because the β parameter measures the deviation from the expected proportion of homozygotes in the sample, given constituent allele frequencies, it can be estimated by

$$\hat{\beta}_i=2(0.5+\hat{\theta}_i)(0.5-\hat{\theta}_i)-\frac{N_{i\mathrm{MS}}}{N}.\tag{2.3}$$

Finally, the α parameter measures the asymmetry in the proportion of the two homozygous genotypes. Again, there only needs to be one α parameter (either MM or SS), and it can be estimated by either

$$\left.\begin{aligned}
\hat{\alpha}_{i\mathrm{MM}} &= \frac{N_{i\mathrm{MM}}}{N}-(0.5+\hat{\theta}_i)^2-\frac{\hat{\beta}_i}{2}\\
\text{or}\quad \hat{\alpha}_{i\mathrm{SS}} &= \frac{N_{i\mathrm{SS}}}{N}-(0.5-\hat{\theta}_i)^2-\frac{\hat{\beta}_i}{2}.
\end{aligned}\right\}\tag{2.4}$$

The two-locus model does not have an analogue of the θ parameter, but does have analogous β and α parameters that represent potential interactions between loci. In this model, β_{ij} represents the deviations from the expected number of double homozygotes (i.e. homozygotes at two loci) relative to heterozygotes, with ij representing any of the combinations X/2L, X/3L or 2L/3L. The parameter α_{ij} represents the deviations from the expected number of any of the 27 female and 21 male two-locus genotypes, i.e. $\mathrm{X}^{\mathrm{MM}}2\mathrm{L}^{\mathrm{MM}}$, $\mathrm{X}^{\mathrm{MS}}2\mathrm{L}^{\mathrm{MS}}$, etc. The frequency of each two-locus genotype is given by two separate expressions, one for double homozygotes and one for all other genotypes. For double homozygotes (e.g. $\mathrm{X}^{\mathrm{MM}}2\mathrm{L}^{\mathrm{MM}}$ or $2\mathrm{L}^{\mathrm{SS}}3\mathrm{L}^{\mathrm{SS}}$):

$$\left.\begin{aligned}
\hat{p}_{i\mathrm{MM},j\mathrm{MM}} &= \hat{p}_{i\mathrm{MM},j\mathrm{MM}}|\beta_{ij}\left(1-\frac{\sum\alpha_{ij}}{\sum_\alpha p}\right)+\alpha_{i\mathrm{MM},j\mathrm{MM}}\\
\text{and}\quad \hat{p}_{i\mathrm{SS},j\mathrm{SS}} &= \hat{p}_{i\mathrm{SS},j\mathrm{SS}}|\beta_{ij}\left(1-\frac{\sum\alpha_{ij}}{\sum_\alpha p}\right)+\alpha_{i\mathrm{SS},j\mathrm{SS},}
\end{aligned}\right\}\tag{2.5}$$

where

$$\hat{p}_{i\mathrm{MM},j\mathrm{MM}}|\beta_{ij}=\hat{p}_{i\mathrm{MM}}\hat{p}_{j\mathrm{MM}}+\frac{\beta_{ij}}{2}$$

$$\text{and}\quad \hat{p}_{i\mathrm{SS},j\mathrm{SS}}|\beta_{ij}=\hat{p}_{i\mathrm{SS}}\hat{p}_{j\mathrm{SS}}+\frac{\beta_{ij}}{2}.$$

For all other genotypes:

$$\hat{p}_{iG,jG}=\hat{p}_{iG,jG}|\beta_{ij}\left(1-\frac{\sum\alpha_{ij}}{\sum_\alpha p}\right)+\alpha_{iG,jG},\tag{2.6}$$

where

$$\hat{p}_{iG,jG}|\beta_{ij}=\hat{p}_{iG}\hat{p}_{jG}\left(1-\frac{\beta_{ij}}{\sum_\beta p}\right).$$

The summations in the denominators are once again necessary to account for the fact that a deficit (excess) of individuals at one or more genotypes must be deposited to (withdrawn from) the other genotypes in proportion to their frequency. The summations are formally defined as

$$\left.\begin{aligned}
\sum\nolimits_\alpha p &= \sum_{\forall i_G j_G:\alpha_{iG,jG}>0}\hat{p}_{ij}|\beta_{ij}\\
\text{and}\quad \sum\nolimits_\beta p &= \sum_{\forall i_G j_G:i_G j_G\neq i_{\mathrm{MM}}j_{\mathrm{MM}}\vee i_{\mathrm{SS}}j_{\mathrm{SS}}}\hat{p}_i\hat{p}_j.
\end{aligned}\right\}\tag{2.7}$$

Note, again, that expectations for genotypes involving the X-linked locus must be adjusted for hemizygosity in males.

The β term tells us if there is an excess/deficit of association between two given alleles, analogous to linkage disequilibrium. Though the sign of this parameter is arbitrary, we have defined positive values of β_{ij} as cases in which there are more MM and SS genotypes than expected. β_{ij} can be estimated by

$$\hat{\beta}_{ij} = \frac{N_{i\mathrm{MM},j\mathrm{MM}}}{N} - \hat{p}_{i\mathrm{MM}}\hat{p}_{j\mathrm{MM}} + \frac{N_{i\mathrm{SS},j\mathrm{SS}}}{N} - \hat{p}_{i\mathrm{SS}}\hat{p}_{j\mathrm{SS}}. \tag{2.8}$$

The α parameter measures any excess or deficit of single two-locus genotypes, given constituent allele frequencies. We can estimate α_{ij} by

$$\hat{\alpha}_{ij} = \frac{N_{ij}}{N} - \hat{p}_{ij}|\beta_{ij}. \tag{2.9}$$

Extensions to the three-locus model are obvious from the above models, adding β_{ijk} and α_{ijk} parameters for three-way interactions among loci. Once again, the β_{ijk} term represents the excess or deficit of triply homozygous genotypes, while the α_{ijk} term represents the excess or deficit of specific three-locus genotypes. Further results for the three-locus model are given in the electronic supplementary material, methods.

