Journal of Heredity 2003:94(2):133–147
DOI: 10.1093/jhered/esg024

Population Structure of *Anopheles gambiae* in Africa

T. LEHMANN, M. LICHT, N. ELISSA, B. T. A. MAEGA, J. M. CHIMUMBWA, F. T. WATSENGA, C. S. WONDJI, F. SIMARD, AND W. A. HAWLEY

From the Entomology Branch, Division of Parasitic Diseases, Centers for Disease Control and Prevention, 4770 Buford Highway, Chamblee, GA 30041 (Lehmann, Licht, and Hawley); Department of Biology, Emory University, Atlanta, GA 30322 (Lehmann and Licht); Unité d'Entomologie Médicale, CIRMF, BP 769, Franceville, Gabon (Elissa); National Institute for Medical Research (NIMR), Box 9653 Dar Es Salaam, Tanzania (Maega); National Malaria Control Center, P.O. Box 32509, Lusaka, Zambia (Chimumbwa); Laboratoire D'Entomologie, Avenue des Huileries, B.P. 1197, Ministry of Health, Kinshasa, Democratic Republic of Congo (Watsenga); and Laboratoire de l'Institut de Recherche pour le Developpement, Organisation de lutte Contre les Grandes Endemies en Afrique Centrale, Yaoundé, Cameroon (Wondji and Simard). We are grateful to Mark Danga, Francis Atieli, George Olang', and Charles Mbogo (Kenya Medical Research Institute) for their help in field collections. We thank Fernando Monteiro, Martin Donnelly, Jean-Pierre Dujardin, and Nora Besansky for discussions on earlier versions of this manuscript. Thanks to the staff of the CDC Core Facility, Computer Support and Data Management for their help, and to Sue Dillard for her dedicated work on Figure 7. This investigation received financial assistance from the UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (TDR).

Address correspondence to Tovi Lehmann at the address above, or e-mail: lbt2@cdc.gov.

Abstract

The population structure of *Anopheles gambiae* in Africa was studied using 11 microsatellite loci in 16 samples from 10 countries. All loci are located outside polymorphic inversions. Heterogeneity among loci was detected and two putative outlier loci were removed from analyses aimed at capturing genome-wide patterns. Two main divisions of the gene pool were separated by high differentiation ($F_{ST} > 0.1$). The northwestern (NW) division included populations from Senegal, Ghana, Nigeria, Cameroon, Gabon, Democratic Republic of Congo (DRC), and western Kenya. The southeastern (SE) division included populations from eastern Kenya, Tanzania, Malawi, and Zambia. Inhospitable environments for *A. gambiae* along the Rift Valley partly separate these divisions. Reduced genetic diversity in the SE division and results of an analysis based on private alleles support the hypothesis that a recent bottleneck, followed by colonization from the NW populations shaped this structure. In the NW division, populations possessing the M rDNA genotype appeared to form a monophyletic clade. Although genetic distance increased with geographic distance, discontinuities were suggested between certain sets of populations. The absence of heterozygotes between sympatric M and S populations in the DRC and the high differentiation in locus 678 ($F_{ST} > 0.28$) contrasted with low differentiation in all other loci ($-0.02 < F_{ST} < 0.09$) and with the persistence of departures from Hardy–Weinberg expectations within each form in the DRC. Neither recent reproductive isolation alone nor selection alone can explain these results, a situation that is compatible with incipient speciation. Because it is possible that the molecular forms play different roles in malaria transmission, future studies should treat them separately.

Malaria claims the lives of more than one million people every year, mostly in Africa (Marshall 2000). As the principal vector of malaria in Africa, *Anopheles gambiae* is arguably that continent's most dangerous animal. More effective methods need to be developed to alleviate malaria's burden. Knowledge of *A. gambiae*'s population structure may be useful for identifying heterogeneity in disease transmission due to distinct vector populations and for tracking and predicting the spread of genes of interest, such as those conferring insecticide resistance.

Many studies of population structure assume migration-drift equilibrium and use F statistics (Wright 1978) to estimate gene flow between populations. The limitations of this approach have been stressed in several recent studies (Whitlock and McCauley 1999), but the lack of alternative analytical tools to interpret genetic variation between populations remains a problem, especially for pest species whose demographics are thought to be historically unstable. Understanding the population structure of *A. gambiae*, the African malaria mosquito, requires consideration of the

historical instability in population size (Donnelly et al. 2001), selection effects on large segments of chromosomal regions contained within inversions, and speciation events (e.g., Coluzzi et al. 1979, 1985; Lanzaro et al. 1998; Powell et al. 1999; Taylor et al. 2001; Toure et al. 1998). It is therefore not surprising that a continental analysis of the population structure of *A. gambiae* has never been published. In this study we describe the population structure of *A. gambiae* on a continental scale and attempt to infer the processes that have shaped it using conventional, adapted, and ad hoc statistics.

Early studies of *A. gambiae* in West Africa revealed heterozygote deficits for certain inversions and suggested that the species is subdivided into partly or fully isolated chromosomal forms termed Mopti, Bamako, Bissau, Forest, and Savanna (Bryan et al. 1982; Coluzzi et al. 1979, 1985). These studies also showed that the frequency of certain inversion arrangements increased every dry season and decreased every wet season, demonstrating selective effects. Most subsequent studies encompassed small geographic scales. Surprisingly high genetic similarity was found between chromosomal forms outside polymorphic inversions, except in ribosomal DNA (rDNA), where two distinct genotypes (molecular forms) have been found (Favia et al. 1997; Gentile et al. 2001; Mukabayire et al. 2001). In Mali and Burkina Faso, one rDNA genotype (M) was solely associated with the Mopti chromosomal form and the other (S) was associated with the Savanna and Bamako forms. In neighboring countries, however, this association broke down, as both genotypes were found in the Savanna and Forest chromosomal forms (della Torre et al. 2001). Consistent with near-complete isolation between these populations, M/S heterozygotes were very rare ($\leq$0.3%), even in sympatric populations (della Torre et al. 2001; Wondji et al. 2001), and a large difference was found in the frequency of the *kdr* mutation (on the second chromosome) between them (Weill et al. 2000). Study of reproductive isolation between *A. gambiae* populations has shifted from chromosomal to molecular forms, but the debate on the taxonomic status of all these populations has yet to be resolved (e.g., Black and Lanzaro 2001; Tripet et al. 2001). In contrast to the situation in west Africa, no indication of reproductive isolation among populations has been found in east Africa (Kamau et al. 1998; Lehmann et al. 1997; Petrarca and Beier 1992; Smits et al. 1996). High differentiation was found between populations separated by the Rift Valley complex (Kamau et al. 1999; Lehmann et al. 1999, 2000), but differentiation was surprisingly low between populations from east and west Africa (Besansky et al. 1997; Kamau et al. 1999; Lehmann et al. 1996, 1999). An analysis of the structure of *A. gambiae* populations from a continental perspective may help resolve some of these enigmas.

In this study we analyze the population structure of *A. gambiae* based on variation at 11 microsatellite loci in populations from Kenya, Tanzania, Malawi, Zambia, Democratic Republic of Congo (DRC), Gabon, Cameroon, Nigeria, Ghana, and Senegal. To minimize confounding genome-wide patterns with locus-specific effects, we selected loci distributed throughout the genome, but located outside known polymorphic inversions. The following questions were addressed: What are the main divisions of the gene pool? How are they related to the rDNA genotypes and to geography? What can be inferred about the processes that shaped this structure?

Materials and Methods

Samples and Collection Methods

Sixteen collections were made between 1992 and 2000 from the locations shown in Figure 1. Sample sizes ranged from 29 to 85 *A. gambiae* specimens per collection (Table 1). The locations included Asembo Bay (1994; hereafter referred to as Asembo), Kisian, Kimilili, and Awendo from western Kenya (1996); Malindi and Jego from eastern Kenya (1996); Mbeya from Tanzania (1997); Mangochi and Nsanje from Malawi (1992); Zambezi from Zambia (1999); Kinshasa from the DRC (2000); Benguia from Gabon (1999); Simbock from Cameroon (1998); Gwamlar from Nigeria (1999); Navrongo from Ghana (1997); and Barkedji from Senegal (1995). Indoor-resting adult mosquitoes (mostly females) were collected by pyrethrum spray or aspiration in Kisian, Kimilili, Awendo, Malindi, Jego, Malawi, Zambia, Nigeria, the DRC, and Ghana. In Tanzania, Cameroon, Gabon, and Senegal, blood-seeking mosquitoes were collected by human-baited night catches. In Asembo, blood-fed and blood-seeking females were collected at dawn by aspiration from net traps hung over the beds of sleeping volunteers. At each site, mosquitoes were collected within 1 week from houses less than 10 km apart.

Karyotypically, populations from Cameroon and Gabon consisted of the Forest form (Fontenille D and Elissa N, unpublished data). Since Kinshasa (DRC) lies within the zone where this type is prevalent, presumably these species are also Forest form. All the other specimens were collected from populations where only the Savanna form was found (Coluzzi et al. 1985; della Torre et al. 2001; Fontenille D, unpublished data; Petrarca et al. 2000).

