Received 7 October 1998, accepted 17 June 1999

Population structure in the malaria vector, *Anopheles arabiensis* Patton, in East Africa

M. J. DONNELLY†, N. CUAMBA†‡, J. D. CHARLWOOD§, F. H. COLLINS¶ & H. TOWNSON†*

†Division of Parasite and Vector Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, U.K., ‡Ministerio da Saude, Instituto Nacional de Saude, C.P. 264 Maputo, Mozambique, §Centro Nacional de Endemias, Caixa Postal 27, Sao Antonio, Principe, West Africa, ¶Department of Biological Sciences, Galvin Life Sciences, University of Notre Dame, PO Box 369, Notre Dame, IN 46556, U.S.A.

The population structure of the malaria vector *Anopheles arabiensis* was investigated using data from six microsatellite loci in samples from localities in Mozambique and Tanzania. Genotype frequencies were neither significantly different between houses in a village in Tanzania nor between villages within a 20-km radius in Mozambique. Thus a deme has an area greater than 20 km in radius. At five of the six loci the heterozygosity of the population from Mozambique was lower than that from Tanzania, implying a lower effective population size (N_e) at this southern edge of the species range. There were significant differences in genotype frequencies between the Tanzanian and Mozambique populations at five of the six loci ($P < 0.05$). Values for both F_{ST} (mean = 0.069) and R_{ST} (mean = 0.025) were significantly different from zero ($P < 0.05$) at four and three out of five loci, respectively, but there was no significant correlation between the two statistics. The wide variation in values of F_{ST} and R_{ST} across loci suggests that care should be taken in interpreting values derived from averaging across loci. Whether the variation results from sampling effects or selectional constraints on some loci is unclear. Although there is evidence for significant differentiation between these populations, estimates of gene flow (*Nm*) calculated from mean F_{ST} and R_{ST} statistics were relatively high, 3.4 and 4.9, respectively. We argue that this is more likely to reflect recent separation of these populations and/or large effective population size rather than large-scale present day migrations.

Keywords: *Anopheles arabiensis,* gene flow, microsatellites, population differentiation.

Introduction

Mosquitoes of the *Anopheles gambiae* complex are the principal vectors of malaria in sub-Saharan Africa. The complex comprises six sibling species that differ in their ecology and epidemiological importance. Two species, *An. gambiae s.s.* and *An. arabiensis*, are regarded as major vectors; another three species, *An. merus*, *An. melas* and *An. bwambae*, are of localized importance, whereas the sixth, *An. quadriannulatus*, is zoophilic and hence not a vector. A knowledge of population structure in the major vector species is fundamental to an understanding of malaria epidemiology and the spread of insecticide resistance. To date most published work on population structure within the complex has focused upon *An. gambiae s.s.* (Coluzzi *et al.*, 1985; Lehmann *et al.*, 1997; and refs. therein). Although the two species exist sympatrically over much of their species range, *An. arabiensis* extends into more arid environments, is more likely to rest in outdoor structures than in the interior of houses and will feed preferentially on animals rather than humans (Charlwood *et al.*, 1995). Such differences in ecology and behaviour between *An. arabiensis* and *An. gambiae* may be reflected in differences in their population structure.

The aim of this study was to determine how the population structure of *An. arabiensis* differs from that previously reported for *An. gambiae*, where possible relating this to known differences in the biology of the two species. Our approach has been to use primers for six polymorphic microsatellite loci, developed for use in *An. gambiae s.s.*, to investigate genetic differences and possible substructure in populations of *An. arabiensis* from Tanzania and Mozambique in East Africa.

*Correspondence. E-mail: htownson@liv.ac.uk

We have investigated mosquito population substructure at three levels: (i) in samples collected from different houses within a village in Tanzania; (ii) in samples collected from two villages, 20 km apart, in Mozambique; (iii) between samples from Mozambique and Tanzania separated by 2000 km. From the derived allele and genotype frequencies we have compared three estimates of population differentiation and we discuss their appropriateness and their significance for our understanding of the biology of this species.

Materials and methods

Sample sites

Mozambique Samples were collected from two villages close to Maputo (25°58′S, 32°36′E), Mozambique. Matola is a coastal peri-urban area on the outskirts of Maputo, whereas Boane (25°2′S, 32°19′E) is approximately 26 km from Maputo and 20 km from Matola. The climate of this region is dry tropical with monthly rainfall fluctuating between 0 and 326 mm and an annual average of 650 mm. The average monthly temperature varies from a low of 19°C in July to a high of 26°C in February. The wet season lasts from November to April with a dry season from May to October and malaria is classified as holoendemic. Mosquitoes were collected inside houses by human landing catches and resting collections between December 1995 and March 1996.

Tanzania Resting collections of mosquitoes were carried out from 18 to 21 July 1995. Samples came from three houses in the village of Kivukoni (8°09′S, 36°24′E), 7 km south of the town of Ifakara on the banks of the Kilombero River, in an area of intense malaria transmission. There are approximately 12 houses in the settlement of Kivukoni all within a radius of 250 m. The area has two rainy seasons. The main rains last from March to May and are followed by a dry season that extends from June to November. A shorter rainy season occurs in December and January and is followed by a brief dry season before the onset of the main rains. The majority of the population are subsistence farmers, predominantly cropping maize and rice (Charlwood *et al.*, 1995).