Once we have our data vector, **N**, and our vector of maximum likelihood parameter estimates, $\hat{\mathbf{N}}$, we can estimate the likelihood of a given model (i.e. a particular combination of different sets of parameters that can take non-zero values) from the density of the multinomial distribution. We used the Akaike information criterion (AIC) to determine which models, after parameterization, best explained the data. To examine whether this model selection method inflates type I error, we simulated data from the multinomial distribution using the same sample sizes as in the experiment. For each type of parameter (θ, α and β), we conducted 10 000 simulations assuming no effect and then assessed the frequency of false positives under different penalty values used to calculate the AIC score. For all parameter types, we found nearly identical results. The standard penalty of 2 was too liberal, allowing a false positive rate of approximately 0.16. A penalty of approximately 4 gave a desired false positive rate of 0.05. Therefore, we opted to implement a penalty of 4 in calculating AIC scores throughout the analysis.

Since we are assuming a forward process, we estimated one-locus parameters first and then estimated two-locus effects based on the new expectations. In this step, we fit 1674 000 total models: 1200 with no two-locus effects (1000 females, 200 males); 614 400 with interactions between 2L and 3L (512 000 females, 102 400 males); 529 200 with interactions between 2L and X (512 000 females, 17 200 males); and 529 200 with interactions between 3L and X (512 000 females, 17 200 males). Note that there are fewer models for the male data because there are only two possible genotypes for the X locus in males, the heterogametic sex. The total number of possible models with three-locus effects was prohibitive; we therefore tested a limited subset of models. First, we parameterized models with only three-locus effects, allowing up to five parameters (e.g. five three-locus genotypes that deviated from the null expectation). We also selected the best models from the previous one- and two-locus analyses and added three-locus effects to see if any improved the fit of the model. Finally, we explicitly tested for an overall deviation in triply homozygous and triply heterozygous genotypes, as these are biologically interesting models which may indicate inbreeding or outbreeding depression.

3. RESULTS

We conducted reciprocal crosses between laboratory strains of M- and S-form mosquitoes. In total, we were able to genotype 2028 F_2 offspring from these crosses, 1008 from the M-female by S-male cross, and 1020 from the S-female by M-male cross; approximately equal numbers of male and female F_2s were scored in each case. Though this experiment was not designed to score total offspring number, there was no apparent difference in offspring number between reciprocal crosses, no bias in offspring sex-ratio and no qualitative difference in offspring number relative to crosses conducted between pure-M and pure-S parents (NJB, unpubl. results). By developing novel PCR-based markers in the three 'genomic islands' that allowed us to differentiate between homozygotes and heterozygotes, we were able to confidently assign three-locus genotypes to more than 99 per cent of all individuals.

We first compared the observed genotypes between the reciprocal crosses to determine whether there was any asymmetry in our results. Comparing male F_2s from the M-female by S-male cross to male F_2s from the S-female by M-male cross (and females to females), we found no significant differences in the observed numbers of one-, two- or three-locus genotypes in either sex after correcting for multiple tests. We therefore combined individuals from the two crosses for all subsequent analyses.

By contrast, though the differences are slight, comparing the combined datasets to the null expectations assuming equal transmission of alleles, Hardy–Weinberg equilibrium, and independent assortment gave highly significant results in both the male and female datasets (these were run separately to account for the different expectations at the X chromosome). One-, two- and three-locus comparisons were all highly significant in both sexes (see the electronic supplementary material, table S1), with the only major difference being a significant deficit of M alleles at the 3L locus in females and not in males. However, males did show some deficit of M alleles at this locus and the difference between males and females was not itself significant ($\chi^2 = 3.16$, $p = 0.075$). Both sexes showed a slight but significant excess of M alleles at the 2L locus. Together, single-locus deviations from expectations at two of the three loci scored can cause deviations in all three two-locus comparisons, which can then cause deviations in the single three-locus comparison. While it is relatively straightforward to account for single-locus deviations in calculating expectations among higher-order interactions, correcting for the effects of deviations in two- or three-locus genotypes is more difficult. Therefore, in order to directly ask whether there were deviations in multi-locus genotypic combinations that were not accounted for by lower-order deviations, we developed a novel likelihood method and applied it to our data (see §2 for details).

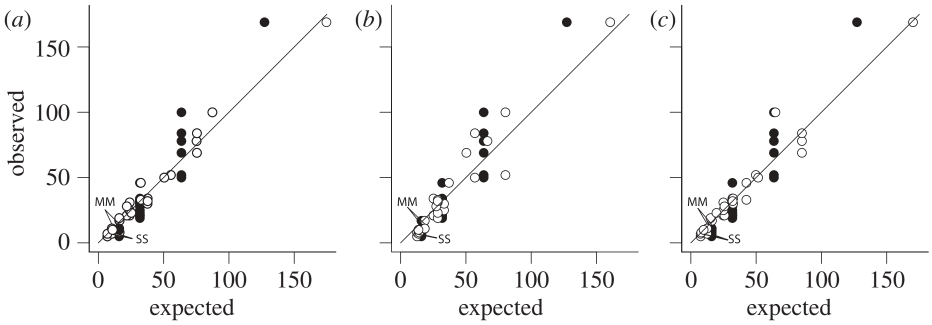

Figure 1. Comparison of the observed and expected number of female genotypes. Each of the 27 possible three-locus genotypes is plotted as a point, with the expected numbers generated by the best (*a*) one-locus, (*b*) two-locus and (*c*) three-locus likelihood model as open circles. The expected numbers for the null model are plotted in each panel for comparison (black circles), and the triply homozygous MM and SS genotypes ($X^{MM}2L^{MM}3L^{MM}$ or $X^{SS}2L^{SS}3L^{SS}$) are indicated.