DNA Extraction and Microsatellite Genotype Scoring

Procedures and loci were previously described in detail (Donnelly et al. 2001; Lehmann et al. 1996, 1997, 1998). Only *A. gambiae* were included in the analysis after species identification was carried out (Scott et al. 1993). Microsatellite loci mapped to regions outside polymorphic chromosomal inversions in *A. gambiae* were selected (Figure 2 and Table 1). Microsatellite alleles were polymerase chain reaction (PCR) amplified with one fluorescent-labeled primer and their size was determined using the ABI 377 sequencing system. Polymerase chain reaction restriction fragment length polymorphism (PCR-RFLP) assays were used to determine the molecular form (Favia et al. 1997). Available specimens from Mangochi (Malawi), Cameroon, and Senegal consisted of one to two legs. DNA extracts of the specimens from Cameroon and Mangochi were

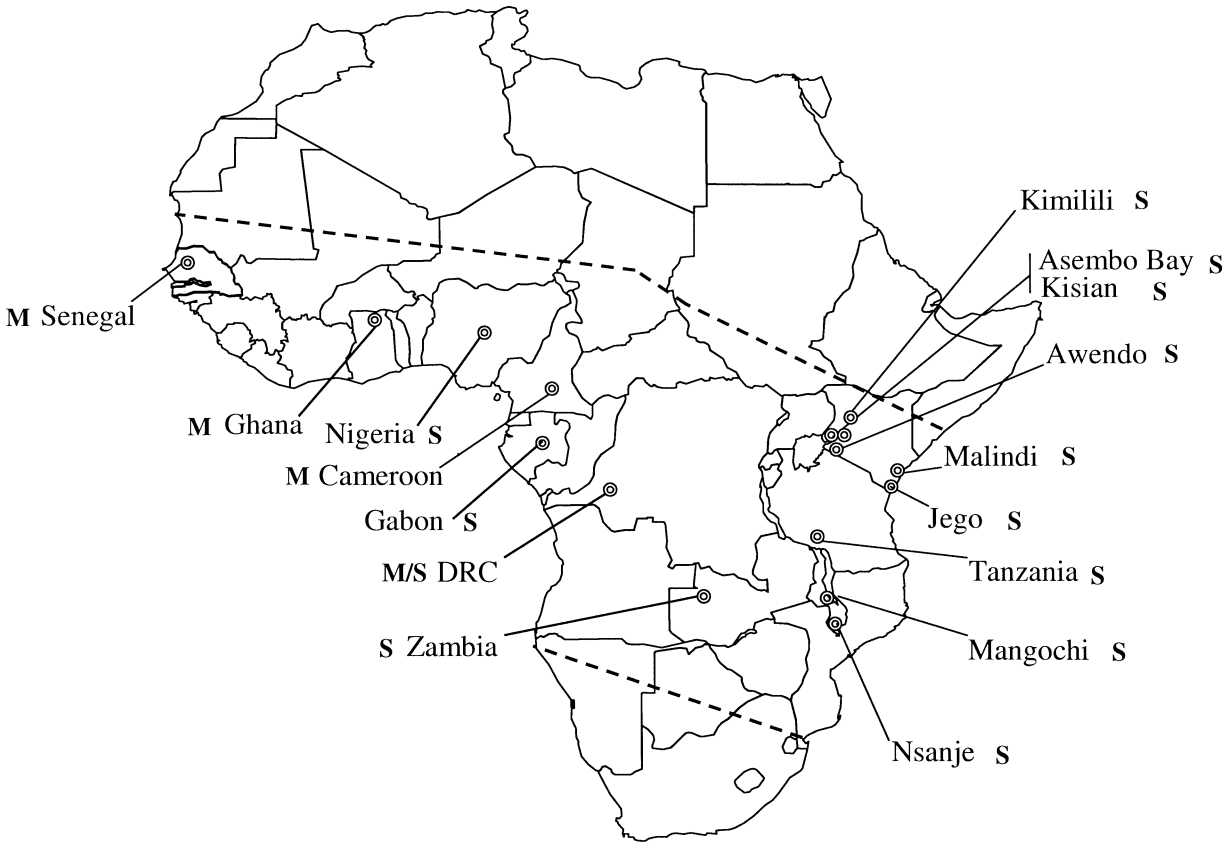

Figure 1. Schematic showing sampling sites of *A. gambiae* in Africa and its approximate range boundaries (dashed lines). "M" and "S" denote molecular forms.

exhausted after genotyping 7 loci, while that of Senegal was exhausted after genotyping 10 loci.

Data Analysis

Goodness-of-fit tests for Hardy–Weinberg (HW) expectations and genotypic linkage disequilibrium were performed using exact tests available in GENEPOP 3.2 (Raymond and Rousset 1995a). Differentiation among populations was assessed by F statistics (Wright 1978) calculated using GENEPOP according to Weir and Cockerham (1984). Deviation of a single locus F_{ST} from zero was tested by exact tests of heterogeneity of gene frequencies (Raymond and Rousset 1995b) using GENEPOP. F_{ST} is a better measure of differentiation than R_{ST} where differentiation is shaped primarily by drift (Slatkin 1995), as was the case for *A. gambiae* (Lehmann et al. 1998, 1999). Global tests were employed to evaluate the significance of multiple tests. The sequential Bonferroni procedure (Holm 1979) can detect a single test-specific departure, such as a locus-specific departure, whereas the binomial test (which estimates the probability of obtaining the observed number of significant tests at the 0.05 level given the total number of tests) can detect weaker departures across multiple tests, such as genome-wide departures. Bootstrapping was performed to calculate 95% confidence intervals (CIs) of parameter estimates using 1000 bootstrap replications, unless specified otherwise. The 95% CI of regression slopes was obtained by bootstrapping residuals (Mooney and Duval 1993). Neighbor-joining population trees were computed using MEGA 2.1 (Kumar et al. 1993) based on matrices of F_{ST} values. The robustness of population trees was evaluated by bootstrapping over loci. The bootstrapped matrices were analyzed using PHYLIP 3.5 (Felsenstein 1989). Bootstrap support values greater than 75% were considered biologically significant (Hillis and Bull 1993). Calculations not available in GENEPOP, MEGA, and PHYLIP were carried out using programs written by T.L. in SAS language (SAS 1990).

Results

Specimens from Senegal, Ghana, and Cameroon consisted of the M genotype. The DNA of eight specimens from Cameroon was exhausted before M/S assays were carried out and two repeatedly failed to produce PCR product. All other populations consisted of the S genotype, except the DRC, where both genotypes coexisted (M:S ratio of 32:15, and 2 which did not yield PCR product). Of importance is that no M/S heterozygotes were observed.

Table 1. Genetic diversity, sample size, and compliance with Hardy–Weinberg expectations (summarized by F_{IS}) across populations and loci

	Seneg	Ghan	Niger	Came	Gabo	DRC	Western Kenya					Tanza	Malawi		Eastern Kenya	
Locus	Barke	Ghan	Gwa	Simb	Beng	Kinsh	Asem	Kisia	Kimil	Awen	Zamb	Mbey	Mang	Nsanj	Jego	Malin
99 (X:3-1)[a]	100	90	104	74	118	93	164	123	78	114	58	74	128	120	106	123
A/A_{boot}[b]	6/5	6/5	5/5	7/6	7/6	7/6	5/4	6/5	5/5	5/4	5/5	4/4	4/4	4/4	5/5	5/5
H_e[c]	73	70	74	78	72	78	74	75	72	73	57	48	53	51	56	44
F_{IS}[d]	.10	.08	.27*	.10	.15	.19*	.09	.19*	.16	.12	−.09	.04	−.10	−.11	−.04	.11
1D (X:1d)	98	88	104	72	118	93	158	124	68	113	58	72	143	106	110	123
A/A_{boot}	3/3	5/4	6/4	4/4	3/3	4/4	6/4	5/4	4/3	5/3	4/3	3/3	5/4	3/3	3/3	3/3
H_e	50	57	57	65	50	56	55	55	49	50	51	29	41	27	29	18
F_{IS}	.12	−.25	.12	.11*	.08	.19	−.11	−.24*	.11	.19	−.01	.03	.11	−.13	.05	−.07
2A (X:2a)	100	90	104	70	116	82	143	121	73	110	58	58	104	116	96	92
A/A_{boot}	10/7	6/6	7/6	5/4	7/5	7/5	10/7	9/7	7/6	7/6	4/4	4/4	3/3	6/5	4/4	4/4
H_e	63	55	60	64	68	71	65	63	66	66	69	68	58	65	60	66
F_{IS}	−.02	.15*	.11	.11	.26*	.31*	.18*	.18	.18*	−.09	**.51*****	.06	.08	.18	**.35****	.24
678 (X:6)	0	84	104	0	118	93	169	124	67	95	58	76	0	128	106	121
A/A_{boot}	—	7/6	21/15	—	10/8	10/8	19/15	22/16	14/12	15/13	13/11	11/10	—	10/9	10/9	9/7
H_e	—	40	89	—	72	74	88	90	87	88	87	88	—	82	81	78
F_{IS}	—	−.02	.03	—	.04	**.36*****	.09	.03	.21*	−.01	.01	.01	—	−.04	.05	−.02
46 (II:7a)	100	90	100	70	118	96	157	124	86	124	58	73	132	123	110	118
A/A_{boot}	13/11	14/11	11/10	9/8	9/8	11/10	12/10	10/9	10/9	11/9	9/8	10/9	8/8	10/9	10/9	9/8
H_e	87	89	87	85	84	88	87	88	86	86	81	85	80	80	76	81
F_{IS}	.01	−.03	.13	−.05	.12**	.00	.04	−.07	−.02	.10	−.07	.09	.08*	−.05	.16	.06
197 (II:7-8)	69	84	104	—	114	95	148	122	72	114	58	72	—	111	110	127
A/A_{boot}	12/10	15/12	12/10	—	16/12	15/12	20/12	21/13	17/13	18/12	10/8	10/10	—	10/8	12/10	10/8
H_e	79	86	84	—	84	83	82	85	87	85	80	81	—	74	81	78
F_{IS}	.01	.04	.17	—	.04	−.03	.06	.05	.14	.16*	.18*	−.11	—	−.01**	−.01	−.02
79 (II:10)	96	84	104	—	118	98	162	124	66	108	58	68	—	127	112	132
A/A_{boot}	10/8	8/7	9/6	—	5/4	13/9	7/5	6/5	6/5	5/5	7/6	6/5	—	5/5	5/4	5/5
H_e	77	79	69	—	49	54	69	73	68	65	73	74	—	75	68	70
F_{IS}	−.09	.09	.09	—	.13	.09	.049	−.04	−.03	.23*	.05	.12	—	.03	.13	−.03
29C (III:29c)	100	90	104	72	118	96	160	120	86	126	58	74	144	124	108	132
A/A_{boot}	2/2	3/2	3/3	2/2	2/2	3/2	2/2	3/2	2/2	2/2	2/2	2/2	2/2	2/2	2/2	2/2
H_e	26	31	41	38	41	35	47	40	44	43	50	50	50	50	43	48
F_{IS}	−.01	−.03	.03	.13	−.07	.13	.20	−.18**	−.12	.08	−.18	−.14	.00	.07	.09	.06
33C (III:33c)	100	90	104	69	118	98	158	124	81	123	58	73	137	126	104	134
A/A_{boot}	11/9	10/10	14/9	11/9	11/8	9/8	18/11	16/10	12/10	14/11	6/5	5/5	9/7	6/5	5/4	6/4
H_e	74	78	82	76	75	74	78	75	82	77	66	50	58	56	41	45
F_{IS}	.02	−.00	.15	.07	.07	.23	.12	.12	−.01	.08	.16	.00	.11*	−.15	.02	.05
577 (III:42)	87	82	104	—	118	98	158	124	82	104	58	72	—	122	80	122
A/A_{boot}	14/10	11/9	10/9	—	8/7	12/8	12/8	10/8	14/11	8/6	6/5	8/6	—	6/5	4/4	5/5
H_e	70	54	61	—	65	63	60	66	76	63	63	62	—	59	61	65
F_{IS}	.04	.09	−.08	—	.03	−.10	.01	−.15	.03	.05	−.10	.02	—	−.10	−.03	−.10
45C (III:45c)	90	90	104	74	118	96	146	22	86	110	58	74	121	122	106	124
A/A_{boot}	9/8	12/9	14/10	7/6	9/8	10/8	12/9	13/10	12/10	10/8	8/7	6/5	7/6	9/7	7/6	8/7
H_e	79	82	82	72	74	76	82	83	84	76	80	65	65	83	69	60
F_{IS}	−.01	.13	.01	.02	−.08	−.05	−.05	.05	−.06	−.11	.00	−.08	.09	−.02	.10	−.19

