Copyright © 1996 by the Genetics Society of America

Inversion Monophyly in African Anopheline Malaria Vectors

Beatriz A. García,*,† Adalgisa Caccone,*,‡ Kostas D. Mathiopoulos§ and Jeffrey R. Powell*

**Department of Biology, Yale University, New Haven, Connecticut 06520-8104, †Cátedra de Química Biológica, Facultad de Ciencias Médicas, Universidad Nacional de Córdoba, Córdoba, Argentina, ‡Dipartimento di Biologia, II Università di Roma "Tor Vergata," Rome 00173, Italy and §Istituto di Parassitologia, Università di Roma "La Sapienza," Rome 00185, Italy*

Manuscript received October 30, 1995
Accepted for publication April 4, 1996

ABSTRACT

The African *Anopheles gambiae* complex of six sibling species has many polymorphic and fixed paracentric inversions detectable in polytene chromosomes. These have been used to infer phylogenetic relationships as classically done with Drosophila. Two species, *A. gambiae* and *A. merus*, were thought to be sister taxa based on a shared *X* inversion designated X^{ag}. Recent DNA data have conflicted with this phylogenetic inference as they have supported a sister taxa relationship of *A. gambiae* and *A. arabiensis*. A possible explanation is that the X^{ag} is not monophyletic. Here we present data from a gene (soluble guanylate cyclase) within the X^{ag} that strongly supports the monophyly of the X^{ag}. We conjecture that introgression may be occurring between the widely sympatric species *A. gambiae* and *A. arabiensis* and that the previous DNA phylogenies have been detecting the introgression. Evidently, introgression is not uniform across the genome, and species-specific regions, like the *X*-chromosome inversions, do not introgress probably due to selective elimination in hybrids and backcrosses.

SINCE the pioneering work of STURTEVANT and DOBZHANSKY (1936), inversions detected in polytene chromosomes of dipteran insects such as Drosophila and mosquitoes have become accepted as standard phylogenetic indicators. The use of these chromosomal variants relies on the assumption that existing inversions are monophyletic in origin, *i.e.*, all carriers of a particular gene order trace their ancestry to a single chromosome. Therefore when two species share an inversion, they must share a common ancestor. Virtually all phylogenetic trees derived from independent data, such as molecular data, have confirmed this assumption by being congruent with the inversion-derived phylogenetic tree. A recent exception was found in the *Anopheles gambiae* complex (BESANSKY *et al.* 1994; MATHIOPOULOS *et al.* 1995; CACCONE *et al.* 1996).

From a medical standpoint, the *A. gambiae* complex is arguably the most important insect group in the world. In sub-Saharan Africa, malaria causes between one and two million deaths each year, most among young children and pregnant women (STÜRCHLER 1989); members of the *A. gambiae* complex are the major vectors in this region. The complex is confined to sub-Saharan Africa and consists of six closely related sibling species (WHITE 1974). *A. gambiae* and *A. arabiensis* are most widespread, are often closely associated with human habitats, and are the most efficient vectors of malaria known. *A. merus* and *A. melas* breed in brackish water on the east and west coasts of Africa, respectively. *A. quadriannulatus* occurs in the southeast of Africa and Zanzibar, and *A. bwambae* is confined to mineral springs of the Semliki Forest in Uganda. These four last-named species are of lesser medical importance due to their less intimate association with human habitats and greater tendency to procure blood meals from nonhuman animals. In the laboratory, hybrids can be formed among almost all species pairs; hybrid males are invariably sterile and females, if formed, are fertile (DAVIDSON *et al.* 1967).

Until recently, the only information on the phylogenetic relationship among these species had been derived from chromosomal inversions detected usually in the polytene chromosomes of nurse cells in adult females. COLUZZI *et al.* (1979) summarize the results that are illustrated in Figure 1. Two inversions are synapomorphic. The X^{ag} indicates the sister status of *A. gambiae* and *A. merus*, and the $3L^{a}$ links *A. melas* and *A. bwambae*. DNA sequence data from the mitochondria (mtDNA) and the nuclear rDNA intergenic spacer (BESANSKY *et al.* 1994; CACCONE *et al.* 1996) conflict with the inversion tree by strongly supporting a sister taxa relationship of *A. gambiae* and *A. arabiensis* to the exclusion of *A. merus*. A simple explanation would be that X^{ag} is not monophyletic. This could come about in two ways: either the inversion arose more than once independently or what is called X^{ag} in the two species is not the same inversion, but rather only sufficiently similar as to be indistinguishable at the level of light microscopic examination of polytene chromosomes. Because inversions protect genes within breakpoints from recombination, if inversions are monophyletic, then DNA sequences within inversions should reflect the evolutionary histories of the inversions, *i.e.*, the single gene copy captured in the

Corresponding author: Jeffrey R. Powell, Department of Biology, Yale University, 165 Prospect St., New Haven, CT 06520-8104.
E-mail: jeffrey.powell@yale.edu

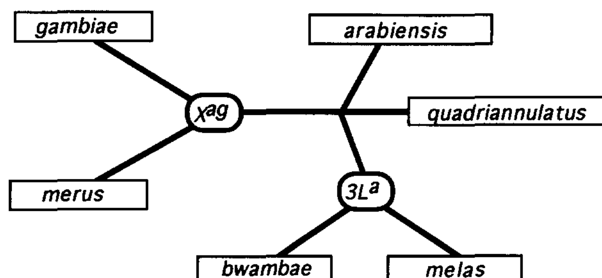

FIGURE 1.—Simplified inversion phylogeny for the *A. gambiae* complex species. The only inversions shown are the synapomorphies (shared derived) implying the sister taxa relationships. Several other polymorphic and autapomorphic inversions exist. From COLUZZI *et al.* (1979).

original chromosome will remain in complete linkage with the inversion. In Drosophila the monophyly of inversions has been confirmed by the fact that DNA sequences within the breakpoints of inversions produce gene trees identical to those deduced for the inversions and all alleles within the same inversion are more similar to one another than to any other alleles (AQUADRO *et al.* 1991; POPADIC and ANDERSON 1994; ROZAS and AGUADÉ 1994). Furthermore, in the only case studied, DNA sequences across inversion breakpoints are identical in several independent copies of the inversion (WESLEY and EANES 1994).