Table 1. Likelihoods of different genetic models.

	females		males	
model	parameters	AIC	parameters	AIC
null		235.54		188.99
one-locus	$\theta_{2L} = 0.041$ $\beta_{2L} = -0.085$ $\beta_{X} = -0.036$ $\alpha_{3L\mathit{MM}} = -0.075$	169.15	$\theta_{2L} = 0.032$ $\beta_{2L} = -0.114$ $\beta_{3L} = -0.066$	121.67
two-locus	$\alpha_{2L\mathit{SS}} = -0.082$ $\alpha_{3L\mathit{MM}} = -0.075$ $\alpha_{X\mathit{SS},2L\mathit{MM}} = -0.022$ $\alpha_{X\mathit{MS},2L\mathit{MS}} = 0.038$	164.96	(no model better than one-locus)	
three-locus	$\beta_{2L} = 0.068$ $\alpha_{2L\mathit{SS}} = -0.082$ $\alpha_{3L\mathit{MM}} = -0.041$ $\alpha_{X\mathit{SS},2L\mathit{MM},3L\mathit{SS}} = -0.008$ $\alpha_{X\mathit{MS},2L\mathit{MS},3L\mathit{SS}} = 0.035$	163.40	$\alpha_{2L\mathit{MS}} = 0.112$ $\alpha_{3L\mathit{MM}} = -0.046$ $\alpha_{X\mathit{S},2L\mathit{MM},3L\mathit{MS}} = 0.024$ $\alpha_{X\mathit{M},2L\mathit{MS},3L\mathit{SS}} = -0.018$ $\alpha_{X\mathit{S},2L\mathit{SS},3L\mathit{SS}} = -0.016$	119.09

We used our likelihood model to parameterize and calculate AIC values for all one- and two-locus parameter combinations; this process allows us to find the model that best fits the data, while minimizing error due to overfitting [50]. For both males and females the model with the lowest likelihood score was a three-locus model (table 1). However, in both cases these models were not significantly better than models with fewer parameters. As an example of how to interpret these results, for males the best overall model (i.e. the one chosen by our model-selection procedure) was a one-locus model with $\theta_{2L} = 0.032$, $\beta_{2L} = -0.114$ and $\beta_{3L} = -0.066$. These parameter estimates indicate that there was a slight excess of M alleles at the 2L locus (such that the allele frequency was 53.2% rather than 50%), and a deficit of homozygous genotypes of both types at the 2L and 3L loci. Overall, our results provide no support for biased co-transmission of alleles because multi-locus models do not explain the data better than single-locus models, and there is therefore no evidence of an interaction among loci with regard to transmission bias.

As a graphical way to depict the fit of the models to the data, figure 1 compares the observed numbers for the 27 possible three-locus female genotypes to the expected numbers under different best-fit models. Figure 2 does the same for the 18 possible three-locus male genotypes (because there was no two-locus model better than a one-locus model, none is included in figure 2 or in table 1). For comparative purposes, we have also plotted the contrast between observed and expected for the null model on each panel. As can be seen for both the female and male data, the likelihood model provides an excellent fit to the data, with the majority of points lying on or close to the diagonal. The best-fit models, for any number of loci, were much more probable than the null model (see also table 1). In addition, the figures make it clear that adding more parameters does not make a qualitative difference in the fit of the data to the expected values. Overall, qualitative and quantitative comparisons indicate that our model provides a much more informative picture of the data than does the simple null model.

Given the complexity of our likelihood model, a detailed power calculation is particularly difficult. However, the fact that we are able to detect significant deviations in multi-locus genotypes as small as 0.8 per cent—and at single loci deviations as small as 3.2 per cent—strongly suggests that we have not missed any major transmission ratio bias.

As well as finding the best-fit likelihood models for our data, we set out to test explicit biological hypotheses about the co-segregation of M and S alleles at all three speciation islands. Overall, there is actually a

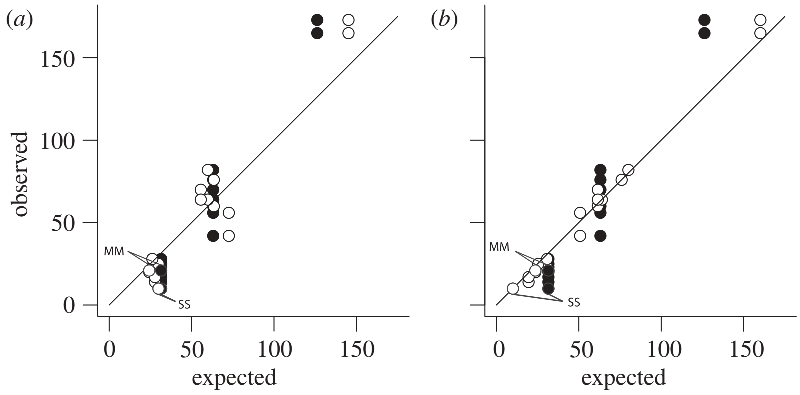

Figure 2. Comparison of the observed and expected number of male genotypes. Each of the 18 possible three-locus genotypes is plotted as a point, with the expected numbers generated by the best (*a*) one-locus and (*b*) three-locus likelihood model as open circles (there was no two-locus model better than the one-locus model). The expected numbers for the null model are plotted in each panel for comparison (black circles), and the triply homozygous MM and SS genotypes ($X^{M}2L^{MM}3L^{MM}$ or $X^{S}2L^{SS}3L^{SS}$) are indicated.

