Semipermeable species boundaries between *Anopheles gambiae* and *Anopheles arabiensis*: Evidence from multilocus DNA sequence variation

N. J. Besansky[†‡], J. Krzywinski[†], T. Lehmann[§], F. Simard[¶], M. Kern[†], O. Mukabayire[†], D. Fontenille[¶], Y. Touré[‖], and N′F. Sagnon[††]

[†]University of Notre Dame, Center for Tropical Disease Research and Training, Department of Biological Sciences, Notre Dame, IN 46556; [§]Centers for Disease Control and Prevention, Division of Parasitic Diseases, Chamblee, GA 30341; [¶]Laboratoire de l'Institut de Recherche pour le Developpement, Organisation de Lutte Contre les Grandes Endemies en Afrique Centrale, BP 288 Yaoundé, Cameroun; [‖]Ecole Nationale de Médicine et de Pharmacie, BP 1805 Bamako, Mali; and [††]Centre National de Recherche et de Formation sur le Paludisme, 01 BP2208 Ouagadougou, Burkina Faso

Communicated by Fotis C. Kafatos, European Molecular Biology Laboratory, Heidelberg, Germany, July 11, 2003 (received for review March 10, 2003)

Attempts to reconstruct the phylogenetic history of the *Anopheles gambiae* cryptic species complex have yielded strongly conflicting results. In particular, *An. gambiae*, the primary African malaria vector, is variously placed as a sister taxon to either *Anopheles arabiensis* or *Anopheles merus*. The recent divergence times for members of this complex complicate phylogenetic analysis, making it difficult to unambiguously implicate interspecific gene flow, versus retained ancestral polymorphism, as the source of conflict. Using sequences at four unlinked loci, which were determined from multiple specimens within each of five species in the complex, we found contrasting patterns of sequence divergence between the X chromosome and the autosomes. The isolation model of speciation assumes a lack of gene flow between species since their separation. This model could not be rejected for *An. gambiae* and *An. arabiensis*, although the data fit the model poorly. On the other hand, evidence from gene trees supports genetic introgression of chromosome 2 inversions between *An. gambiae* and *An. arabiensis*, and also points to more broad scale genetic exchange of autosomal sequences between this species pair. That such exchange has been relatively recent is suggested not only by the lack of fixed differences at three autosomal loci but also by the sharing of full haplotypes at two of the three loci, which is in contrast to several fixed differences and considerably deeper divergence on the X. The proposed acquisition by *An. gambiae* of sequences from the more arid-adapted *An. arabiensis* may have contributed to the spread and ecological dominance of this malaria vector.

The role of introgressive hybridization in evolution remains contentious, especially among zoologists. As defined by the dominant biological species concept, species are groups of actually or potentially interbreeding natural populations that are reproductively isolated from other such groups (1). Under a strict interpretation of this model, gene flow between "good" species, to the extent that it occurs at all, is inconsequential to species evolution, because hybrids are presumed less fit. Yet recent studies of *Drosophila* and *Anopheles* species (e.g., refs. 2–5), suggest that introgression is not necessarily rare or inconsequential; indeed, it may be advantageous (2). This alternative viewpoint is not new. In a comprehensive treatment of the systematics of mosquito fauna of the South Pacific, published just one year before Mayr's *Animal Species and Evolution* (1), Belkin (ref. 6, p. 60) hypothesized that hybridization was an important mechanism in the speciation, the formation of new types, and the evolution of mosquitoes. Evidence for the importance of introgressive hybridization in mosquito evolution comes from a group of sibling species, the *Anopheles gambiae* complex.

This complex, considered a single polytypic species before 1956 (7), is now known to consist of at least seven partially intersterile cryptic species found in Africa south of the Sahara (8, 9). Three halophilic species are minor malaria vectors because of limited density and distribution: *Anopheles melas* and *Anopheles merus*, from western and eastern coastal mangroves, respectively (the latter species extending inland in South Africa), and *Anopheles bwambae*, from mineral springs in the Semliki Forest of Uganda. *Anopheles quadriannulatus* species A and B, freshwater species patchily distributed in the south and east, respectively, pose no threat to public health, owing to zoophily. These five are mutually allopatric, except where *An. quadriannulatus* A contacts *An. merus* in South Africa, but their ranges overlap those of *An. gambiae* and *Anopheles arabiensis*. The latter two species, both widespread and extensively sympatric, are major malaria vectors. These two species also were the subject of recent controversy, because conflicting evidence supported either *An. arabiensis* or *An. merus* because the true sister taxon to *An. gambiae* (2, 10, 11).

At the center of the conflict, the X chromosome is the source of fixed chromosome inversion and sequence differences by which these isomorphic taxa are recognized. Inversion *Xag*, a paracentric inversion on the X that is monophyletic in origin (10), differentiates *An. gambiae* and *An. merus* from other species, and identifies them as sister taxa (8). (*An. arabiensis* possesses an independent autapomorphic compound inversion on the X.) Also X-linked but outside of the inversion in pericentromeric heterochromatin, intergenic spacer sequences of ribosomal DNA distinguish each species but strongly support an alternative phylogenetic hypothesis in which *An. arabiensis* is allied with *An. gambiae* (2). Resolution of the conflict requires invoking either lineage sorting or introgression of X chromosome sequences between *An. gambiae*, and either *An. merus* or *An. arabiensis*. In nature, important premating isolating mechanisms exist between sympatric populations of these species, as F1 hybrids are detected only rarely (0.02–0.76%; refs. 12 and 13). Nevertheless, postmating isolating mechanisms are incomplete, because F1 hybrid females are fertile and potentially broker interspecific gene exchange. The *An. gambiae–An. arabiensis* relationship inferred from other markers, especially the paraphyletic, intertwining mtDNA (2, 14), has been interpreted as reflecting the flow of nuclear and mtDNA sequences across species boundaries, rather than the persistence of ancestral alleles.