Based on this, we set out to test the monophyly of the *A. gambiae* X^{ag} inversion since it is the one that links *A. gambiae* and *A. merus* (Figure 1) and thus is responsible for the discordance with the molecular phylogenies. We studied a guanylate cyclase gene known to be within the breakpoints (Figure 2A). We sequenced multiple strains from different geographic origins belonging to five of the six species of the *A. gambiae* complex; *A. bwambae* was not available to us.

MATERIALS AND METHODS

Mosquito strains: The strains used in this study are listed in Table 1 together with their commonly used acronyms, their source, and the geographic origin. We used four strains of *A. gambiae* (SUA, G3, GMMKG and MU), two strains each of *A. merus* (V12 and ZULU), *A. melas* (BAL and BRE), and *A. quadriannulatus* (CHIL and SQUAD), and three strains of *A. arabiensis* (AB1, ARZAG and AJ).

DNA extraction, cloning: Total DNA was extracted from individual frozen adults or pupae using a modified Drosophila extraction protocol (LIVAK 1984). We chose the DNA fragment studied here from a set of random cDNA clones from *A. gambiae* that had been mapped by hybridization to microdissected divisional probes (MATHIOPOULOS and LANZARO 1995); exact location was confirmed by *in situ* hybridization to polytene chromosomes (A. DELLA TORRE, personal communication). We selected a 240-bp clone (pKM42) that had been mapped to division 3 within the X^{ag} inversion (FIGURE 2A; COLUZZI and SABATINI 1967). Using this clone as a probe, we screened an *A. gambiae* cosmid library according to the Screening of SuperCos 1 Cosmid Libraries protocol (Stratagene). The cosmid DNA from the positive colony was used to construct a restriction map. Digestion of this DNA with the restriction enzymes *Eag*I/*Pst*I and *Eco*RI/*Hin*dIII produced two fragments (1 and 2.4 kb long), which were subcloned into pBluescript II SK (Stratagene). Both fragments were sequenced by manual double-stranded sequencing using ^{35}S and the Sequenase version 2.0 DNA sequencing kit (U.S. Biochemical).

TABLE 1

Strains of the *A. gambiae* complex used in this study

Species	Strain abbreviation	Source	Geographic origin
A. gambiae	G3	CDC	Gambia
A. gambiae	MU	CDC	Tanzania
A. gambiae	SUA	RM1	Liberia
A. gambiae	GMMKG	CDC	Burkina Faso
A. arabiensis	ARZAG	CDC	Burkina Faso
A. arabiensis	AJ	RM1	Madagascar
A. arabiensis	AB1	RM1	Eritrea
A. merus	V12	CDC	Kenya
A. merus	ZULU	CDC	Zululand
A. melas	BAL	CDC	Gambia
A. melas	BRE	CDC	Gambia
A. quadriannulatus	CHIL	CDC	Zimbabwe
A. quadriannulatus	SQUAD	CDC	

CDC refers to the Center for Disease Control, Atlanta, courtesy of Drs. NORA BESANSKY and FRANK COLLINS. RM1 refers to Istituto di Parassitologia, I Università di Roma, Italy, courtesy of Prof. MARIO COLUZZI.

PCR and sequencing: Using the sequence information from the cloned *A. gambiae* fragments, we designed five primers that together amplify ~770 bp of the guanylate cyclase gene for all the Anopheles species used in this study. We used two external primers (primer 1: 5′-TCGGGTTCGGCCAGTACTGT, primer 4: 5′-ATGGTGGAGCTGCCGTACAA) and three internal primers (primer 6c: 5′-GTGGTTTCGCTGATGTTGATG, primer 6: the reverse complement of primer 6c, and primer 7: 5′-GACATGCTCGATATGGCGAAG). These primers were used for both double- and single-stranded DNA PCR amplifications and for direct sequencing. PCR amplifications were carried out in a Hybaid thermal cycler (Omnigene) in 50 μl of a solution containing 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, each dNTP at 2.5 mM, each primer at 1 μM, genomic DNA (10–100 ng) and 2 units of Amplitaq (Perkin-Elmer Cetus). After a first step of 45 sec of denaturation (94°), the amplification parameters for each of the 30 subsequent cycles of the PCR were as follows: 94° for 25 sec (denaturation), 55° for 60 sec (annealing), and 71.5° for 70 sec (extension). Double-stranded amplified products were isolated by electrophoresis in 1% agarose gels and used as templates to generate single-stranded DNA for direct sequencing by the unbalanced primer method (GYLLENSTEN and ERLICH 1988). The single-stranded DNA products were purified by electrophoresis in 1.2% low melting point agarose gels followed by gelase treatment (Epicentre Technologies). Both strands were sequenced with an automatic sequencer (Applied Biosystems Model 373A) following the manufacturer's protocols.