Table 1. Continued

Locus	Seneg	Ghan	Niger	Came	Gabo	DRC	Western Kenya					Tanza	Malawi		Eastern Kenya	
	Barke	Ghan	Gwa	Simb	Beng	Kinsh	Asem	Kisia	Kimil	Awen	Zamb	Mbey	Mang	Nsanj	Jego	Malin
Mean 11 loci	94	88	104	72	118	94	157	123	77	113	58	72	130	121	104	123
A_{boot}	6.6	6.6	7.0	4.9	5.8	6.7	7.1	7.3	7.0	6.5	5.4	5.4	4.4	5.1	4.9	4.8
H_e	67	66	71	65	66	68	71	71	72	69	67	63	57	62	60	59
F_{IS}	.02	.02	.10	.07	.07	.13	.06	−.02	.05	.07	.04	.01	.06	−.03	.08	.04
Mean 7 loci	98	90	103	72	118	93	155	123	80	117	58	71	130	120	106	121
A_{boot}	5.9	6.0	6.1	5.2	5.3	5.9	6.2	6.2	5.8	5.7	4.5	4.3	4.4	4.6	4.2	4.3
H_e	64	66	68	67	66	68	69	68	68	67	64	56	57	57	53	51

[a] Abbreviated locus name followed by its cytological location following Zheng et al. (1996). This row denotes the number of chromosomes scored. Odd numbers resulted from scoring males in X-linked loci and from rare cases (<2%) where reliable scores were made only for one allele.

[b] The number of alleles per sample (A) and the bootstrapped mean number of alleles (A_{boot}), which refers to the mean number of alleles in a randomly selected (with replacement) sample of 20 individuals (40 chromosomes). One hundred pseudosamples were taken to calculated the mean.

[c] Unbiased expected heterozygosity (%), calculated following Nei (1987).

[d] The inbreeding coefficient, calculated according to Weir and Cockerham (1984). The significance level represents significance at the individual level. Bold values represent significant tests at the multitest level under the sequential Bonferroni procedure (see Materials and Methods) considering all tests.

*,**,*** = $P < .05$, $P < .01$, $P < .001$.

Figure 2. Schematic showing the position of microsatellite loci (Table 1) in relation to polymorphic inversions of *A. gambiae* s.s. Cytogenetic divisions bounding each chromosomal arm are shown. Note that there is no clear separation between the centromere and division 6 of the X chromosome, after Coluzzi et al. (1985).

Population Homogeneity and Independence of Loci

Significant departures from HW expectations within populations were found in 23 of 167 tests ($P < .001$, binomial test). As in previous microsatellite studies (e.g., Donnelly and Townson 2000; Kamau et al. 1998; Lehmann et al. 1998, 1999; Simard et al. 2000), most deviations (19) were associated with positive F_{IS} values, reflecting heterozygote deficits (Table 1). Locus-specific departures based on the sequential Bonferroni test were detected in locus *678* in the DRC and in locus *2A1* in Zambia and Jego. Heterozygote deficits in multiple loci indicate that samples may consist of several pooled subpopulations (Wahlund effect), the effects of inbreeding, or the presence of null alleles (Callen et al. 1993). Unlike the Wahlund effect and inbreeding, which affect the entire genome, observed deficits were clustered on locus *2A1* (seven populations). A few individuals repeatedly failed to produce a PCR product at *2A1*, suggesting that they were homozygotes for null alleles. Excluding locus *2A1* left 12 significantly positive F_{IS} values ($P > .12$, binomial test) and no population had a significant heterozygote deficit at the multitest level.

Assuming that the molecular forms represent two panmictic units, we expected that the deficit of heterozygotes observed in the total DRC population would disappear if tests were performed separately in each form. Contrary to our expectation, the deficit of heterozygotes across loci observed for the pooled DRC population did not disappear when M and S populations were analyzed separately. In the total DRC sample ($n = 49$), significant F_{IS} values were observed at three of the four X-linked loci (*2A1*, *99*, and *678*; Table 1). In the M subpopulation ($n = 32$), significantly positive F_{IS} values were still seen at these loci; positive values were also detected at loci *29C1* and *79*, on the third and the second chromosomes, respectively. Even in the smaller S subsample ($n = 15$), a significantly positive F_{IS} was observed at locus *678* and its value remained high (0.31). Locus *678* is mapped to division 6 of the X chromosome (Zheng et al. 1996), near the rDNA (Figure 2). The persistence of heterozygote deficits in the separate M and S subpopulations

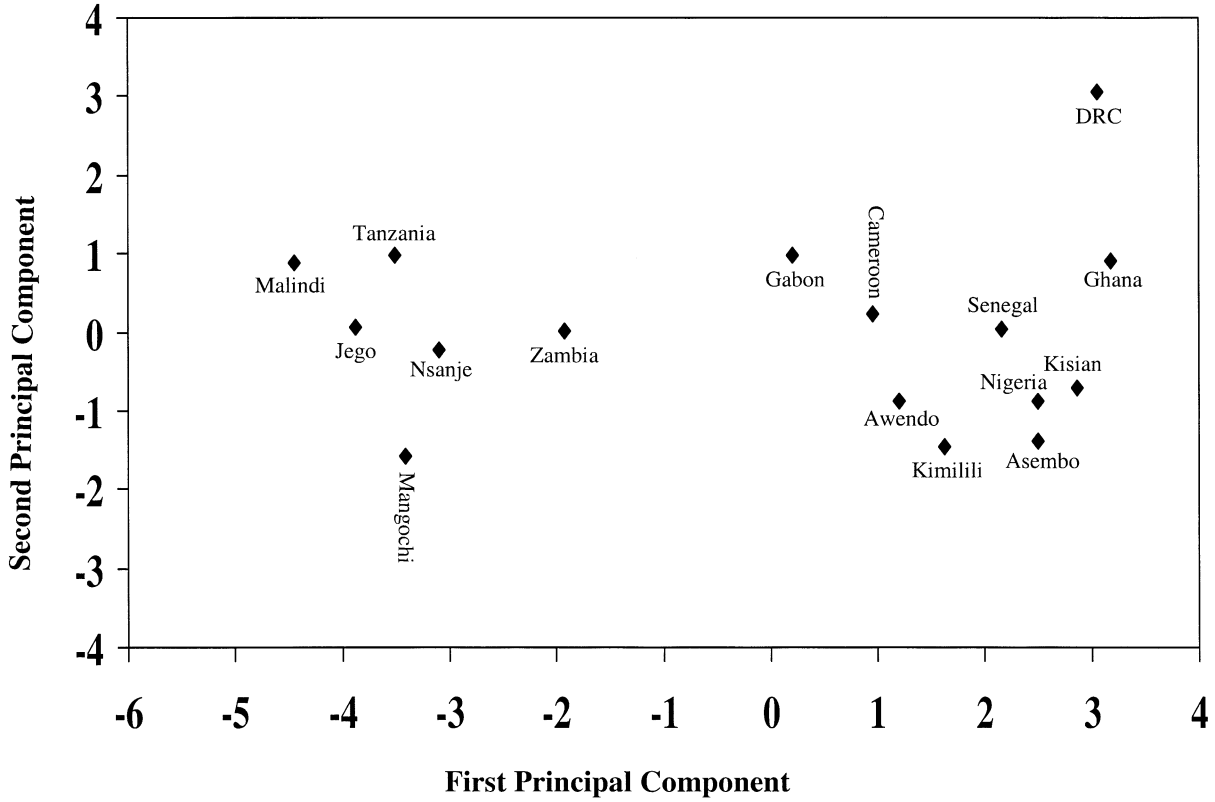

Figure 3. Clustering of populations using principal component analysis (PCA) based on their genetic diversity. Per locus estimates of the bootstrapped mean number of alleles and the mean expected heterozygosity across 10 loci were subject to this analysis. Locus *678* was not included because genotype scores were unavailable for three populations. Coordinates are the first and second principal components derived by PCA. Because they were genotyped at only seven loci, PCA scores for Cameroon and Mangochi (Malawi) (labeled vertically) were calculated by replacing the missing values for three loci with corresponding overall mean values. See the text for details.

is consistent with locus-specific selection rather than with two reproductively isolated populations.

Linkage disequilibrium (LD) tests were performed between all pairs of loci for all populations. The Wahlund effect or inbreeding should cause within-population LD because members of the different subpopulations will have different probabilities of carrying certain combinations of alleles. However, if null alleles cause the heterozygote deficits, LD is not expected, because all individuals are equally likely to carry a null allele, and the association between alleles from different loci is not disturbed. The fraction of significant ($P < .05$, single test) LD tests ranged from zero (Kimilili, Cameroon, and Senegal) to 11% (Malindi), but none were significant at the multitest level ($P > .06$, binomial tests), or by the sequential Bonferroni method. No evidence for LD was detected in the DRC (6 of 66 tests; $P = .11$, binomial test). These results suggest that loci are independent and that null alleles were the likely cause of the heterozygote deficits. Nevertheless, the insignificant LD tests may reflect the limited power of this test.