There is no defined hybrid zone between *An. gambiae* and *An. arabiensis*; their ranges coincide even at a microgeographic scale. An alternative approach to studying the role of introgressive hybridization in the history of these species is to examine patterns of genetic variation at multiple loci within and between sibling species (e.g., refs. 5 and 15). Previously (14, 16), we found extensive mtDNA haplotype sharing across Africa between *An. gambiae* and *An. arabiensis*. Here, by using the same and additional population samples, as well as three additional sibling species (*An. merus*, *An. melas*, and *An. quadriannulatus*), we extend this approach to four loci representing each chromosome ($2N = 6$). These loci include the *white* gene on the X, the tryptophan oxygenase (*tox*) gene on

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AY117771–AY118063).

[‡]To whom correspondence should be addressed. E-mail: besansky.1@nd.edu.

Table 1. Source of *An. gambiae sensu lato* specimens

Country	Species	No. of specimens			
		white	*tox*	*G6pd*	*xdh*
Kenya	*gam*	32	18	9	7
	ara	20	12	3	4
	mer	4	4	4	2
Mali	*gam*	20	20	0	0
	ara	6	5	0	0
Burkina Faso	*gam*	0	0	0	4
Liberia	*gam*	1	1	0	0
Senegal	*gam*	11	15	5	1
	ara	15	16	5	8
	mel	5	0	3	0
The Gambia	*mel*	2	2	2	2
South Africa	*ara*	4	2	5	0
	mer	1	0	1	0
	qua	1	0	1	0
Zimbabwe	*qua*	2	3	3	2

gam, An. gambiae; ara, An. arabiensis; mer, An. merus; qua, *An. quadriannulatus; mel, An. melas.*

chromosome 2, and two genes on the left and right arms of chromosome 3: glucose-6-phosphate dehydrogenase (*G6pd*) and xanthine dehdrogenase (*xdh*), respectively. The *white* gene, located inside inversion *Xag* common to *An. gambiae* and *An. merus*, and the *tox* gene, located within inversion *2Rb* common to *An. gambiae* and *An. arabiensis*, were previously studied in Mali, West Africa (17). Not previously studied in the *An. gambiae* complex, the *G6pd* and *xdh* genes are located outside of any chromosomal inversions. We analyze these sets of sequences within population genetic and phylogenetic frameworks, and test whether the data are consistent with a model of reproductively isolated species sharing ancestral polymorphism, or whether introgression must be invoked.

Materials and Methods

Mosquito Sampling. Table 1 lists the geographic origins and designations for all specimens. Anophelines were identified as members of the *An. gambiae* complex by using morphological keys (18), and species identification was performed with a ribosomal DNA-PCR assay (19). With the exception of the Malian specimens, the karyotype with respect to inversion *2Rb* could not be determined, as ovarian nurse cell polytene chromosomes were unavailable. The *An. gambiae* and *An. arabiensis* specimens from Kenya, Senegal, and South Africa are derived from samples previously analyzed for mtDNA sequence variation (14); specimens from Mali were characterized previously at *white* and *tox* (17). Further sampling details can be obtained from the authors.

DNA Methods. DNA was extracted from individual specimens according to protocol 48 (20). Standard procedures were used for amplification and sequencing of *white*, *tox*, *G6pd*, and *xdh* (17, 21); detailed protocols and oligonucleotides used are available from the authors. Sequences from the guanylate cyclase (*gua*) and mtDNA ND5 genes were obtained from GenBank (accession nos. U42609–U42614 and U42619-U42621; ref. 10) and (AF020983–AF020987, AF020991–AF020993, AF020998–AF020999, AF021004–AF021008, and AF021019–AF021025; ref. 14).

Because male anophelines are hemizygous for X-linked genes, *white* PCR products from males (available only from Kenya and Zimbabwe samples) were purified by using the Wizard PCR Preps kit (Promega), and were directly cycle sequenced. All other PCR products were cloned by using pGEM-T Easy Vector (Promega) before sequencing. Sequences are available under GenBank accession nos. AY117771–AY117894 (*white*), AY117895–AY117992 (*tox*), AY117993–AY118033 (*G6pd*), and AY118034–AY118063 (*xdh*).

Sequences were checked for accuracy on both strands by using the FRAGMENT ASSEMBLY programs of the Genetics Computer Group [(GCG), Madison, WI] software package, or by using SEQUENCE NAVIGATOR (PE Applied Biosystems, Foster City, CA), and multiple alignment was performed with the PILEUP program of the GCG.