Data analysis: The sequences were aligned by eye since they showed high similarity among species. Sequences have been deposited in GenBank under accession numbers U42609–U42623 and can also be obtained from the authors. Genetic divergence within and between species was measured by maximum likelihood (ML) distances (FELSENSTEIN 1981) and four-parameter distances (TAKAHATA and KIMURA 1981). The latter

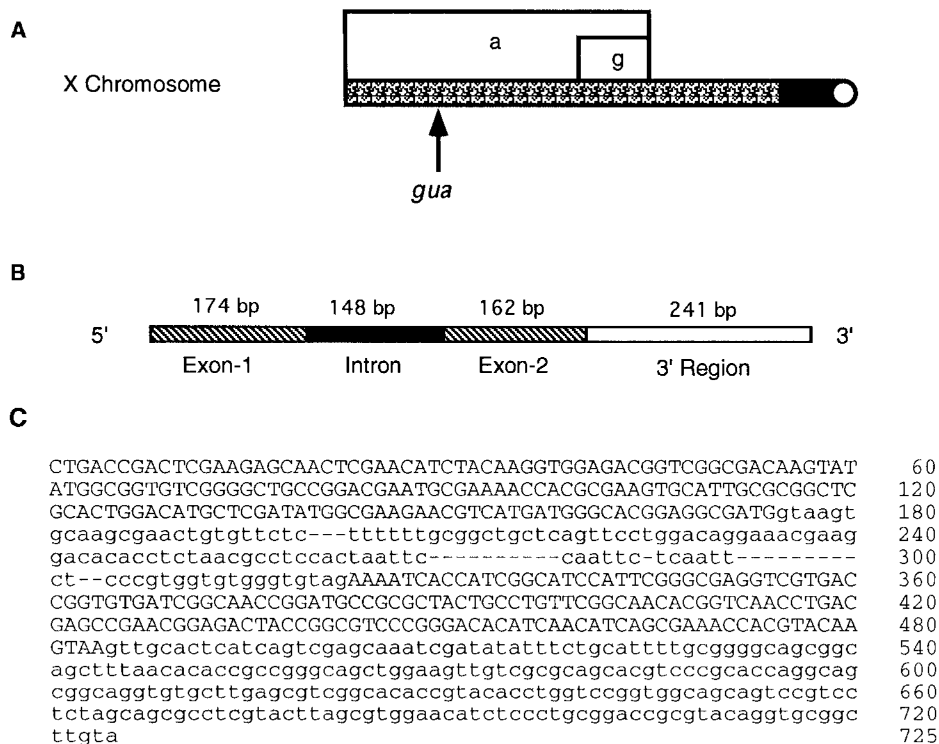

FIGURE 2.—The guanylate cyclase gene in *A. gambiae*. (A) Approximate location of *gua* inside the "ag" inversion on the *A. gambiae X* chromosome. The X^{ag} chromosome differs from the "standard" by two inversions, the "g" inversion being included within the "a", with one presumed common breakpoint. The light and dark shading represents euchromatin and heterochromatin, respectively. The relative length of the heterochromatin is much smaller in polytene chromosomes and larger in mitotic and meiotic chromosomes (GATTI *et al.* 1977). (B) Schematic representation of the different functional regions in the *gua* fragment. (C) Nucleotide sequence of *A. gambiae*, strain SUA. Nucleotide 1 is in the first coding position. Uppercase, exons 1 and 2; lowercase, intron and 3′ noncoding sequence. Gaps relative to other sequences in this study are denoted by a dash.

distances were used in the comparison of divergence of different regions.

Phylogenetic trees were inferred by neighbor-joining (NJ) (SAITOU and NEI 1987), maximum parsimony (MP), and ML (FELSENSTEIN 1981), and the robustness of the phylogenetic hypothesis was tested by bootstrapping (FELSENSTEIN 1985). MP analyses were performed using PAUP 3.1.1 (SWOFFORD 1993) with the following options: branch and bound search, MULPARS in effect, MAXTREES = 100. Indels (insertion/deletions) were coded as single characters, irrespective of their length. When indels of different length overlapped, each size class was a different character state. We ran parsimony analyses on all characters unweighted, using only transversions, and including and excluding indels. We used the entire region without differential weighting of coding and noncoding regions. NJ and ML analyses were carried out using PHYLIP 3.56c (FELSENSTEIN 1994). Three NJ analyses were performed using distance matrices based on ML (FELSENSTEIN 1981), the six-parameter model of GOJOBORI *et al.* (1982), and TAJIMA and NEI's (1984) method. The robustness of each node was evaluated by bootstrapping 100 times the original data set (MP analysis). The bootstrap values for the NJ and ML analyses were obtained by generating 100 distance matrices from the original data set (programs SEQBOOT and DNADIST in PHYLIP) and producing a majority rule consensus tree for the 100 NJ trees obtained from the distance matrices (programs NEIGHBOR and CONSENSE in PHYLIP).

RESULTS

Gene structure and homology: A schematic representation of the DNA fragment studied and a single sequence for the *A. gambiae* strain SUA are presented in Figure 2, B and C. Comparisons between the sequences of cDNA and genomic DNA for the SUA strain allowed us to identify three distinct regions: 336 bp of amino-acid coding sequence (organized into two exons), a 148-bp putative intron, and 241 bp of 3′ noncoding DNA. The two exons are very similar to the human (GIULI 1992) and Drosophila (SHAH and HYDE 1995) soluble guanylate cyclase gene, being 77.6/56.9% (exon 1) and 88.7/83% (exon 2) identical at the amino-acid level to the Drosophila/human genes, respectively. We tentatively designate this gene *gua*. The 3′ DNA region was considered noncoding because it does not have any similarity to any region of the cDNA sequence from Drosophila or vertebrates and has two stop codons (positions 614–616 and 680–682, Figure 2C), which occur only in some of the strains. A fuller description of location and structure of this gene will be provided in a future publication.