Within-Population Diversity

Genetic diversity within populations was moderate to high for all loci (Table 1). Variation in sample size between populations (58–164 chromosomes per locus per population) can have a substantial effect on the observed number of alleles (A). To compare populations in an unbiased way we randomly sampled 20 individuals (40 chromosomes) from each population, with replacement (100 times), and calculated the number of alleles and unbiased expected heterozygosity (H_e; Nei 1987) for each locus. The bootstrapped H_e mean remained within 2% of the original value in all cases and is thus not shown in Table 1. To summarize between-population differences in genetic diversity, we applied principal component analysis (PCA) based on the correlation matrix of the bootstrapped mean A and mean H_e of each locus. This analysis was based on 10 loci (excluding Mangochi and Cameroon, and locus *678*, which was not genotyped in Senegal) and captured 65% of the total variation (across 20 original variables) in two principal components (Figure 3). The first principal component ac-

counted for 53% of the total variation and separated two groups of populations. One group included Malindi (Kenya), Jego (Kenya), Nsanje (Malawi), Tanzania, and Zambia, while the other group included the remaining nine populations. The first group had lower scores, indicating lower diversity. Geographically the groups corresponded to the southeastern (SE) region and northwestern (NW) region of sub-Saharan Africa (Figure 1). The second principal component accounted for 12% of the variation and appeared to separate the DRC from all the rest. Analysis of all 16 populations based on seven loci yielded the same groups based on the first principal component (accounting for 58% of the variation), and placed Mangochi (Malawi) in the SE group and Cameroon in the NW group, as expected. The second principal component (accounting for 10% of the variation) did not separate any population (not shown). To check whether the explanation that genetic diversity is lower in the SE region is consistent with the original variables, the differences between regions of bootstrapped A and H_e means were tested in a multivariate analysis of variance and found to be highly significant (A, 6.6 versus 5.1; H_e, 69 versus 62; $P < .001$). Clearly this is not a test of the difference between regions, as populations were grouped into regions based on the PCA, but the results of the test are consistent with our interpretation of the PCA.

Differentiation Between Populations: Heterogeneity Among Loci

If populations represent distinct demes, significant differentiation is expected across the genome. Heterogeneity in allele frequency between pairwise population comparisons was found in 117 of 120 pairs using Fisher multisample tests. Insignificant heterogeneity was found in three comparisons between western Kenyan populations (Asembo versus Kisian, Kimilili versus Kisian, and Kimilili versus Awendo) located within 200 km of each other.

F_{ST} values among loci varied widely (range −0.02 to 0.40). Tables of F_{ST} values between all population pairs for every locus are available from T.L. upon request. To assess whether this variation reflected locus heterogeneity, we calculated jackknifed means for each locus for every population pair (excluding one locus at a time and calculating 11 means, each without a single locus) and calculated an influence index: $(F_{original} - F_{jackknifed})/F_{original}$. This index estimates the relative change in F_{ST} after excluding a given locus. Figure 4A plots the F_{ST} value for all population pairs more than 200 km apart and genotyped at all 11 loci against the influence index for each locus. This index is not intended as a statistical test, as we do not know the limits of the distribution of the influence index under locus homogeneity. Rather it was used as an exploratory tool to identify putatively influential loci, defined as those whose removal of their single F_{ST} value (of 11 values) changed the mean F_{ST} by 60–95% (upper right corner of Figure 4, top panel). The same values, all with $F_{ST} > 0.2$, are located 5 to 15 interquartile range units (IQR) away from their corresponding

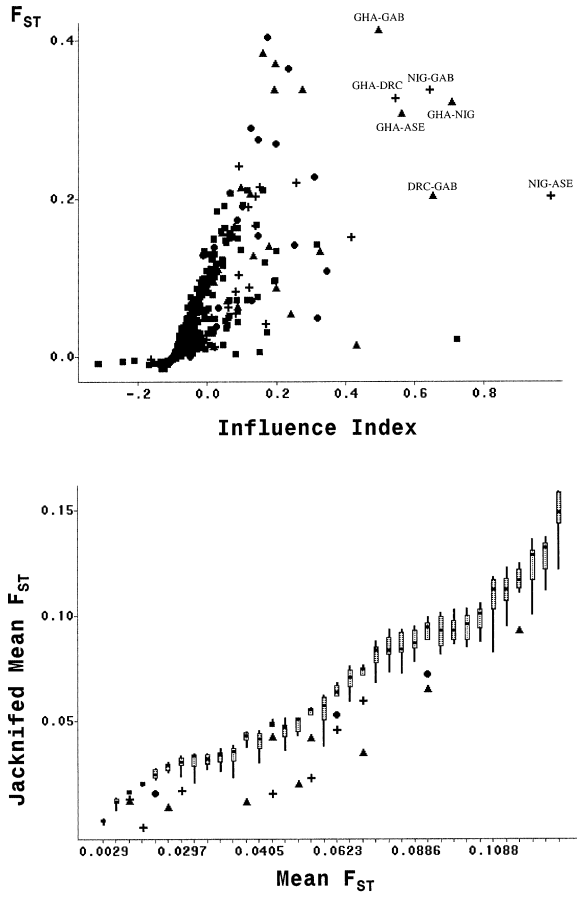

Figure 4. Detecting loci associated with **(top)** influential and **(bottom)** outlier F_{ST} values for populations more than 200 km apart, which were genotyped in all 11 loci. Loci *678, 79*, and *29* are labeled with ▲, +, and ●, respectively. Population pairs associated with outlier and influential values (defined as described in the text) are shown based on the first three letters of their name (e.g., GHA denotes Ghana). Note: the point at the lower right caused a trivial change in the mean differentiation from 0.0008 to 0.003. In the box whisker plot, the box extends between the 25th and the 75th percentile, that is, across one IQR, and the whiskers extend up to the most extreme value, but not beyond 1.5 times the IQR. Values located more than two IQR from the median are shown.

median (Figure 4, bottom panel). Values located more than three IQR from the median are typically considered as extreme outliers (SAS 1999). Notably, all 13 points (out of 396) which fell more than three IQR from their median (all values with influence larger than 35%) belonged to locus *678* or locus *79*, which are mapped near the rDNA and inversion 2Rb, respectively (Figure 2; see below).

The extreme outlier values of locus *678* were found in comparisons between the M and S populations. Allele distributions at locus *678* (Figures 5 and 6) show the basis for this pattern. In Ghana (M genotype), the allele size range was 149–163, with a 153 bp allele predominating (77%). In

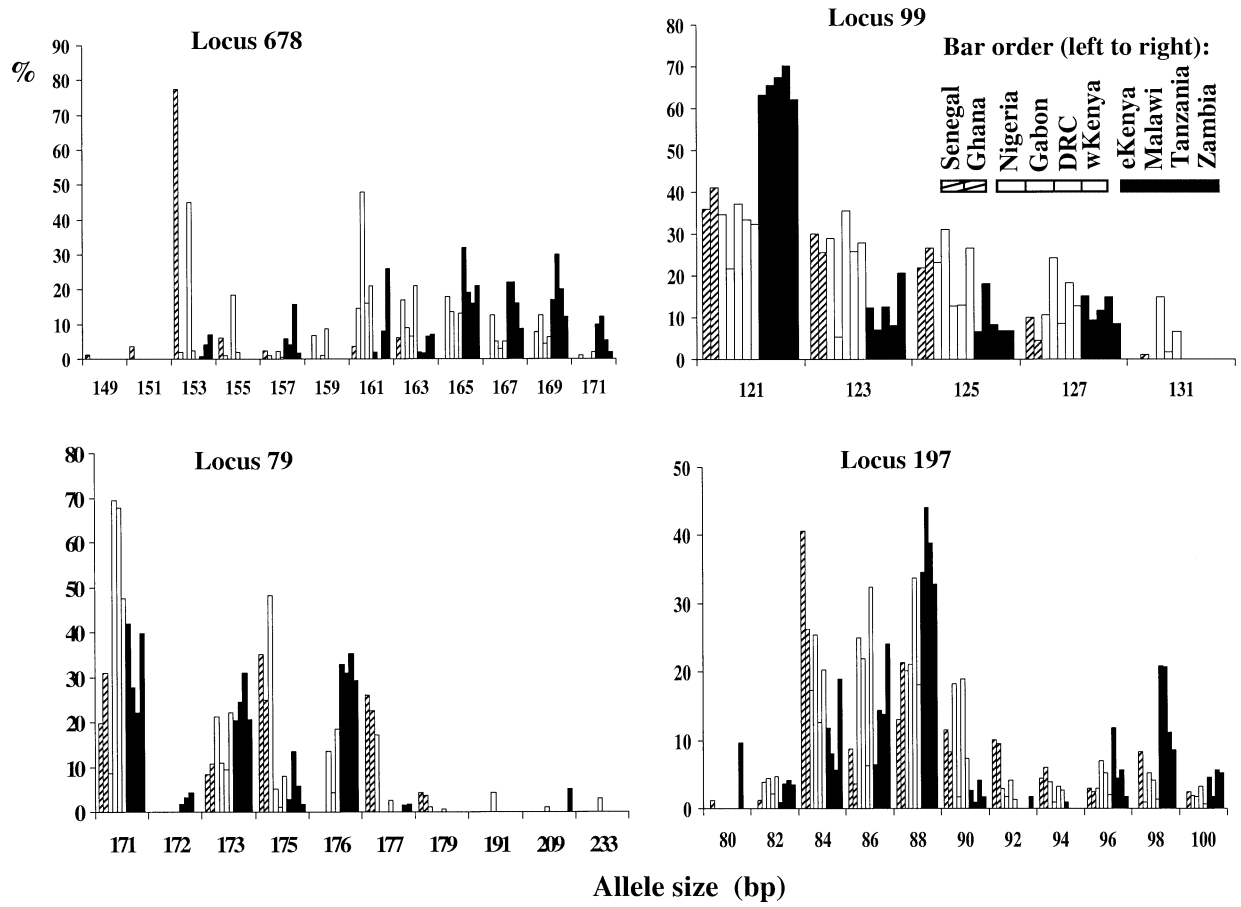

Figure 5. Allele distributions across selected populations for loci *678* and *79* and the loci nearest to each of them (10 cM apart) to each of them. Bar order denotes different populations as shown in the legend. Rare alleles (<5%) are not depicted unless they signify important differences.

neighboring Nigeria, as for all S populations, the allele size range was 153–211, and a 165 bp allele predominated (18%; Figure 5). Allele 153 was rare or absent in all S-genotype populations. In the DRC, the overall frequency of allele 153 was 45%, but was 65% in its M subpopulation and only 4% in its S subpopulation (Figure 6). There were 11 heterozygotes for this allele, of which 10 belonged to the M form. The distinct allelic profile at locus *678* of each molecular form resulted in extremely high F_{ST} values (0.2–0.4) and in the departure from HW expectations reported above.