PCR Error. In this article, direct sequencing was not possible in most instances due to extensive insertion/deletion polymorphism and, for the *white* locus, the lack of males from most samples. To adjust for the effect of PCR error on diversity indices and statistics relying on these indices, we used a statistical approach (by using a program written in SAS language by T.L., which is available from the authors). The original set of aligned sequences (minus those obtained by direct sequencing) was used to generate multiple error-adjusted datasets in which the level and pattern of genetic variability was corrected to the level expected without polymerase error. Adjustments to the original dataset were conditioned on the probability that a given base in a column of aligned sequences was misincorporated, assuming an overall error rate of 0.15% per base (estimated by sequencing five clones from each of three field specimens). The conditional probability of PCR error at each site was calculated based on (*i*) whether the position was invariant, singleton, or polymorphic; (*ii*) the frequency of the base at that position; and (*iii*) the expected number of invariant, singleton, and polymorphic sites in the original data. Further, the algorithm considered the markedly different probabilities of observing a particular misincorporated nucleotide given an inferred source nucleotide. The performance of the algorithm (to be described in detail elsewhere) was evaluated in two pairs of experimental datasets, where one set contained PCR errors and the matching set was empirically corrected (by direct sequencing or by sequencing multiple clones per specimen). The program was found to reduce >85% of the bias in the polymorphism statistics (F.S., M. Licht, and T.L., unpublished data). Here, we analyzed each of six replicate error-adjusted datasets per locus, and compared the results to those obtained from the original data. With the exception of tests of the isolation model of speciation (see below), all trends statistically supported in the original dataset were supported in the error-adjusted datasets.

Data Analysis. DNA polymorphism statistics, tests of neutrality, and measures of divergence were computed with DNASP 3.5 (22). SITES 1.1, which can be accessed at: http://lifesci.rutgers.edu/~heylab, was used to estimate the recombination parameter, γ, and to generate data for input into HKA and WH, programs distributed by J. Hey (Rutgers University, Piscataway, NJ). HKA, which implements the method described in ref. 23, was used to test for neutral molecular evolution. WH, which implements methods described in refs. 24 and 25, was used to fit an isolation model of speciation. This model assumes that the two species from which data have been sampled arose from a single ancestral species T generations ago, without subsequent gene flow between descendant species. Using γ and the observed numbers of shared, fixed, and exclusive polymorphic sites summed across multiple loci, under the assumption of constant effective population sizes and neutral mutation, the program returns estimates of θ and T that most closely fit the data. The quality of fit of the data to this model was assessed by comparing χ^2 and WWH (Wang–Wakeley–Hey) test statistic values (15, 25) to a distribution of values generated by 10,000 computer simulations of the coalescent process based on parameters estimated from the data.

The genetic structure of *An. gambiae* and *An. arabiensis* populations was examined by a hierarchical analysis of molecular variance, and by population pairwise F_{ST} statistics, with each being tested for significance by using 10,000 permutations with ARLEQUIN 2.0 (26). Both analyses were performed on data in which gapped alignment positions were either excluded, or treated as missing data with zero weight. Because there was no significant difference in the outcome, data reported are from the latter treatment.

Genealogical relationships among sequences were estimated by using maximum parsimony as implemented in PAUP* 4.d65 (27), by

using heuristic searches and TBR branch swapping. Gaps were treated as missing data, but were scored as additional characters in the *tox* analysis. Analyses were done by stepwise random addition of taxa with 1,000 replications. Bootstrapping was performed over 100–500 replicates, each with 10 random additions of sequences. Significance of character incongruence was evaluated by implementing the incongruence length difference (ILD) test in PAUP*, by using 10,000 replicates, each with 10 random additions of sequences.

Results

DNA Polymorphism. Samples of *An. gambiae* and *An. arabiensis* were collected concurrently from villages in East Africa (Kenya) and West Africa (Mali, Senegal) up to 6,000 km apart, to maximize the level of diversity represented. Sampling of *An. merus*, *An. quadriannulatus*, and *An. melas* was more limited in scope (Table 1). Wherever possible, sequences at each locus were determined from the same specimens. Physical locations of the four loci (*white*, *tox*, *G6pd*, and *xdh*) with respect to chromosome arm, relevant paracentric inversions, and previously studied loci are indicated in Fig. 1, which is published as supporting information on the PNAS web site, www.pnas.org. Multiple sequence alignments for each locus from five species (which are available in the PopSet database of GenBank) revealed substantial insertion/deletion variation as well as base pair substitutions. A summary of sequence variation and other basic statistics is given in Table 5, which is published as supporting information on the PNAS web site. For *An. gambiae* and *An. arabiensis* autosomal loci, nucleotide diversity adjusted for PCR error averaged 2.6% and 1.9%, respectively, whereas values for the X-linked *white* locus were lower by at least half. These two species are both polymorphic across their ranges for the cytologically identical chromosomal inversion *2Rb*, which includes the *tox* locus. The total *tox* samples of *An. gambiae* and *An. arabiensis* included individuals of unknown karyotype. For the *An. gambiae* and *An. arabiensis* specimens from Mali with known karyotype, statistics were computed separately for subsamples containing the homokaryotypic inverted (*2Rb/b*) and homokaryotypic standard (uninverted: $2R+^{b}/+^{b}$) arrangements. No apparent difference in polymorphism level between alternative arrangements was noted.

Natural Selection in *An. gambiae* and *An. arabiensis*. The pattern of sequence variation was consistent with neutral evolution at all loci except for that of *white*. Tajima's *D*, a statistic that measures the difference between two estimators of $4N\mu$ (π and θ), is expected to be zero under a neutral model with constant population size (28). The *D* statistic was not significantly different from zero except for *white* sequences in both species. At *white*, *D* values were significantly negative, indicating an excess of low frequency mutations. In principle, this result could be due to a population bottleneck followed by demographic expansion, or a selective sweep at *white* or a locus nearby. Because the results of the Tajima test appeared to be locus-specific rather than genome-wide; they are more consistent with purifying selection at or near *white* than recent demographic expansions in these species. Nevertheless, a second test (HKA) was unable to detect departures from neutral evolution.