small deficit of all triply homozygous genotypes, $X^{MM}2L^{MM}3L^{MM}$ or $X^{SS}2L^{SS}3L^{SS}$ for females and $X^{M}2L^{MM}3L^{MM}$ or $X^{S}2L^{SS}3L^{SS}$ for males (see electronic supplementary material, table S1). There is a slight excess of triply heterozygous genotypes ($X^{MS}2L^{MS}3L^{MS}$) for females, but our most probable three-locus model does not indicate that this is significant. In fact, none of the estimated parameters (i.e. parameters whose value is different from 0) in the best-fit two- and three-locus models include positive values for doubly or triply homozygous genotypes, though the best-fit three-locus model for males includes a parameter with a deficit of the $X^{S}2L^{SS}3L^{SS}$ genotype (table 1). Note that, because we could not exhaustively search every possible three-locus model, we may not have found the globally best-fit likelihood model. However, given the fact that we are most interested in testing specific hypotheses about co-transmission of all three loci and no overall excess is detected, we believe our results provide biological insight into this system.

4. DISCUSSION

In this paper, we investigated patterns of inheritance at three 'speciation islands' in *A. gambiae*. In its native range in Africa this species is divided into two incipient species, M and S. Wild-caught mosquitoes are almost always triply homozygous for the M allele at the three speciation islands ($X^{MM}2L^{MM}3L^{MM}$) or triply homozygous for the S allele ($X^{SS}2L^{SS}3L^{SS}$). Only approximately 1 per cent of all wild-caught individuals are hybrids (at least at the X island, the only one genotyped in the vast majority of studies), based on studies totaling more than 10 000 samples [8,23,30,39–41]. The earliest microarray studies found little population structure outside of these three regions—which together comprise only 3 per cent of the genome [2,8]—though more recent resequencing and genotyping of whole genomes have revealed additional small regions of high differentiation outside the previously identified islands [34,35].

(a) *A test for transmission ratio distortion*

Assuming that the relatively high levels of observed hybridization imply commensurately high levels of gene flow, it had been thought that the level of introgression between M and S was high enough to homogenize regions of the genome outside the three islands [2]. In order to maintain the strong association between alleles at unlinked markers in the face of recombination and gene flow, however, it must be the case that either a large fraction of offspring with recombinant genotypes do not survive, or that some form of transmission ratio distortion favours triply M ($X^{M}2L^{M}3L^{M}$) or triply S ($X^{S}2L^{S}3L^{S}$) gametes such that fully homozygous individuals at all three islands are more likely to be formed. Here, we have tested this latter possibility in a laboratory cross between M and S individuals.

Though we do find evidence for transmission ratio distortion at two of the three islands, at least in females (see electronic supplementary material, table S1), the over-representation of alleles is small and in opposite directions at the two loci. At the 2L locus M alleles are passed on at significantly higher levels than expected, while at the 3L locus S alleles are over-represented in F_2 offspring (both approx. 53% observed versus 50% expected); there is no distortion at the X locus in either males or females. Transmission ratio bias in opposite directions should lead to increased mixing among M and S alleles, exactly the opposite pattern as to that observed in nature. These results also provide little support for the centromeric drive hypothesis [43,44]: we do see transmission ratio distortion (a predicted outcome of centromeric conflict), but it is not present at all three loci, and it is in different directions for the two loci at which it occurs in females. While the distortion we do observe could be due to a conflict between centromeric satellite DNA and DNA-binding proteins targeted to the centromere (or other changes to the centromeres), there is no evidence that this particular conflict plays any role in the isolation between M and S mosquitoes. A similar situation may be occurring in hybrids between the monkeyflowers *Mimulus guttatus* and *M. nasutus*, where evidence for strong meiotic drive is found at a single centromere [51] even though none of the mapped reproductive isolation loci co-occur with this centromere [52–54].

Previous studies have found no detectable intrinsic postzygotic reproductive isolation between M and S in F_1 or backcross individuals raised in the laboratory [55], and we also do not find any decrease in the reproductive success of F_1s or the survival of F_2s here. This result suggests that there is likely to be some form of extrinsic reproductive isolation, either in the survival or reproductive success of hybrid individuals in natural populations. It has been found that male and female mosquitoes of *A. gambiae* match the vibration frequency of their antennae in order to mate, and that M and S form mosquitoes will preferentially flight-tone match with individuals of their own form [31]. It is possible that these interactions—which happen at close range—or interactions that bring mosquitoes into the same swarm in the first place, are disrupted in the laboratory environment, such that hybrids deficient in flight-tone matching are not at a competitive disadvantage. There are also known to be differences between forms in the survival of larvae in temporary versus permanent pools of water, largely due to differences in the time to develop and the ability to avoid predators [27,56]. If hybrids represent unfit intermediates in either of these traits, such extrinsic deficiencies may only be manifested in the field. Finally, it may also be the case that the laboratory strains used here differ in some unknown way from M and S individuals in the wild, although this concern is somewhat alleviated when testing for intrinsic, as opposed to extrinsic, incompatibilities. Further crosses, among many different wild-caught strains, will be necessary to determine whether there is anything unique about the particular cross we have carried out.

(b) *Implications for speciation in* A. gambiae

Models of speciation with gene flow require that different loci have different abilities to introgress after initial hybrid matings. Alleles at loci conferring higher fitness in one environment or genetic background are not expected to introgress between species, while any region of the genome that does not determine differential fitness between populations may freely introgress, as long as recombination uncouples these regions from selected loci. These models therefore suggest that gene flow is most probable in regions farther away from selected loci, where recombinant gametes are most probable.