DNA extraction, species identification and microsatellites

All mosquitoes were stored dry over silica gel prior to DNA extraction. DNA from all specimens was extracted using a modification of the phenol:chloroform technique of Ballinger-Crabtree *et al.* (1992) and resuspended in 100 (L of TE buffer. Each specimen was then identified to species level using species-specific PCR primers, following the technique of Scott *et al.* (1993). Of 13 locus-specific primers (Primers 1D1, 2A1, 29C1, 33C1 from Lehmann *et al.*, 1996a; Ag2H26, Ag2H46, Ag2H102, Ag2H175, Ag2H1010, Ag3H83, Ag3H88, Ag3H93, Ag3H249, from Zheng *et al.*, 1996) designed to amplify regions of the *An. gambiae s.s.* genome, 10 were found to amplify consistently in *An. arabiensis* (Primers 1D1, 29C1, 33C1 Ag2H26, Ag2H46, Ag2H102, Ag2H175, Ag2H1010, Ag3H88, Ag3H249) and eight in *An. merus* (Primers 1D1, 29C1, 33C1, Ag2H26, Ag2H102, Ag2H175, Ag2H1010, Ag3H249). Six primers were selected for use in a more detailed study of *An. arabiensis* population substructure and gene flow (Table 1). All of these loci, except locus 33C1, lie outside

Table 1 Microsatellite loci studied in more detailed survey of population differentiation and gene-flow in *Anopheles arabiensis*

Locus	Cytological location	Repeat motif	Repeat no. & allele size	T_A(°C)	Primer sequence (5–3′)
29C1[1]	IIIR:29C	(TGA)	$(TGA)_3$ –145	55°C	ATG TTC CAG AGA CGA CCC AT TGT TGC CGG TTT GTT GCT GA
33C1[2]	IIIR:33C	(AGC, TGG)	$(AGC)_9$–148	55°C	TTG CGC AAC AAA AGC CCA CG ATG AAA CAC CAC GCT CTC GG
Ag2H46	IIR:7A	(GT)	$(GT)_8$–138	55°C	CGC CCA TAG ACA ACG AAA GG TGT ACA GCT GCA GAA CGA GC
Ag3H88	IIIL:	(GT)	$(GT)_9$ –176	60°C	TGC GGC GGT AAA GCA TCA AC CCG GTA ACA CTG CGC CGA C
Ag2H175	IIR:	(CA)	$(CA)_8$–97	60°C	AGG AGC TGC ATA ATT CAC GC AGA AGC ATT GCC CGC ATT CC
Ag3H249	IIIR:	(GT)	$(GT)_{15}$–128	60°C	ATG TTC CGC ACT TCC GAC AC GCG AGC TAC AAC AAT GGA GC

[1] Locus 29C1 is from the *xanthine dehydrogenase* gene. [2] Locus 33C1 is associated with the *dopa decarboxylase* (*Ddc*) gene.

the major known chromosomal inversions (Zheng *et al.*, 1996).

PCR using fluorescent markers

PCR products were labelled fluorescently using one of two methods. Either one of the primer pair was labelled with a 5′ fluorescent dye (6-FAM, HEX or TET; GIBCO, Palo Alto, CA) or fluorescent UTPs were added to reactions with unlabelled oligonucleotide primers. PCR reactions were carried out in 25 μL in 0.5 mL Eppendorf tubes with a mineral oil overlay. Each reaction contained approximately 1/100th of the genomic DNA of an individual mosquito, 1 × reaction buffer, 200 μM of each dNTP, 1.5 mM $MgCl_2$, 0.5 Units of DNA Polymerase (DyNAzymeII DNA Polymerase, Finnzymes Oy) and 12.5 pmols of each primer. In reactions where fluorescently labelled primers were not used, fluorescently labelled UTP was included in the reaction mix at a concentration of 6.4 nM. All amplifications were performed in a Hybaid Omnigene thermal cycler using the following amplification programme: a 5-min 94°C denaturation step followed by 30 cycles of 30 s at 94°C, 30 s at 55–60°C (primer dependent) and 30 s at 72°C; there was a final extension step of 5 min at 72°C. Products were run on a Perkin-Elmer ABI PRISM 377 Automatic Sequencer using the default settings. Data were automatically collected and analysed using GENESCAN and GENOTYPER software (Applied Biosystems). The alleles were sized using a 'local southern' option. This technique estimates allele sizes by using the three molecular weight markers closest in size to the product to determine a best fit estimate.

Data analysis

Linkage disequilibrium Linkage equilibrium was tested using a contingency table test for genotypic linkage disequilibrium between pairs of loci in a population, based upon the null hypothesis that genotypes at one locus are independent of genotypes at the other locus. Calculations were performed using the GENEPOP V3–1 program (Raymond & Rousset, 1995) which performs a significance test using Markov chain procedures.

Hardy–Weinberg equilibrium (HWE) Tests of deviation from Hardy–Weinberg proportions were performed using the GENEPOP V3–1 program (Raymond & Rousset, 1995) and based on the score test (*U*-test) with a null hypothesis of random union of gametes and an alternative hypothesis of heterozygote deficiency.

Nei's unbiased estimates of heterozygosity The within-population heterozygosity at each locus was estimated using Nei's (1978) unbiased estimator $h_e = 2n(1 - \Sigma p_i^2)/(2n - 1)$, where p_i is the frequency of the *i*th allele and n is the number of individuals in the sample. Standard errors were calculated based upon a variance of $\Sigma(h_e - H_e)^2/n(n - 1)$, where $H_e = \Sigma(h_e)/n$.