The HKA test is premised on a prediction of the neutral model of molecular evolution, that levels of interspecific divergence and intraspecific polymorphism should be correlated (23). HKA testing was conducted by using multiple sequences per species pair at each of six loci: four from this article, mtDNA ND5 (14), and X-linked guanylate cyclase (10). Tests were carried out both on the original and PCR error-adjusted data. The significance of the test statistic was evaluated against a distribution generated by 10,000 coalescent simulations, by using parameters estimated from the data. In no instance was there a significant deviation from neutral expectation: *An. gambiae–An. merus*, $\chi^2 = 3.23$, <95% ($\chi^2_{adj} = 4.47$, <84%) of simulated values; *An. arabiensis–An. merus*, $\chi^2 = 10.53$, <25% ($\chi^2_{adj} = 15.88$, <6%) of simulated values; and *An. arabiensis–An. gambiae*, $\chi^2 = 7.69$, <29% ($\chi^2_{adj} = 11.02$, <14%) of simulated values.

Interspecific Divergence. Locus-by-locus estimates of differentiation between *An. gambiae* and other species in the complex are given in Table 2. In the absence of gene flow between species, an equilibrium-neutral model predicts that interspecific differentiation will increase with time since speciation. The most closely related species pair should show the smallest net divergence (D_a), the fewest fixed differences, and the most shared polymorphisms across all loci (after adjusting for X versus autosome linkage). Based on the shared *Xag* inversion, the most closely related species pair is expected to be *An. gambiae* and *An. merus*. Although species could not be ranked consistently according to levels of divergence from *An. gambiae*, some striking trends emerged that proved robust to PCR error. For each species pair, the highest levels of net divergence and fixed differences were measured at *white*, whereas those at *xdh* were notably lower. In general, these measures were usually higher between *An. gambiae* and halophilic (*An. melas* and *An. merus*) rather than freshwater (*An. arabiensis* and *An. quadriannulatus*) species. In particular, divergence and fixed differences between *An. gambiae* and *An. merus* were always higher, usually much higher, than between *An. gambiae* and *An. arabiensis*, which was contrary to expectation. It is especially interesting to note that this trend held even at the *white* locus, within the *Xag* inversion. This trend is in contrast to the *gua* locus within *Xag*, at which *An. gambiae* and *An. merus* had the most shared polymorphisms and relatively few fixed differences (10), suggesting a complex history for the *Xag* inversion. Neither introgression nor ancestral polymorphism alone can account for these results. The data exclude neither possibility, and it is likely that both have left their signature across loci and between species. Other evolutionary forces, possibly selection, may also be operating.

An interesting pattern was also revealed at the *tox* locus, within the *2Rb* inversion shared by *An. gambiae* and *An. arabiensis*. F_{ST} values indicated significant differentiation between all species pairs at all loci, except for *tox* between *An. gambiae* and *An. arabiensis*. The apparent differentiation at *tox* was because of alternative chromosomal arrangements, not because of interspecific differences (Table 2). When samples were partitioned by inverted arrangement, only interspecific (and even intraspecific) comparisons between ($2Rb/b - 2R+^{b}/+^{b}$) but not within ($2Rb/b - 2Rb/b$) arrangements were significant. As proposed by Mukabayire *et al.* (17), this finding suggests a common origin for the *2Rb* inversions in both species. However, the sharing of this *2Rb* arrangement between species cannot explain the sharing of so many polymorphisms at *tox*, of which 12 derive from uninverted *An. gambiae* and inverted *An. arabiensis* arrangements.

Beyond merely sharing polymorphic sites at these four loci, *An. gambiae* and *An. arabiensis* share identical haplotypes at both *G6pd* and *tox* loci across Africa. For *G6pd*, this sharing includes *An. arabiensis* haplotypes from Kenya, South Africa, and Senegal that are identical to *An. gambiae* haplotypes from Senegal (A_A7_10 = G_17385; A_SA4 = A_17376 = G_17390). For *tox*, this sharing includes *An. gambiae* haplotypes from Kenya and Senegal that are identical to *An. arabiensis* haplotypes from Senegal (G_A4_11 = A_14145; G_14817 = A_1536 = A_14153), and if gapped alignment positions are completely excluded from consideration, *An. gambiae* and *An. arabiensis* haplotypes from Mali (G_S_b6 = A_b93). Given the rates of recombination estimated at these loci (Table 5), it seems unlikely that entire haplotypes would be preserved intact between species since their splitting, although this event has not been dated and is presumed relatively recent.

The interspecific sharing of entire haplotypes at freely recombining loci is consistent with secondary contact and introgressive hybridization, rather than with retained ancestral polymorphism. On the other hand, because any instance of introgression is local, finding introgressed haplotypes distributed across 6,000 km of

Table 2. Differentiation between *An. gambiae* and other species in the *An. gambiae* complex

