Insect Molecular Biology (2005) **14**(4), 353–363

SINE insertion polymorphism on the X chromosome differentiates *Anopheles gambiae* molecular forms

M. J. Barnes*†, N. F. Lobo*, M. B. Coulibaly*, N′F. Sagnon‡, C. Costantini‡ and N. J. Besansky*

**Center for Tropical Disease Research and Training, Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA; and ‡Centre National de Recherche et Formation sur le Paludisme, Ouagadougou, Burkina Faso*

Abstract

Polymorphic SINE insertions can be useful markers for assessing population structure and differentiation. *Maque* is a family of SINE elements which, based on bioinformatic analysis, was suggested to have been active recently in *Anopheles gambiae*, the major vector of malaria. Here, we report the development of polymorphic *Maque* insertions as population genetic markers in *A. gambiae*, and the use of these markers to better characterize divergence on the X chromosome between *A. gambiae* M and S molecular forms in populations from Burkina Faso and Mali. Our data are consistent with the recent activity of *Maque*. Phylogenetic analysis suggests that at least two recently active lineages may have a role in mediating genome evolution. We found differences in element insertion frequency and sequence between the M and S populations analysed. Significant differentiation was observed between these two groups across a 6 Mb region at the proximal (centromeric) end of the X chromosome. Locus-specific F_{ST} values ranged from 0.14 to 1.00 in this region, yet were not significantly different from zero in more distal locations on the X chromosome; the trend was consistent in populations from both geographical locales suggesting that differentiation is not due to local adaptation. Strong differentiation between M and S at the proximal end of the X chromosome, but not outside this region, suggests the action of selection counteracting limited gene flow between these taxa and supports their characterization as incipient species.

Keywords: *Anopheles gambiae*, incipient species, *Maque* transposable elements, molecular forms, short interspersed repetitive elements (SINEs).

Introduction

Transposable elements (TEs) are well known for their ubiquitous presence in most eukaryotic genomes and their often substantial contribution to total nuclear DNA content (> 45% in humans). Growing evidence points to TEs as a potent force in genome plasticity and evolution (Kidwell & Lisch, 2002). TEs are responsible for a considerable amount of polymorphism and differentiation between populations and species, caused directly by new element insertions, as well as by mediating the generation of chromosomal inversions, deletions, duplications, heterochromatin, gene conversion, alternative splicing and other processes that can alter gene expression (Kidwell & Lisch, 2002). Because of this, TE polymorphisms can be powerful markers for inferring phylogenetic relationships, population structure, and intraspecific demographics (Shedlock & Okada, 2000; Batzer & Deininger, 2002; Biedler *et al.*, 2003).

TEs are classified into two broad groups: those whose transposition occurs through a cut-and-paste mechanism at the DNA level (transposons), and those that use an RNA intermediate (retrotransposons) (Craig, 2002). Among the latter are short (< 500 bp) interspersed elements (SINEs) that lack long-terminal repeats and coding potential. As a consequence of these deficiencies, SINEs have special advantages over most other TEs as population genetic markers. First, SINEs are considered 'homoplasy-free characters' under two assumptions: that the chance of independent insertions into the same site is remote and, as there is no known mechanism for SINE excision, the frequency of deletion is practically nil (Shedlock & Okada, 2000; Batzer & Deininger,

doi: 10.1111/j.1365-2583.2005.00566.x

Received 16 November 2004; accepted after revision 1 February 2005. Correspondence: Nora J. Besansky, Department of Biological Sciences, 317 Galvin Life Sciences Bldg, University of Notre Dame, Notre Dame, IN 46556–0369, USA. Tel.: +1 574 631 9321; fax: +1 574 631 3996; e-mail: nbesansk@nd.edu

†Present address: The Scripps Research Institute, La Jolla, CA, USA.

2002). Thus, at the population level, shared insertions at a locus are identical by descent and the absence of an element at that locus can be considered the ancestral state. Second, because SINEs are short, insertion polymorphisms at individual genomic locations can be assayed easily and rapidly by PCR. In genomes where SINE elements are both abundant and actively amplifying, the most recent SINE insertions are polymorphic among individuals in a population or species, as not enough time has passed for fixation (or loss) to occur (Hamada *et al*., 1998; Batzer & Deininger, 2002). Thus, the pattern of SINE insertion polymorphism within species can help illuminate recent evolutionary events, such as the origin of human ethnic groups (e.g. Novick *et al*., 1998; de Pancorbo *et al*., 2001; Comas *et al*., 2004). Similarly, SINE polymorphisms may help resolve complexities in the population genetic structure of the mosquito *Anopheles gambiae*.

A. gambiae is the principal vector of human malaria in Africa south of the Sahara, which suffers 90% of the fatal cases worldwide (Breman *et al*., 2004). In West Africa there are populations of *A. gambiae* that are ecologically differentiated and at least partially reproductively isolated, suggestive of ongoing speciation (Powell *et al*., 1999; Tripet *et al*., 2001; della Torre *et al*., 2002). When defined on the basis of fixed sequence differences in ribosomal DNA (rDNA), *A. gambiae* can be divided into two groups, known as M and S molecular forms, that roughly correspond to ecologically differentiated populations in parts of West Africa (reviewed in della Torre *et al*., 2002). As no other fixed differences (morphological, chromosomal or molecular) have been discovered despite intensive screens, spirited debate has centred on the taxonomic status of the molecular forms and, more importantly, whether or not they should be considered as entities on independent evolutionary trajectories (Gentile *et al*., 2002; della Torre *et al*., 2002). This question is of fundamental importance to the study of speciation, but it also has applied relevance, as it bears on vector-based strategies to control malaria. One roadblock to progress is the difficulty in finding strong enough signals to distinguish alternative hypotheses – molecular forms as polymorphic components of a single species, or as emerging species – above a common background of shared ancestral polymorphisms and limited ongoing gene flow. To date, strong molecular evidence for population subdivision that correlates with evidence from other sources (ecological measurements and chromosomal inversion polymorphism), is limited to four loci (but see Wondji *et al*., 2002). One of these loci is on chromosome 2L, associated with insecticide resistance (Gentile *et al*., 2004; and references therein). The other three loci (including the rDNA) are at the proximal (centromeric) end of the X chromosome (Wang *et al*., 2001; Lehmann *et al*., 2003). We hypothesized that this pattern is not due to random chance, but rather that this region of the X chromosome may carry loci directly involved in behavioural and ecological differences between M and S. If true, differentiation may be significant near these loci, but not necessarily in other regions of the X chromosome. SINE insertion polymorphism could provide a powerful and convenient indicator of differentiation. As part of an ongoing effort to explore the extent and nature of sex-linked differentiation between M and S (Stump *et al*., 2005), we screened for polymorphic insertions of the *A. gambiae* SINE element *Maque* (Tu, 2001) on the X chromosome in natural populations. Here we report the successful development of seven independent *Maque* loci as markers, and demonstrate their utility in differentiating populations of M and S from Burkina Faso and Mali.