Estimates of population differentiation and gene flow The differences between populations were investigated by three methods: (i) exact tests of genotype homogeneity among populations (Goudet *et al.*, 1996); (ii) Weir & Cockerham's (1984) estimates of *F* statistics; (iii) Slatkin's (1995) R_{ST} statistic.

Genotype-based exact tests of population differentiation were calculated using the GENEPOP V3–1 program (Raymond & Rousset, 1995). Tests of between-population genotype frequency homogeneity were calculated using an unbiased estimate of the *P*-value of a log-likelihood based exact test. A Markov chain method was used to generate an unbiased estimate of the exact probability. The parameters of the Markov chain were the default values of the program. If more than two sample populations were present, all population pairs were tested. A simple comparison of genotypic frequencies can be used to compare populations or loci even when they are not in HWE. Although exact tests of genotype and allele frequencies may be the most sensitive detectors of population differentiation, they provide no estimate of the magnitude of the differences: to assess population structure in a quantitative manner it is necessary to use either F_{ST} or R_{ST} estimators.

F-statistics (F_{IS}, F_{IT} and F_{ST}) were calculated using the FSTAT program (Goudet, 1995). This program calculates unbiased estimates of *F*-statistics and performs numerical re-sampling by bootstrap and jack-knife procedures in order to estimate confidence intervals and the significance of values. The average F_{ST} value was calculated following Weir & Cockerham (1984).

Slatkin's (1995) R_{ST} statistic is similar to F_{ST} but developed specifically for microsatellite loci to account for the assumed differences in mutational processes at these loci. F_{ST}-based measures of differentiation are based on an infinite allele model (IAM) or *k*-allele model (KAM) and may not be applicable to microsatellite loci which are thought to follow a stepwise mutation model (SMM) more closely. Calculation of R_{ST}, based upon the repeat numbers of alleles, assumes that there is no restriction on allele size and that the mutation rate is relatively constant across all alleles. In practice there is emerging evidence to suggest that both these assumptions may sometimes be violated (Amos *et al.*, 1996; Lehmann *et al.*, 1996b).

R_{ST} statistics were calculated by hand. Average, across loci, values for R_{ST} were calculated following Slatkin's procedure of averaging the denominators and numerators for all loci prior to calculation. The significance of R_{ST} values was determined by a simple ANOVA test. In our discussion we consider the most appropriate measure of population differentiation with reference to both the possible underlying mutation processes at individual loci and the methods of calculating each statistic.

As an estimate of gene flow, the number of migrants per population per generation (*Nm*), was calculated for F_{ST} according to the equation $Nm = (1 - F_{ST})/4F_{ST}$. Calculations of *Nm* for R_{ST} followed Slatkin (1995), $Nm = (d - 1)\ (1 - R_{ST})/4d\ R_{ST}$, where *d* is the number of populations in the study. These calculations assume an island model of population structure and mutation–drift equilibrium. As discussed below, this may not be the most appropriate model for these populations and, indeed, there may be no simple relationship between estimators of genetic differentiation and gene flow.

Mutation model The underlying mutation model influences which method of calculating gene flow is most appropriate for the data set. The appropriateness of the mutation model for these loci was examined following the example of Garcia de Leon *et al.* (1997), based on the simulations of Shriver *et al.* (1993) who demonstrated that when values of heterozygosity are greater than 0.5, the number of alleles for a given heterozygosity can be significantly larger for a locus following the Infinite Alleles Model (IAM) than for one following the Stepwise Mutation Model (SMM). This permits a simple empirical test of which model is more likely to fit the data, although no estimates of statistical confidence are possible by this test.

In all cases where multiple tests were performed, significance levels were adjusted following Bonferroni procedures (Holm, 1979).

Results and discussion

All *An. gambiae s.l.* from Mozambique were identified as either *An. arabiensis* or *An. merus*, the relative proportions of *An. arabiensis* being 103/122 in Matola and 44/85 in Boane. In Kivukoni, Tanzania all specimens were identified as either *An. gambiae* (54) or *An. arabiensis* (89). Subsets of the *An. arabiensis* samples were taken and used in the following analysis of population substructure and gene flow.

All loci save one (29C1) were highly polymorphic with between four and 14 alleles per locus, excluding null alleles (see later discussion). For the pooled samples, excluding locus 29C1, the average number of alleles per locus was 9.6 and the average unbiased estimate of heterozygosity was 0.739 (Table 2). There were no significant associations, indicative of linkage disequilibrium, between any

Table 2 Observed (upper value) and expected (lower value) estimates of heterozygosity of *Anopheles arabiensis* for each sample by loci, the number of alleles found at each locus for pooled populations and details of which mutation model the loci most closely fit: the Infinite Allele Model (IAM) or the Stepwise Mutation Model (SMM)

