Population Structure and Population History of *Anopheles dirus* Mosquitoes in Southeast Asia

*Catherine Walton,* Jane M. Handley,* Willoughby Tun-Lin,† Frank H. Collins,‡§ Ralph E. Harbach,§ Visut Baimai,|| and Roger K. Butlin* 

*School of Biology, University of Leeds, Leeds, England; †Department of Medical Research, Ministry of Health, Yangon, Myanmar; ‡Division of Parasitic Diseases, Centers for Disease Control, Atlanta, Georgia; §Department of Entomology, The Natural History Museum, London, England; ||Department of Biology, Mahidol University, Bangkok, Thailand.

Separating the confounding effects of long-term population history from gene flow can be difficult. Here, we address the question of what inferences about gene flow can be made from mitochondrial sequence data in three closely related species of mosquitoes, *Anopheles dirus* species A, C, and D, from southeast Asia. A total of 84 sequences of 923 bp of the mitochondrial cytochrome oxidase I gene were obtained from 14 populations in Thailand, Myanmar, and Bangladesh. The genealogy of sequences obtained from two populations of *An. dirus* C indicates no contemporary gene flow between them. The $F_{ST}$ value of 0.421 therefore probably represents a recent common history, perhaps involving colonization events. *Anopheles dirus* A and D are parapatric, yet no differentiation was seen either within or between species. The starlike genealogy of their haplotypes, smooth unimodal mismatch distributions, and excess of low frequency mutations indicate population expansion in *An. dirus* A and D. This, rather than widespread gene flow, explains their low within-species $F_{ST}$ values (0.018 and 0.022). The greater genetic diversity of *An. dirus* D suggests that expansion occurred first in species D and subsequently in species A. The current geographical separation and low hybrid fitness of these species also argue against ongoing interspecific gene flow. They suggest instead either historical introgression of mtDNA from *An. dirus* D into species A followed by independent range expansions, or a selective sweep of mtDNA that originated in *An. dirus* D. While not excluding contemporary gene flow, historical population processes are sufficient to explain the data in *An. dirus* A and D. The genealogical relationships between haplotypes could not be used to make inferences of gene flow because of extensive homoplasy due to hypervariable sites and possibly also recombination. However, it is concluded that this approach, rather than the use of fixation indices, is required in the future to understand contemporary gene flow in these mosquitoes. The implications of these results for understanding gene flow in another important and comparable group of malaria vector mosquitoes in Africa, the *An. gambiae* complex, are also discussed.

**Introduction**

Mitochondrial DNA (mtDNA) variation is an important and widely implemented tool in indirect studies of gene flow. Inferences of contemporary migration events are often made from the geographical pattern of genetic variation, although this can also be influenced by a population history of colonization events or range expansions or contractions (Templeton 1998 and references therein). In this paper, we consider what inferences about gene flow can be drawn from mtDNA sequence data for a group of anopheline mosquitoes. Many anopheline mosquitoes are vectors of human malaria, and one of the most effective means of controlling malaria is controlling vector populations. Understanding gene flow in anopheline mosquitoes is therefore important, particularly given the worldwide spread of insecticide resistance. In addition, large research efforts are being directed toward the goal of genetically modifying mosquitoes of the *Anopheles gambiae* complex, which contains several major African vectors, such that they are incapable of transmitting malarial parasites (Collins 1994). An appreciation of the nature and extent of the barriers to gene flow in anopheline mosquitoes is therefore critical if refractory genes are to be spread effectively through vector populations. In southeast Asia, some of the major malaria vectors are members of the *Anopheles dirus* complex, the focus of this study. An additional interest in studying the *An. dirus* complex is that it is largely restricted to areas of tropical evergreen rainforest in southeast Asia that are rapidly diminishing. Members of the *An. dirus* complex may therefore serve as useful indicator species of the effects of forest loss and fragmentation on biodiversity.

The *An. dirus* complex comprises at least seven species (Peyton 1989), but this study concerns only species A, C, and D. *Anopheles dirus* A and D have widespread distributions with a narrow zone of overlap along the Thai-Myanmar border (fig. 1). From this area, *An. dirus* A extends eastward across Thailand to Laos, Vietnam, and Cambodia, and *An. dirus* D extends westward through Myanmar toward Bangladesh and Assam (Rosenberg and Maheswary 1982; Tun-Lin et al. 1985; Baimai et al. 1988a). *Anopheles dirus* C appears to be associated with rocky limestone outcrops that occur in the central region of the Thai-Myanmar border and southward into peninsular Thailand (Baimai et al. 1988a). Consequently, *An. dirus* C has a patchy distribution. Limited allozyme evidence from the *An. dirus* complex in Thailand suggests little genetic population
structure (Green et al. 1992). In An. dirus D, there is a cline of chromosomal inversion frequencies from Thailand through Myanmar into Bangladesh (Baimai et al. 1988b), but it is not known whether or not this is maintained by selection.