Results and Discussion

Maque *insertion polymorphism*

Initially, primers were designed to flank fifteen distinct *Maque* insertions detected *in silico* on the X chromosome of the sequenced *A. gambiae* PEST strain. Of these fifteen loci, seven were polymorphic for the presence/absence of a *Maque* element in subsamples (twenty of each molecular form) from our study population in Mali. For easy reference, these seven loci (Table 1) were designated '*Mq*' followed by

Table 1. *Maque* loci investigated

Locus*	Scaffold	Chromosome coordinate†	Forward primer (5′–3′)	Reverse primer (5′–3′)	Annealing temp (T_m)	Length (+ *Mq*) (bp)
*Mq*4A-1	8846	3386165	AACAATAGCAGTTGTCTGTTCGAG	TGAATTGAAACAATAGTGTTGGTG	57	241
*Mq*1C-1	8847	11534424	CTGTTTCTGCGATCCTGTTA	AGTTCCATTGAAACATGTGC	57	433
*Mq*1A-1	8847	12400551	GATAGCGGTTCGAATGGTAG	TCCAGTTCGCATTGAAATTA	57	459
*Mq*1A-2	8847	12702249	AAGCTCGGATAGCACCAAAA	ACCTCAGTCGTGGGTCGTAT	57	484
*Mq*5B-1	8963	16166049	CGGTTAATCAGTGGTTGTCA	AAAGGCGACAACAGAAAATC	57	451
*Mq*5C-1	8811	18227328	AGATTGCAATATTTTTACATTTGA	CGACCCATGATGATTTTGT	57	595
*Mq*6–1	8975	21500827	AAAGATATTGTCCTGATTTTTGC	TTCAAATCGCAAACTTTCAC	54	440

†Ensembl Mosquito v24.2b.1 (1 September 2004)

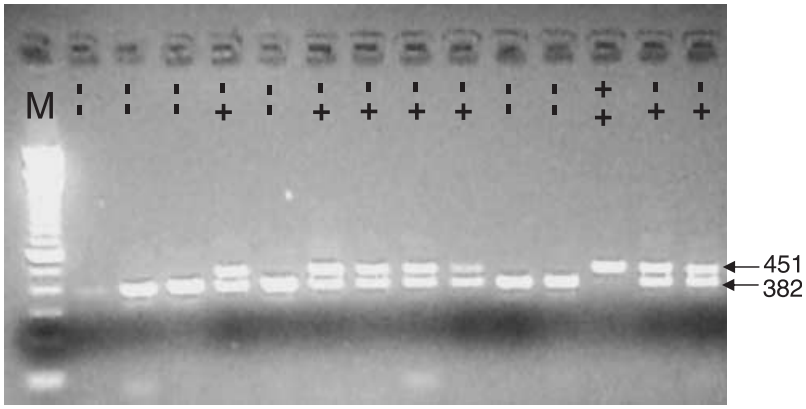

Figure 1. Example of PCR results from locus *Mq*5B-1 indicating the presence (+) or absence (–) of a *Maque* insertion on each X chromosome of *Anopheles gambiae* females collected from Mali. Arrows point to 'filled sites' (451 bp including target site duplication), and 'empty sites' (382 bp). M, 100 bp molecular weight marker.

Table 2. *Maque* insertion polymorphism on the X chromosome of *Anopheles gambiae* M and S

	Gene diversity				*Maque* allele frequency				Null frequency	
	M form		S form		M form		S form		M	S
Locus	Mali $N = 222$	Burkina $N = 49$	Mali $N = 344$	Burkina $N = 55$	Mali $N = 222$	Burkina $N = 49$	Mali $N = 344$	Burkina $N = 55$	Burkina	
*Mq*4A-1	0.19	0.20	0.22	0.35	0.90	0.89	0.88	0.78	0.06	0.04
*Mq*1C-1	0.23	0.35	0.15	0.07	0.13	0.22	0.08	0.04	0.06	0.04
*Mq*1A-1	0.37	0.34	0.33	0.35	0.24	0.21	0.21	0.22	0.06	0.07
*Mq*1A-2	0.39	0.37	0.30	0.22	0.74	0.77	0.82	0.88	0.12	0.13
*Mq*5B-1	0.44	0.49	0.18	0.24	0.32	0.39	0.10	0.14	0.04	0.09
*Mq*5C-1	0.03	0.00	0.47	0.51	0.01	0.00	0.64	0.51	0.18	0.20
*Mq*6-1	0.00	0.00	0.00	0.07	1.00	1.00	0.00	0.04	0.00	0.00

N, sample size in number of chromosomes typed.

the number/letter corresponding to their chromosomal location on the cytogenetic map (Coluzzi *et al*., 2002), a dash and a unique number. Each of the seven loci were scored for *Maque* insertion polymorphism in the total population samples from Burkina Faso (M = 49; S = 55) and Mali (M = 111; S = 172). A representative result from locus *Mq*5B-1 in Mali is shown in Fig. 1.

Southern analysis was used to verify interpretation of alleles as containing or lacking a *Maque* insertion. With very few exceptions, *Maque* insertions behaved like dimorphic markers in these populations, with one smaller allele lacking an insertion ('empty site') and one larger allele containing an insertion ('*Maque*-filled site'). In scoring for insertion polymorphism, rare alleles differing in size from the canonical filled-site allele but confirmed by Southern analysis to be *Maque*-filled were pooled with the canonical allele. Unexpected results were obtained for loci *Mq*6-1 and *Mq*5C-1. Although only two allelic size classes were found, *Maque* insertions were detected in both. Sequence analysis (see below) revealed that these alleles differed by the copy number of *Maque* insertions (one vs. two copies for *Mq*6-1; two vs. three copies for *Mq*5C-1). For the purposes of population genetic analysis (reported in Tables 2 and 3), alternative alleles at these loci were treated as equivalent to filled and empty sites.