The other extreme outlier values were associated with locus *79*, in comparisons between Nigeria and three other populations (Gabon, DRC, and Asembo). Populations from Ghana, Senegal, and Nigeria had moderate to high frequencies of alleles 175 and 177 (25%–48% and 17%–26%, respectively), which were low elsewhere (<15% and <4%, respectively). These populations also lacked allele 176, which was at least moderately frequent elsewhere. Notably, allele 176 does not fit a series predicted by a 2 bp repeat motif and a constant flanking sequence (Figure 5). Alleles 175, 176, and 177 were absent from the M population in the DRC, although allele 176 was found in the S population there (13%). Unlike locus *678*, locus *79* did not separate the M and S genotypes. Its location (subdivision 11A) near inversion 2Rb (Figure 2), which is under selection (Coluzzi et al. 1985; Toure et al. 1994, 1998), suggests an inversion-specific effect in accordance with Lanzaro et al. (1998).

High F_{ST} values were also observed for locus *29*, but influence indices were less than 40% (Figure 4, top panel) and their location within the whiskers of the box plots (Figure 4, bottom panel) indicate no exceptional deviations. With only two alleles segregating, F_{ST} in locus *29* attains high values when one allele dominates, that is, is closer to fixation (Wright 1978). Besides the influential/outlier values of loci *678* and *79*, the distribution of the jackknifed F_{ST} means of the remaining nine loci appears homogeneous, as only 3 of 326 values are located beyond the "whiskers" (Figure 4, bottom panel). These results strongly caution that variation between certain populations at loci *678* and *79* has been shaped not only by drift, migration, and mutation, but also by selection operating presumably on a linked locus. Because the homogeneity of loci was probably violated by including a locus with $F_{ST} = 0.31$ with 10 loci whose mean is 0.02 (range -0.006 to 0.075), as exemplified by locus *678*

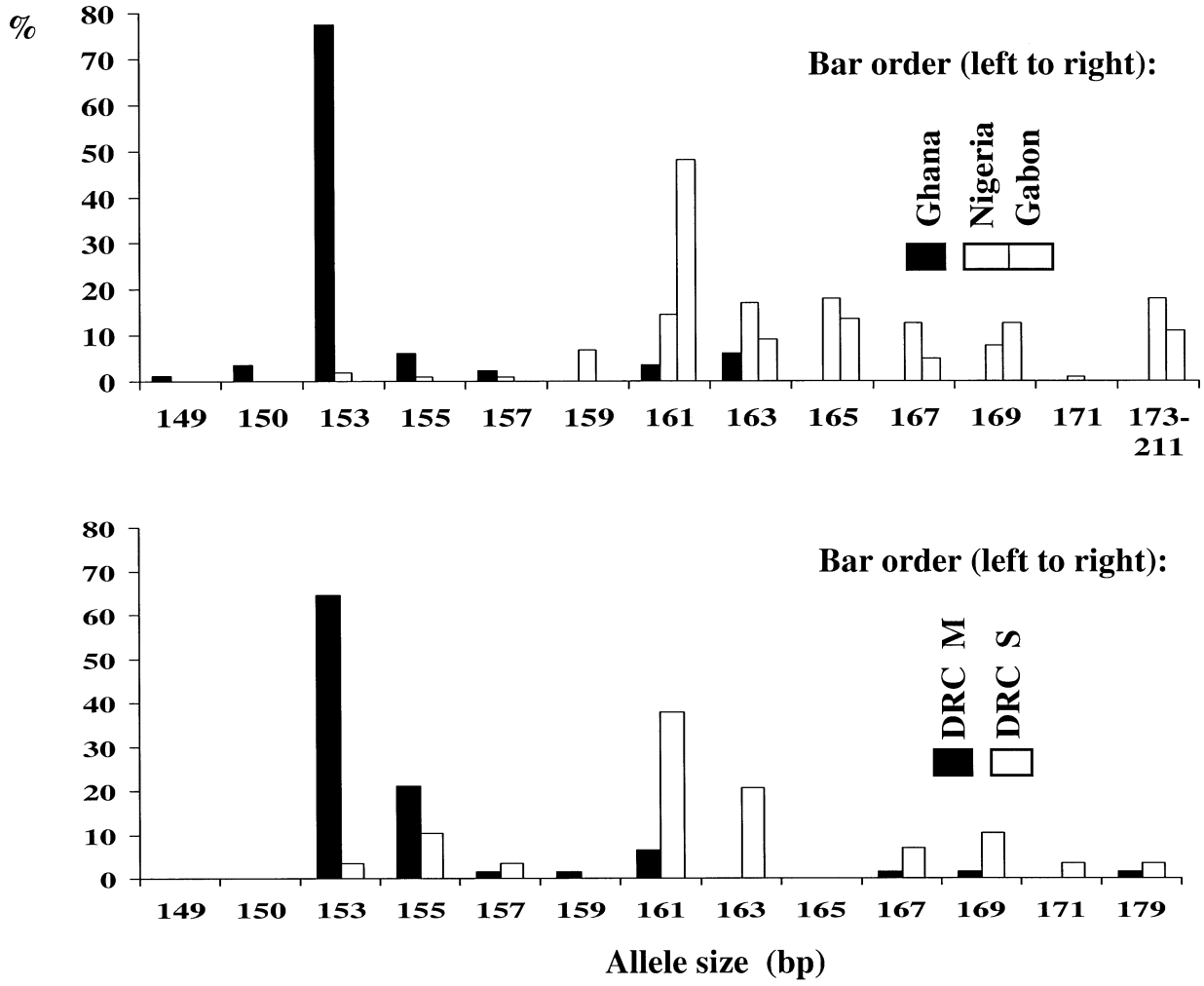

Figure 6. Allele distributions for locus *678* of sympatric **(bottom)** and neighboring allopatric **(top)** M and S populations. Bar order denotes different populations as shown in legend.

between Asembo and Ghana (influence index 56%), we removed loci *678* and *79* from analyses aimed at capturing genome-wide patterns.

Differentiation Between Populations: Phylogeography

An unrooted neighbor-joining population tree was constructed based on a matrix of the mean F_{ST} across nine loci between all pairs of populations (Figure 7). Large F_{ST} distances with high bootstrap values (BV $>$ 97) distinguished two groups of populations. One group included populations from eastern Kenya, Malawi, Tanzania, and possibly Zambia. The other group included all other populations. These groups correspond to those discerned by PCA (NW and SE; Figure 3) with the exception of Zambia, which was separated by equal F_{ST} distances from each division (branch length 0.03, BV $>$ 95; Figure 7).

Within each division, populations separated by shorter geographic distances appear to be more related genetically, but this pattern was broken by the cluster of M populations (Figure 7). Of importance is that the M subpopulation from the DRC was clustered with the other more northerly M populations, a long genetic distance from its sympatric S subpopulation. As locus *678* was excluded, this cluster represents a genome-wide pattern consistent with a common descent for all M populations. Despite the clade's low bootstrap support (BV = 51), the assembly of all M populations regardless of geographic distance is unlikely to be a result of noise. Of interest is that the grouping of Senegal and Ghana apart from all the other populations (BV = 87; Figure 7) suggests a split between populations inhabiting the savanna and the forest.

Substructure was greater in the geographically smaller SE subdivision, in which three nested clusters were discerned (Figure 7). These included two populations from eastern Kenya nested in a cluster containing Nsanje, which were nested in a cluster containing Mangochi. All populations in this cluster were nested within Tanzania.

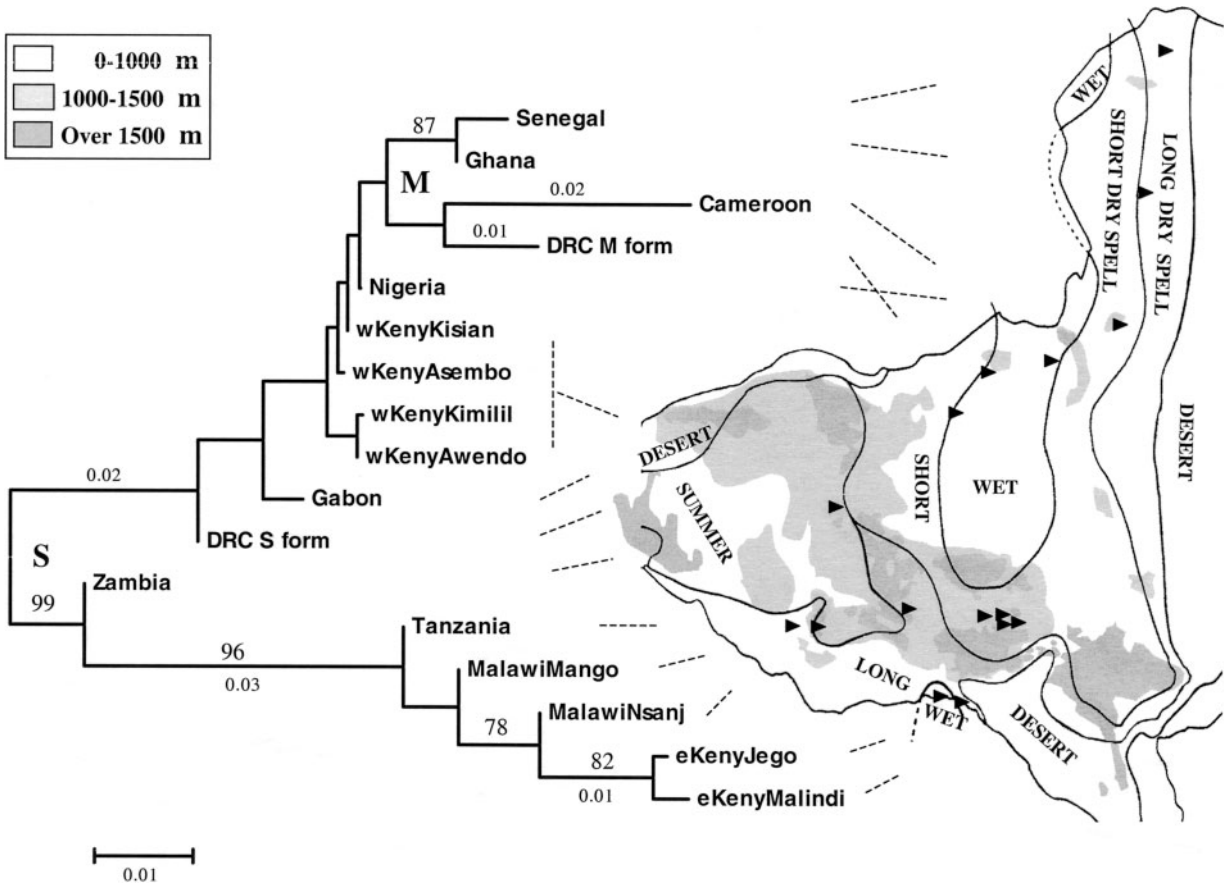

Figure 7. Unrooted neighbor-joining population tree based on mean F_{ST} across nine loci (excluding loci *678* and *79*) superimposed on a schematic map showing collection sites (solid triangles), altitude, and climatic zones. M and S populations are denoted at the bases of the clades. Dotted lines point to collection sites. Populations from Cameroon and Mangochi (Malawi) were genotyped in only seven loci and corresponding mean F_{ST}s were used at loci *197* and *577*. Fractions denote branch length (over 0.01) and integers denote biologically significant bootstrap support values. See the text for details. Map is based on various sources including Goudie (1996) and Kingdon (1989).