Cytogenetic studies of mitotic and polytene chromosomes (Baimai, Harrison, and Nakavachara 1980; Baimai, Poopittayasataporn, and Kijchaloo 1988), allozyme data (Green et al. 1992), and crossing studies (Baimai et al. 1987) indicate that An. dirus A and C are very closely related. In their polytene chromosomes, the banding pattern is almost completely homosequential, and in F1 hybrids there is >90% synapsis. The F1 progeny are fertile and viable, with the exception of sterile males that are produced by the cross of female An. dirus A with male An. dirus C. By the same criteria, An. dirus D appears to be more distantly related. In the crossing studies between An. dirus D and An. dirus A or C, the only viable F1 progeny produced were females from crosses of male An. dirus A × female An. dirus D. However, experimental numbers were fairly low, and it is difficult to extrapolate from these laboratory crosses what might happen in nature. The only hybrid recorded in natural populations is that of An. dirus C-D, which was not formed in the small-scale laboratory crosses (Walton et al. 1999b).

**Materials and Methods**

Mosquito Collection and Identification

Mosquitoes were collected from 11 sites in Thailand, Myanmar, and Bangladesh in an attempt to represent the geographic distributions of the species (fig. 1). Along the Thai-Myanmar border, there were sympatric populations of An. dirus A and D at two sites (4 and 6) and sympatric populations of species C and D at site 5, making a total of 14 populations. With the exception of Myanmar, adult females were collected from all sites and assumed to be unrelated. In Myanmar, mosquito larvae were collected from small pools of water and reared to adulthood. Since broods of larvae are temporally and spatially structured, sibling relationships between the individuals cannot be excluded. Mosquitoes were preserved by desiccation, and total DNA was extracted from individuals using a “salting-out” protocol (Sunnucks and Hales 1996; Walton et al. 1999a). Since these species of An. dirus are essentially isomorphic, species identification was performed using an allele-specific PCR method (Walton et al. 1999a).

Amplification and Sequencing of mtDNA

A 1,230-bp fragment of the mitochondrial cytochrome oxidase I (COI) gene was amplified using An. dirus–specific primers D1664-86 (5′-GTTCCTTATA-TATTAGGACCC-3′) and D2893-70 (5′-AATAGATGAAAGATAATTGTAGG-3′). The primer names indicate the homologous positions on the An. gambiae mitochondrial genome (Beard, Hamm, and Collins 1993). The PCR reactions were performed in 50-μl volumes on a Hybaid Touchdown thermal cycler. Each reaction mix included DNA equivalent to ~1/100 of a mosquito; 1 μM each primer, 200 μM dNTP, 1.5 mM MgCl₂, 20 mM (NH₄)₂SO₄, 75 mM Tris-HCl (pH 9.0), 0.01% (w/v) Tween, and 2.5 U AmpliTaq DNA polymerase (Applied Biosystems). Initial denaturation was for 5 min at 94°C and was followed by 35 cycles of amplification (20 s at 94°C, 30 s at 45°C, and 1 min at 72°C) and a final incubation of 3 min at 72°C. The products were purified on columns (Promega) and 923 bp products were sequenced with D1908-1932 (5′-TTATTACTACTGTAAATATTACG-3′) and DCOI-2376-99 (5′-TTTTTAGTGTAGTAGCAACATTAC-3′), and two reverse primers, D2451-29 (5′-ACAAATCCAAATGCTCAAAGTAT-3′) and the reverse PCR primer, D2893-70. Sequencing reactions were performed with TaqFS dye-terminator fluorescent chemistry (Applied Biosystems) on an Applied Biosystems Model 373 automated sequencer. DNA sequences were assembled using the SeqEd multiple-sequence editor program (ABI 1992).

Analysis of mtDNA Variation

The DNA sequences were aligned using PILEUP and translated into amino acid sequences using the *Drosophila* mtDNA genetic code in TRANSLATE (Genetics Computer Group 1994). The sequence was
confirmed as that of mitochondrial cytochrome oxidase I by comparison with the *An. gambiae* sequence (Beard, Hamm, and Collins 1993). Pairwise distances between haplotypes were estimated in ARLEQUIN 1.1 (Schneider et al. 1997). Due to the low level of diversity, no corrections were made for multiple substitutions. Analysis of population genetic structure was carried out using analysis of molecular variance (AMOVA) (Beard, Hamm, and Collins 1993). Pairwise distances between haplotypes were estimated in ARLEQUIN 1.1. Slatkin and Hudson (1991) showed that the mismatch distribution of stable populations have a ragged profile due to stochastic lineage loss. In contrast, an exponentially growing population has a smooth unimodal distribution approaching a Poisson distribution. This reflects an underlying starlike genealogy in which all of the coalescent events occurred within a narrow time window.

The population parameter $\theta = 2N\mu$, where $N$ is the effective population size of females and $\mu$ is the neutral mutation rate per site per generation, was estimated by $\theta_s$ and $\theta_m$. $\theta_s$ is based on the average pairwise number of differences between sequences (Tajima 1983), and $\theta_m$ is estimated from the number of segregating sites in a population (Watterson 1975). A neutral infinite-sites model and a constant-sized population that has reached drift–mutation equilibrium are assumed for both estimators. Within-population estimates were made in ARLEQUIN 1.1. Within-species estimates and tests of neutrality (Fu 1997) were made on the web page of Y.-X. Fu (http://hgc.sph.uth.tmc.edu/cgi-bin/upblue.pl). Tests of neutrality detect deviations from the pattern of polymorphism expected from a neutral model of evolution. Among the most commonly used of these is Tajima’s (1989) test statistic, which compares the two estimators of $\theta_s$ and $\theta_m$. The other statistics used (with the exception of $F_{st}$) are similar in form but use the information in the sample differently. Fu’s (1997) $F_{st}$ statistic is negative if there is an excess of alleles over that expected for a given estimate of $\theta_m$. The program FLUCTUATE (Kuhner, Yamato, and Felsenstein 1998) was used to make simultaneous, phylogenetic estimates of present-day $\theta_s$ and the population growth rate, $g$, assuming an exponential model of growth using a maximum-likelihood approach.