	Kivukoni, Tanzania					Mozambique			
	House 1 $2n = 48$	House 2 $2n = 40$	House 3 $2n = 32$	Total $2n = 120$	No. of alleles	Boane $2n = 46$	Matola $2n = 126$	No. of alleles	All mutation mode
Locus 29C1	0.208	0.100	0.071	0.137	2	0.087	0.097	2	SMM/IAM
	0.189	0.097	0.067	0.130		0.091	0.092		
Locus 33C1	0.625	0.600	0.500	0.583	7	0.652	0.500	7	SMM/IAM
	0.751	0.734	0.683	0.728		0.567	0.625		
Locus Ag2H46	0.708	0.500	0.333	0.544	10	0.682	0.613	10	SMM
	0.893	0.856	0.884	0.884		0.852	0.844		
Locus Ag3H88	0.368	0.526	0.300	0.417	13	0.250	0.242	10	IAM
	0.706	0.861	0.821	0.846		0.693	0.811		
Locus Ag2H175	0.522	0.526	0.467	0.509	5	0.545	0.339	4	IAM
	0.567	0.747	0.663	0.605		0.591	0.553		
Locus Ag3H249	0.350	0.450	0.538	0.434	7	0.350	0.194	9	IAM
	0.692	0.653	0.782	0.695		0.662	0.735		
Mean ± SE	0.464 ± 0.08	0.450 ± 0.07	0.368 ± 0.07	0.437 ± 0.05	7.3	0.428 ± 0.10	0.331 ± 0.08	7	
	0.633 ± 0.10	0.658 ± 0.12	0.650 ± 0.10	0.660 ± 0.11		0.576 ± 0.10	0.610 ± 0.11		

pair-wise combination of alleles across loci at any population level ($P > 0.05$; 120 pair-wise comparisons comprising 15 pair-wise comparisons between loci for eight samples; significance level adjusted for multiple tests).

Deviations from HWE

Five, of the six loci (except 29C1) showed significant deviation ($P < 0.05$) from HWE as a result of heterozygote deficit. In Table 3 values of F_{IS} for each locus at each population level are given together with the significance of the test of deviation from HWE based on an alternative hypothesis of heterozygote deficit. It can be seen that deviation from HWE as a result of heterozygote deficit is detected at all population levels although not necessarily within all sample/loci combinations at a population level.

Deviation from HWE is a common finding in studies that utilize microsatellite loci. This deficit is usually attributed to null alleles (Callen *et al.*, 1993; Lehmann *et al.*, 1996a; Garcia de Leon *et al.*, 1997), selection (Garcia de Leon *et al.*, 1997), or grouping of gene pools (Wahlund effect) (Gibbs *et al.*, 1997). In addition, inbreeding or nonrandom mating may also result in heterozygote deficits. Individuals in the Tanzanian sample were collected from three houses and it was therefore possible to investigate whether there was indication of grouping at the village level. Only one locus (Ag3H88) gave evidence of between-house differences but this result was no longer significant when significance levels were adjusted to account for multiple testing. Thus any heterozygote deficit is unlikely to result from pooling samples from different houses into village collections.

Null alleles, that is alleles that are not amplified because of mutations at the primer binding sites, can result in underestimation of heterozygosity because they are detected only in the homozygous state. An estimate of the expected frequency of the null allele can be made by assuming that the observed heterozygote deficit is attributable to null alleles. Following Chakraborty *et al.* (1992) the estimated frequency of the null allele, $r = (H_{\mathrm{e}} - H_{\mathrm{o}})/(H_{\mathrm{e}} + H_{\mathrm{o}})$, where $H_{\mathrm{e}} =$ expected frequency of heterozygotes and $H_{\mathrm{o}} =$ the observed frequency of heterozygotes. From this figure it is possible to calculate the expected number of homozygous null individuals likely to be observed in a sample of a given size. Table 4 details both the observed and expected number of null allele homozygotes. A specimen which was shown to have amplifiable DNA but would not produce a product for a given locus on three occasions was regarded as a null allele homozygote. It can be seen that in general, the frequency of the putative null allele homozygotes agrees with the expected value indicating that in these cases null alleles are the likely cause of heterozygote deficit. However, for two loci (Ag3H88 and Ag3H249) there are higher than expected numbers of putative null allele homozygotes. Examination of the raw data showed that in 37% of cases, an individual that was a putative null allele homozygote at locus Ag3H88 was also an apparent null allele homozygote at locus Ag3H249. This suggests that a number of these supposed null homozygotes are caused by a failure of amplification because of variations in template quality. Exclusion of null allele homozygotes from the analysis will affect the allele and genotype frequency distributions and this could lead to inaccuracies in measures of population structure. Previous cytological studies have found no evidence for deviation from HWE in east African populations of *An. arabiensis* (Petrarca & Beier, 1992), which suggests that there is unlikely to be nonrandom mating in our samples. This is also consistent with the observation that populations of *An. gambiae s.l.* are normally in HWE (Petrarca & Beier, 1992; Lehmann *et al.*, 1996a). The latter authors found null alleles in locus Ag2H46 in

Table 3 Estimates of F_{IS} and significance level for goodness of fit tests to HWE, based upon a null hypothesis of heterozygote deficit, for all *Anopheles arabiensis* populations. Also F_{IT} value for pooled population

	Kivukoni, Tanzania				Mozambique		All
Locus	House 1	House 2	House 3	Total	Boane	Matola	F_{IT}
29C1	−0.095NS	−0.027NS	0.000NS	−0.065NS	−0.022NS	−0.043NS	−0.06NS
33C1	0.171*	0.186**	0.275NS	0.200***	−0.154NS	0.15**	0.373***
H46	0.207***	0.424***	0.632***	0.387***	0.204*	0.275***	0.344***
H88	0.485***	0.396***	0.647***	0.510***	0.647***	0.672***	0.609***
H175	0.082NS	0.301**	0.302NS	0.238**	0.079NS	0.390**	0.326***
H249	0.501**	0.316*	0.320*	0.379***	0.477***	0.669***	0.529***