Table 3. Locus-by-locus differentiation between *Anopheles gambiae* M and S

	F_{ST}	
Locus	Mali	Burkina
*Mq*4A-1	0.00^{ns}	0.03^{ns}
*Mq*1C-1	0.01*	0.13**
*Mq*1A-1	0.00^{ns}	-0.02^{ns}
*Mq*1A-2	0.02*	0.02^{ns}
*Mq*5B-1	0.15***	0.14**
*Mq*5C-1	0.57***	0.49***
*Mq*6-1	1.00***	0.96***

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ^{ns}, not significant.

Table 2 provides estimates of polymorphism in M and S forms based on gene diversity (expected heterozygosity) and *Maque* allele frequency at each locus in both locales. These data revealed an intermediate level of polymorphism in both molecular forms at most loci, suggesting that they could be informative as population genetics markers. Gene

diversity and *Maque* frequency varied relatively little between locales within a form. Larger and more consistent differences were observed between forms.

No departures from Hardy–Weinberg equilibrium were detected in the samples from Mali (data not shown), except for one significantly positive F_{IS} value of 0.48 (indicating heterozygote deficit) at the *Mq*5C-1 locus in the S form ($P < 0.05$ after Bonferroni adjustment). This locus-specific result is most likely due to a high rate of null alleles at *Mq*5C-1 (see below), rather than genome-wide processes such as non-random mating or pooling of subpopulations (Wahlund effect). Because we sampled only males from Goundri village in Burkina Faso, we could not test for random mating in this population. However, no significant linkage disequilibrium was detected in either the Mali or the Burkina Faso samples for any pair-wise comparison of loci, consistent with random mating.

Null (non-amplifying) alleles are commonly described in microsatellite-based population studies of anophelines and other species (Lehmann *et al*., 1996; Dakin & Avise, 2004), and are due to polymorphism in the primer binding site(s). In a recent study of *Anopheles funestus* population structure in Burkina Faso, average null allele frequencies across sixteen microsatellite loci were estimated at 8% (Michel *et al*., 2005). Because anopheline males are hemizygous for the X chromosome, X-linked null alleles are immediately apparent and need not be estimated based on deviation from expected heterozygosity, as must be done for females. In a study of microsatellite variation on the X chromosome that employed the same male specimens from Burkina Faso as studied here, we observed average rates for null alleles of 5–6% across seventeen loci, ranging from 0–31% (Stump *et al*., 2005). Given this amount of nucleotide variation in flanking sequences, it is not surprising that detection of SINE insertions by PCR is subject to a similar rate of null alleles, an average of 7–8% across seven loci in both M and S molecular forms in the Burkina Faso sample (Table 2). Locus *Mq*5C-1 had a particularly high rate of null alleles, 18% in M and 20% in S, whereas no null alleles were detected at *Mq*6-1 in either form.

Differentiation inferred from Maque *insertion polymorphism*

In previous studies, significant and large estimates of differentiation between M and S (e.g. $F_{ST} = 0.31$; Lehmann *et al*., 2003) were found only at two microsatellite loci from the proximal end of the X chromosome (divisions 5D-6). We predicted that levels of *Maque* insertion polymorphism in M and S would follow the same trend. After verifying that F_{ST} estimates of differentiation were not significantly different from zero within molecular forms from the two sampling locations in Mali (between Banambani and Moribabougou villages), we pooled each molecular form across these villages. Locus-by-locus pair-wise F_{ST} estimates of differentiation between M and S were derived independently for the Mali and Burkina Faso populations (Table 3). In both locales, differentiation was significant and very large at loci in divisions 5–6, as predicted. Furthermore, in this region F_{ST} values increased from ~0.15 to near fixation (i.e. ~1.0) moving proximally. At other loci, M and S were either undifferentiated or marginally differentiated, with the exception of locus *Mq*1C-1 in the Burkina Faso population.

Maque *sequence similarity at orthologous loci*

Assuming that all *Maque*-filled alleles at a given locus in M and S descended from a unique insertion event in a common ancestor, all *Maque* sequences at that locus from different individuals in the population would have been identical initially. As random nucleotide substitutions and insertion-deletion (indel) events steadily create new alleles at the locus, the amount of sequence variation should reflect time since insertion (or alternatively, time since the last selective sweep, bottleneck or gene conversion). Therefore, the pattern in which *Maque* sequence variation is distributed within/between M and S at a given locus may be informative about the degree of isolation between these molecular forms. Are most polymorphisms shared or private? How much sequence divergence exists between forms relative to the level of polymorphism within forms?

To gain further insight about differentiation between M and S at orthologous loci, sequences were determined directly from several randomly chosen 'filled-site' alleles from both molecular forms, using the hemizygous male specimens from Burkina Faso. Of the six loci that were sequenced (all except *Mq*1A-1), four contained a single *Maque* insertion. Two loci, *Mq*6-1 and *Mq*5C-1, contained as many as two or three tandem *Maque* insertions (Fig. 2). Their structure was unexpectedly complex, but informative about processes affecting *Maque* evolution. As these results bear on our interpretation of differentiation between M and S, we first describe their insertion structure before presenting a summary of sequence polymorphism and divergence at all loci.