Intraspecific variation: We sequenced two individuals from each of two to four strains per species for a total of 23 sequences from 13 of the strains; for SUA, SQUAD, and AB1, only one individual was sequenced. All the analyses were carried out on a 725-bp fragment in common among all the individuals studied. Table 2 summarizes the levels of sequence variation found in this fragment. Including insertions and deletions (indels) a total of 89 variable sites were observed across all species. It is clear that the majority of variation is in the intron, with the exons showing the least. In the exons, six variable sites were synonymous and one was a replacement. The location of polymorphisms within strains and gene regions are shown in Figure 3.

Two types of intraspecific variation were detected: variation among individuals from the same strain and differences between conspecific strains. No variation occurred among individuals within the strains of *A. gam-*

TABLE 2

Summary of sequence variation in the *gua* gene among 13 strains of Anopheles

Gene region	Percentage variable sites	Length (bp)
Exon 1	2.9 (5)	174
Exon 2	1.2 (2)	162
Intron	20.9 (31) 22.3 (33)	148
3′ Noncoding	7.5 (18)	241

For the intron, upper figure is base pair substitutions and lower figure is indels. No indels occur in the other regions. Total number of variable sites is shown in parentheses.

biae, A. arabiensis, and *A. quadriannulatus.* Two alleles of the ZULU strain of *A. merus* differed at 10 sites and two alleles of the BRE strain of *A. melas* differed at 11 sites. Variation among conspecific strains of both *A. quadriannulatus* and *A. arabiensis* was quite limited (only one position). Conspecific strains of *A. gambiae, A. melas* and *A. merus* showed more variation. Nine, 13, and 17 sites were polymorphic within each of the three species, and the variation included both base-pair substitutions and indels.

Interspecific divergence: Because all methods of calculating distances produced very similar results, we present only the results for maximum likelihood distances (Table 3). This table also gives fixed differences and the number of base positions at which all sequences of one species differ from all sequences of the second species. Intraspecific differences range between 0.09 and 1.24%, the highest being for *A. melas. A. gambiae* strains are about half as divergent as the *A. melas* strains despite the fact that the sample of the former species was from a much wider geographic area. The other three species show very little intraspecific divergence (0.09–0.14%). Interspecific distances varied from 1.88 to 3.81%. The smallest average interspecific distances are between *A. gambiae* and *A. merus* (1.89%) and *A. arabiensis* and *A. quadriannulatus* (1.88%). On average *A. gambiae* and *A. merus* strains are about 3.1% divergent from all the other species.

There are 13 fixed differences between *A. gambiae* and *A. merus* strains. Among any other species pair the number of fixed differences ranges from 14 (*A. arabiensis vs. A. quadriannulatus*) to 35 (*A. melas vs. A. merus*), including both nucleotide substitutions and indels (Table 3). The number of "unique" fixed differences, or autapomorphies, is quite different among species. *A. gambiae* and *A. arabiensis* have only five and three autapomorphies, respectively. *A. merus* and *A. quadriannulatus* have seven and eight, respectively, while *A. melas* has a total of 16 autapomorphies.

Phylogenetic analyses: Over the entire 725-bp fragment 87.7% of the nucleotide positions were constant, 1.9% were variable but phylogenetically uninformative, and 10.3% were informative. Most of the informative variation occurs in the intron. Most variation within the coding region is silent and occurs in third positions.

Maximum parsimony with branch and bound search weighting transversions (TV) and transitions (TS) equally and including indels produced six equally parsimonious trees. These trees are 78 steps long with a consistency index (CI) of 0.936 (CI excluding uninformative characters = 0.875) and retention index (RI) of 0.956. Assigned branch lengths (ACCTRAN option in PAUP) for one of the six MP trees are shown in Figure 4. Other branch and bound searches excluding indels or using only transversions produced trees that are identical in topology to the one shown in Figure 4, except for minor topological rearrangements at the in-

```
Base       1111111111222222222222222222222222222222222222222222222223333333333444555555666666666666
position  55266788889000011222223333344445556667777777778888899999999999900000111100834455801344567789
          17659801563012918347890124602595791390123456783569012345678902349234603840658655615191517

Comment   ssssr                                                              ss

SUA       CGGGGatggct---ggcccgacaggacgactgccaa----------t-tatt---------t--ttgtgCCgcctggcagcctcccgac
GMMKG     ....................g.....a......................a...........................t......t
G3        ..............a....g.....r.......................w....................c......w......
MU        ....................g.....a.....................................................c.......t
ARZAG     ......a.......a.t.t-----acgag..ta...........................cg.ca.....c.c..c.....g.ct
AJ        ......a.......a.t.t-----acgag..ta...........................cg.ca.....c.c..c.....g.ct
ABI3      ......a.......a.t.t-----acgag..ta...........................cg.ca..a..c.c..c.....g.ct
BAL       ..ACA.a.t.....a.t.t-----acga.t.t....aactcctccaa.................c..a.G.ttc.cttct..gt...t
BRE1      .AAC..a.......a.t.t-----acga...t....aactcctccaa.................c..aT....c.c..ct...t.a.t
BRE2      ..ACA.a.t.....a.t.t-----acga...t....aactcctccaa..c..............c..a.G.ttc.ct.ct..gt...t
V12       T.....ac.....t...tt.g......a...t...............ac---cccccccc.cc..t.......c.c..c.......ct
ZULU1     T.....ac...ttt...tt.g......a...t..........................c..c..t.......c.c..c.......ct
ZULU2     T.....ac.....t...tt.g......a...t..........................c-cc..t.......c.c..c.......ct
CHIL      .....ga..ac....at.t-----acgag.ctat-c..................................c.c..c.......ct
SQUAD     .....ga..ac....at.t-----acgag.ctat-c.........................c......c.c..c.......ct
```

FIGURE 3.—Polymorphic sites among strains of *A. gambiae, A. arabiensis, A. melas, A. merus* and *A. quadriannulatus.* First three rows refer to base position relative to Figure 2C. "Comment" row: s, synonymous substitutions in an exon; r, amino acid replacement substitution; absence of a letter, nucleotide substitution within the intron or the 3′ noncoding region. The sequence of SUA (*A. gambiae*) is used as reference; deletions relative to other sequences are denoted by a dash. Nucleotides identical to the reference in the remaining 14 lines are indicated by a dot. Uppercase, exon sites; lowercase, intron and 3′ noncoding region.