### Results

**Mitochondrial Sequence Data**

The analysis is based on 923 bp of the mitochondrial cytochrome oxidase I gene sequence obtained from 84 individuals (approximately six individuals per species per site; table 1). Six individuals (three each from sites 2 and 3 in Myanmar) that were duplicates of other haplotypes at the same site were excluded in case they were related individuals, in order to make population structure estimation conservative. The *An. dirus* se-

### Table 1: Summary Data for Haplotypes Among Sites and Species

<table>
<thead>
<tr>
<th>Collection Site</th>
<th>Species</th>
<th>$N$</th>
<th>mtDNA Haplotypes</th>
<th>$S$</th>
<th>$\theta_s \pm SE$</th>
<th>$\theta_m \pm SE$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bangladesh</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Chaklapunjee</td>
<td>D</td>
<td>6</td>
<td>13, 31, 57, 67, 69, 70</td>
<td>16</td>
<td>0.00642 ± 0.00168</td>
<td>0.00759 ± 0.00160</td>
</tr>
<tr>
<td>Myanmar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Myaing</td>
<td>D</td>
<td>6*</td>
<td>10, 11, 12, 24, 34, 65</td>
<td>16</td>
<td>0.00707 ± 0.00184</td>
<td>0.00759 ± 0.00160</td>
</tr>
<tr>
<td>3. Kyauk</td>
<td>D</td>
<td>5*</td>
<td>11, 12, 17, 24, 64</td>
<td>12</td>
<td>0.00607 ± 0.00183</td>
<td>0.00624 ± 0.00155</td>
</tr>
<tr>
<td>Thailand</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Maesod</td>
<td>A</td>
<td>6</td>
<td>15, 44, 45, 48, 50, 53</td>
<td>9</td>
<td>0.00398 ± 0.00111</td>
<td>0.00427 ± 0.00096</td>
</tr>
<tr>
<td>5. Kanchanaburi</td>
<td>C</td>
<td>6</td>
<td>4, 3(5), 5(2), 6</td>
<td>3</td>
<td>0.00130 ± 0.00046</td>
<td>0.00142 ± 0.00040</td>
</tr>
<tr>
<td>6. Ratclaburi</td>
<td>A</td>
<td>5</td>
<td>23, 38, 39, 40, 59</td>
<td>8</td>
<td>0.00412 ± 0.00130</td>
<td>0.00416 ± 0.00108</td>
</tr>
<tr>
<td>7. Tung Song</td>
<td>C</td>
<td>6</td>
<td>14, 21, 36, 38, 46, 54</td>
<td>18</td>
<td>0.00679 ± 0.00177</td>
<td>0.00854 ± 0.00179</td>
</tr>
<tr>
<td>8. Loei</td>
<td>A</td>
<td>8</td>
<td>16, 27, 28, 30, 37, 41, 43, 61</td>
<td>15</td>
<td>0.00480 ± 0.00106</td>
<td>0.00543 ± 0.00095</td>
</tr>
<tr>
<td>9. Sukhon Nakhon</td>
<td>A</td>
<td>6</td>
<td>17, 25, 42, 56, 60, 62</td>
<td>10</td>
<td>0.00382 ± 0.00106</td>
<td>0.00475 ± 0.00105</td>
</tr>
<tr>
<td>10. Ubon</td>
<td>A</td>
<td>6</td>
<td>15, 19, 21, 49, 55, 60</td>
<td>12</td>
<td>0.00455 ± 0.00124</td>
<td>0.00570 ± 0.00124</td>
</tr>
<tr>
<td>11. Chantaburi</td>
<td>A</td>
<td>6</td>
<td>26, 29, 32, 52, 59, 63</td>
<td>10</td>
<td>0.00412 ± 0.00114</td>
<td>0.00475 ± 0.00105</td>
</tr>
</tbody>
</table>

**Note:**—Haplotypes in bold occur more than once. A number in parentheses indicates the frequency of a haplotype at a particular site. $S$ is the number of segregating sites.
Table 2
Variable Positions in the 923 bp of the COI Gene for the 70 Haplotypes of *Anopheles dirus*Species A, C, and D

<table>
<thead>
<tr>
<th>Position</th>
<th>Haplotype 1</th>
<th>Haplotype 2</th>
<th>Haplotype 3</th>
<th>Haplotype 4</th>
<th>Haplotype 5</th>
<th>Haplotype 6</th>
<th>Haplotype 7</th>
<th>Haplotype 8</th>
<th>Haplotype 9</th>
<th>Haplotype 10</th>
<th>Haplotype 11</th>
<th>Haplotype 12</th>
<th>Haplotype 13</th>
<th>Haplotype 14</th>
<th>Haplotype 15</th>
<th>Haplotype 16</th>
<th>Haplotype 17</th>
<th>Haplotype 18</th>
<th>Haplotype 19</th>
<th>Haplotype 20</th>
<th>Haplotype 21</th>
<th>Haplotype 22</th>
<th>Haplotype 23</th>
<th>Haplotype 24</th>
<th>Haplotype 25</th>
</tr>
</thead>
</table>