At locus *Mq*6-1, in accord with the *A. gambiae* genome sequence derived from the PEST strain, the sequences of five individuals from the M molecular form all contained two tandem *Maque* elements, both flanked by distinct 10 bp direct repeats inferred to be target site duplications (Fig. 2). These five M sequences were invariant. The tandem elements, here referred to as #1 and #2, were apparently derived from independent insertions, as their sequence at the 5′- and 3′-ends and the target duplications differ. The sequences of five individuals from the S form were identical and contained only a single *Maque* insertion. In the absence of sequence information, a reasonable expectation might have been that the S allele of the *Mq*6-1 locus, with only a single insertion, was ancestral. However, careful sequence analysis revealed that the S allele was derived from the M allele by a deletion event associated with unequal crossing over between insertions #1 and #2. Evidence

Figure 2. Sequence alignment indicating structural and sequence variation at loci containing multiple *Maque* insertions in *Anopheles gambiae* M and S. Dashes represent gaps introduced to improve the alignment; dots (or vertical lines) represent sequence identity. Sequences in bold marked by asterisks are target duplications flanking *Maque* insertions. Arrows labelled #1, #2 and #3 indicate the location and numerical designation of a *Maque* insertion. The terminal CAA_n of *Maque* sequences are underlined. (A) Structure of locus *Mq*6-1. Below is a multiple alignment of *Maque* insertions #1 and #2 sequenced from locus *Mq*6–1. (B) Structure of locus *Mq*5C-1. Sequence in lower case lies outside *Maque* elements and their target duplications. Below is a multiple sequence alignment of *Maque* insertions #1, #2 and #3 sequenced from *Mq*5C-1.

for this can be seen in the lower part of Fig. 2A, in which insertions #1 and #2 are aligned to each other and to the S insertion. The single S insertion appears chimeric, in that it is identical to insertion #1 at the 5′ end and to insertion #2 at the 3′ tail. Because insertions #1 and #2 do not differ in the interior two-thirds of their sequence, it is not possible to precisely locate the breakpoint of the recombination event. What is clear is that the deletion involved the equivalent of one complete copy of *Maque* and one copy of each of the two 10 bp target duplications. Although we only sequenced five individuals from both M and S, taken together with the pattern revealed from PCR and Southern analysis, our data suggest that in the Burkina Faso and Mali populations, M and S are fixed (or nearly so) for these alternative alleles.

Ten S specimens were sequenced at locus *Mq*5C-1, representing two different allele size classes. Five identical sequences representing the larger size class (the S2 allele in Fig. 2B) contained three complete copies of *Maque* in tandem, flanked by the same 13 bp target duplications. The other five sequences representing the smaller S1 allele were identical except for one apparent recombinant with the larger S2 allele. The S1 allele contained only two complete *Maque* elements in tandem (corresponding to #2 and #3 in Fig. 2B), each flanked by the same 13 bp target duplications. Sequences determined from five M individuals at locus *Mq*5C-1 match the sequenced *A. gambiae* PEST genome (Holt *et al.*, 2002). These sequences, identical except for a singleton substitution, reveal two tandem *Maque* insertions: one complete (corresponding to #2 in Fig. 2B) and a second truncated at the 3′-end with respect to both the CAA_n tail and the expected target duplication (corresponding to #3). Prior to sequencing, we had interpreted S1 and M alleles to be identical, based on their size. In light of the sequence data, it appears that M and S1 alleles have several fixed differences. Therefore, extrapolating these results to the rest of the samples from Burkina Faso and Mali, differentiation at *Mq*5C-1 between M and S may be complete and the results reported in Table 3 at this locus are conservative underestimates.

The history of locus *Mq*5C-1 is more elusive than that of *Mq*6-1, but it clearly involved multiple recombination events (including insertions, gene conversion, unequal recombination and/or normal crossing over). Among *Maque* elements from the same position in the tandem array (#1, #2 or #3), the sequences are identical or nearly so, especially within molecular forms. However, sequence comparisons between *Maque* elements from different positions in the tandem array (e.g. #1 vs. #2) show correlated differences among copies that seem to reflect three different ancestries (Fig. 2B, bottom). Yet, the target duplications flanking each element are not only perfectly conserved in all sequences at this locus, they also are identical among tandem element copies. This suggests that the reiteration of *Maque* at this locus is not due to multiple *de novo* insertions, but may be the result of passive duplications by an unknown mechanism, one that included the element plus flanking target duplications. A possible explanation for the apparent lack of common ancestry among *Maque* elements at this locus is gene conversion, an alternative mechanism of recombination whose importance is becoming increasingly recognized, especially in SINE evolution (e.g. Kass *et al.*, 1995; Roy *et al.*, 2000). In the evolution of human *Alu* elements, gene conversions that occur only within an element have been shown to convert original insertions representing one *Alu* subfamily into sequences of another subfamily (Kass *et al.*, 1995). Alternative scenarios involving multiple *de novo* insertions cannot be ruled out definitively in the absence of sequence information at this locus from close relatives of *A. gambiae*. Further complexity is indicated by the truncation of one copy of *Maque* in M (#3).

Table 4. Nucleotide polymorphism and divergence of *Maque* insertions in *Anopheles gambiae* M and S molecular forms at four loci

Locus	π_M (%)	π_S (%)	Dxy (%)	Da (%)	Shared	Fixed
*Mq*4A-1	1.3	2.3	1.80	0.03	1	0
*Mq*1C-1	0.0	0.0	0.00	0.00	–	–
*Mq*1A-2	1.0	2.1	2.30	0.55	1	0
*Mq*5B-1	1.0	0.0	0.48	0.00	0	0

π_M and πS, average pair-wise nucleotide divergence per site in M and S, respectively; Dxy, average nucleotide substitutions between forms; Da, net average nucleotide substitutions between forms; Shared, number of polymorphisms shared between forms; Fixed, number of polymorphisms fixed between forms.

A general result from these two complex loci, *Mq*6-1 and *Mq*5C-1, is that variation between forms was far greater than variation within forms. Indeed, there were fixed differences between forms at these loci, which also happen to be nearest the centromere. Of the remaining four loci that were sequenced (from five individuals from both M and S per locus), there were no fixed differences. The little variation found was distributed equally between M and S; when divergence between forms (Dxy) was adjusted to account for polymorphism within each form, the net difference (Da) was zero. These results, summarized in Table 4, are consistent with the results based on allele size differences.