When applying these models to systems in which multiple loci determine differential fitness between populations, however, even the probability of introgression for loci unlinked to the selected ones can be quite low because they can only introgress on gametes that have the correct combination of alleles [42]. This is especially true in systems such as *A. gambiae*, where near-perfect associations between M and S alleles at all three islands are maintained. For example, if an F_1 hybrid mates with a parental individual of either type, in a one-locus model we expect that only 50 per cent of gametes will pass along the 'correct' allele (i.e. the one matching the parent's allele at that locus). In a two-locus model, this proportion goes down to 25 per cent, and in a three-locus (autosomal) model it goes down to 12.5 per cent. Assuming a rate of hybridization of 1 per cent—meaning that F_1 hybrids are formed in 1 per cent of all matings—the effective rate of gene flow at even unlinked markers could be as low as 0.125 per cent ($= 0.01 \times 0.125$) in a three-locus model if the 'incorrect' multi-locus genotypes are lethal, which is a very low rate. If there are more than three loci that show perfect association with one another [34,35], this makes the effective rate of gene flow even lower, as the number of recombinant individuals containing the correct combination of M or S alleles at this many loci will be vanishingly small (cf. [9]). It should also be noted that the available evidence suggests that there is no observed difference in the frequency of hybrid individuals among larvae sampled from pools and adults [57]. These calculations assume very strong selection against recombinant genotypes, but this would have to be the case in order to explain the observed patterns of disequilibrium among alleles in the islands if hybridization occurs approximately 1 per cent of the time. Strong transmission ratio distortion at multiple islands—in the same direction—would allow for biased co-transmission of the islands in the face of gene flow and a lessening of the selective load. However, the only bias we observe is small and in different directions for two of the three islands.

Together, these results suggest a viable alternative hypothesis for the observed patterns of heterogeneous differentiation across the M and S genomes. To be explicit, we take the original model (i.e. speciation-with-gene-flow) to posit that the low levels of differentiation seen across the vast majority of the genome are due to ongoing gene flow, with regions of high differentiation containing loci refractory to introgression [2]. The alternative hypothesis (i.e. low-gene-flow) posits that M and S have largely stopped exchanging genes, with little to no gene flow [7,8]. In this scenario, the lack of differentiation across much of the genome is due to shared ancestral polymorphisms and not introgression. Regions of high differentiation represent loci at which advantageous alleles have arisen and fixed in the two sub-species, but these differences may or may not be directly involved in reproductive isolation between the two. Although our data cannot by themselves distinguish between the widely invoked model of speciation-with-gene-flow and an alternative model with low-gene-flow, below we reconsider multiple lines of evidence in light of these two models.

The main obstacle to the alternative model is the relatively high rate of hybridization found across most of the range in which M and S co-occur in nature (approx. 1%; [8,23,39–41]), as the low-gene-flow model implies that F1s must have very low fitness, i.e. speciation between M and S is nearly complete. However, it is important to recognize that all but one [8] of these previous studies identified hybrids only by genotyping the island found on the X chromosome, which means that they were unable to distinguish F_1s from later-generation recombinants. If inter-form mating occurs at an appreciable frequency, but F_1 hybrid individuals have low fitness, then there could be hybridization without gene flow. Of the five hybrid individuals found by White *et al.* [8], three were F_1s, with the two non-F_1s homozygous at two of three loci (they were $X^{MM}2L^{MS}3L^{MM}$ and $X^{SS}2L^{SS}3L^{MS}$). These numbers are too small to make any statistical conclusions, but they at least suggest an over-representation of F_1 individuals; larger samples of hybrid individuals will have to be collected in order to make conclusions about the possible low fitness of F_1s.

On the other hand, highly advantageous insecticide resistance alleles definitely appear to introgress between M and S [58–60]. Whether introgression at universally advantageous loci—that may endow normally less-fit hybrids with high fitness—is representative of the rest of the genome is unknown and will have to await careful analysis of many more loci.

Both models are also consistent with patterns of divergence across the genome. In the low-gene-flow model, the speciation islands represent regions at which new, possibly linked, advantageous mutations have arisen independently in the two sub-species. These mutations arose on different haplotypes drawn from the same ancestral pool of variation, driving them to fixation. Thus, though the number of fixed differences in such regions will be higher than at loci not under selection, the absolute level of divergence between the two haplotypes will be approximately equal to the level of divergence between any two random haplotypes in the ancestral population (cf. [61]). Under ongoing speciation-with-gene-flow, absolute variation in the islands is expected to be proportional to the time since the two species split. Though ancestral levels of variation are not known, there is not greater single-nucleotide divergence between M and S in the speciation islands relative to divergence between any two particular M or S haplotypes taken from neighbouring regions [2,8]. Under a low-gene-flow model there is nothing particularly special about the three speciation islands detected by previous microarray studies, and many smaller ‘islands’ may have been missed due to technical limitations. Because recombination appears to be lower in the centromeres of *A. gambiae* [62]—where the islands are found—individual selective sweeps will affect longer stretches of the genome, making them easier to detect given the relatively low resolution of the *A. gambiae* microarray (which was not designed as a tiling array). Consistent with a low-gene-flow model, recent whole-genome sequencing has been able to detect many smaller regions of high differentiation [34,35].

The nature of divergence between M and S forms of *A. gambiae* appears to be more complicated than was implied by initial whole-genome studies [2,32]. Though the split between the two forms appears to have been relatively recent [26,63], the initial events contributing to differences in niche preference, niche adaptation and assortative mating have been difficult to identify [64]. In the process that starts with two completely inter-fertile populations and ends with two distinct species, M and S may be much closer to the ‘finish line’ than the starting line. One of the most important outstanding questions revolves around the amount of current realized gene flow across the M and S genomes. Though a growing literature has focused on how ‘genomic islands of speciation’ are generated, the models all used speciation-with-gene-flow, despite the difficulties inherent to such models [42]. It may be that few of these systems are in migration-drift equilibrium because there has not been sufficient time, making it hard to distinguish between models with and without gene flow [61]. While the data presented here are consistent with a low-gene-flow model, further determining the processes by which these lineages split will have to be done by analyses that can distinguish ancestral polymorphism from migration, and that can provide a historical time-frame for these processes [65,66].