**Note.**—Dots indicate identity with the reference haplotype.

quences were very AT-rich (71.0% A and T bases). Nucleotide substitutions were identified at 74 of the 923 sites, the majority of which were transitions (80.2%). At sites 84, 327, 777, and 912, there were three segregating bases, and at site 630, all four bases were segregating. Only two substitutions resulted in amino acid changes: those at position 40 (V→M) in haplotype 34 and at position 912 (Q→H) in haplotype 19. Sequence differences among the haplotypes are shown in table 2. The reference sequence for haplotype 1 is deposited in EMBL with the accession number AJ271384.
Table 2
Continued

\[\begin{array}{cccccccc}
26 & \ldots A\ldots A\ldots C\ldots A\ldots C\ldots C\ldots C\ldots
27 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots C\ldots T\ldots
28 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots C\ldots T\ldots
29 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots T\ldots C\ldots T\ldots
30 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots G\ldots T\ldots C\ldots T\ldots
31 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots G\ldots T\ldots C\ldots T\ldots
32 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots G\ldots T\ldots C\ldots T\ldots
33 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
34 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
35 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
36 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
37 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
38 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
39 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
40 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
41 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
42 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
43 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
44 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
45 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
46 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
47 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
48 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
49 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
50 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
51 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
52 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
53 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
54 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
55 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
56 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
57 & \ldots A\ldots A\ldots C\ldots A\ldots A\ldots C\ldots T\ldots A\ldots G\ldots G\ldots C\ldots T\ldots
\end{array}\]

\(\theta_y = 0.00424 \pm 0.000152\) and \(\theta_z = 0.00484 \pm 0.00025\) for An. dirus A and \(\theta_y = 0.00637 \pm 0.00022\) and \(\theta_z = 0.00720 \pm 0.00035\) for An. dirus D.

Population Genetic Structure

The within-species \(F_{ST}\) value for An. dirus C (table 3) indicates that the two An. dirus C populations are genetically quite distinct. In contrast, in An. dirus A and D, there appears to be little genetic differentiation among the populations within each species, even over large geographic distances. In a hierarchical AMOVA, pairwise comparisons were made between four groups: species A and D each comprised a group of six populations, and the genetically distinct species C populations were treated as separate groups. The comparisons of species A or D with either of the species C popula-
Table 2
Continued

| 58 | ...G.A.A.TA...C..............C...AG.T........T...AA...C...G...T...C...T... |
| 59 | ...A.A.TA...C..............C...AG........T...AA...C...G...G...C...T... |
| 60 | ...A.A.TA...C..............C...AG........T...AA...C...G...T...C...T... |
| 61 | ...A.A.TA...C..............C...AG........T...AA...C...G...T...C...T... |
| 62 | ...A.A.TA...C..............C...AG........T...AA...C...G...T...C...T... |
| 63 | ...A.A.TA...C..............C...AG........T...AA...C...G...T...C...T... |
| 64 | ...A.A.TA...C..............C...AG........T...AA...C...G...T...C...T... |
| 65 | ...A.A.TA...C..............C...AG........T...AA...C...G...T...C...T... |
| 66 | ...A.A.TA...C..............C...AG........T...AA...C...G...T...C...T... |
| 67 | ...A.A.TA...C..............C...AG........T...AA...C...G...T...C...T... |
| 68 | ...A.A.TA...C..............C...AG........T...AA...C...G...T...C...T... |
| 69 | ...A.A.A..TC...G...A...A...C...AG........T...GA...C...G...T...C...T... |
| 70 | ...G.A.A.A..TC...G...A...A...C...AG........T...GA...C...G...T...C...T... |

Genealogical Relationships Among An. dirus Haplotypes

The minimum spanning haplotype network for An. dirus C (fig. 2A) is well resolved and implies that only one mutation has occurred twice within the network. Population 7C has two quite divergent lineages. The An. dirus C network is linked to the combined network of species A and D (fig. 2B) through one of the missing intermediate haplotypes via seven fixed differences.

The relationships of species A and D haplotypes (fig. 2B) form a starburst pattern that is highly characteristic of a population expansion. (In the sense used here, population expansion could apply equally well to a demographic expansion of the mosquitoes themselves or to a selective sweep of their mtDNA.) Based on the number of haplotypes and the level of variation, the probability of parsimonious connections (i.e., no unobserved mutations) across the network is estimated to be >97% (Templeton, Crandall, and Sing 1992). Despite this, many mutational events being inferred repeatedly and sometimes resulting in closed loops. Although there are many equally or slightly longer parsimonious alternatives, the overall starlike topology of the network remains unchanged, as does the level of homoplasy. A root to the An. dirus C network cannot be inferred with any reliability. Many lineages are represented by only a single haplotype, and many missing intermediates have been inferred, since only a fraction of the haplotypes that exist have been detected. Three noncore haplotypes, 21, 38, and 17, are shared between An. dirus A and D. There are also several lineages (those leading to haplotypes 21, 46, 24, 47, 54, and 51) that contain haplotypes from both species. There are no clear-cut geographic associations among haplotypes on the same lineages. Possible exceptions are haplotypes 38 (species A and D), 39 (A), and 40 (A) from site 5; haplotypes 11 (D) and 12 (D), both of which are found at sites 2 and 3 in Myanmar; and haplotypes 48 (A), 50 (A), and 51 (D) from site 4.