The phylogeographic analysis supports three clear patterns: (1) a continental split between the NW and SE divisions, (2) a shallower, secondary split between the M and the S populations within the NW division, and (3) a series of nested population clusters in the SE division. Differentiation between the M and S populations was lower than that between several allopatric populations of the same molecular form, and the distinction between Forest (Cameroon, Gabon, DRC) and Savanna (all others) chromosomal forms added little to the classification of populations.

Differentiation Between Populations: Isolation by Distance

The distribution of *A. gambiae* across its range in sub-Saharan Africa is more or less continuous (Coetzee et al. 2000). Large areas inhospitable to this species that can isolate populations, such as the Rift Valley complex (Kamau et al. 1999; Lehmann et al. 1999, 2000) are few and migration around them may counteract their effect. The isolation by distance model (Wright 1951) is therefore a plausible one for *A. gambiae*. It assumes that migration is equally likely in all directions and that distance is the main determinant of differentiation.

The positive slope of the regression of $F_{ST}/(1 - F_{ST})$ on the logarithm of the distance (Rousset 1997), based on the mean F_{ST} (excluding loci *678* and *79*), is consistent with the isolation by distance model (slope = 0.013, $P < .001$; Table 2). To determine whether this pattern also occurs within divisions and subdivisions, we repeated this analysis within and between these units (Table 2). Positive slopes ($P < .05$) were found in all analyses except between molecular forms in the NW division, within the M form, and when both forms were included in the NW division (Table 2). Rousset (1997) pointed out that the slope depends on the range of the logarithm of distance values. Large differences between the estimates of the slopes (Table 2) suggest heterogeneity of the slopes, but we must accommodate the disparity in distance ranges. Even after truncating the distance range of the S population in the NW division to match that of the SE

Table 2. The intercept, slope, and significance of isolation by distance models in which $F_{ST}/(1 - F_{ST})$ was regressed on the logarithm of distance* (km)

Division	Molecular form	*N*	Range (km)	Intercept	Slope	P^a	(r^2)
All	S & M	136	1–6,600	−.0393	0.0127	**.001**	(.11)
All within division[b]	S & M	70	1–5,730	0.0034	0.0021	**.030**	(.09)
All between division[c]	S & M	66	650–6,600	−.0551	0.0193	**.004**	(.11)
SE	S only	15	190–2,180	−.0842	0.0151	**.004**	(.36)
NW (all)	S & M	55	1–5,730	0.0068	0.0015	.089	(.08)
NW	S only	21	30–3,060	−.0049	0.0018	**.006**	(.25)
NW	M only	6	1,000–4,170	−.0535	0.0104	.39	(.11)
NW between forms[d]	S vs. M	28	1–5,730	0.0344	.0014	—	(.17)

* The shortest overland distance was used (not across ocean or lakes over 10 km width). Comparison between sympatric molecular forms in the DRC (distance = 0) was not done. Distance between populations across the eastern Rift Valley complex or the Adamawa Mountains (Cameroon) was estimated as the shortest distance around them.

[a] Significance determined using a Mantel test. Bold values represent significant test after applying the sequential Bonferroni test.

[b] Excluding comparisons of populations between subdivisions.

[c] Including only comparisons of populations between divisions.

[d] Including only comparisons of populations between molecular forms (in the NW division).

division (NW, 30–2230 km; SE, 190–2180 km), the slope for the S populations in the NW division (0.0024, 95% CI 0.001–0.003) was significantly smaller than that in the SE division (0.0151, 95% CI 0.005–0.025).

Discussion

Distinguishing locus-specific from genome-wide patterns is a central prerequisite for the description of *A. gambiae* population structure because of the effects of selection and inversions (della Torre et al. 2001; Gentile et al. 2001; Lanzaro et al. 1998; Mukabayire et al. 2001; Taylor et al. 2001). Locus homogeneity with respect to differentiation between populations was probably violated unless loci *678* and *79* were removed from the analyses of genome-wide patterns. The deepest split in the gene pool was measured between the SE and NW divisions (Figure 7). Other clusters of populations, such as the molecular forms, were separated by smaller genetic distances. Understanding the processes that have shaped this structure starts with the question of what might constitute the barriers among these units.

Between the NW and SE divisions there are a near-continuous series of highlands, deserts, and lakes along the Rift Valley system that are inhospitable for *A. gambiae* (Figure 7). In addition, an inhospitable dry savanna extends from northern Kenya between the narrow coastal plain and the highlands and spreads into central Tanzania (not depicted in Figure 7). This series of ecogeographic features probably constitutes the northern barrier between divisions, as was shown previously (Lehmann et al. 1999, 2000). No such geographic features are found between the DRC and Zambia or between Nsanje (Malawi) and Zambia (Figure 7). Alternatively, a recent bottleneck in the SE region explains both the high differentiation between divisions and the lower genetic diversity within the SE division. In the absence of geographical barriers, distance alone determines differentiation. For instance, the isolation by distance model (SE, S form; Table 2) is consistent with the differentiation between Zambia and both the DRC (S) and Nsanje (1340 km apart; expected $F_{ST} = 0.024$; observed mean $F_{ST} = 0.017$ and 0.031, respectively). Consistent with a northern barrier, however, distance fails to predict the differentiation between eastern and western Kenya (expected = 0.014, observed = 0.11; see Lehmann et al. 1999). These results suggest that migration between divisions occurs mostly in the south and are consistent with a bottleneck that shaped this structure.

The Role of Colonization in Establishing the SE Subdivision

The nested clusters of populations in the SE division may reflect a series of founder events associated with a population spreading from the northwest and progressing eastward and northward along the coast. Under a mutation-drift equilibrium with similar effective population size, each population of a given population pair has the same likelihood of possessing private alleles. Colonization events, particularly if associated with founder effects, will result in loss of private alleles in colony populations. Accordingly, most alleles of the colony population are expected to exist in the source population, but many alleles of the source population are absent in the colony due to the founder effect. To test for the existence of such a pattern in the SE division, we calculated an index of inclusiveness of alleles and used it as an ad hoc test statistic. The inclusiveness index (I) was calculated as the proportion of alleles in the colony (c) population that are shared with a putative source (s) population ($Fc[s]$) times the difference between this and the "reverse" fraction: $Ic[s] = Fc[s] \times (Fc[s] - Fs[c])$. Unlike differentiation indices, inclusiveness is typically asymmetric ($Ic[s] \neq Is[c]$) and is asymptotically bounded between 1 and −1. Large positive values require that $Fc[s]$ is large and that $Fs[c]$ is considerably smaller, indicating that the data are consistent with the hypothesis that c and s represent a colony and a source, respectively. Near-zero values indicate that the two populations possess nearly the same alleles or very different alleles,

Table 3. Inclusiveness index (*Ic*[*s*]) and significance tests to assess different colonization hypotheses (see text for details)

Hypothesis	Colony (*c*) and source (*s*) populations[a]	Individual pair *Ic*[*s*] (mean)	Random 95% CI[b]	*P*
Sequential colonization	Md/Jo[Mw]	0.05/0.02 (0.03)	−0.16–0.22	ns
	Md/Jo/Mw[Tz]	0.09/0.05/0.06 (0.07)	−0.13–0.19	ns
	Md/Jo/Mw/Tz[Zb]	0.19/0.16/0.15/0.09 (0.15)	−0.11–0.16	ns
	Overall mean	(0.095)	−0.08–0.11	ns
One source followed by drift	Md/Jo/Mw/Tz/Zb[WK]	0.37/0.30/0.30/0.22/0.06 (0.26)	−0.10–0.14	$<10^{-3}$
	Md/Jo/Mw/Tz/Zb[DRCs]	0.18/0.16/0.13/0.05/0.02 (0.11)	−0.09–0.16	ns
	Md/Jo/Mw/Tz/Zb[DRCm]	0.19/0.16/0.16/0.12/0.08 (0.14)	−0.10–0.15	ns

[a] Md, Jo, Mw, Tz, and Zb denote Malindi, Jego, Nsanje (Malawi), Tanzania, and Zambia, respectively, WK denotes Asembo (western Kenya), which was arbitrarily selected to represent this region (see text). Mangochi (Malawi) and Cameroon were excluded as they were genotyped at only seven loci.

[b] 95% CI was adjusted to the number of tests performed by dividing α (.05) by 5.

but in either case the data do not conform to the "source-colony" hypothesis. Small negative values indicate that the relationships may be reversed (i.e., the putative colony is more likely to be the source population). To avoid bias due to differences in sample size, we used the same sample size (per locus) for each pair of populations by sorting specimens in order of collection and excluding those in excess. A priori hypotheses were tested by bootstrapping the matrix of inclusiveness values between all pairs of populations, after excluding the values under the hypothesis to determine whether they were higher than 95% of the bootstrapped values.

Our first hypothesis is based on the structure of the population tree (Figure 7): a series of colonization events took place starting with migrants from Zambia into Tanzania; migrants from Tanzania into Malawi; and last, migrants from Malawi into eastern Kenya. An alternative hypothesis is that all SE populations were colonized from one origin in a rather short time (not sequentially). Since western Kenya and the DRC are closest to the SE division, they were considered the most plausible source populations. In these tests, alleles across all 11 loci were used (a total of 219 alleles), the population from Asembo was arbitrarily selected to represent western Kenya (Kisian, Kimilili, and Awendo were excluded, although if their values are averaged the same patterns were obtained), and the S and M populations of the DRC were considered separately (see below).