TABLE 3

Differences within and between species

Species 1	Species 2				
	gambiae	*merus*	*arabiensis*	*melas*	*quadriannulatus*
gambiae	0.0055	0.0189	0.0301	0.0381	0.0326
merus	13 (3)	0.0009	0.0245	0.0354	0.0263
arabiensis	21 (4)	24 (8)	0.0014	0.0279	0.0188
melas	30 (14)	35 (18)	23 (10)	0.0124	0.0373
quadriannulatus	24 (5)	26 (9)	14 (1)	29 (11)	0.0014

Above diagonal average pairwise ML distances between species. Average intraspecific ML distances are on the diagonal. Below the diagonal is the number of fixed differences between each species pair, in parentheses, the number of those fixed differences that are indels.

traspecific level. Using TV only, the *A. gambiae* G3 strain becomes the sister taxon of the GMMKG strain instead of SUA. When indels were excluded, the relationship between the *A. merus* strains collapses to a trichotomy. NJ trees based on ML distances, six-parameter distances, and TAJIMA and NEI distances had the same topology as the one shown in Figure 4, as did the ML tree. The robustness of each node was evaluated by bootstrap analyses for the MP, ML, and NJ trees, and their respective values, top to bottom, for each node are reported on Figure 4.

In all analyses (MP, ML, and NJ) all intraspecific strains cluster together, and the support for the monophyly of each species is strong both in terms of branch lengths and bootstrap values, indicating that intraspecific variation does not cloud interspecific differences. Interestingly, *A. melas* strains are quite different from one another, both in terms of genetic distances (Table 3) and phylogenetically informative characters changes (Figure 3), even though they both come from The Gambia. This contrasts with the low genetic differentiation within all the other species and in particular within *A. gambiae,* whose four strains came from regions scattered throughout Central Africa.

The node clustering the *A. gambiae* strains with the *A. merus* strains is the most strongly supported node connecting any two pairs of species. The relationship of the other species to the *gambiae-merus* clade and to

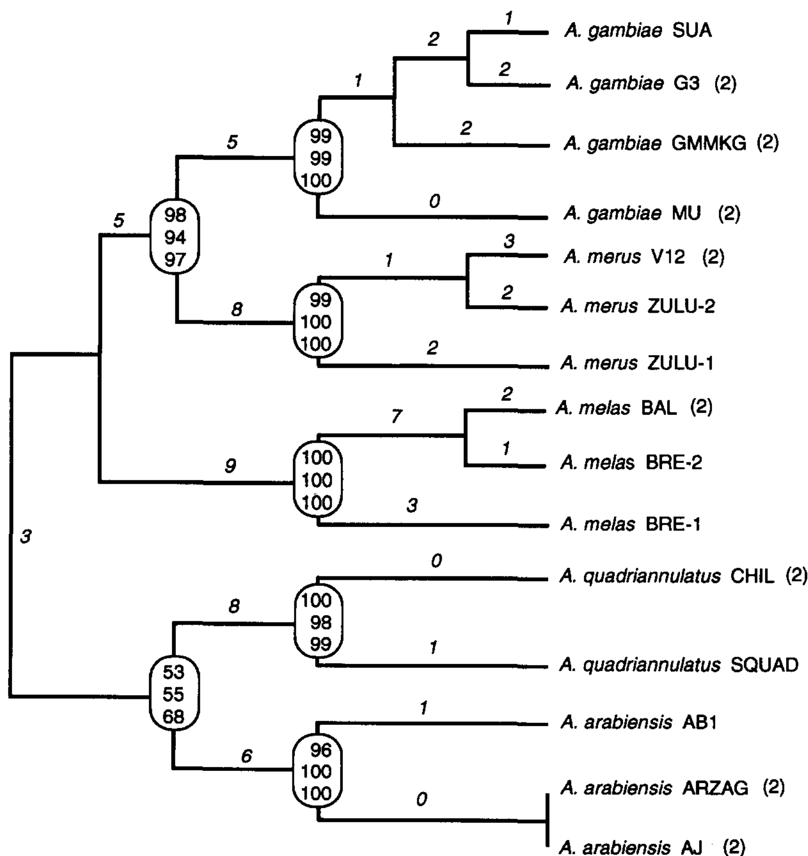

FIGURE 4.—Majority rule bootstrap consensus tree based on DNA sequences inside the inversion X^{ag} in five species of the *A. gambiae* complex; the tree is unrooted. Numbers at nodes are bootstrap values (100 replications) for parsimony (upper figure), maximum likelihood (middle figure), and neighbor-joining (bottom figure). Italicized numbers above branches are branch lengths for the parsimony tree (ACCTRAN option in PAUP, unweighted). Numbers next to the strain names at the tips of the tree mean two individuals for that strain were sequenced; when they had identical sequences a (2) appears; if they had different sequences the numbers 1 and 2 are shown.