Pattern of Segregating Sites in An. dirus A and D

Anopheles dirus A and D both have smooth unimodal mismatch distributions (fig. 3), consistent with

Table 3
FST Estimates for Within- and Between-Species Comparisons

<table>
<thead>
<tr>
<th>Species (population)</th>
<th>C (5)</th>
<th>C TS (7)</th>
<th>A (all)</th>
<th>D (all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C (7)</td>
<td>0.421*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (all)</td>
<td>0.798**</td>
<td>0.741**</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td>D (all)</td>
<td>0.737**</td>
<td>0.680**</td>
<td>0.048*</td>
<td>0.022</td>
</tr>
</tbody>
</table>

Note.—Within-species FST values are on the diagonal, in bold.
* P < 0.01.
** P < 0.001.
Fig. 2.—Haplotype networks for *Anopheles dirus* species C (A) and species A and D (B). The size of a circle is proportional to the number of haplotypes observed once, twice, or three times, and a unit branch length represents one mutation. 0 = postulated haplotypes that were not observed. A, Haplotype network for population 5C (open circles) and population 7C (filled circles) of species C. The network is connected to that of species A and D through the postulated ancestral haplotype by seven fixed mutational differences. B, Haplotype network for species A (open circles) and D (filled circles). Gray circles represent haplotypes shared by *An. dirus* A and D. Numbers 1–5 along branches identify sites with homoplastic substitutions along the closed loops discussed in the text: 1—780 A/G; 2—72 A/G; 3—756 T/A; 4—528 A/G; 5—159 C/T.
population expansion. No simple statistical test can be made of a fit of the distributions to a Poisson distribution, as not only are the pairwise comparisons not independent, but the alleles share a common history. Despite this proviso, there are strong deviations from a Poisson distribution for both species A and species D (fig. 3), with the means being considerably larger than the variances. The frequency spectrum of segregating sites, equivalent to mutations in the genealogy, is shown in figure 4. There is an excess of low-frequency mutations compared with that expected for a constant-size population.

Tests of Neutrality

An excess of low-frequency mutations, as observed in figure 4, can be the result of demographic expansion, genetic hitchhiking, or background selection. Several tests of neutrality designed to detect these evolutionary forces (Fu 1997) were performed on An. dirus A and D (table 4). The effect of multiple hits was minimized by inferring the minimum number of mutations that must have occurred at sites that are segregating for more than two bases. Multiple mutations to the same base at a site have not been inferred from the parsimony network due to the uncertainty there. Since the ability to detect deviations from neutral expectations increases with sample size, the individuals of all the populations from each species were pooled. This is reasonable given that there is essentially no genetic differentiation within An. dirus A or D and that the effect of population structure, if any did exist, would be an apparent excess of older mutations (Fu 1996). However, all the test statistics reported here have negative values, indicating an excess of low-frequency mutations. Most of the values were highly significant in An. dirus D but not in An. dirus A. However, the $F_S$ test was extremely significant in both cases.

Estimates of the Effective Population Size

When there has been population growth, conventional estimates of $\theta$ ($\theta_S$ and $\theta_u$) will underestimate the current effective population size. For An. dirus A and D, where there is evidence of population growth, we therefore used the program FLUCTUATE (Kuhner, Yamato, and Felsenstein 1998) to estimate the present-day value of $\theta_0$, called here $\theta_0$. The algorithm is sensitive to the starting values of $\theta_0$, so to obtain reasonable initial estimates, several preliminary runs were made using five short chains of 400 steps and two long chains of 2,000 steps with a wide range of starting $\theta_0$ values. Three longer searches were then made for each species using 20 short chains of 400 steps and 5 long chains of 20,000 steps using initial $\theta_0 = 0.4$ for An. dirus A and initial $\theta_0 = 0.7$ for An. dirus D. Different starting genealogies were used for each long search. The median values of $\theta_0$ and the corresponding growth rates are shown in table 5. Consistently lower estimates of $\theta_0$ were obtained for An. dirus A than for species D, and the estimates of growth rates were consistently higher for An. dirus A than for An. dirus D. Although the $\theta_0$ confidence intervals of each species overlap, there is no overlap of the $g$ confidence intervals. The values for $\theta_0$ and $g$ are both very large. Consequently, the estimates of current female effective population sizes from $\theta_0$ are extremely large, approximately two to three orders of magnitude larger than those
for *An. dirus* C. Assuming a constant size for species C, the female effective population size, $N$, is $6.50 \times 10^5$ or $7.10 \times 10^5$ (from $\theta_s$ and $\theta_S$, respectively) for population 5, and $N$ is $3.32 \times 10^6$ or $2.85 \times 10^6$ (from $\theta_s$ and $\theta_S$, respectively) for population 7.