Maque *sequence diversity at paralogous loci*

SINE evolutionary dynamics are driven by one or more 'master' or source elements capable of retrotransposition (Shedlock & Okada, 2000; Batzer & Deininger, 2002). As mutations accumulate in a source element of a SINE family, these are propagated in all of its copies, allowing each subfamily to be recognized by diagnostic mutation(s). Given that we have focused on polymorphic (thus presumably relatively recent) *Maque* insertion events, we asked whether these independent insertions could be attributed to more than one source element or subfamily. From a multiple sequence alignment (Fig. 3) that included the *Maque*

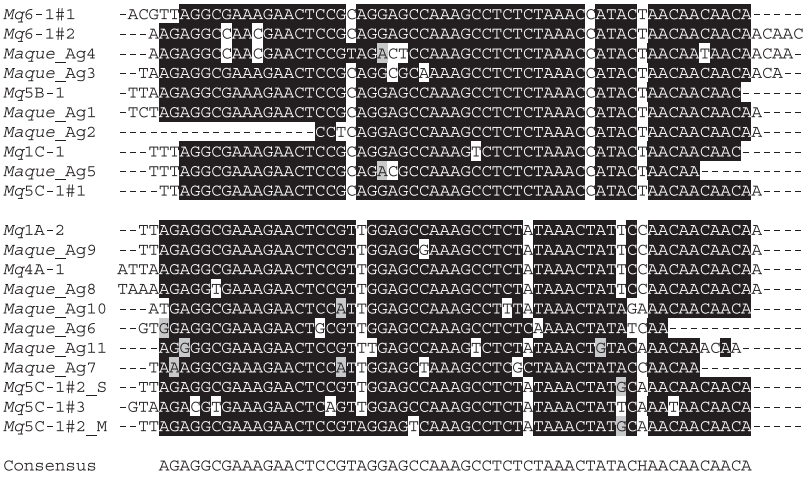

Figure 3. Multiple alignment of *Maque* elements sequenced from *Anopheles gambiae* M and S forms in the present study, and those reported by Tu (2001). Sequences used in the alignment represent the consensus of individual *Maque* insertions at each locus. Sequences comprising the two main clades identified in the phylogenetic analysis are separated by a blank line. Positions shaded in black are identical in > 50% of the aligned sequences.

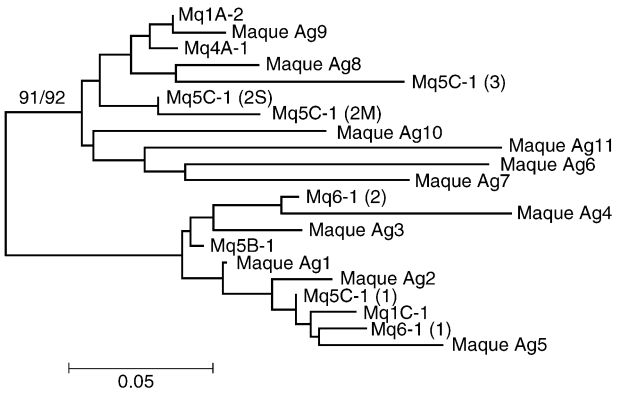

Figure 4. Phylogenetic relationships among *Maque* sequences inferred from the alignment in Fig. 3. Bootstrap support (maximum parsimony/ neighbor joining) is shown only at those nodes that were supported by > 75% of the replicates for both methods of inference. Branch length corresponds to number of substitutions.

insertions studied here and *Maque* sequences reported previously by Tu (2001), we reconstructed an unrooted tree from which two main clades could be inferred, corresponding to those identified by Tu (2001) (Fig. 4). Although there are sequence polymorphisms within each clade, the two clades are distinguished by fixed or strong frequency differences at 5 positions in the alignment. Unlike the present study, Tu (2001) neither specifically targeted the X chromosome nor focused on polymorphic *Maque* insertions. Therefore, the concordant results of these two studies suggest that two main source elements, or two families of source elements, may be responsible for most (all?) insertions of *Maque* in the *A. gambiae* genome. These source elements predate evolutionary divergence of M and S, as *Maque* sequences from M and S are interspersed in each clade.

Recent Maque *activity*

In human populations, the vast majority of the more than one million *Alu* SINE insertions are no longer polymorphic; sufficient time has passed since retroposition that they have become fixed for the presence of the insertion in human genomes. Only 0.1% of *Alu* elements in humans (~1200) have mobilized so recently that insertion polymorphism can be useful for inference of population structure (Batzer & Deininger, 2002). Based on evidence from the X chromosome of *A. gambiae*, a much higher proportion of *Maque* SINE elements – nearly 50% of those screened in this study – are polymorphic for presence/absence at a locus. Such insertion polymorphism is, by itself, indicative of recent amplification of *Maque* elements. Additional evidence is provided by perfectly conserved target duplications, with the exception of one polymorphic position in one of the direct repeats of *Mq*1A-2 and *Mq*4A-1 (data not shown). Recent activity can also be reflected in the number of identical copies dispersed throughout the genome, with the caveat that *Maque* is prone to 5′-truncation during the retroposition or integration process. When any one of the sequenced *Maque* insertions was used to query the *A. gambiae* genome using Blastn under low stringency, ~100 hits were returned each time by the Ensembl genome browser (data not shown). When the stringency was increased to 'exact match' and the results were parsed for matches that extended the full-length of the query sequence, most searches yielded one to four hits, two yielded eight to ten hits (*Mq*5B-1, *Mq*5C-1#2S), and another two yielded thirty-four to forty-one hits (*Mq*6-1#1, #2) distributed across the genome (including unmapped scaffolds). We interpret these results to mean that the most recently active, or the most highly transcribed, elements among those studied here are likely represented by the two from locus *Mq*6-1.

Limitations of Maque *elements as markers*

The relatively recent activity and intermediate levels of polymorphism exhibited by most of the *Maque* loci employed in this study are positive features for population genetics markers. However, *Maque* markers are not without potential limitations,