TABLE 4

Average genetic distances for nuclear and mitochondrial DNA regions between species of the *Anopheles gambiae* complex

Gene Region	Exons *gua*	AT-rich mtDNA	3′ *gua*	ND5 mtDNA	3′ EST	ND4 mtDNA	ITS rRNA	Intron *gua*
Ave. D	0.017	0.020	0.021	0.027	0.036	0.074	0.080	0.128
Ratio	1	1.2	1.2	1.6	2.1	4.4	4.7	7.5

Ave. D is the average distance calculated by the four-parameter method of TAKAHATA and KIMURA (1981). Only third codon positions were used for coding regions. Exons *gua*, coding region of the nuclear guanylate cyclase gene; AT rich mtDNA, noncoding mtDNA; 3′ *gua*, 3′ noncoding nuclear DNA (*gua*); ND4 and ND5 mtDNA, coding regions of mtDNA; 3′ EST, 3′ noncoding region of nuclear esterase gene; ITS rRNA, intergenic transcribed spacer of nuclear rDNA. Data for ITS, 3′EST, ND4, and ND5 are from BESANSKY *et al.* (1994). Data for the AT-rich region are from CACCONE *et al.* (1996). Ratio is the relative rate of change standardized to the slowest evolving region, Exons *gua*.

each other is much less well supported. *A. arabiensis* strains cluster with the *A. quadriannulatus* strains, but this topology is weak both in terms of bootstrap values and branch lengths, even though the same relationship is obtained in all phylogenetic analyses, and the average ML distance value between the two species (0.0188, Table 3) is the smallest for interspecific comparisons.

Evolutionary rates between different genomic regions: In using the level of DNA differences among different parts of the genome to infer differences in rates of evolution, an assumption is made that the different parts have been evolving independently for an equal amount of time. We have reason to believe that introgression is occurring for parts of the genome between *A. gambiae* and *A. arabiensis* (BESANSKY *et al.* 1994; CACCONE *et al.* 1996; and see DISCUSSION). Therefore, we have excluded all comparisons with *A. arabiensis* in the following.

Four-parameter distances (TAKAHATA and KIMURA 1981) were calculated for pairwise species comparisons for different genomic regions. Table 4 shows these values for the guanylate cyclase fragment and compares them with the same distances for other parts of the genome: the intergenic transcribed spacer of the nuclear rRNA cluster, ITS; the 3′ noncoding region of an esterase gene, 3′ EST; two protein-coding regions of the mtDNA, ND4 and ND5 (all from BESANSKY *et al.* 1994); and the AT-rich noncoding region of the mtDNA (CACCONE *et al.* 1996). For coding regions, distances were computed only on third codon positions. In Table 4, distances between all pairwise combinations of the four species are averaged.

DISCUSSION

Before addressing the inversion monophyly issue, we address two other issues The first is the great heterogeneity of *A. melas* compared to the other species. *A. melas* breeds in brackish water on the west coast of Africa; the two strains we analyzed came from the small country Gambia. Yet the nucleotide difference among alleles from these two strains is more than twice that between alleles within any other species. The samples of *A. gambiae* and *A. arabiensis*, in particular, came from a wide geographic range, from west and east Africa and Madagascar, some 6000 km distant. Yet their nucleotide differences are very slight. We have no reason to believe there might be cryptic taxa within *A. melas* especially considering that two of the most different alleles came from individuals of the same strain, BRE (Figure 4). Previous studies of mtDNA from these same strains did not reveal any unusual degree of heterogeneity compared to the other species (BESANSKY *et al.* 1994; CACCONE *et al.* 1996). We have no explanation for this seeming anomaly seen with the *gua* gene fragment.

Second, in considering the relative rates of evolution of different parts of the genome (Table 4), the most surprising finding is the slow rate of evolution of the AT-rich region of the mtDNA, about the same rate as the third positions of the nuclear protein-coding gene, *gua*. We showed elsewhere (CACCONE *et al.* 1996) that the slow rate of the AT-rich region cannot be attributed to the base composition bias. The 3′ noncoding region of *gua* is evolving at about the same rate as third positions in the exons, a pattern not unexpected. However, the intron is evolving much more rapidly, a rather surprising finding. In comparing nuclear genes to mitochondrial genes, we see that the third position of protein-coding regions are evolving two to four times faster in the mtDNA, a difference consistent with other insects, Drosophila being the best-studied (POWELL 1997).

With regard to the main purpose of this study, the results presented here are strongly consistent with the X^{ag} being a monophyletic inversion in the *A. gambiae* complex. Evidently the gene sequences within X^{ag} have remained associated with this inversion, *i.e.*, neither gene conversion nor rare double crossovers have destroyed the association, at least not in the sample of chromosomes we have studied. This is not surprising as this inversion is fixed in *A. gambiae* and *A. merus* and thus, unlike a floating polymorphism, there is less chance for disassociation. Presumably the inversion went to fixation in a lineage common to the two species and, if this occurred by selection, the sojourn time when the association could be destroyed, would be short. This

supports the phylogenetic inferences shown in Figure 1, *i.e., A. gambiae* and *A. merus* are sister taxa. The node connecting these two species is well supported by the *gua* DNA sequence data; in fact it is the only well-supported interspecific node (Figure 4). There is some indication that *A. quadriannulatus* and *A. arabiensis* are sister taxa, a result consistent with MATHIOPOULOS *et al.* (1995), although the statistical support (bootstrap values) for this relationship is not strong in either study and must be considered tentative. The sister taxa relationship of *A. melas* and *A. bwambae* indicated by the synapomorphic $3L^a$ (Figure 1) was supported by the DNA sequences of the AT-rich region of mtDNA (CACCONE *et al.* 1996). Therefore, to date all the DNA sequence data have been consistent with the monophyletic origin of inversions in the *A. gambiae* complex.