*$P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

Table 4 The numbers of observed (upper value) and expected (lower value) null allele homozygotes for each locus at each *Anopheles arabiensis* population level

	Kivukoni, Tanzania				Mozambique	
Locus	House 1	House 2	House 3	Total	Boane	Matola
29C1	0	0	0	0	0	0
	0.00	0.00	0.00	0.00	0.00	0.00
33C1	0	0	0	0	0	0
	0.19	0.2	0.38	0.72	0.00	0.76
Ag2H46	0	2	1	3	1	0
	0.32	1.38	3.26	3.42	0.28	1.58
Ag3H88	5	1	6	12	7	6
	2.38	1.16	3.46	6.9	5.08	18.40
Ag2H175	1	1	1	3	1	0
	0.05	0.60	0.48	0.42	0.46	3.65
Ag3H249	4	0	3	7	3	13
	2.59	0.67	0.54	3.18	2.18	21.36

The observed numbers of homozygotes are based upon the number of individuals which did not produce any discernible products upon repeated amplification. The calculation of expected numbers of null allele homozygotes is based upon Chakraborty *et al.* (1992) with adjustment for sample size.

An. gambiae s.s. (Lehmann *et al.*, 1996a) and concluded that they were likely to be present also at locus 33C1 (Lehmann *et al.*, 1997). It seems likely that the presence of null alleles at these loci is the cause of the heterozygote deficit we see in our *An. arabiensis* populations.

Population substructure between villages in Mozambique

Allele frequencies at the six loci were similar for both populations from Mozambique. There was no evidence for population differences between the samples from either F_{ST} or R_{ST} (mean $F_{ST} = 0.010$; mean $R_{ST} = 0.008$) or from analysis of genotypic frequencies in each population ($P > 0.05$), although it should be noted that the sample from Boane is relatively small ($n = 23$). Thus the indications are that populations of *An. arabiensis* are effectively panmictic in this region within an area of radius 20 km Lehmann *et al.* (1997) also found no evidence for between-house differentiation in *An. gambiae* from western Kenya and concluded that the area occupied by a deme of *An. gambiae* was in excess of a radius of 25 km.

Estimation of genetic differentiation between Mozambique and Tanzania

The sample from Matola in Mozambique was compared with the population from Kivukoni in Tanzania. At only three of the six loci were the most common alleles the same in both populations (Fig. 1) and population-specific alleles were found in four of the loci examined. For only three of the 14 population-specific alleles was the allele frequency per locus per population greater than 5%. For all but one locus the sample from Tanzania produced higher estimates of heterozygosity than did the sample from Matola. This suggests that the Mozambique sample may be drawn from a parent population that is genetically depauperate, perhaps as a result of lower effective size. The Mozambique samples are from the extreme southern end of the geographical range of *An. arabiensis* where habitat fragmentation may reduce effective population size. Apparent decreases in genetic diversity towards the geographical edges of a species' range have been observed in *Drosophila* (Dobzhansky, 1970 and refs. therein).

Exact tests showed that the complement of genotypes at five out of six loci was highly significantly different between Tanzania and Mozambique ($P < 0.05$ for genotypes; Table 5); the locus that showed no difference (locus 29C1) was close to fixation for the same allele in both populations. Because it contributes nothing to the study of population structure, this locus has been excluded from further analysis and discussion. The F_{ST} values of four of the five remaining loci, were significantly different from zero ($P < 0.05$–$P < 0.001$) (Table 5). However, there was considerable heterogeneity between loci, with estimates of F_{ST} ranging from 0.015 to 0.239. R_{ST} values also showed a large variation across loci (0.002–0.098, mean 0.027) and were only significant at three out of five loci (Table 5). There was no detectable correlation between F_{ST} and R_{ST} values at each locus or across all loci averaged.