Contrast	*white* (*X*: 2*A*)					*tox* (2: 12*E*)				*G6pd* (3:39*C*)				*xdh* (3: 29*C*)				
	D_{xy}	D_a	F_{ST}, *P* value	Shared	Fixed	D_{xy}	D_a	F_{ST}, *P* value	Shared	Fixed	D_{xy}	D_a	F_{ST}, *P* value	Shared	Fixed	D_{xy}	D_a	F_{ST}, *P* value	Shared	Fixed
gam–qua	4.98	2.95	0.702 (0.000)	1 (4.1)	18	3.55	1.52	0.306 (0.000)	4 (2.3)	2	3.63	1.50	0.380 (0.000)	1 (1.3)	1	1.17	0.20	0.260 (0.027)	1 (0.3)	0
gam–mel	6.86	5.84	0.818 (0.000)	0 (1.6)	22	5.10	3.59	0.498 (0.001)	0 (0.2)	9	2.82	1.35	0.372 (0.000)	0 (0.5)	2	1.80	0.91	0.507 (0.010)	2 (0.3)	6
gam–mer	6.30	5.05	0.791 (0.000)	2 (2.6)	16	5.69	3.20	0.495 (0.000)	5 (3.9)	8	4.29	2.82	0.574 (0.000)	0 (0.5)	8	1.53	0.90	0.483 (0.011)	0 (0.1)	5
gam–ara	4.40	3.39	0.763 (0.000)	6 (9.1)	8	3.15	0.32	0.110 (0.000)	30 (12.4)	0	3.11	1.23	0.391 (0.000)	8 (2.1)	0	1.17	0.25	0.212 (0.000)	11 (1.0)	0
gam 2Rb/b–ara 2Rb/b	—	—	—	—	—	2.58	−0.08	−0.024^{NS}	17 (2.0)	0	—	—	—	—	—	—	—	—	—	—
gam $2R^{+b}/+^{b}$*–ara 2Rb/b*	—	—	—	—	—	3.00	0.37	0.140 (0.020)	12 (2.2)	0	—	—	—	—	—	—	—	—	—	—
gam $2R^{+b}/+^{b}$*–gam 2Rb/b*	—	—	—			2.94	0.38	0.128 (0.006)			—	—	—			—	—	—		

D_{xy}, D_a values have been multiplied by 100; shared polymorphisms are reported as observed values and, in parentheses, the number expected by recurrent mutation (35). NS, not significant.

Africa is unexpected, even for a flying insect. To shed light on this paradox, we examined the partitioning of sequence variation within and between *An. gambiae* and *An. arabiensis* at each locus by performing a hierarchical analysis of molecular variance (Table 3). For this analysis, populations were defined as species-specific samples from Kenya, Senegal, South Africa, Burkina Faso, or Mali; groups were defined as species (and alternative *2R* arrangements, for *tox*). At each locus except for *white*, most variation was distributed within populations, with nearly all of the remaining variation distributed between species. This pattern was reversed for *white*, the only locus for which the fixation index F_{CT} was significant by permutation testing. In agreement with other analyses of *tox* (Table 2), a significant amount of variation was distributed between alternative *2R* arrangements, but not between species. With respect to the paradox of broadly distributed shared haplotypes, it is noteworthy that for each of the four loci, the amount of variation distributed among populations within species was either insignificant (*white*, *tox*, *2R* arrangements, and *xdh*) or very minor (*G6pd*). This finding is consistent with previous mtDNA and microsatellite studies that have suggested shallow population structure (14, 29, 30). Without well defined geographic population structure, localized introgression events cannot readily be perceived as such (16).

Gene Trees. Traditional phylogenetic methods that produce bifurcating gene trees cannot accurately portray genealogical relationships at the intraspecific level, due to recombination. However, our purpose here was to assess whether sequences from a given species are monophyletic, and to examine interspecific relationships. Fig. 2, which is published as supporting information on the PNAS web site, shows maximum parsimony trees constructed from sequences at each locus from five members of the *An. gambiae* complex. In general, sequences derived from the same species are monophyletic, with the notable exception being those from *An. gambiae* and *An. arabiensis* at autosomal loci. In the *tox*, *G6pd*, and *xdh* gene trees, sequences from this pair of species not only cluster together, but also intertwine, which is reminiscent of previous findings based on mtDNA and esterase gene sequences (2, 14). By contrast, the *white* gene tree is congruent with the traditional chromosomal inversion-based phylogeny of the group (8), as well as that of another gene (*guanylate cyclase*) located inside the *Xag* inversion (10). In this case, monophyletic *An. gambiae* and *An. merus* sequences cluster as sister taxa. Attempts to root the tree with the *white* gene sequence from the closely related species *Anopheles christyi* were unsuccessful, because the intron was too divergent in length and sequence to permit reliable alignment to the ingroup sequences. Although our trees are unrooted, the conclusion of (*An. gambiae* + *An. merus*) is supported by a previous phylogenetic analysis of mtDNA that included *An. christyi* as an outgroup (31). The root could not be placed unambiguously in that mtDNA study, but the important point is that it did not lie in the clade consisting of *An. gambiae* or *An. merus* (31). Incongruence between trees constructed from the *white* and autosomal gene sequences is not a trivial result of weak phylogenetic signal, as suggested by significant ILD tests ($P < 0.001$ in all cases).