Given these results, how can we reconcile them with the strong indication from mtDNA that *A. gambiae* and *A. arabiensis* are sister taxa (BESANSKY *et al.* 1994; CACCONE *et al.* 1996)? One possibility is that it is stochastic, *i.e.*, for chance reasons the two different types of molecules produce different trees. While we cannot rule this out, the strong bootstrap values associated with the two contrasting phylogenies would seem to argue against this. Each data set contains strong and clear phylogenetic signal. Another possibility is lineage sorting of ancestral polymorphisms. We tend to discount this as multiple sequences from each species always cluster as expected if sufficient time since last common ancestor has been long enough to eliminate misleading ancestral polymorphism. This is true for both the nuclear gene studied here and for the mtDNA sequences (BESANSKY *et al.* 1994; CACCONE *et al.* 1996), with one important exception noted below.

We favor the hypothesis that introgression for mtDNA is occurring (or has occurred in the recent past) between *A. gambiae* and *A. arabiensis.* The only exception to all alleles within a species clustering into a single lineage is for mtDNA for these two species; multiple alleles interdigitate in a phylogenetic analysis (BESANSKY *et al.* 1994) as expected for introgressing sequences. *A. gambiae* and *A. arabiensis* are widely sympatric and hybrids are known from nature at a rate of ~0.1–0.2% (WHITE 1971; COLUZZI *et al.* 1979; PETRARCA *et al.* 1991). Hybrid females are fertile. In other groups there is indication that mtDNA may introgress more easily than nuclear DNA (FERRIS *et al.* 1983; POWELL 1983; SOLIGNAC *et al.* 1986). Elsewhere (CACCONE *et al.* 1996) we present evidence that the likely direction of introgression has been from *A. gambiae* to *A. arabiensis.*

More difficult to reconcile with these results is the fact that the intergenic transcribed spacer region (ITS) of the nuclear rRNA cluster also supports the sister taxa status of *A. gambiae* and *A. arabiensis* (BESANSKY *et al.* 1994). The nuclear rRNA cluster resides in the centromeric heterochromatin of the *X* chromosome in *A. gambiae* (COLLINS *et al.* 1989) and is therefore linked to the X^{ag} inversion. If introgression is the explanation for the ITS sequences being very similar in *A. gambiae* and *A. arabiensis,* then clearly there must be sufficient recombination to allow the rRNA region to introgress while the X^{ag} is selected against. About half of the *X* chromosome is heterochromatic in meiotic cells (GATTI *et al.* 1977), so there is room for recombination.

An alternative explanation is that the X^{ag} has introgressed between *A. merus* and *A. gambiae* and the mtDNA and nuclear ITS are indicating the true phylogenetic relationship. Evidence against this explanation comes from laboratory experiments on hybrid populations. A. DELLA TORRE (personal communication) maintained hybrid populations for several generations in the laboratory and found that *X*-chromosome inversions are selected against in hybrid populations and are eliminated within two generations; introgressed autosomal inversions could persist for up to 20 generations, the duration of the experiments.

Clearly, the resolution of the question of species phylogeny *vs.* gene phylogenies in this group of mosquitoes is a complex problem. The view is emerging of a mosaic genome consisting partly of species-specific regions and partly of introgressed regions. In such cases it would seem most likely that the species-specific regions would be the best indicators of species phylogenetic relationships. The results of DELLA TORRE mentioned in the previous paragraph would argue that *X*-chromosome inversions might be the best candidates for species-specific regions. In fact, the gene arrangements of the *X* chromosome have long been used as the species-specific diagnostic feature distinguishing *A. gambiae* from *A. arabiensis* (COLUZZI 1966). The data from *gua* presented here are the first DNA sequences for a gene within an *X* chromosome inversion in the *A. gambiae* complex.

Regardless of the final resolution of the conflicting data sets, the results here are strongly indicative of introgression between these two most important vectors of malaria in the world. The practical implications of this could be substantial. For example, insecticide-resistance genes could pass between species. COLUZZI *et al.* (1979) have evidence that introgression of an autosomal inversion has led to the adaptive expansion of *A. arabiensis* into a new ecological niche. Finally, any attempts to genetically manipulate *A. gambiae* or *A. arabiensis* in a disease control program may be compromised by introgression.

We thank NORA BESANSKY, FRANK COLLINS, ALESSANDRA DELLA TORRE, and MARIO COLUZZI for supplying mosquitoes, and ETSUKO MORIYAMA for help with the analysis. This work was supported by the U.S. National Institutes of Health, Grant RO1-AI 35215. B.A.G. was partially supported by the Latin American Fellowship Program of the Pew Charitable Trusts.

LITERATURE CITED

AQUADRO, C. F., A. L. WEAVER, S. W. SCHAEFFER and W. W. ANDERSON, 1991 Molecular evolution of inversions in *Drosophila pseu-*

doobscura: the amylase gene region. Proc. Natl. Acad. Sci. USA **88:** 305–309.

BESANSKY, N. J., J. R. POWELL, A. CACCONE, D. MILLS HAMM, J. A. SCOTT, *et al.*, 1994 Molecular phylogeny of the *Anopheles gambiae* complex suggests genetic introgression between principal malaria vectors. Proc. Natl. Acad. Sci. USA **91:** 6885–6888.

CACCONE, A., B. A. GARCÍA and J. R. POWELL, 1996 Evolution of the mitochondrial DNA control region in the *Anopheles gambiae* complex. Insect Mol. Biol. **5:** 51–59.

COLLINS, F. H., S. M. PASKEWITZ and V. FINNERTY, 1989 Ribosomal RNA genes of the *Anopheles gambiae* complex, pp. 1–28 in *Advances in Disease Vector Research,* Vol. 6, edited by K. F. HARRIS. Springer-Verlag, New York.

COLUZZI, M., 1966 Osservazioni comparative sul cromosoma *X* nelle specie A e B del complesso *Anopheles gambiae.* Rend. Acad. Naz. Lincei **40:** 671–678.