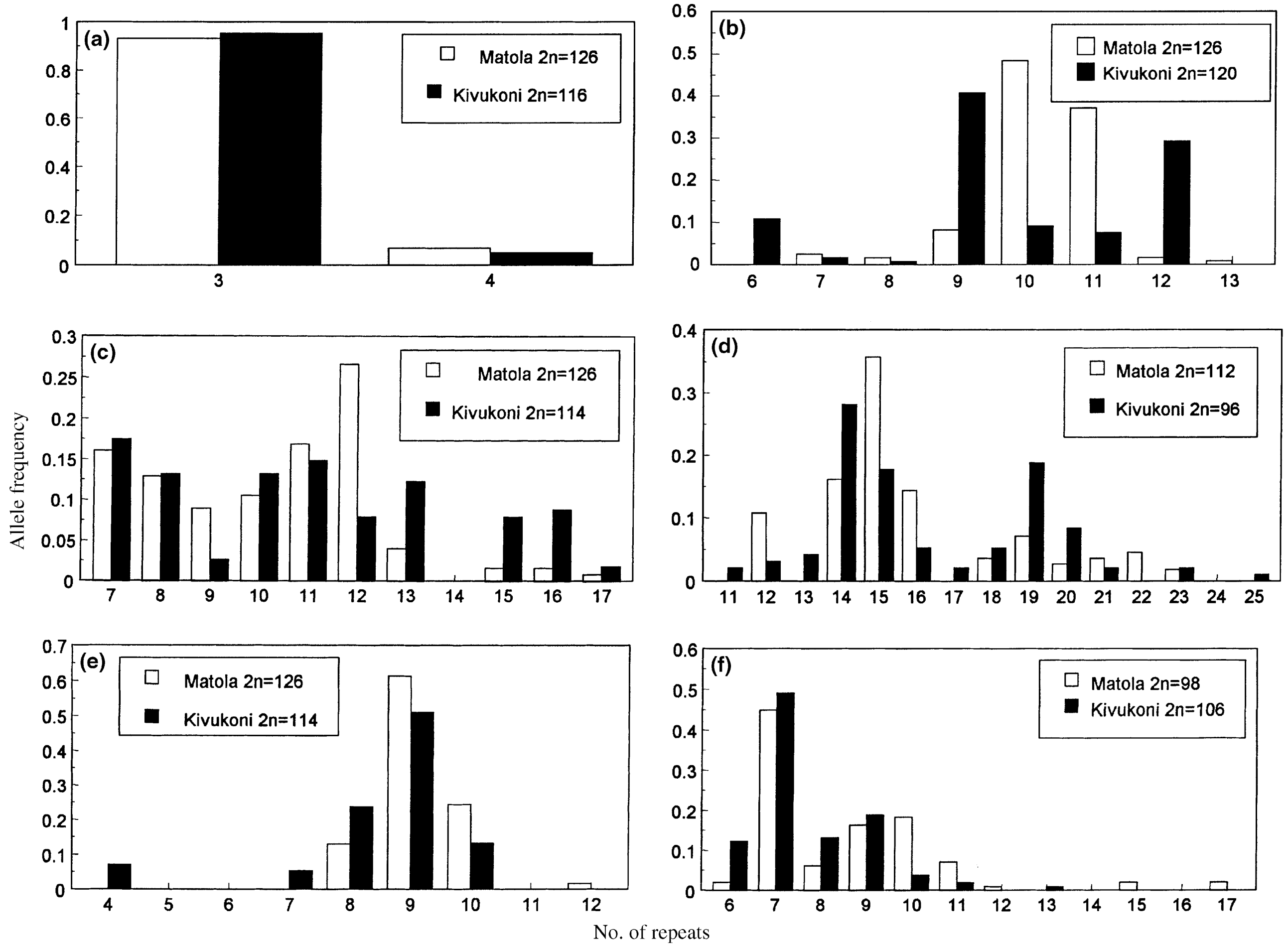

Fig. 1 Allele frequency arrays in *Anopheles arabiensis* for (a) locus 29C1, (b) locus 33C1, (c) locus Ag2H46, (d) locus Ag3H88, (e) locus Ag2H175, (f) locus Ag3H249.

Because the genotype frequencies at five of the six loci were shown to be significantly different, the nonsignificant values of F_{ST} and R_{ST} statistics at some loci raises the question whether F_{ST} and R_{ST} and the tests of their significance are sufficiently sensitive. The large between-locus variation in values of both F_{ST} (0.015–0.239) and R_{ST} (0.002–0.098) may simply reflect sampling effects but may be caused by selection upon some of the loci; possibly a result of linkage effects with nearby coding regions or chromosomal inversions. Lehmann *et al.* (1996b) inferred that locus Ag2H46 was under mutational constraint in *An. gambiae* but it is not known whether this locus, or others, are similarly constrained in *An. arabiensis*.

This study highlights the variations in sensitivity of F_{ST} and R_{ST} for detecting evidence of differentiation. For example, the allele frequency distribution for locus 33C1 (Fig. 1b) shows a large difference between populations, with the modal and the next most frequent allele being different in each population. This results in high estimates of population differentiation and low estimates of gene flow ($\theta = 0.239$, $Nm = 0.79$) when the F_{ST} statistic is used. However, the R_{ST} value for this locus is low, implying high levels of gene flow ($R_{ST} = 0.022$, $Nm = 5.55$). The converse is observed at locus Ag2H175 ($\theta = 0.026$, $Nm = 9.39$; $R_{ST} = 0.098$, $Nm = 1.15$) with allele arrays that are apparently very similar. F_{ST}, which measures variation across the allele array, is more sensitive to variation in allele frequencies across the array than is R_{ST} which may give undue weight to alleles furthest from the median class. In addition, only one of the loci studied here conformed primarily to a SMM, consistent with the growing evidence for deviation from a strict SMM for microsatellite loci (Di Rienzo *et al.*, 1994; Garcia de Leon *et al.*, 1997) (Table 2). Furthermore, we have found a number of polymorphisms involving insertions or deletions that result in the formation of electrophoretically identical but nonhomologous alleles (Donnelly & Townson, in prep.). This homoplasy may affect both F_{ST} and R_{ST} estimators

Table 5 Estimates of genetic differentiation and geneflow (*Nm*) between samples of *Anopheles arabiensis* from Kivukoni, Tanzania and Matola, Mozambique

	Exact tests		F_{ST} statistics (Weir & Cockerham, 1984)		R_{ST} statistics (Slatkin, 1995)	
Locus†	Alleles	Genotypes	θ	*Nm*	R_{ST}	Nm
Locus 29C1	0.59	0.58	0.000 NS	∝	0.000 NS	∝
Locus 33C1	$P < 0.001$	$P < 0.001$	0.239***	0.79	0.022*	5.6
Locus Ag2H46	$P < 0.001$	$P = 0.007$	0.021***	11.6	0.012 NS	10.3
Locus Ag3H88	$P < 0.001$	$P = 0.009$	0.033***	7.3	0.002 NS	62.4
Locus Ag2H175	$P < 0.001$	$P = 0.003$	0.026*	9.4	0.098*	1.2
Locus Ag3H249	$P < 0.001$	$P = 0.011$	0.015 NS	16.4	0.074*	1.6
Mean all six loci			0.069***	3.4	0.025*	4.9
95% Bootstrapped CI			0.017–0.156			
Mean five loci excluding			0.022***	11.1		
33C1 95% CI			0.015–0.029			

†Sample sizes: Matola, Mozambique $2n = 124$; Kivukoni, Tanzania $2n = 120$.
$*P < 0.05$; $**P < 0.01$; $***P < 0.001$.