**Discussion**

**Homoplasy**

Extensive homoplasy was observed in *An. dirus* A and D and places some limitations on interpretation of the data. The homoplasy could indicate the presence of hypervariable sites, particularly given the strong AT bias. The presence of five sites that segregate at more than two bases is clear evidence of some site hypervariability. However, to account for all of the observed homoplasy by site hypervariability, each of the 39 apparently nonsingleton mutational events would have to occur on average about twice (to make 78 mutational connections in the network). It is reasonable to suppose that the closely related species *An. dirus* C would have a similar distribution of hypervariable sites, yet only 4 of its 16 observed mutations were in common with those of species A and D. Even if this low number of shared polymorphic sites were all due to hypervariable sites, rather than to retained ancestral polymorphism, this finding suggests that hypervariable sites cannot be the sole cause of homoplasy. As an example, mutations 1–4, which occur throughout the network (fig. 2B), do not segregate at more than two bases, and none of these mutations occurs in *An. dirus* C. Another potential cause

**Table 4**

**Tests of Neutrality**

<table>
<thead>
<tr>
<th>$\theta$ Values and Tests</th>
<th><em>An. dirus</em> A</th>
<th><em>An. dirus</em> D</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_S$ ...............</td>
<td>0.00805</td>
<td>0.01394</td>
</tr>
<tr>
<td>$\theta_g$ ...............</td>
<td>0.00438</td>
<td>0.00634</td>
</tr>
<tr>
<td>Tajima’s $T$ .............</td>
<td>-1.517*</td>
<td>-1.893***</td>
</tr>
<tr>
<td>Fu’s $F_s$ ................</td>
<td>-20.468****</td>
<td>-20.352****</td>
</tr>
<tr>
<td>Fu and Li’s $D$ ..........</td>
<td>-1.443</td>
<td>-2.769**</td>
</tr>
<tr>
<td>Fu and Li’s $F$ ..........</td>
<td>-1.637</td>
<td>-2.715***</td>
</tr>
<tr>
<td>Fu and Li’s $D^*$ ........</td>
<td>-1.200</td>
<td>-2.276*</td>
</tr>
<tr>
<td>Fu and Li’s $F^*$ ........</td>
<td>-1.445</td>
<td>-2.367*</td>
</tr>
</tbody>
</table>

*Note.*—The values of $\theta_S$ and $\theta_g$ used in the tests (calculated inferring additional mutations at sites segregating for more than two bases) are also shown.

* $P < 0.1$.
** $P < 0.02$.
*** $P < 0.01$.
**** $P < 0.001$.

**Table 5**

**FLUCTUATE Estimates of $\theta_0$ and the Exponential Growth Rate**

<table>
<thead>
<tr>
<th>Species</th>
<th>$\theta_0$</th>
<th>$\theta_0 - 95%$ CI</th>
<th>$N_0$</th>
<th>$g$</th>
<th>$g - 95%$ CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Anopheles dirus</em> A</td>
<td>0.412</td>
<td>0.260–0.564</td>
<td>$2.06 \times 10^4$</td>
<td>2.324</td>
<td>2.068–2.580</td>
</tr>
<tr>
<td><em>An. dirus</em> D</td>
<td>0.757</td>
<td>0.377–1.137</td>
<td>$5.00 \times 10^4$</td>
<td>1.734</td>
<td>1.497–1.972</td>
</tr>
</tbody>
</table>

*Note.*—Point estimates of $\theta_0$ and $g$ are shown with approximate 95% confidence intervals (CIs). The exponential growth rate, $g$, is in units of $1/\mu$ generations. $N_0$, the current effective female population size, is estimated from $\theta_0$. 

**Fig. 4.**—Frequency spectrum for the numbers of segregating sites in *An. dirus* A and D. Diamonds and triangles show the number of expected sites in each frequency class under a model of constant population size (see Harpending et al. 1998). There is uncertainty in the ancestral state at a site; for example, a mutation that occurred early in the genealogy and currently occupies all but one site cannot be distinguished from a mutation that has occurred recently and occupies only one site. The expected distribution was therefore folded at one half the sample size. The resulting number was multiplied by the function $1 - (x^{n-1} - 1)$ (where $x$ = frequency of mutant sites) to remove ascertainment bias due to the small sample size relative to the number of alleles likely to be present in the population (Harpending et al. 1998). Due to the interdependence of sites, there is no simple test of the constant-size hypothesis.
of homoplasy is recombination. Although mitochondrial sequences are usually considered strictly maternally inherited and undergo no recombination, evidence has been put forward recently of mtDNA recombination in humans (Eyre-Walker, Smith, and Maynard Smith 1999). A minimum of three recombination events would be required to account for the closed loops in the network (fig. 2B).

Population Expansion in An. dirus A and D

There was considerable deviation from a Poisson fit in the mismatch distributions (fig. 3) due, at least in part, to homoplasy. Nevertheless, the smooth, unimodal distributions are consistent with some sort of population expansion, as simulations have shown that stable populations almost never produce this type of profile (Harpending 1994). Nor can such a profile be the result of pooling data from structured subpopulations (Rogers et al. 1996). This expansion can be visualized in the starlike haplotype network (fig. 2B). The six core haplotypes (56, 55, 31, 60, 30, and 59) that have many mutational connections, tend to be at a slightly higher frequency, and come from a wide geographical area are probably ancestral (Crandall and Templeton 1993).