COLUZZI, M., and A. SABATINI, 1967 Cytogenetic observations on species A and B of the *Anopheles gambiae* complex. Parassitologia **9:** 73–88.

COLUZZI, M., A. SABATINI, V. PETRARCA and M. A. DI DECO, 1979 Chromosomal differentiation and adaptation to human environments in the *Anopheles gambiae* complex. Trans. R. Soc. Trop. Med. Hyg. **73:** 483–497.

DAVIDSON, G., H. E. PATERSON, M. COLUZZI, G. F. MASON and D. W. MICKS, 1967 The *Anopheles gambiae* complex, pp. 211–250 in *Genetics of Insect Vectors of Disease,* edited by J. W. WRIGHT and R. PAL. Elsevier, Amsterdam.

FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

FELSENSTEIN, J., 1985 Confidence limits on phylogenies: an approach using bootstrap. Evolution **39:** 783–791.

FELSENSTEIN, J., 1994 PHYLIP, version 3.56c. University of Washington, Seattle.

FERRIS, S. D., R. D. SAGE, C.-M. HUANG, J. T. NIELSEN, U. RITTE *et al.*, 1983 Flow of mitochondrial DNA across a species boundary. Proc. Natl. Acad. Sci. USA **80:** 2290–2294.

GATTI, M., G. SANTINI, S. PIMPINELLI and M. COLUZZI, 1977 Fluorescence banding techniques in the indentification of sibling species of the *Anopheles gambiae* complex. Heredity **38:** 105–108.

GIUILI, G., U. SCHOLL , F. BULLE and G. GUELLAEN, 1992 Molecular cloning of the cDNAs coding for the two subunits of soluble guanylyl cyclase from human brain. FEBS Lett. **304:** 83–88.

GOJOBORI, T., K. ISHII and M. NEI, 1982 Estimation of average number of nucleotide substitutions when the rate of substitutions varies with nucleotide. J. Mol. Evol. **18:** 414–423.

GYLLENSTEN, U. B., and H. A. ERLICH, 1988 Generation of single-stranded DNA by the polymerase chain reaction and its application to direct sequencing of the HLA-DQA locus. Proc. Natl. Acad. Sci. USA **85:** 7652–7656.

LIVAK, K. J., 1984 Organization and mapping of a sequence on the *Drosophila melanogaster X* and *Y* chromosomes that is transcribed during spermatogenesis. Genetics **107:** 611–634.

MATHIOPOULOS, K. D., and G. C. LANZARO, 1995 Distribution of genetic diversity in relation to chromosomal inversions in the malaria mosquito *Anopheles gambiae.* J. Mol. Evol. **40:** 578–584.

MATHIOPOULOS, K. D., J. R. POWELL and T. F. MCCUTCHAN, 1995 An anchored restriction-mapping approach applied to the genetic analysis of the *Anopheles gambiae* Malaria Vector Complex 1. Mol. Biol. Evol. **12:** 103–112.

PETRARCA, V., J. BEIER, C. ONYANGO, F. KOROS, J. ASIAGO *et al.*, 1991 Species composition of the *Anopheles gambiae* complex (diptera: Culicidae) at two sites in western Kenya. J. Med. Entomol. **28:** 307–313.

POPADIC A., and W. W. ANDERSON, 1994 The history of a genetic system. Proc. Natl. Acad. Sci. USA **91:** 6819-6823.

POWELL, J. R., 1983 Interspecific cytoplasmic gene flow in the absence of nuclear gene flow: evidence from *Drosophila.* Proc. Natl. Acad. Sci. USA **80:** 492–495.

POWELL, J. R., 1997 *Progress and Prospects in Evolutionary Biology: The Drosophila Model.* Oxford University Press, New York.

ROZAS, J., and M. AGUADÉ, 1994 Gene conversion is involved in the transfer of genetic information between naturally occurring inversions of *Drosophila.* Proc. Natl. Acad. Sci. USA **91:** 11517–11521.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

SHAH, S., and D. R. HYDE, 1995 Two *Drosophila* genes that encode the alpha and beta subunits of the brain soluble guanylyl cyclase. J. Biol. Chem. **270:** 15368-15376.

SOLIGNAC, M., MONNEROT, M. and J. C. MOUNOLOU, 1986 Concerted evolution of sequence repeats in *Drosophila* mitochondrial DNA. J. Mol. Evol. **24:** 53-60.

STURTEVANT, A. H., and T. DOBZHANSKY, 1936 Inversions in the third chromosome of wild races of *Drosophila pseudoobscura* and their use in the study of the history of the species. Proc. Natl. Acad. Sci. USA **22:** 448–452.

STÜRCHLER, D., 1989 How much malaria is there worldwide? Parasitol. Today **5:** 39–40.

SWOFFORD, D., 1993 PAUP: phylogenetic analysis using parsimony, version 3.1.1s: program and documentation. Illinois Natural History Survey, Urbana.

TAJIMA, F., and M. NEI, 1984 Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1:** 269–285.

TAKAHATA, N., and M. KIMURA, 1981 A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics **98:** 641–657.

WESLEY, C. S., and W. F. EANES, 1994 Isolation and analysis of the breakpoint sequences of chromosome inversion In(3L)Payne in *Drosophila melanogaster.* Proc. Natl. Acad. Sci. USA **91:** 3132–3136.

WHITE, G. B., 1971 Chromosomal evidence for natural interspecific hybridization by mosquitoes of the *Anopheles gambiae* complex. Nature **231:** 184–185.

WHITE, G. B., 1974 The *Anopheles gambiae* complex and disease transmission in Africa. Trans. R. Soc. Trop. Med. Hyg. **68:** 278–301.

Communicating editor: W. F. EANES