Shared Variation and the Isolation Model of Speciation. At issue is whether the amount of shared nucleotide variation between *An. gambiae* and *An. arabiensis* is consistent with reproductive isolation since species splitting (the null hypothesis), or whether secondary contact and introgressive hybridization must be invoked. Put another way, do the contrasting levels of shared polymorphisms and fixed differences found across multiple loci fall within a range that could be expected of isolated populations evolving neutrally, given the stochastic nature of lineage sorting? This problem was approached by fitting the data from six loci (four from this article, *gua*, and mtDNA ND5) to an isolation model of speciation (25), in which *An. gambiae* and *An. arabiensis* are assumed to have arisen from a common ancestral species at some point in the past, without

EVOLUTION

Table 3. Hierarchical AMOVA for *An. gambiae* and *An. arabiensis*

Source of variation	*white*	*tox*: (*gam–ara*)	*tox*: ($2R+^b/+^b$–*2Rb*/*b*)	*G6pd*	*xdh*
			Percent variation		
Among groups*	76.24	−9.03	14.03	36.79	19.81
Among populations† within groups	0.12	14.24	−1.38	5.83	−0.04
Within populations	23.63	94.79	87.35	57.39	80.23
			Fixation indices		
F_{SC} (population/group)	0.005^{NS}	0.131	−0.016^{NS}	0.092*	−0.001^{NS}
F_{ST} (population/total)	0.764***	0.052*	0.126*	0.426***	0.198**
F_{CT} (group/total)	0.762*	−0.090^{NS}	0.140^{NS}	0.368^{NS}	0.198^{NS}

NS, not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

*Groups are defined as species (*gam–ara*) or as *2R* arrangements ($2R+^b/+^b$–*2Rb*/*b*).

†Populations are defined as species-specific samples from Kenya, Senegal, South Africa, Burkina Faso, and Mali; or as *2R* arrangements partitioned by species, as applicable.

subsequent gene flow between them. Four tests were carried out on different datasets: (*i*) the original data; (*ii*) the original data corrected for chance occurrence of shared polymorphisms, as in Table 2; (*iii*) PCR error-adjusted data; and (*iv*) PCR error-adjusted data corrected for chance occurrence of shared polymorphisms. Estimates of the population mutation rate parameter for *An. gambiae*, *An. arabiensis* and the ancestral species (θ_{gam}, θ_{ara}, and θ_A, respectively), and time of separation derived under this model for all four tests are given in Table 4, together with statistics for goodness of fit. The relatively large value of θ_A, a consequence of accommodating large numbers of fixed differences as well as shared polymorphisms into the model, suggests a poor fit (25), yet values of the test statistics generally do not allow rejection of the model. The χ^2 statistic measures departures from the expected numbers in four categories of sites: (*i*) polymorphisms exclusive to species 1; (*ii*) those exclusive to species 2; (*iii*) shared polymorphisms; and (*iv*) fixed differences. The WWH statistic measures the variance in the number of fixed differences and shared polymorphisms. That only tests 1 and 2 (not adjusted for PCR error) allowed for rejection of the model, and only by the χ^2 statistic, suggests that PCR error was responsible, by inflating the number of exclusive polymorphisms. There may be at least three reasons, not mutually exclusive, for failure to reject the model in the other tests. First is the fact that most loci included had large amounts of shared polymorphism; only two (*white* and *gua*) had fixed differences. Second, the isolation model relies strongly on two assumptions: neutral evolution, possibly violated at *white*; and constant effective population size, also likely violated (32). Third, the simulation results are sensitive to the amount of recombination, which is underestimated by γ (33). This result raises the variance around the simulated numbers of exclusive, shared, and fixed sites, broadening the distribution of test statistics of the fit of the model to the data and making the test more conservative (WH documentation; which can be accessed at: http://lifesci.rutgers.edu/~heylab). With this conclusion in mind, the results of test 4, marginally significant by the χ^2 statistic, are provocative.

Discussion

If *An. gambiae* and *An. arabiensis* were species with deeply structured populations across Africa, it should have been possible to test for interspecific gene flow by comparing levels of sequence divergence in sympatric and allopatric samples. This was the premise underlying the design of a previous mtDNA study of the same samples (14), where it was predicted that within-species differentiation between samples from Kenya and Senegal might exceed interspecific differentiation in sympatric samples from those locales. However, that and other studies (32) using mtDNA and microsatellites showed surprisingly shallow population structure across Africa in both species, not only due to potentially high vagility but probably also due to relatively recent population expansions in response to human population expansion and environmental alterations. Moreover, *An. gambiae* and *An. arabiensis* were found to share not only mtDNA polymorphisms but also mtDNA haplotypes across 6,000 km of Africa. This result was interpreted in terms of mtDNA gene flow across species boundaries, but incomplete sorting of ancestral polymorphism could not be ruled out without evidence from additional loci.

This article used a multilocus approach to explore patterns of DNA polymorphism and divergence at four nuclear genes within and among five species in the *An. gambiae* complex. These data present a compelling argument for introgressive hybridization between *An. gambiae* and *An. arabiensis*. Taken together, they reveal genomes that are mosaics with respect to gene flow, as described previously for *Drosophila pseudoobscura* and its close relatives (ref. 5 and references therein). In contrast to shared polymorphisms and even full haplotype sharing of mtDNA as well as autosomal sequences, X-linked sequences showed fixed differences and relatively deep divergences when compared between these species. It has been argued that radiation of the *An. gambiae* complex has been quite recent, but a notable feature of gene trees reconstructed from mtDNA as well as nuclear genes is monophyly of sequences from taxa other than *An. gambiae* and *An. arabiensis*. Thus, retention of ancestral polymorphism is not generally observed between species in the complex, other than between *An. gambiae* and *An. arabiensis*. In mtDNA gene trees, as well as trees from the autosomal genes *tox*, G*6pd*, and *xdh*, sequences from this species pair were not only clustered but were intertwined (this study and refs. 2, 3, and 31), which was in contrast to the sister taxa relationship between *An. gambiae* and *An. merus* specified by genes

Table 4. Fitting *An. gambiae* and *An. arabiensis* to the isolation model of speciation