The low inclusiveness values do not support the "sequential colonization" hypothesis, but high and significant values support the "one source" hypothesis (Table 3). These high values highlight the lack of private alleles in the SE populations and that most of the SE alleles are found in western Kenya (Table 3). To clarify the sources of the SE division, we tested Senegal, Ghana, Cameroon, Gabon, and Nigeria as possible source populations. They were all rejected as source populations ($0.02 \leq I \leq 0.08$, $P > .1$), except Nigeria ($I = 0.14$, $P < .01$), which has an allele profile very similar to western Kenya (Figure 7; mean $F_{ST} = 0.0003$). While the true origin cannot be determined, a colonization event from a population(s) most similar to western Kenya is supported by these results.

The Extent of Isolation Between the M and S Populations

The complex status of the molecular forms is highlighted by our results, some of which support recent reproductive isolation, but others support selection (Table 4). If reproductive isolation alone has generated the high differ-

Table 4. Summary of conflicting results that defy a simple explanation for the status of the molecular forms

Recent reproductive isolation[a]	Locus-specific selection[b]
No M/S heterozygotes were observed in DRC	Deviations from HW equilibrium persisted within each molecular form in the DRC
$F_{ST} > 0.2$ at locus *678* (and rDNA)	Mean F_{ST}[c] was 0.018 (0.024 in the DRC)
Statistically significant differentiation was found genome-wide[c] between sympatric M and S populations (DRC)	Loci of high differentiation between forms (rNDA and *678*) are found only in one chromosomal region (division 6)
Monophyletic origin for the M form is suggested based on genome-wide differentiation	The mean differentiation between molecular forms is lower than that between allopatric populations within each form

[a] Important predictions: observing no hybrids, agreement with HW expectations within form, and agreement among multiple independent loci showing high differentiation even between sympatric populations.

[b] Important predictions: loci of high differentiation are exceptional and that they are not independent. Selection is compatible with departures from HW expectations. Selection effect on a particular locus can spread to linked loci.

[c] Mean F_{ST} refers to the mean over nine loci, excluding loci *678* and *79*. The same nine loci are referred to when reference is made to genome-wide differentiation.

entiation at the rDNA and locus *678* ($F_{ST} > 0.2$), why is the genome-wide differentiation so much lower ($F_{ST} = 0.018$)? Heterogeneity among loci is a strong signature of selection. Further, the colocalization of the two high-differentiation loci in division 6 contrasts with other X-linked loci (*1D1, 2A1*, and *99*) that show no differentiation between M and S forms in the DRC ($-0.01 < F_{ST} < 0.018$, $P > .2$, exact tests) and indicate that loci *678* and the rDNA are not independent. Notably the genome-wide differentiation between the molecular forms is lower than that between some allopatric populations within forms, albeit the results above suggest that the differentiation between divisions is the result of a founder effect (see more below). Finally, contrary to expectations based on two panmictic units in the DRC, departures from HW expectations persisted on locus *678* in both forms, and additional departures, not seen in the pooled population, were detected in each form despite the lower sample size. Nevertheless, at least partial reproductive isolation between sympatric M and S populations was evident even from several autosomal loci associated with significant differentiation in the DRC (locus *46*, $F_{ST} = 0.043$; locus *197*, $F_{ST} = 0.038$; locus *577*, $F_{ST} = 0.083$; $P < .005$, exact tests) and in other localities (della Torre et al. 2001; Wondji et al. 2001). Further, a low frequency (<2%) of M/S matings was measured in Mali (Tripet et al. 2001) and a large difference in the frequency of the *kdr* mutation between sympatric M and S populations was reported (Weill et al. 2000). Neither selection alone nor recent isolation alone can explain these findings.

An explanation that accounts for all these findings involves fluctuating gene flow and selection. According to this hypothesis, the combined effects of assortative mating and selection against M/S heterozygotes severely limits gene flow between them under typical conditions. However, if selection is relaxed, in as yet unobserved conditions, then gene flow will reduce the differentiation genome-wide. The traces of gene flow will be quickly obscured in chromosomal regions affected by selection and hitchhiking. A different explanation is that heterozygote females (for an X-linked locus) may be at a disadvantage under natural conditions, but gene flow could continue via males; a situation which cannot be resolved with available samples, as they consist almost exclusively of females. Fluctuating selection and gene flow is a hypothesis that postulates unobserved events, but some support for it is found in the allele distributions at locus *678*. Unlike the population in Ghana that is very distinct from its neighboring S population (Nigeria), the M population in the DRC shares alleles 167, 169, and 179 (all absent in Ghana) with the sympatric S population (Figure 6). Separated by a large number of repeats (three to nine) from the nearest M allele (161 bp), they probably originated by gene flow from the sympatric S population, although a common ancestral origin cannot be ruled out. Nucleotide data are needed to determine whether the M form is indeed monophyletic, as suggested by Figure 7.

Recent studies have suggested that differentiation between species can be high in several genes despite persistent gene flow and lower differentiation in other genes (e.g., Machado et al. 2002) and that divergence leading to incipient speciation can occur between populations that exchange genes at high rates (e.g., Beheregaray and Sunnucks 2001; Michalak et al. 2001). These conclusions are backed by male sterility (e.g., Machado et al. 2002) or by extensive phenotypic differences (e.g., Beheregaray and Sunnucks 2001; Michalak et al. 2001), which are not paralleled in the present knowledge of *A. gambiae* molecular forms. However, these findings strengthen the view that the molecular forms represent incipient species (della Torre et al. 2002 and references therein). Additional studies exploring the phenotypic differences between the forms are clearly needed. Regardless of the taxonomic debate, the molecular forms may represent distinct epidemiological entities, so treating them separately is operationally justified, at least until differences in their vectoral capacity and susceptibility to means of control are evaluated.

Gene Flow and Other Processes Shaping the Population Structure

Consistent with the isolation by distance model, differentiation increased with geographic distance, but the effect of distance was rather small. For example, the expected F_{ST} across 6,000 km using the steepest slope (of the SE division) is only 0.045. The large differences between slopes suggest that factors other than distance are involved in shaping differentiation. An accurate estimate of gene flow is beyond our grasp because of departure from the mutation-drift-migration equilibrium (Donnelly et al. 2001). Isolation by distance and the inferred colonization process further complicate the interpretation of gene flow.

The colonization of the SE division probably does not represent the original introduction of *A. gambiae* into this area, but a colonization following a large-scale bottleneck. Traces of a historical bottleneck were detected in the SE division (Jego), but not in the NW division (Asembo) based on the imbalance index (Donnelly et al. 2001). What could cause such a large-scale bottleneck? Severe droughts occur repeatedly in Africa, with the most profound effects in long dry spell regions, where the dry season normally lasts 6 months or longer (Goudie 1996). This climate prevails in large parts of the SE region and in the Sahel (Figure 7). The impact of severe drought is exemplified in the drought of 1770–1830, which totally dried up Lake Rukwa (Tanzania) and Lake Chiuta (Malawi), reduced the levels of Lake Malawi by more than 120 m, and brought famine that led to human migration (Nicholson 1998a,b; Owen et al. 1990). Such events probably reduced both the number of available breeding sites and hosts. The areas least affected by droughts are the equatorial forests and the belt of wet savanna (short dry spell) surrounding it (Goudie 1996). It is plausible therefore that the ultimate source populations of *A. gambiae* reside in the forest and the wet savannas, whereas populations residing in the dry savanna and the Sahel are doomed to experience extinction/bottlenecks every few hundred years. This process might help maintain the genetic homogeneity of *A. gambiae* across its range and reduce the

"effective range" (loosely defined as the range of permanent populations) of the species to nearly half of its presently observed range. This effective range may be even smaller, because droughts "shrink" the forest and the ring of wet savanna around it. Why do populations in the Sahel (e.g., Barkedji, Senegal) show no sign of a bottleneck compared with those in the SE division (e.g., Jego)? The Sahel is a narrow belt, less than 500 km from the wet savanna, a relatively short distance that allows rapid colonization and replenishment of genetic diversity. In addition, the Sahel is populated by the relatively drought-tolerant M molecular form. The SE region, on the other hand, is partly separated by the Rift Valley and its associated highlands and by a longer distance, and so requires a longer period to replenish its genetic diversity. If periodic droughts shaped the deepest division of the structure of *A. gambiae* populations, then other drought-sensitive species in Africa with moderate mobility should present similar patterns. A remarkably similar pattern was observed in *Anopheles funestus* (Yan G, personal communication) and in *Drosophila teissieri*, showing reduced genetic diversity in three Tanzanian populations (SE) compared with those from Gabon and Ivory Coast (NW) (Cobb et al. 2000). Additional studies on relevant species will permit a more rigorous test of the effects of severe drought on *A. gambiae* in the context of regional phylogeography.

References

Beheregaray LB and Sunnucks P, 2001. Fine-scale genetic structure, estuarine colonization and incipient speciation in the marine silverside fish *Odontesthes argentinensis*. Mol Ecol 10:2849–2866.

Besansky NJ, Lehmann T, Fahey GT, Fontenille D, Braack LE, Hawley WA, and Collins FH, 1997. Patterns of mitochondrial variation within and between African malaria vectors, *Anopheles gambiae* and *An. arabiensis*, suggest extensive gene flow. Genetics 147:1817–1828.

Black WC and Lanzaro GC, 2001. Distribution of genetic variation among chromosomal forms of *Anopheles gambiae* s.s: introgressive hybridization, adaptive inversions, or recent reproductive isolation? Insect Mol Biol 10:3–7.

Bryan JH, Di Deco MA, and Petrarca V, 1982. Inversion polymorphism and incipient speciation in *Anopheles gambiae* s.s in the Gambia, West Africa. Genetica 59:167–176.

Callen DF, Thompson AD, Shen Y, Phillips HA, Richards RI, Mulley JC, and Sutherland GR, 1993. Incidence and origin of "null" alleles in the (AC)n microsatellite markers. Am J Hum Genet 52:922–927.

Cobb M, Huet M, Lachaise D, and Veuille M, 2000. Fragmented forests, evolving flies: molecular variation in African populations of *Drosophila teissieri*. Mol Ecol 9:1591–1597.