depending upon the application. As first reported by Tu (2001), these SINEs do not seem to insert randomly into the genome; instead, they are found disproportionately often near genes. Of the seven loci, four are within known or predicted transcripts, less than 2 kb from the nearest exon. Due to the incomplete annotation of the *A. gambiae* genome, this number is underestimated by at least one. We found by tBlastx searches of GenBank that *Mq*6-1 is located within the intron of a homologue of the *Drosophila forked* gene, although no gene was predicted in this region. The propensity to insert near genes may weaken the assumption that *Maque* insertions generally behave as neutral population genetic markers. In this regard, it is worth noting that locus *Mq*1C-1, at which an unexpectedly large and significant F_{ST} value was measured between M and S in Burkina Faso, is only 2.5 kb away from a predicted exon whose potential product contains a peritrophin A domain characteristic of chitin binding proteins (data not shown). Further evidence for non-random distribution of *Maque* in the *A. gambiae* genome was the discovery of two loci with tandem *Maque* elements. It is likely that 'coclustering' of *Maque* elements is an outcome of unequal crossing over and gene conversion (Jurka *et al.*, 2004), as opposed to independent insertions into virtually identical positions in the genome. Another potential limitation to *Maque* markers is their dimorphic character, reducing power to detect differentiation relative to microsatellite markers that have multiple alleles at each locus. In this respect, they resemble single nucleotide polymorphisms (SNPs) which are typically dimorphic in a population. As with SNPs, this limitation can be overcome by using an adequately large number of loci. It can also be overcome, as demonstrated here, by focusing on recently active elements. Finally, although there is no known excision mechanism analogous to that of retrotransposons with LTRs or transposons with inverted repeats, the two loci characterized by coclustered *Maque* elements were both subject to deletions that seem to have occurred secondarily. Such complex dynamics were not misleading in our population genetic study of M and S, where we did not infer ancestral-descendent relationships. However, determining the polarity of characters is important in phylogenetic reconstruction. Based on size alone (without accompanying sequence information), the deleted alleles could have been mistaken for the ancestral state. This suggests that caution should be exercised in making assumptions about the ancestral state at a locus in the absence of sequencing a subset of alleles. This problem is of special concern if the pattern of *Maque* insertions is used as a basis for phylogenetic reconstruction of *A. gambiae* populations or species in the *A. gambiae* complex.

Conclusions

In this study, we set out to develop polymorphic SINE loci as markers for studying the pattern of differentiation between *A. gambiae* molecular forms M and S on the X chromosome. Despite their limitations, *Maque* markers proved very sensitive in this application. *Maque* insertion polymorphism and sequence data revealed a trend of increasing differentiation toward fixed differences between M and S forms, from distal to proximal locations on the X chromosome. It is worth emphasizing that the differentiation between M and S is due to differences in the frequency of insertions common to M and S, and also to structural differences in alleles circulating in these forms. These structural changes were caused by secondary deletions or passive duplications (i.e. at *Mq*5C-1 and *Mq*6-1), rather than by form-specific insertions. As an indication of the superior resolution offered by recently active *Maque* elements, variation in the sequence flanking the elements in this study yielded no sign of differentiation, even at the proximal end of the X chromosome, with the exception of two fixed differences flanking locus *Mq*6-1.

The data from Mali and Burkina Faso are consistent, indicating that differentiation between M and S at the base of the X chromosome (divisions 5B-6) is not specific to a single geographical locale. Indeed, a genome-wide study of *A. gambiae* from a larger region including Gabon, Ghana, Mali and the Democratic Republic of Congo found strong differentiation between M and S in X chromosome division 5D at microsatellite locus H678 (Lehmann *et al.*, 2003). Taken together with this previous study, our results also demonstrate that the elevated level of divergence seen on the X chromosome is not specific to a single class of markers with a given set of evolutionary constraints; a parallel study using X-linked microsatellite markers in Burkina Faso revealed the same pattern of differentiation between M and S (Stump *et al.*, 2005). From the available data, the region of the 22 Mb X chromosome that shows large and significant differentiation between M and S extends at least from division 5B (*Mq*5B-1) through to division 6 (*Mq*6-1 and beyond to rDNA), a region encompassing 6 Mb. Outside of this region of the X chromosome, and on the autosomes, little differentiation between M and S has been detected. This pattern is consistent with the hypothesis that at the base of the X chromosome, one or more genes may be under relatively strong differential selection in the face of limited ongoing gene flow between M and S. Given that the X chromosome, even at its proximal end, has been shown to recombine at a moderate rate (Zheng *et al.*, 1996), and that evidence is lacking for linkage disequilibrium between *Maque* loci or between microsatellite loci in this chromosome region (Stump *et al.*, 2005), the size of this region suggests that multiple genes may underlie the genetic and ultimately the ecological and behavioural differentiation between these taxa.

The presence of strong differentiation between M and S across 6 Mb at the base of the X chromosome does not by itself resolve the taxonomic debate surrounding *A. gambiae* M and S, nor the larger issue as to whether these taxa are

on independent evolutionary trajectories, although it does add considerable weight to the argument that M and S are incipient species. The almost complete absence of differentiation distal to this 6 Mb region conflicts with a whole-genome concept of speciation, in which there is complete reproductive isolation between nascent species. Instead, the juxtaposition of strongly differentiated and undifferentiated regions is consistent with a genic view of speciation (Wu, 2001; Wu & Ting, 2004). According to this view, residual gene flow between diverging species allows exchange of all genes except the small fraction that are directly contributing to differential adaptation or functional divergence, through their effects on behaviour or physiology. Further progress in understanding the forces underlying the divergence of M and S will require a more detailed unravelling of the population genomics of the X chromosome, carried out in the context of ecological and behavioural studies that can shed light on the functional significance of genetic divergence at the organismal and population level.

Experimental procedures

Mosquito samples

In Mali, collections were made in August 2000 from Moribabougou and Banambani, villages located near Bamako and about 20 km apart (for detailed description see Toure *et al*., 1998). Adult *A. gambiae s.l.* mosquitoes were collected by aspiration from the inside walls of houses. In Burkina Faso, indoor resting *A. gambiae s.l.* were collected by insecticide spray catches in September 2001 in Goundri village, Burkina Faso (for detailed description, see Costantini *et al*., 1996). In each of these West African villages, morphologically indistinguishable *A. gambiae* (M and S molecular forms) and *Anopheles arabiensis* are sympatric and often are present within the same samples. All specimens identified as *A. gambiae s.l.* using the keys of Gillies & De Meillon (1968) were placed in tubes and preserved at room temperature over desiccant. DNA was isolated from individual specimens using DNeasy Tissue Kits (Qiagen, Valencia, CA) or the Wizard SV 96 Genomic DNA Purification System (Promega, Madison, WI) and resuspended in 50 μl eluent buffer. Mosquitoes were identified to species and molecular form using a PCR-RFLP (restriction fragment length polymorphism) assay based on ribosomal DNA (Fanello *et al*., 2002).