An excess of low-frequency mutations relative to that expected in a population of constant size, particularly in An. dirus D (fig. 4), was detected by the tests of neutrality. Mutations present at one or a few copies are likely to be of recent origin, whereas those that occur early in the genealogy are likely to be present at higher frequencies. Fu (1997) has shown that his $F_s$ test is particularly powerful at detecting population growth and hitchhiking, whereas the $D$ and $F$ tests of Fu and Li (1993), which essentially test for an excess of singleton versus nonsingleton mutations, are more powerful than $F_s$ and Tajima’s (1989) test at detecting background selection. The highly significant $F_s$ statistics for An. dirus A and D suggest that population expansion (demographic or selective sweep) is more likely to be the cause of deviation from neutral expectations than background selection. However, high $F_s$ values could also indicate the presence of recombination, as this would increase the number of alleles relative to the number of pairwise differences between sequences.

The estimates of exponential growth rates, $g$, from the FLUCTUATE program also indicate a large population expansion. However, exponential growth is only one possible mode of growth, and there are many others that could have produced the observed unimodal mismatch distributions (Slatkin and Hudson 1991). The presence of sites 72, 528, 756, and 780 segregating at high frequencies (13–24 out of 72 sequences) in species A and D is not consistent with a simple model of exponential growth, as this is unlikely to result in high-frequency mutations. This implies that the corresponding mutations at these sites (numbered 1–4 in fig. 2B) existed in the population prior to expansion. This is consistent with these older mutations linking together core haplotypes. If a smaller, more stable population existed for some time before expansion, rare recombination would become less improbable as an explanation for homoplasy involving these sites. Expansion from a population that was sufficiently large to contain some variation already could also contribute to the deviation from the Poisson fit in the mismatch distributions. In estimating $\theta_0$ and $g$, FLUCTUATE assumes a constant rate of exponential growth over the entire length of the genealogy. The fact that An. dirus D has, if anything, a higher value of $\theta_0$ but a lower growth rate than species A most likely indicates a longer genealogy for species D in which population expansion occurred earlier, enabling greater diversity to accumulate in species D, rather than implying a higher growth rate for species A.

The time to coalescence in an exponentially expanding population is $t/2u$ (Slatkin and Hudson 1991), where $i$ is the average number of pairwise differences between sequences and $u$ is the per-sequence mutation rate per generation. Omitting the sites with mutations 1–4 that are likely to have been present before a rapid growth phase, this relation gives the start of population expansion very approximately as 135,000 years ago in An. dirus A and 250,000 years ago in An. dirus D (assuming 10 generations per year and a mutation rate of $10^{-8}$ per site per year; Powell et al. 1986). The human population may also have increased in size, and estimates from mtDNA data put the start of expansion at 60,000–120,000 years ago (Rogers and Harpending 1992) or 262,500 years ago (Slatkin and Hudson 1991). If the mosquito expansions are due to demographic expansion rather than a selective sweep, they may have occurred at approximately the same time as the human population expansion. This could be due to a habitat change that allowed both human and mosquito populations to expand, or, alternatively, the expansion of the human population could have resulted in a habitat change for the mosquitoes that allowed them to expand.

Gene Flow and Population History

When gene flow is inferred indirectly from $F$ statistics, the assumption of equilibrium between drift and migration is often reasonable, since migration is a relatively fast homogenizing force. However, the genetic signature of a population expansion can remain for a long time (Rogers and Harpending 1992), obscuring true ecological population structure that may exist. For An. dirus A and D, there is compelling evidence of population expansion. In addition, since the rate of approach to equilibrium is dependent on effective population size (Nei and Feldman 1972), it would take an extremely long time for the very large population sizes of An. dirus A and D to reach migration-drift equilibrium. It would therefore be misleading to infer migration rates from the $F_{ST}$ values.

In contrast to An. dirus A and D, the two C populations show considerable population structure. The fixed difference between the two species C populations indicates that the population 5 sample has a distinct genealogy in which all of the alleles coalesce more recently than the time of any connection they have to alleles in the other population. Hey (1991) has shown that
for these sample sizes even a single fixed difference is significant and implies an absence of gene flow between these two populations. An $F_{ST}$ value of 0.421 therefore probably reflects a recent shared ancestry rather than indicating low levels of gene flow. The genetic structure evident within and between populations of An. dirus C indicates that this species has a population history very different from that of species A and D. This may be due to the patchy distribution of An. dirus C, making it more likely to have been influenced by colonization events. If species A and D have undergone demographic expansion, this is more likely to have been due to contiguous range expansion.

Genetic Introgression

The genetic similarity between An. dirus A and D, along with their recent, common genealogy, was completely contrary to the expectation of a sister relationship between species A and C, with species D being more distant. This could be explained by the retention of ancestral polymorphism between species A and D if our current understanding of the species relationships is incorrect. However, in addition to evidence already cited (cytological, hybrid fitness, and allozyme) to support a close An. dirus A and C relationship, sequence data from nuclear loci also concur with this (unpublished data). An alternative explanation of these observations is that mtDNA has introgressed from An. dirus D into species A. This is also consistent with laboratory cross-mating studies in which female hybrids were formed from species A (male) × species D (female) crosses but not from the reverse pairing of the same number of individuals (Baimai et al. 1987). The low fitness of these hybrids suggests that any route for interspecific gene flow must be severely restricted.