Test	θ_{gam}	θ_{ara}	θ_A	T	$P\chi^2$	P_{WWH}
1	44.4 (0–482.2)	27.7 (0–156.6)	140.7 (38.9–327.5)	0.271 (0.029–0.538)	0.033	0.283
2	52.4 (0–370.3)	32.0 (0–133.8)	120.5 (28.4–281.9)	0.287 (0.034–0.562)	0.008	0.309
3	21.1 (0–601.5)	12.2 (0–116.9)	148.4 (35.8–395.0)	0.249 (0.017–0.541)	0.095	0.446
4	35.7 (0–1080.6)	14.7 (0–111.1)	137.2 (31.6–360.8)	0.257 (0.007–0.543)	0.058	0.466

Parameter estimates and their 99% confidence intervals are given. Time of and separation (T) is measured in $2N_{gam}$ generations, and *P* values represent the proportion of simulations giving and values greater than observed. Tests 1 and 2, data without and with correction for chance sharing of polymorphisms; tests 3 and 4, data adjusted for PCR error without and with correction for chance sharing of polymorphisms.

inside inversion *Xag*. These data are consistent with a model in which there are semipermeable barriers to gene flow between *An. gambiae* and *An. arabiensis*, with *Xag*-linked sequences barred from exchange in the face of introgression at *tox*, *xdh*, and possibly *G6pd*. It is noteworthy that in the face of this evidence, the isolation model of Wang *et al.* (25) was not rejected, but the strong dependence of the model on both recombination rates and neutral evolution may have compromised its applicability here. These limitations point to the need for more powerful analytical approaches for testing an isolation model of speciation.

If *An. gambiae* and *An. merus* are indeed sister taxa, as suggested by (*i*) sequences within the *Xag* inversion, and (*ii*) autosomal sequences within the uninverted arrangement of *2La*, and (*iii*) the degree to which other genome regions support the conflicting relationship of *An. gambiae* plus *An. arabiensis* and by inference, introgressive hybridization, is striking. Consider the list of such regions: mtDNA, chromosome 2 inversions *2Rb*, *2La*, and sequences within or tightly linked to these, genes and microsatellite sequences on chromosome 3, and X chromosome genes and microsatellite sequences outside of the *Xag* inversion (this study, refs. 2, 3, 17, and 31, and A. Wang, C. della Torre, C. Schwager, G. Blass, Y. T. Dolo, F. Collins, G. Lanzaro, F. Kafatos, and L. Zheng, unpublished data). This situation is sobering in terms of its implications for phylogeny reconstruction. Without the relevant background information about the biology and genetics of these species, it is possible that the unwitting molecular systematist would become increasingly confident in the "wrong" answer as more loci were sequenced. The risk of being either misled or of coming to the "right" conclusion by lucky accident would surely increase by adopting a "total-evidence" approach. At the level of closely related species, it is only through a careful locus-by-locus assessment of sequence divergence that history can be deciphered.

The apparent mixing of *An. gambiae* and *An. arabiensis* genomes also raises the fundamentally important question of what proportion of the genome, and which locations, are reproductively isolated. Despite some mixing, the species have not fused, nor would any vector biologist familiar with these species in Africa be prepared to question or revoke their status as good species, owing to characteristically different behaviors. Given their extensive sympatry in the narrow sense, that hybrids are found so rarely (despite intense sampling efforts) is testament to strong premating isolation barriers. There are fixed ribosomal DNA and chromosomal inversion differences on the X, and polymorphic inversions on chromosome 2 (e.g., *2Rd*, *2Rd′*, and *2Ra*) that are exclusive to one species. Those regions of the genome that resist introgression (34) are expected to include the genes directly involved in speciation. The availability of the complete *An. gambiae* genome sequence should help guide the discovery these "speciation genes."

Whereas genes involved in reproductive isolation may be protected against gene flow, this is not necessarily true for genes that could have an important bearing on malaria vectorial capacity. Candidate genes, not yet identified, would be those underlying traits that increase the probability of contact between humans and vectors, such as preference for human bloodmeals, indoor resting behavior, oviposition in sites created by human activity, and tolerance of aridity which would allow range expansion into formerly inhospitable environments. With respect to the latter trait, overwhelming evidence supports the role of inversions *2Rb* and *2La* in conferring tolerance to arid climatic conditions. In West Africa, frequencies of these inversions are strongly correlated with humidity and rainfall, both seasonally and spatially. Thus, the transitions from rainy to dry season within a locale, or from rainforest locales in the south to Sahel in the north, are both accompanied by prominent increases in the frequency of *2Rb* and *2La* inversions (8). As this and other studies have shown, these inversions are identical by descent in *An. gambiae* and *An. arabiensis* (3, 17), and ecological, biogeographical, and molecular phylogenetic evidence suggests that they have most likely been captured by *An. gambiae* from *An. arabiensis* (11). Acquisition of these inversions is thought to have conferred upon *An. gambiae* the ecological and genetic flexibility that allowed it to invade the savannas beyond its ancestral range in the rainforest, thereby helping to establish this species as the foremost malaria vector (11). That semipermeable barriers to gene flow remain between *An. gambiae* and *An. arabiensis* is both an invitation to the study of speciation, and a threat to controlling the vectorial potential of these disease vectors.