We thank M. Kern for assistance in the laboratory and L. Moyle, P. Nosil, J. Feder, M. Noor and two anonymous reviewers for comments that helped to improve the manuscript. This work was funded by National Institutes of Health grant R01-AI63508 to N.J.B.

REFERENCES

1 Nosil, P., Funk, D. J. & Ortiz-Barrientos, D. 2009 Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* **18**, 375–402. (doi:10.1111/j.1365-294X.2008.03946.x)

2 Turner, T. L., Hahn, M. W. & Nuzhdin, S. V. 2005 Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* **3**, e285. (doi:10.1371/journal.pbio.0030285)

3 Wu, C. I. 2001 The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865. (doi:10.1046/j.1420-9101.2001.00335.x)

4 Wu, C. I. & Ting, C. T. 2004 Genes and speciation. *Nat. Rev. Genet.* **5**, 114–122. (doi:10.1038/nrg1269)

5 Barton, N. H. 2006 Evolutionary biology: how did the human species form? *Curr. Biol.* **16**, R647–R650. (doi:10.1016/j.cub.2006.07.032)

6 Wakeley, J. 2008 Complex speciation of humans and chimpanzees. *Nature* **452**, E3–E4. (doi:10.1038/nature06805)

7 Turner, T. L. & Hahn, M. W. 2010 Genomics islands *of* speciation or genomics islands *and* speciation? *Mol. Ecol.* **19**, 848–850. (doi:10.1111/j.1365-294X.2010.04532.x)

8 White, B. J., Cheng, C. D., Simard, F., Costantini, C. & Besansky, N. J. 2010 Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Mol. Ecol.* **19**, 925–939. (doi:10.1111/j.1365-294X.2010.04531.x)

9 Barton, N. & Bengtsson, B. O. 1986 The barrier to genetic exchange between hybridizing populations. *Heredity* **57**, 357–376. (doi:10.1038/hdy.1986.135)

10 Harrison, R. G. 1990 Hybrid zones: windows on evolutionary processes. In *Oxford surveys in evolutionary biology* (eds J. Antonovics & D. Futuyma), pp. 69–128. Oxford, UK: Oxford University Press.

11 Rieseberg, L. H., Whitton, J. & Gardner, K. 1999 Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics* **152**, 713–727.

12 Gompert, Z., Parchman, T. L. & Buerkle, C. A. 2012 Genomics of isolation in hybrids. *Phil. Trans. R. Soc. B* **367**, 439–450. (doi:10.1098/rstb.2011.0196)

13 Payseur, B. A., Krenz, J. G. & Nachman, M. W. 2004 Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. *Evolution* **58**, 2064–2078.

14 Teeter, K. C. *et al.* 2008 Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Res.* **18**, 67–76. (doi:10.1101/gr.6757907)

15 Geraldes, A., Ferrand, N. & Nachman, M. W. 2006 Contrasting patterns of introgression at X-linked loci across the hybrid zone between subspecies of the European rabbit (*Oryctolagus cuniculus*). *Genetics* **173**, 919–933. (doi:10.1534/genetics.105.054106)

16 Baxter, S. W. *et al.* 2010 Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in the *Heliconius melpomene* clade. *PLoS Genet.* **6**, e1000794. (doi:10.1371/journal.pgen.1000794)

17 Counterman, B. A. *et al.* 2010 Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in *Heliconius erato*. *PLoS Genet.* **6**, e1000796. (doi:10.1371/journal.pgen.1000796)

18 Kronforst, M. R., Young, L. G., Blume, L. M. & Gilbert, L. E. 2006 Multilocus analyses of admixture and introgression among hybridizing *Heliconius* butterflies. *Evolution* **60**, 1254–1268.

19 Nadeau, N. J. *et al.* 2012 Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Phil. Trans. R. Soc. B* **367**, 343–353. (doi:10.1098/rstb.2011.0198)

20 della Torre, A., Fanello, C., Akogbeto, M., Dossou-yovo, J., Favia, G., Petrarca, V. & Coluzzi, M. 2001 Molecular evidence of incipient speciation within *Anopheles gambiae* s.s. in West Africa. *Insect Mol. Biol.* **10**, 9–18. (doi:10.1046/j.1365-2583.2001.00235.x)

21 Favia, G., Lanfrancotti, A., Spanos, L., Siden-Kiamos, I. & Louis, C. 2001 Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of *Anopheles gambiae* s.s. *Insect Mol. Biol.* **10**, 19–23. (doi:10.1046/j.1365-2583.2001.00236.x)

22 Gentile, G., Slotman, M., Ketmaier, V., Powell, J. R. & Caccone, A. 2001 Attempts to molecularly distinguish cryptic taxa in *Anopheles gambiae* s.s. *Insect Mol. Biol.* **10**, 25–32. (doi:10.1046/j.1365-2583.2001.00237.x)

23 della Torre, A., Tu, Z. & Petrarca, V. 2005 On the distribution and genetic differentiation of *Anopheles gambiae* s.s. molecular forms. *Insect Biochem. Mol. Biol.* **35**, 755–769. (doi:10.1016/j.ibmb.2005.02.006)

24 Lehmann, T. & Diabate, A. 2008 The molecular forms of *Anopheles gambiae*: a phenotypic perspective. *Infect. Genet. Evol.* **8**, 737–746. (doi:10.1016/j.meegid.2008.06.003)