Karyotype information was not available for these mosquitoes, and thus the proportion of Bamako chromosomal form potentially present together with the Savanna chromosomal form in the Mali samples of *A. gambiae* S is unknown. However, the lack of significant heterogeneity within or between samples of S from Mali, and the correspondence of results between Mali and Burkina Faso (which lacks the Bamako form) suggests that this did not pose a serious problem (see Results and Discussion).

Identification and detection of Maque *insertion polymorphism*

Maque elements on the X chromosome were located *in silico* by BLASTN searches of the *A. gambiae* genome using the *Maque* consensus sequence (Tu, 2001) as a query. For each insertion, the 500 bp flanking each side of the element were used to query the *A. gambiae* genome. For those loci in which at least one of the flanking regions returned a single hit to the genome (suggesting unique sequence), primers were designed to amplify across the element using Primer 3 software (Rozen & Skaletsky, 2000).

PCR was performed using a GeneAmp 9600 thermal cycler (Applied Biosystems, Foster City, CA). Each 25 μl reaction contained 1 pmol of each primer, 0.2 mM of each dNTP, 1.5 mM $MgCl_2$, 2.5 U Taq polymerase, and 1 μl of a 1 : 10 dilution of template DNA extracted from a single mosquito. Thermocycling conditions were: 94 °C denaturation for 10 min followed by thirty-five cycles of 94 °C for 30 s, 54–57 °C for 30 s and 72 °C for 30 s, with a final 72 °C extension of 7 min. Products were separated by electrophoresis on a 1.5% agarose gel.

For the Malian samples, presence or absence of the *Maque* insertion in PCR products was confirmed by Southern blot analysis: PCR products were transferred from the gel by capillary blotting to Hybond N+ membranes (Amersham Biosciences, Piscataway, NJ) in 20× standard saline citrate (SSC) buffer as recommended by the manufacturer, followed by cross-linking of DNA to membrane using an optimized UV Cross-linker (Fisher Biosciences). Two alternative detection methods were used. For loci *Mq*1A-1, *Mq*1C-1, *Mq*5B-1 and *Mq*6-1, separate probes were designed for the element (5′-biotin-labelled) and the flanking sequence (5′-fluoroscein labelled). Probe sequences are available on request. Biotin or fluoresceinlabelled probes (Invitrogen, Carlsbad, CA), at a concentration of 125 pmol/ml, were hybridized to the membrane at 45 °C and detected using streptavidin-AP or antifluorescein, respectively, followed by treatment with colorimetric substrate according to protocols included in the Multicolor Detection Kit (Roche, Basel, Switzerland). For loci *Mq*5C-1, *Mq*4A-1, and *Mq*1A-2, the CPD-Star Detection Kit (Roche) was used as recommended. Probes were designed for the 'empty' site of each locus by 3′-end labelling a PCR product lacking the *Maque* insertion (3′-DIG Labeling Kit, Roche). To detect the *Maque* element in each 'filled' site, a PCR product from locus *Mq*1A-1 containing a *Maque* insertion was 3′-end labelled and used as a probe.

Indices of polymorphism (*Maque* insertion frequency, gene diversity) and differentiation (F_{ST}) were computed using *MSA* 3.12 (Dieringer & Schlotterer, 2003). *Fstat* 2.9.3.2 (Goudet, 2001) was used to assess deviation from Hardy–Weinberg equilibrium at each locus in samples of M and S from Mali, as indicated by the inbreeding coefficient F_{IS}, as well as linkage disequilibrium between pairs of loci within M or S in Mali and Burkina Faso. Significance was tested using the randomization approach implemented in *Fstat*, with Bonferroni-adjusted *P*-values.

Sequence analysis

Up to five randomly chosen subsamples of PCR products representing 'empty' and 'filled' sites (where applicable) of each locus from hemizygous male Burkina Faso specimens were directly sequenced on both strands using ABI Big Dye Terminator v.2 chemistry and an ABI Prism 3700 DNA Analyser. Sequences were trimmed and evaluated with SEQMAN II software (DNASTAR), and aligned in CLUSTALX 1.83 (Thompson *et al*., 1997). They have been deposited in GENBANK under accession numbers AY871108–AY871174. Basic polymorphism and divergence statistics (π, Dxy, Da) (Nei, 1987) were estimated using DnaSP4 (Rozas *et al*., 2003). Phylogenetic analysis using neighbor-joining and maximum parsimony algorithms was carried out in Mega 2.1 (Kumar *et al*., 2001).

Acknowledgements

We thank the inhabitants of Goundri, Burkina Faso and Banambani/Moribabougou, Mali for their collaboration and the Directors and entomological staff of CNFRP and MRTC for their support. We thank Peter Andolfatto for helpful discussion. The suggestions of two anonymous reviewers improved the manuscript. This study was funded by grants from the NIH (AI44003) to NJB and the UNDP/World Bank/World Health Organization Special Program for Research and Training in Tropical Diseases (00892) to N'F.S. MJB was supported by a grant from the University of Notre Dame College of Science Honors Program; this work is the outcome of his honours thesis.

References

Batzer, M.A. and Deininger, P.L. (2002) Alu repeats and human genomic diversity. *Nat Rev Genet* **3**: 370–379.

Biedler, J., Qi, Y., Holligan, D., della Torre, A., Wessler, S. and Tu, Z. (2003) Transposable element (TE) display and rapid detection of TE insertion polymorphism in the *Anopheles gambiae* species complex. *Insect Mol Biol* **12**: 211–216.

Breman, J.G., Alilio, M.S. and Mills, A. (2004) Conquering the intolerable burden of malaria: what's new, what's needed: a summary. *Am J Trop Med Hyg* **71**: 1–15.

Coluzzi, M., Sabatini, A., della Torre, A., Di Deco, M.A. and Petrarca, V. (2002) A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science* **298**: 1415–1418.

Comas, D., Schmid, H., Braeuer, S., Flaiz, C., Busquets, A., Calafell, F., Bertranpetit, J., Scheil, H.G., Huckenbeck, W., Efremovska, L. and Schmidt, H. (2004) Alu insertion polymorphisms in the Balkans and the origins of the Aromuns. *Ann Hum Genet* **68**: 120–127.