(Viard *et al.*, 1998). Recent data suggest that mutation rates at microsatellite loci are lower than has previously been assumed. Thus Schug *et al.* (1997) have observed much lower rates of mutation at microsatellite loci in *Drosophila melanogaster* (6.3×10^{-6}) than in mammalian species (10^{-3}–10^{-5}). If this lower mutation rate holds for *An. arabiensis*, populations that have recently become isolated may not be detected using R_{ST} estimators. Constraints upon mutation at each locus would also exert a similar effect in reducing the accumulation of mutations in each population. A lower rate of mutation at these loci would increase the influence of drift on allele frequencies and thereby make F_{ST} a more appropriate estimator.

There is a growing body of evidence that R_{ST} may not be an appropriate estimator of population substructure for all microsatellite data. Forbes *et al.* (1995) found that F_{ST} was more sensitive to differences between allopatric populations and Perez-Lezaun *et al.* (1997) observed that genetic distance measures such as F_{ST}, which do not consider mutational relationships among alleles and which have known relationships to differentiation by drift, better reflect currently understood patterns of human evolution than do mutation-based distance measures, such as R_{ST}.

Estimates of gene flow

Large estimates of geneflow (*Nm*) were derived from both F_{ST} and R_{ST}. The calculation of gene flow from F_{ST} statistics assumes that migration occurs at random among all populations (an island model), that the populations are in migration–drift equilibrium and that the loci are neutral. Our estimates of gene flow based on F_{ST} statistics may be too large as they account solely for drift and do not account for mutation rates at these loci. However, notwithstanding our caveat about using values of gene flow averaged across loci, the values of *Nm* in excess of 3, as determined both by F_{ST} and R_{ST} ($Nm = 3.37$ and 4.9, respectively), appear to suggest that there are apparently few barriers to gene flow over 2000 km. However, we would argue that these high values of *Nm* are more likely to reflect large effective population sizes and recent separation than large-scale migrations. This result is in broad agreement with studies based upon multi locus allozyme electophoresis and chromosome polymorphisms which showed low levels of population differentiation in *An. arabiensis* from south of the Rift Valley (Petrarca & Beier, 1992). In comparison, Lehmann *et al.* (1996a) obtained similar values for gene flow between populations of *An. gambiae* from east and west Africa separated by a distance of approximately 6000 km (*Nm* 7.7 and 3.4 for F_{ST} and R_{ST}, respectively). The lower levels of gene flow observed in this study are heavily influenced by locus 33C1. Removal of this locus from the analysis (Table 5) results in higher estimates of gene flow (mean $Nm = 11.11$). Locus 33C1 is within the 3Ra chromosomal inversion strongly suggesting that loci within inversions may behave differently.

Conclusion

Appetitive flights in anopheline species are usually 1–2 km and this is not readily reconciled with the apparently high levels of gene flow observed in this and other studies. It has been proposed that the low values of F_{ST} and R_{ST} for *An. gambiae* are attributable to either constraints on microsatellite loci or a large amount of human-mediated transport (Lehmann *et al.*, 1996a). Although there are numerous examples of long-range transport of members of the *An. gambiae* complex by land, sea and air (for review see Gillies & De Meillon, 1968), the large estimates of the numbers of migrants per year implicit in our calculated values of F_{ST} and R_{ST} seem too great to be attributable to such a cause. *Anopheles arabiensis* commonly feeds on cattle and larvae are often found in rice-fields. It seems likely that this species has extended its range relatively recently, aided by the expansion of human settlement and intensive agriculture. We would argue that the population structure we observe in *An. arabiensis* reflects the relatively recent expansion of its population range and the large effective size of its populations, rather than a large amount of contemporary gene flow.

Acknowledgements

The authors would like to acknowledge the comments of Tovi Lehmann on an earlier version of the manuscript, Anthony Cornel for details of the location of chromosomal inversions and Margaret Hughes for advice on the ABI 377 analysis. This work was funded by a Wellcome Trust Biodiversity Prize Studentship.

References

AMOS, W., SAWCER, S. J., FEAKES, R. W. AND RUBINSZTEIN, D. C. 1996. Microsatellites show mutational bias and heterozygote instability. *Nature (Genetics)*, **13**, 390–391.

BALLINGER-CRABTREE, M. E., BLACK, W. C., V. I. AND MILLER, B. R. 1992. Use of genetic polymorphisms detected by the Random-Amplified Polymorphic DNA Polymerase Chain Reaction (RAPD-PCR) for differentiation and identification of *Aedes aegypti* subspecies and populations. *Am. J. Trop. Med. Hyg.*, **47**, 893–901.