25 della Torre, A., Costantini, C., Besansky, N. J., Caccone, A., Petrarca, V., Powell, J. R. & Coluzzi, M. 2002 Speciation within *Anopheles gambiae*—the glass is half full. *Science* **298**, 115–117. (doi:10.1126/science.1078170)

26 Crawford, J. E. & Lazzaro, B. P. 2010 The demographic histories of the M and S molecular forms of *Anopheles gambiae* s.s. *Mol. Biol. Evol.* **27**, 1739–1744. (doi:10.1093/molbev/msq070)

27 Diabate, A. *et al.* 2005 Larval development of the molecular forms of *Anopheles gambiae* (Diptera: Culicidae) in different habitats: a transplantation experiment. *J. Med. Entomol.* **42**, 548–553. (doi:10.1603/0022-2585(2005)042[0548:LDOTMF]2.0.CO;2)

28 Diabate, A., Dabire, R. K., Heidenberger, K., Crawford, J., Lamp, W. O., Culler, L. E. & Lehmann, T. 2008 Evidence for divergent selection between the molecular forms of *Anopheles gambiae*: role of predation. *BMC Evol. Biol.* **8**, 5. (doi:10.1186/1471-2148-8-5)

29 Diabate, A., Dao, A., Yaro, A. S., Adamou, A., Gonzalez, R., Manoukis, N. C., Traore, S. F., Gwadz, R. W. & Lehmann, T. 2009 Spatial swarm segregation and reproductive isolation between the molecular forms of *Anopheles gambiae*. *Proc. R. Soc. B* **276**, 4215–4222. (doi:10.1098/rspb.2009.1167)

30 Tripet, F., Toure, Y. T., Taylor, C. E., Norris, D. E., Dolo, G. & Lanzaro, G. C. 2001 DNA analysis of transferred sperm reveals significant levels of gene flow between molecular forms of *Anopheles gambiae*. *Mol. Ecol.* **10**, 1725–1732. (doi:10.1046/j.0962-1083.2001.01301.x)

31 Pennetier, C., Warren, B., Dabire, K. R., Russell, I. J. & Gibson, G. 2010 'Singing on the wing' as a mechanism for species recognition in the malarial mosquito *Anopheles gambiae*. *Curr. Biol.* **20**, 131–136. (doi:10.1016/j.cub.2009.11.040)

32 Stump, A. D., Fitzpatrick, M. C., Lobo, N. F., Traore, S., Sagnon, N. F., Costantini, C., Collins, F. H. & Besansky, N. J. 2005 Centromere-proximal differentiation and speciation in *Anopheles gambiae*. *Proc. Natl Acad. Sci. USA* **102**, 15 930–15 935. (doi:10.1073/pnas.0508161102)

33 Turner, T. L. & Hahn, M. W. 2007 Locus- and population-specific selection and differentiation between incipient species of *Anopheles gambiae*. *Mol. Biol. Evol.* **24**, 2132–2138. (doi:10.1093/molbev/msm143)

34 Lawniczak, M. K. N. *et al.* 2010 Widespread islands of divergence in *Anopheles gambiae* revealed by whole genome sequences. *Science* **330**, 512–514. (doi:10.1126/science.1195755)

35 Neafsey, D. E. *et al.* 2010 Complex gene flow boundaries among sympatric *Anopheles* vector mosquito populations revealed by genome-wide SNP genotyping. *Science* **330**, 514–517. (doi:10.1126/science.1193036)

36 Lanzaro, G. C., Toure, Y. T., Carnahan, J., Zheng, L. B., Dolo, G., Traore, S., Petrarca, V., Vernick, K. D. & Taylor, C. E. 1998 Complexities in the genetic structure of *Anopheles gambiae* populations in west Africa as revealed by microsatellite DNA analysis. *Proc. Natl Acad. Sci. USA* **95**, 14 260–14 265. (doi:10.1073/pnas.95.24.14260)

37 Wang, R., Zheng, L. B., Toure, Y. T., Dandekar, T. & Kafatos, F. C. 2001 When genetic distance matters: measuring genetic differentiation at microsatellite loci in whole-genome scans of recent and incipient mosquito species. *Proc. Natl Acad. Sci. USA* **98**, 10 769–10 774. (doi:10.1073/pnas.191003598)

38 Wondji, C., Simard, F. & Fontenille, D. 2002 Evidence for genetic differentiation between the molecular forms M and S within the forest chromosomal form of *Anopheles gambiae* in an area of sympatry. *Insect Mol. Biol.* **11**, 11–19. (doi:10.1046/j.0962-1075.2001.00306.x)

39 Costantini, C. *et al.* 2009 Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in *Anopheles gambiae*. *BMC Ecol.* **9**, 16. (doi:10.1186/1472-6785-9-16)

40 Gentile, G., della Torre, A., Maegga, B., Powell, J. R. & Caccone, A. 2002 Genetic differentiation in the African malaria vector, *Anopheles gambiae* ss, and the problem of taxonomic status. *Genetics* **161**, 1561–1578.