Since there are no highly divergent lineages of what could be original species A mtDNA within An. dirus A, this implies that species D mtDNA has replaced all of the species A mtDNA. It is unlikely that all of the An. dirus D haplotypes would be passed into species A, and this effective genetic bottlenecking could have resulted in the lower diversity observed in An. dirus A. Given the restricted route to gene flow, a selective sweep of mtDNA that originated from An. dirus D seems more plausible than demographic expansion. This is also in accord with the indication of a population expansion occurring in An. dirus D before species A. However, the earlier suggestion that the population expansion in An. dirus D, as well as that in An. dirus A, may have stemmed from more than one haplotype would seem to counter a selective sweep. An alternative argument is that introgression could have occurred when An. dirus A had a small population size, such that drift would have a strong effect, and shared a restricted distribution with species D, followed by independent range expansion.

Explanations of low between-species differentiation in terms of ongoing, rather than historical, introgression seem highly implausible. First, the scale of hybridization events would have to be very large to overcome the low hybrid fitness. Second, opportunities for introgression would be limited to the narrow area of overlap of species A and D (coupled with extensive within-species gene flow) and/or arise as a result of long-distance migration. Third, despite their apparent intermingling, there does seem to be some separation between species A and D haplotypes in the network, suggestive of independent recent evolutionary histories: lineages arising from core haplotypes 30 and 55 are composed entirely of species D haplotypes, and 9 of the 10 lineages arising from core haplotype 31 are found in species A.

Comparison with the African Malaria Vectors Anopheles gambiae and Anopheles arabiensis

Findings strikingly similar to those reported here for An. dirus A and D have been reported in a similar mtDNA study on a pair of closely related mosquito species, An. gambiae and An. arabiensis, from east, west, and south Africa (Besansky et al. 1997). The very low levels of genetic differentiation within and between these species were interpreted in terms of both extensive gene flow and recent range expansion from relatively large stable populations. It was suggested that it is at least theoretically possible that equilibrium exists between gene flow and genetic drift across Africa within each species. Lehmann et al. (1996) also suggested that intraspecific gene flow on a continental scale is plausible via the considerable active dispersal ability of An. gambiae and long-distance migration facilitated by humans. However, given the evidence for population expansion (Besansky et al. 1997) and the large effective population sizes of these mosquitoes (Lehmann et al. 1998), migration-drift equilibrium seems to us unlikely. As in An. dirus A and D, this would preclude the inference of gene flow from population structure.

The mtDNA data of An. gambiae and An. arabiensis were also interpreted as suggesting extensive interspecific gene flow. Anopheles gambiae and An. arabiensis differ from An. dirus A and D in that they are largely sympatric across sub-Saharan Africa. In addition, their female hybrids are fertile (Coluzzi et al. 1979). Extensive ongoing interspecific gene flow is therefore more feasible between An. gambiae and An. arabiensis. In the starlike network of the An. gambiae and An. arabiensis haplotypes, haplotypes of one species connect to apparently younger haplotypes in another species that could imply interspecific gene flow (Besansky et al. 1997). However, as in An. dirus A and D, homoplasy is also evident in the An. gambiae data with many alternative connections in the network, so it is difficult to make any convincing inferences from the relationships between particular haplotypes.

In conclusion, historical introgression appears to be sufficient to explain the observations in An. dirus A and D without the need to infer large-scale contemporary interspecific gene flow and, indeed, some reason to think that it is unlikely. Ongoing extensive interspecific gene flow is plausible in An. gambiae and An. arabiensis but unlikely in An. dirus. The similarity of the two mtDNA
data sets therefore suggests that the hypothesis of only historical introgression (without ongoing introgression) between species cannot as yet be rejected in either case. Even within species, the genetic data for An. dirus A and D, and, by analogy, An. gambiae and An. arabensis, seem to be adequately explained by historical population processes without any need to infer extensive gene flow. Even if they have not reached migration-drift equilibrium, information on gene flow may be contained within the genealogical relationships of the haplotypes. However, before a phylogenetic approach can be used, much greater certainty is required with regard to the relationships between haplotypes. The problem of homoplasy encountered here is familiar in the analysis of human mtDNA, in which there is both site hypervariability (Wakeley 1993) and possibly also recombination (Eyre-Walker, Smith, and Maynard Smith 1999). In addition to obtaining more mtDNA sequences to gain greater certainty about the haplotype relationships, it is clearly essential to obtain data from the same populations on nuclear loci to be able to distinguish between the hypotheses of a selective sweep and demographic expansion. Achieving a greater understanding of the evolutionary history of these mosquitoes is important if reliable estimates of gene flow, which are critical to the control of malaria, are to be made in the future.

Acknowledgments

The mosquito collections on which this work was based would not have been possible without the helpful support of the staff of the malaria centers of Thailand and Mr. Sawadwongporn and Mr. Somsak. We would also like to thank Professor Sornchai Looareesuwan for the samples from Ratchaburi in Thailand, N. P. Maheswary for the samples from Bangladesh, Brian Holloway and the staff of the Biotechnology Core Facility at the Centers for Disease Control and Prevention in Atlanta, Ga., for synthesizing oligonucleotide primers, and Alan Templeton for providing his Mathematica package. This work was funded by a Wellcome Trust (London) fellowship in Biodiversity to C. Walton with the sponsorship of R. K. Butlin.

LITERATURE CITED


Population Structure of Anopheles dirus 973


ELEFTHERIOS ZOULROS, reviewing editor

Accepted March 1, 2000