Costantini, C., Li, S.G., della Torre, A., Sagnon, N., Coluzzi, M. and Taylor, C.E. (1996) Density, survival and dispersal of *Anopheles gambiae* complex mosquitoes in a west African Sudan savanna village. *Med Vet Entomol* **10**: 203–219.

Craig, N.L. (2002) Mobile DNA: an introduction. In *Mobile DNA II.* (Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M., eds.), pp. 3–11. ASM Press, Washington, DC.

Dakin, E.E. and Avise, J.C. (2004) Microsatellite null alleles in parentage analysis. *Hered* **93**: 504–509.

Dieringer, D. and Schlotterer, C. (2003) Microsatellite analyzer (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol Ecol Notes* **3**: 167–169.

Fanello, C., Santolamazza, F. and 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.

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

Gentile, G., Santolamazza, F., Fanello, C., Petrarca, V., Caccone, A. and della Torre, A. (2004) Variation in an intron sequence of the voltage-gated sodium channel gene correlates with genetic differentiation between *Anopheles gambiae* s.s. molecular forms. *Insect Mol Biol* **13**: 371–377.

Gillies, M.T. and De Meillon, B. (1968) *The Anophelinae of Africa South of the Sahara*. South African Institute for Medical Research, Johannesburg.

Goudet, J. (2001) FSTAT, a program to estimate and test gene diversities and fixation indices (Version 2.9.3). Available from http://www.unil.ch/izea/softwares/fstat.html

Hamada, M., Takasaki, N., Reist, J.D., Decicco, A.L., Goto, A. and Okada, N. (1998) Detection of the ongoing sorting of ancestrally polymorphic SINEs toward fixation or loss in populations of two species of charr during speciation. *Genetics* **150**: 301–311.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**: 129–149.

Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V.V. and Jurka, M.V. (2004) Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci USA* **101**: 1268–1272.

Kass, D.H., Batzer, M.A. and Deininger, P.L. (1995) Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. *Mol Cell Biol* **15**: 19–25.

Kidwell, M.G. and Lisch, D.R. (2002) Transposable elements as sources of genomic variation. In *Mobile DNA II.* (Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M., eds.), pp. 59–90. ASM Press, Washington, DC.

Kumar, S., Tamura, K., Jakobsen, I.B. and Nei, M. (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**: 1244–1245.

Lehmann, T., Hawley, W.A. and Collins, F.H. (1996) An evaluation of evolutionary constraints on microsatellite loci using null alleles. *Genetics* **144**: 1155–1163.

Lehmann, T., Licht, M., Elissa, N., Maega, B.T., Chimumbwa, J.M., Watsenga, F.T., Wondji, C.S., Simard, F. and Hawley, W.A. (2003) Population structure of *Anopheles gambiae* in Africa. *J Hered* **94**: 133–147.

Michel, A.P., Guelbeogo, W.M., Grushko, O., Schemerhorn, B.J., Kern, M., Willard, M.B., Sagnon, N'F., Costantini, C. and Besansky, N.J. (2005) Molecular differentiation between chromosomally defined incipient species of *Anopheles funestus*. *Insect Mol Biol* doi: 10.1111/j.1365-2583.2005.00568.x.

Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Novick, G.E., Novick, C.C., Yunis, J., Yunis, E., Antunez De Mayolo, P., Scheer, W.D., Deininger, P.L., Stoneking, M., York, D.S., Batzer, M.A. and Herrera, R.J. (1998) Polymorphic Alu insertions and the Asian origin of Native American populations. *Hum Biol* **70**: 23–39.

de Pancorbo, M.M., Lopez-Martinez, M., Martinez-Bouzas, C., Castro, A., Fernandez-Fernandez, I., De Mayolo, G.A., De Mayolo, A.A., De Mayolo, P.A., Rowold, D.J. and Herrera, R.J. (2001) The Basques according to polymorphic Alu insertions. *Hum Genet* **109**: 224–233.

Powell, J.R., Petrarca, V., della Torre, A., Caccone, A. and Coluzzi, M. (1999) Population structure, speciation, and introgression in the *Anopheles gambiae* complex. *Parassitologia* **41**: 101–113.

Roy, A.M., Carroll, M.L., Nguyen, S.V., Salem, A.H., Oldridge, M., Wilkie, A.O., Batzer, M.A. and Deininger, P.L. (2000) Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res* **10**: 1485–1495.

Rozas, J., Sanchez-Delbarrio, J.C., Messeguer, X. and Rozas, R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.

Rozen, S. and Skaletsky, H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. In *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. (Krawetz, S. and Misener, S., eds.), pp. 365–386. Humana Press, Totowa, NJ.

Shedlock, A.M. and Okada, N. (2000) SINE insertions: powerful tools for molecular systematics. *Bioessays* **22**: 148–160.

Stump, A.D., Shoener, J.A., Costantini, C., Sagnon, N.F. and Besansky, N.J. (2005) Sex-linked differentiation between incipient species of *Anopheles gambiae*. *Genetics* **169**: 1509–1519.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **24**: 4876–4882.

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

Toure, Y.T., Petrarca, V., Traore, S.F., Coulibaly, A., Maiga, H.M., Sankare, O., Sow, M., Dideco, M.A. and Coluzzi, M. (1998) The distribution and inversion polymorphism of chromosomally recognized taxa of the *Anopheles gambiae* complex in Mali, West Africa. *Parassitologia* **40**: 477–511.

Tripet, F., Toure, Y.T., Taylor, C.E., Norris, D.E., Dolo, G. and 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.

Tu, Z. (2001) Maque, a family of extremely short interspersed repetitive elements: characterization, possible mechanism of transposition, and evolutionary implications. *Gene* **263**: 247–253.

Wang, R., Zheng, L., Toure, Y.T., Dandekar, T. and 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 U S A* **98**: 10769–10774.

Wondji, C., Simard, F. and 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.

Wu, C.-I. (2001) The genic view of the process of speciation. *J Evol Biol* **14**: 851–865.

Wu, C.-I. and Ting, C.-T. (2004) Genes and speciation. *Nat Rev Genet* **5**: 114–122.

Zheng, L., Benedict, M.Q., Cornel, A.J., Collins, F.H. and Kafatos, F.C. (1996) An integrated genetic map of the African human malaria vector mosquito, *Anopheles gambiae*. *Genetics* **143**: 941–952.