41 Simard, F. *et al.* 2009 Ecological niche partitioning between *Anopheles gambiae* molecular forms in Cameroon: the ecological side of speciation. *BMC Ecol.* **9**, 17. (doi:10.1186/1472-6785-9-17)

42 Felsenstein, J. 1981 Skepticism towards Santa Rosalia, or why are there so few kinds of animals? *Evolution* **35**, 124–138. (doi:10.2307/2407946)

43 Henikoff, S., Ahmad, K. & Malik, H. S. 2001 The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102. (doi:10.1126/science.1062939)

44 Henikoff, S. & Malik, H. S. 2002 Selfish drivers. *Nature* **417**, 227. (doi:10.1038/417227a)

45 Lee, C. E. & Frost, B. W. 2002 Morphological stasis in the *Eurytemora affinis* species complex (Copepoda: Temoridae). *Hydrobiologia* **480**, 111–128. (doi:10.1023/A:1021293203512)

46 Fanello, C., Santolamazza, F. & della Torre, A. 2002 Simultaneous identification of species and molecular forms of the *Anopheles gambiae* complex by PCR-RFLP. *Med. Vet. Entomol.* **16**, 461–464. (doi:10.1046/j.1365-2915.2002.00393.x)

47 Guo, Z., Liu, Q. H. & Smith, L. M. 1997 Enhanced discrimination of single nucleotide polymorphisms by artificial mismatch hybridization. *Nat. Biotechnol.* **15**, 331–335. (doi:10.1038/nbt0497-331)

48 Wilkins, E. E., Howell, P. I. & Benedict, M. Q. 2006 IMP PCR primers detect single nucleotide polymorphisms for *Anopheles gambiae* species identification, Mopti and Savanna rDNA types, and resistance to dieldrin in *Anopheles arabiensis*. *Malaria J.* **5**, 125. (doi:10.1186/1475-2875-5-125)

49 Santolamazza, F., della Torre, A. & Caccone, A. 2004 Short report: a new polymerase chain reaction-restriction

fragment length polymorphism method to identify *Anopheles arabiensis* from *An. gambiae* and its two molecular forms from degraded DNA templates or museum samples. *Am. J. Trop. Med. Hyg.* **70**, 604–606.

50 Burnham, K. P. & Anderson, D. R. 2002 *Model selection and multimodel inference: a practical information-theoretic approach*. New York, NY: Springer.

51 Fishman, L. & Saunders, A. 2008 Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science* **322**, 1559–1562. (doi:10.1126/science.1161406)

52 Case, A. L. & Willis, J. H. 2008 Hybrid male sterility in *Mimulus* (Phrymaceae) is associated with a geographically restricted mitochondrial rearrangement. *Evolution* **62**, 1026–1039. (doi:10.1111/j.1558-5646.2008.00360.x)

53 Fishman, L. & Willis, J. H. 2006 A cytonuclear incompatibility causes anther sterility in *Mimulus* hybrids. *Evolution* **60**, 1372–1381.

54 Sweigart, A. L., Fishman, L. & Willis, J. H. 2006 A simple genetic incompatibility causes hybrid male sterility in *Mimulus*. *Genetics* **172**, 2465–2479. (doi:10.1534/genetics.105.053686)

55 Diabate, A., Dabire, R. K., Millogo, N. & Lehmann, T. 2007 Evaluating the effect of postmating isolation between molecular forms of *Anopheles gambiae* (Diptera: Culicidae). *J. Med. Entomol.* **44**, 60–64. (doi:10.1603/0022-2585(2007)44[60:ETEOPI]2.0.CO;2)

56 Gimonneau, G., Bouyer, J., Morand, S., Besansky, N. J., Diabate, A. & Simard, F. 2010 A behavioral mechanism underlying ecological divergence in the malaria mosquito *Anopheles gambiae*. *Behav. Ecol.* **21**, 1087–1092. (doi:10.1093/beheco/arq114)

57 Edillo, F. E., Toure, Y. T., Lanzaro, G. C., Dolo, G. & Taylor, C. E. 2002 Spatial and habitat distribution of *Anopheles gambiae* and *Anopheles arabiensis* (Diptera: Culicidae) in Banambani Village, Mali. *J. Med. Entomol.* **39**, 70–77. (doi:10.1603/0022-2585-39.1.70)

58 Djogbenou, L., Chandre, F., Berthomieu, A., Dabire, R., Koffi, A., Alout, H. & Weill, M. 2008 Evidence of introgression of the *ace-1*^{R} mutation and of the *ace-1* duplication in West African *Anopheles gambiae* s. s. *PLoS ONE* **3**, e2172. (doi:10.1371/journal.pone.0002172)

59 Etang, J. *et al.* 2009 Polymorphism of intron-1 in the voltage-gated sodium channel gene of *Anopheles gambiae* s.s. populations from Cameroon with emphasis on insecticide knockdown resistance mutations. *Mol. Ecol.* **18**, 3076–3086. (doi:10.1111/j.1365-294X.2009.04256.x)

60 White, B. J. *et al.* 2011 Adaptive divergence between incipient species of *Anopheles gambiae* increases resistance to *Plasmodium*. *Proc. Natl Acad. Sci. USA* **108**, 244–249. (doi:10.1073/pnas.1013648108)

61 Noor, M. A. F. & Bennett, S. M. 2009 Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity* **103**, 439–444. (doi:10.1038/hdy.2009.151)

62 Pombi, M., Stump, A. D., della Torre, A. & Besansky, N. J. 2006 Variation in recombination rate across the X chromosome of *Anopheles gambiae*. *Am. J. Trop. Med. Hyg.* **75**, 901–903.

63 Coluzzi, M., Sabatini, A., della Torre, A., Di Deco, M. A. & Petrarca, V. 2002 A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science* **298**, 1415–1418. (doi:10.1126/science.1077769)

64 Cassone, B. J., Mouline, K., Hahn, M. W., Pombi, M., Simard, F., Costantini, C. & Besansky, N. J. 2008 Differential gene expression in incipient species of *Anopheles gambiae*. *Mol. Ecol.* **17**, 2491–2504. (doi:10.1111/j.1365-294X.2008.03774.x)

65 Hey, J. & Nielsen, R. 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**, 747–760. (doi:10.1534/genetics.103.024182)

66 Hey, J. & Nielsen, R. 2007 Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl Acad. Sci. USA* **104**, 2785–2790. (doi:10.1073/pnas.0611164104)