Coetzee M, Craig M, and le Sueur D, 2000. Distribution of African malaria mosquitoes belonging to the *Anopheles gambiae* complex. Parasitol Today 16:74–77.

Coluzzi M, Petrarca V, and Di Deco MA, 1985. Chromosomal inversion intergradation and incipient speciation in *Anopheles gambiae*. Boll Zool 52: 45–63.

Coluzzi M, Sabatini A, Petrarca V, and Di Deco MA, 1979. Chromosomal differentiation and adaptation to human environments in the *Anopheles gambiae* complex. Trans R Soc Trop Med Hyg 73:483–497.

della Torre A, Costantini C, Besansky NJ, Cacccone A, Petrarca V, Powell JR, and Coluzzi M, 2002. Speciation within *Anopheles gambiae*—the glass is half full. Science 298:115–117.

della Torre A, Fanello C, Akogbeto M, Dossou-yovo J, Favia G, Petrarca V, and Coluzzi M, 2001. Molecular evidence of incipient speciation within *Anopheles gambiae* s.s. in West Africa. Insect Mol Biol 10:9–18.

Donnelly MJ, Licht MC, and Lehmann T, 2001. Evidence for recent population expansion in the evolutionary history of the malaria vectors *Anopheles arabiensis* and *Anopheles gambiae*. Mol Biol Evol 18:1353–1364.

Donnelly MJ and Townson H, 2000. Evidence for extensive genetic differentiation among populations of the malaria vector *Anopheles arabiensis* in eastern Africa. Insect Mol Biol 9:357–367.

Favia G, della Torre A, Bagayoko M, Lanfrancotti A, Sagnon N, Toure YT, and Coluzzi M, 1997. Molecular identification of sympatric chromosomal forms of *Anopheles gambiae* and further evidence of their reproductive isolation. Insect Mol Biol 6:377–383.

Felsenstein J, 1989. PHYLIP—phylogeny inference package (version 3.2). Cladistics 5:164–166.

Gentile G, Slotman M, Ketmaier V, Powell JR, and Caccone A, 2001. Attempts to molecularly distinguish cryptic taxa in *Anopheles gambiae* s.s. Insect Mol Biol 10:25–32.

Goudie AS, 1996. Climate: past and present In: The physical geography of Africa (Adams WM, Goudie AS, and Orme AR, eds). Oxford: Oxford University Press; 35–59.

Hillis DM and Bull JJ, 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenic analysis. Syst Biol 42:182–192.

Holm S, 1979. A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70.

Kamau L, Lehmann T, Hawley WA, Orago AS, and Collins FH, 1998. Microgeographic genetic differentiation of *Anopheles gambiae* mosquitoes from Asembo Bay, western Kenya: a comparison with Kilifi in coastal Kenya. Am J Trop Med Hyg 58:64–69.

Kamau L, Mukabana WR, Hawley WA, Lehmann T, Irungu LW, Orago AA, and Collins FH, 1999. Analysis of genetic variability in *Anopheles arabiensis* and *Anopheles gambiae* using microsatellite loci. Insect Mol Biol 8:287–297.

Kingdon J, 1989. Island Africa. Princeton, NJ: Princeton University Press.

Kumar S, Tamura K, and Nei M, 1993. MEGA: molecular evolutionary genetics analysis, version 1.01 University Park, PA: Pennsylvania State University.

Lanzaro GC, Toure YT, Carnahan J, Zheng L, Dolo G, Traore S, Petrarca V, Vernick KD, and Taylor CE, 1998. Complexities in the genetic structure of *Anopheles gambiae* populations in west Africa as revealed by microsatellite DNA analysis. Proc Natl Acad Sci USA 95:14260–14265.

Lehmann T, Besansky NJ, Hawley WA, Fahey TG, Kamau L, and Collins FH, 1997. Microgeographic structure of *Anopheles gambiae* in western Kenya based on mtDNA and microsatellite loci. Mol Ecol 6:243–253.

Lehmann T, Blackston CR, Besansky NJ, Escalante AA, Collins FH, and Hawley WA, 2000. The Rift Valley complex as a barrier to gene flow for *Anopheles gambiae* in Kenya: the mtDNA perspective. J Hered 91:165–168.

Lehmann T, Hawley WA, Grebert H, and Collins FH, 1998. The effective population size of *Anopheles gambiae* in Kenya: implications for population structure. Mol Biol Evol 15:264–276.

Lehmann T, Hawley WA, Grebert H, Danga M, Atieli F, and Collins FH, 1999. The Rift Valley complex as a barrier to gene flow for *Anopheles gambiae* in Kenya. J Hered 90:613–621.

Lehmann T, Hawley WA, Kamau L, Fontenille D, Simard F, and Collins FH, 1996. Genetic differentiation of *Anopheles gambiae* populations from east and west Africa: comparison of microsatellite and allozyme loci. Heredity 77(pt 2):192–200.

Machado CA, Kliman RM, Markert JA, and Hey J, 2002. Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. Mol Biol Evol 19:472–488.

Marshall E, 2000. Malaria. A renewed assault on an old and deadly foe. Science 290:428–430.

Michalak P, Minkov I, Helin A, Lerman DN, Bettencourt BR, Feder ME, Korol AB, and Nevo E, 2001. Genetic evidence for adaptation-driven incipient speciation of *Drosophila melanogaster* along a microclimatic contrast in "Evolution Canyon," Israel. Proc Natl Acad Sci USA 98:13195–13200.

Mooney CZ and Duval RD, 1993. Bootstrapping: a nonparametric approach to statistical inference New Berry, CA: Sage.

Mukabayire O, Caridi J, Wang X, Toure YT, Coluzzi M, and Besansky NJ, 2001. Patterns of DNA sequence variation in chromosomally recognized taxa of *Anopheles gambiae*: evidence from rDNA and single-copy loci. Insect Mol Biol 10:33–46.

Nei M, 1987. Molecular evolutionary genetics New York: Columbia University Press.

Nicholson SE, 1998a. Fluctuations of Rift Valley Lakes Malawi and Chilwa during historical times: a synthesis of geological, archeological and historical information In: Environmental change and response in east African lakes (Lehman JT, ed). Amsterdam: Kluwer; 207–231.

Nicholson SE, 1998b. Historical and modern fluctuations of Lakes Tanganyika and Rukwa and their relationship to rainfall variability Amsterdam: Kluwer; 1–19.

Owen R, Crossley R, Johnson TC, Tweddle D, Kornfield I, Davison S, Eccles DH, and Engstrom DE, 1990. Major low levels of Lake Malawi and their implications for speciation rates in cichlid fishes. Proc R Soc Lond B Biol Sci 240:519–553.

Petrarca V and Beier JC, 1992. Intraspecific chromosomal polymorphism in the *Anopheles gambiae* complex as a factor affecting malaria transmission in the Kisumu area of Kenya. Am J Trop Med Hyg 46:229–237.

Petrarca V, Nugud AD, Ahmed MA, Haridi AM, Di Deco MA, and Coluzzi M, 2000. Cytogenetics of the *Anopheles gambiae* complex in Sudan, with special reference to *An. arabiensis*: relationships with east and west African populations. Med Vet Entomol 14:149–164.

Powell JR, 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.

Raymond M and Rousset F, 1995a. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Hered 86:248–249.

Raymond M and Rousset F, 1995b. Testing heterozygote excess and deficiency. Genetics 140:1413–1419.

Rousset F, 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145:1219–1228.

SAS, 1990. SAS language: references, version 6. Cary, NC: SAS Institute.

Scott JA, Brogdon WG, and Collins FH, 1993. Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. Am J Trop Med Hyg 49:520–529.

Simard F, Lehmann T, Lemasson JJ, Diatta M, and Fontenille D, 2000. Persistence of *Anopheles arabiensis* during the severe dry season conditions in Senegal: an indirect approach using microsatellite loci. Insect Mol Biol 9:467–479.

Slatkin M, 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462.

Smits A, Roelants P, Van Bortel W, and Coosemans M, 1996. Enzyme polymorphisms in the *Anopheles gambiae* (Diptera:Culicidae) complex related to feeding and resting behavior in the Imbo Valley, Burundi. J Med Entomol 33:545–553.

Taylor C, Toure YT, Carnahan J, Norris DE, Dolo G, Traore SF, Edillo FE, and Lanzaro GC, 2001. Gene flow among populations of the malaria vector, *Anopheles gambiae*, in Mali, West Africa. Genetics 157:743–750.

Toure YT, Petrarca V, Traore SF, Coulibaly A, Maiga HM, Sankare O, Sow M, Di Deco MA, and Coluzzi M, 1994. Ecological genetic studies in the chromosomal form Mopti of *Anopheles gambiae* s.str. in Mali, West Africa. Genetica 94:213–223.

Toure YT, Petrarca V, Traore SF, Coulibaly A, Maiga HM, Sankare O, Sow M, Di Deco MA, 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 YT, Taylor CE, Norris DE, Dolo G, and Lanzaro GC, 2001. DNA analysis of transferred sperm reveals significant levels of gene flow between molecular forms of *Anopheles gambiae*. Mol Ecol 10: 1725–1732.

Weill M, Chandre F, Brengues C, Manguin S, Akogbeto M, Pasteur N, Guillet P, and Raymond M, 2000. The *kdr* mutation occurs in the Mopti form of *Anopheles gambiae* s.s. through introgression. Insect Mol Biol 9: 451–455.

Weir BS and Cockerham CC, 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.

Whitlock MC, and McCauley DE, 1999. Indirect measures of gene flow and migration: $F_{ST} = 1/(4N_m + 1)$. Heredity 82:117–125.

Wondji C, Simard F, and Fontenille D, 2001. Evidence for genetic differentiation between molecular forms M and S within the forest chromosomal form of *Anopheles gambiae* in an area of sympatry. Insect Mol Biol.

Wright S, 1951. The genetical structure of populations. Ann Eugen 15: 323–354.

Wright S, 1978. Evolution and the genetics of populations: variability within and among natural populations Chicago: University of Chicago Press.

Zheng L, Benedict MQ, Cornel AJ, Collins FH, and Kafatos FC, 1996. An integrated genetic map of the African human malaria vector mosquito, *Anopheles gambiae*. Genetics 143:941–952.

Recieved August 23, 2002
Accepted December 31, 2002

Corresponding Editor: Stephen Schaeffer