CALLEN, D. F., THOMPSON, A. D., SHEN, Y., PHILLIPS, H. A., RICHARDS, R. I., MULLEY, J. C. AND SUTHERLAND, G. R. 1993. Incidence and origin of "Null alleles in the $(AC)_n$ microsatellite markers. *Am. J. Hum. Gen.*, **22**, 1–10.

CHAKRABORTY, R. D. E., ANDRADEM., DAIGER, S. P. AND BUDOWLE, B. 1992. Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. *Ann. hum. Genet.*, **56**, 45–57.

CHARLWOOD, J. D., SMITH, T., KIHONDA, J., BILLINGSLEY, P. F. AND TAKKEN, W. 1995. Density independent feeding success of malaria vectors (Diptera: Culicidae) in Tanzania. *Bull. ent. Res.*, **85**, 29–35.

COLUZZI, M., PETRARCA, V. AND DI DECO, M. A. 1985. Chromosomal inversion intergradation and incipient speciation in *Anopheles gambiae*. *Boll. Zool.*, **52**, 45–63.

DI RIENZO, A., PETERSON, A. C., GARZA, J. C., VALDES, A. M., SLATKIN, M. AND FREIMER, N. B. 1994. Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 3166–3170.

DOBZHANSKY, T. 1970. *Genetics of the Evolutionary Process*. Columbia University Press, New York.

FORBES, S. H., HOGG, J. T., BUCHANAN, F. C., CRAWFORD, A. M. AND ALLENDORF, F. W. 1995. Microsatellite evolution in congeneric mammals: domestic and bighorn sheep. *Mol. Biol. Evol.*, **12**, 1106–1113.

GARCIA DE LEON, F. J., CHIKHI, L. AND BONHOMME, F. 1997. Microsatellite polymorphism and population subdivision in natural populations of European Sea Bass *Dicentrarchus labrax* Linnaeus, 1758). *Mol. Ecol.*, **6**, 51–62.

GIBBS, H. L., PRIOR, K. A., WEATHERHEAD, P. J. AND JOHNSON, G. 1997. Genetic structure of the threatened eastern massauga rattlesnake *Sistrurus c. catenatus*: evidence from microsatellites DNA markers. *Mol. Ecol.*, **6**, 1123–1132.

GILLIES, M. T. AND DE MEILLON, B. 1968. *The Anophelinae of Africa South of the Sahara (Ethiopian Zoogeographical Region)* 2nd edn. *Publ. S. Afr. Inst. Med. Res.*, **54**, 1–343.

GOUDET, J. 1995. Fstat, Version 1.2: a computer program to calculate F statistics. *J. Hered.*, **86**, 485–486.

GOUDET, J., RAYMOND, M., DE MEEUS, T. AND ROUSSET, F. 1996. Testing differentiation in diploid populations. *Genetics*, **144**, 1933–40.

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

LEHMANN, T., BESANSKY, N. J., HAWLEY, W. A., FAHEY, T. G., KAMAU, L. AND COLLINS, F. H. 1997. Microgeographic structure of *Anopheles gambiae* in western Kenya based on mtDNA and microsatellite loci. *Mol. Ecol.*, **6**, 243–253.

LEHMANN, T., HAWLEY, W. A. AND COLLINS, F. H. 1996b. An evaluation of evolutionary constraints on microsatellite loci using null alleles. *Genetics*, **144**, 1155–1163.

LEHMANN, T., HAWLEY, W. A., KAMAU, L., FONTENILLE, D., SIMARD, F. AND COLLINS, F. H. 1996a. Genetic differentiation of *Anopheles gambiae* populations from East and West Africa: comparison of microsatellite and allozyme loci. *Heredity*, **77**, 192–200.

NEI, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583–590.

PEREZ-LEZAUN, A., CALAFELL, F., MATEU, E., COMAS, D., RUIZ-PACHECO, R. AND BERTRANPETIT, J. 1997. Microsatellite variation and the differentiation of modern humans. *Hum. Genet.*, **99**, 1–7.

PETRARCA, V. AND BEIER, J. C. 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.

RAYMOND, M. AND ROUSSET, F. 1995. GENEPOP (Version 1.2): population genetics software for exact tests and ecumenicism. *J. Hered.*, **86**, 248–249.

SCHUG, M. D., MACKAY, T. F. C. AND AQUADRO, C. F. 1997. Low mutation rates of microsatellite loci in *Drosophila melanogaster*. *Nature (Genetics)*, **15**, 99–102.

SCOTT, J. A., BROGDON, W. G. AND COLLINS, F. H. 1993. Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. *Am. J. Trop. Med. Hyg.*, **49**, 520–529.

SHRIVER, M. D., JIN, L., CHAKRABORTY, R. AND BOERWINKLE, E. 1993. VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics*, **134**, 983–993.

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

VIARD, F., FRANCK, P., DUBOIS, M.-P., ESTOUP, A. AND JARNE, P. 1998. Variation of microsatellite size homoplasy across electromorphs, loci and populations in three invertebrate species. *J. Mol. Evol.*, **47**, 42–51.

WEIR, B. S. AND COCKERHAM, C. C. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

ZHENG, L., BENEDICT, M. Q., CORNEL, A. J., COLLINS, F. H. AND KAFATOS, F. C. 1996. An integrated genetic map of the African human malaria vector mosquito, *Anopheles gambiae*. *Genetics*, **143**, 941–952.