We thank J. Feder, H. Hollocher, and anonymous reviewers for constructive comments on the manuscript; Thomas Fahey, Jane Carter, and Cristina Rafferty for technical assistance; and F. Collins, L. Braack, S. Koenraadt, and W. Takken for samples. This work was supported by National Institutes of Health Grant R01 AI44003 (to N.J.B.) and the United Nations Development Program/World Bank/World Health Organization Special Program for Research and Training in Tropical Diseases.

1. Mayr, E. (1963) *Animal Species and Evolution* (Harvard Univ. Press, Cambridge, MA).
2. Besansky, N. J., Powell, J. R., Caccone, A., Hamm, D. M., Scott, J. A. & Collins, F. H. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 6885–6888.
3. Caccone, A., Min, G. S. & Powell, J. R. (1998) *Genetics* **150,** 807–814.
4. Walton, C., Handley, J. M., Collins, F. H., Baimai, V., Harbach, R. E., Deesin, V. & Butlin, R. K. (2001) *Mol. Ecol.* **10,** 569–580.
5. Machado, C. A., Kliman, R. M., Markert, J. A. & Hey, J. (2002) *Mol. Biol. Evol.* **19,** 472–488.
6. Belkin, J. N. (1962) *Mosquitoes of the South Pacific* (Univ. of California Press, Berkeley).
7. Davidson, G. (1956) *Nature* **178,** 861–863.
8. Coluzzi, M., Sabatini, A., Petrarca, V. & Di Deco, M. A. (1979) *Trans. R. Soc. Trop. Med. Hyg.* **73,** 483–497.
9. Hunt, R. H., Coetzee, M. & Fettene, M. (1998) *Trans. R. Soc. Trop. Med. Hyg.* **92,** 231–235.
10. Garcia, B. A., Caccone, A., Mathiopoulos, K. D. & Powell, J. R. (1996) *Genetics* **143,** 1313–1320.
11. Powell, J. R., Petrarca, V., della Torre, A., Caccone, A. & Coluzzi, M. (1999) *Parassitologia (Rome)* **41,** 101–113.
12. Temu, E. A., Hunt, R. H., Coetzee, M., Minjas, J. N. & Shiff, C. J. (1997) *Ann. Trop. Med. Parasitol.* **91,** 963–965.
13. Toure, Y. T., Petrarca, V., Traore, S. F., Coulibaly, A., Maiga, H. M., Sankare, O., Sow, M., DiDeco, M. A. & Coluzzi, M. (1998) *Parassitologia (Rome)* **40,** 477–511.
14. Besansky, N. J., Lehmann, T., Fahey, G. T., Fontenille, D., Braack, L. E., Hawley, W. A. & Collins, F. H. (1997) *Genetics* **147,** 1817–1828.
15. Kliman, R. M., Andolfatto, P., Coyne, J. A., Depaulis, F., Kreitman, M., Berry, A. J., McCarter, J., Wakeley, J. & Hey, J. (2000) *Genetics* **156,** 1913–1931.
16. Donnelly, M. J., Pinto, J., Girod, R., Besansky, N. J. & Lehmann, T. (2003) *Heredity*, in press.
17. Mukabayire, O., Caridi, J., Wang, X., Toure, Y. T., Coluzzi, M. & Besansky, N. J. (2001) *Insect Mol. Biol.* **10,** 33–46.
18. Gillies, M. T. & De Meillon, B. (1968) *The Anophelinae of Africa South of the Sahara* (South African Institute for Medical Research, Johannesburg).
19. Scott, J. A., Brogdon, W. G. & Collins, F. H. (1993) *Am. J. Trop. Med. Hyg.* **49,** 520–529.
20. Ashburner, M. (1989) *Drosophila: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY).
21. Krzywinski, J., Wilkerson, R. C. & Besansky, N. J. (2001) *Syst. Biol.* **50,** 540–556.
22. Rozas, J. & Rozas, R. (1999) *Bioinformatics* **15,** 174–175.
23. Hudson, R. R., Kreitman, M. & Aguade, M. (1987) *Genetics* **116,** 153–159.
24. Wakeley, J. & Hey, J. (1997) *Genetics* **145,** 847–855.
25. Wang, R. L., Wakeley, J. & Hey, J. (1997) *Genetics* **147,** 1091–1106.
26. Schneider, S., Kueffer, J.-M., Roessli, D. & Excoffier, L. (1997) ARLEQUIN (Genetics and Biometry Laboratory, University of Geneva, Geneva), Version 2.0.
27. Swofford, D. L. (1999) PAUP*: *Phylogeny Analysis Using Parsimony (*and Other Methods), Version 4.0b2* (Sinauer, Sunderland, MA).
28. Tajima, F. (1989) *Genetics* **123,** 585–595.
29. Lehmann, T., Hawley, W. A., Kamau, L., Fontenille, D., Simard, F. & Collins, F. H. (1996) *Heredity* **77,** 192–200.
30. Kamau, L., Mukabana, W. R., Hawley, W. A., Lehmann, T., Irungu, L. W., Orago, A. A. & Collins, F. H. (1999) *Insect Mol. Biol.* **8,** 287–297.
31. Caccone, A., Garcia, B. A. & Powell, J. R. (1996) *Insect Mol. Biol.* **5,** 51–59.
32. Donnelly, M. J., Licht, M. C. & Lehmann, T. (2001) *Mol. Biol. Evol.* **18,** 1353–1364.
33. Hey, J. & Wakeley, J. (1997) *Genetics* **145,** 833–846.
34. della Torre, A., Merzagora, L., Powell, J. R. & Coluzzi, M. (1997) *Genetics* **146,** 239–244.
35. Clark, A. G. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 7730–7734.