Population Structure Among African and Derived Populations of *Drosophila simulans*: Evidence for Ancient Subdivision and Recent Admixture

Martha T. Hamblin and Michel Veuille

Laboratoire d'Ecologie-EPHE, Université Pierre-et-Marie Curie, 75252 Paris Cedex 05, France

Manuscript received November 11, 1998
Accepted for publication May 24, 1999

ABSTRACT

Previous studies based on allozyme variation have found little evidence for genetic differentiation in *Drosophila simulans*. On the basis of DNA sequence variation at two nuclear loci in four African populations of *D. simulans*, we show that there is significant structure to *D. simulans* populations within Africa. Variation at one of the loci, vermilion, appears to be neutral and supports an eastern African origin for European and American populations. Samples from the West Indies, Europe, and North America had a nucleotide diversity lower than that of African populations at vermilion and show nonequilibrium haplotype distributions at both vermilion and G6pd, consistent with a hypothesis of recent bottleneck and possibly also admixture in the history of these populations. Directional selection, previously documented at G6pd, appears to have occurred within the coalescence time of the species, obscuring deep population history.

The closely related species *Drosophila melanogaster* and *D. simulans* are widely used in studies of evolution at the phenotypic, genetic, and molecular levels. In spite of many broad similarities, there are important differences between these species, the causes of which are not well understood (Parsons 1975). Both species evolved in tropical Africa and have spread worldwide as human commensals in historical times (Lachaise et al. 1988), although *D. melanogaster* has apparently inhabited temperate regions for sufficient time to develop significant genetic differentiation between populations from different continents (Choudhary and Singh 1987; Begun and Aquadro 1993; Benassi and Veuille 1995). In contrast, little population structure has been detected in *D. simulans* as assessed by studies of phenotypic variation, allozymes, and mtDNA (Hytyia et al. 1985). However, a recent study of microsatellite variation among one African and three non-African populations of *D. simulans* (Irvin et al. 1998) revealed more genetic differentiation between populations of *D. simulans* than had been observed using allozymes.

A discordance between DNA-level and protein-level evolution is also observed in comparisons between these species: while allozyme variation is slightly higher in *D. melanogaster* (Choudhary and Singh 1987), DNA sequence variation is two to three times higher in *D. simulans* (Aquadro et al. 1988; Moriyama and Powell 1996). Similarly, the ratio of replacement to silent DNA variation is significantly higher for *D. melanogaster* than *D. simulans* (Moriyama and Powell 1996). This genome-wide discrepancy between protein and (presumably largely neutral) DNA evolution implies that selection on protein variation has been different in these species. Patterns of codon usage also vary between the species, with a higher proportion of preferred codons having been fixed in the *D. simulans* lineage (Akashi 1996).

Several nonexclusive hypotheses have been proposed to explain these different patterns of molecular variation: *D. simulans* has a different “adaptive strategy” than *D. melanogaster* and/or has become a cosmopolitan species more recently than *D. melanogaster* (Choudhary and Singh 1987); *D. simulans* has a larger effective population size than *D. melanogaster*, leading to stronger effects of weak selection (Aquadro et al. 1988; Akashi 1996). Testing of these hypotheses should also consider possible differences in population history and structure between the species. Recent demographic events, in particular, are expected to affect genome-wide patterns of variation and may cause departures from the predictions of models assuming that populations are at equilibrium with respect to migration, mutation, and drift. Recovery from such events is expected to take on the order of $4N_e$ (where $N_e$ is effective population size) generations, much longer than the few hundred years since the establishment of many non-African populations of *D. melanogaster* and *D. simulans*.

Progress in elucidating population history and structure for these species has been hampered by a dearth of studies of DNA sequence variation from true population samples representing geographically diverse populations, especially African populations that are potentially ancestral. This is particularly true for *D. simulans*, for which DNA sequence variation has been examined almost exclusively in North American and European population samples or in worldwide collections of alleles that are inappropriate for many population genetic tests. There is only one published report of DNA se-
quence variation in more than one population sample of D. simulans, a study of the vermilion locus by Begun and Aquadro (1995), which showed significant differentiation (\(F_{ST} = 0.25\)) at silent sites between a North American and Central African population. Furthermore, most of the sequence variants were not shared between populations, implying that the Central African population was not ancestral to the North American one. This, in turn, implies that African populations of D. simulans are genetically differentiated if one assumes that some other African population is ancestral.

The North American vermilion dataset showed a strong haplotype structure (i.e., alleles fell into very divergent classes for which few intermediate haplotypes were observed), a pattern that has also been observed at a number of other unlinked loci in North American samples of D. simulans alleles (Begun and Aquadro 1994; V. L. Bauer and C. F. Aquadro, personal communication). Such a pattern might result from mixing of alleles from genetically differentiated founder populations in the recent past. The presence of highly divergent, geographically restricted mitochondrial DNA (mtDNA) lineages in this species (Solignac and Monnerot 1986; Sat t a and Takahata 1990) lends plausibility to this hypothesis. If admixture has occurred, the observed level of variation might not be an accurate indicator of the long-term effective population size of the ancestral populations. Because differences in effective population size have frequently been invoked to explain differences in patterns of molecular variation between D. melanogaster and D. simulans from derived populations (Aquadro et al. 1988; Ohta 1992; Akashi 1996), the question of population admixture needs to be resolved.

The goal of this study was to use nuclear DNA sequence data to assess the level of population differentiation of African and non-African populations of D. simulans and to begin to reconstruct their evolutionary histories. In particular, we wanted to test the hypothesis that divergent alleles from derived populations reflect population structure in the ancestral populations from which they were founded. The possibility of population admixture has been raised before to explain divergent lineages in worldwide collections of alleles of D. simulans (e.g., Hasson et al. 1998), but this question can be directly addressed only by studying true population samples. We surveyed two unlinked regions on the X chromosome for which some sequence polymorphism data were already available: vermilion (Begun and Aquadro 1995) and glucose-6-phosphate dehydrogenase (G6pd; Eanes et al. 1996). For each locus, we sequenced a relatively short region (~700 bp) containing a reasonable number of segregating sites and showing the pattern of divergent haplotypes described above.

**MATERIALS AND METHODS**

**Population samples:** Table 1 shows the dates and locations of the collections of D. simulans used in this study. Samples were obtained using attractive baits. Wild-collected flies were used to establish isofemale lines (Kenya, Tanzania, Antilles, Zimbabwe) or extract chromosomes using attached-X lines (Cameroon) or were frozen immediately in the laboratory (Italy). Flies from isofemale lines were frozen within 3 mo after trapping, except for Kenya (8 mo) and Antilles (1 yr). For vermilion, we included the sequences from Raleigh, North Carolina (United States), published by Begun and Aquadro (1995), because this sample comes from a geographic region not represented in our data.

**DNA methods:** DNA was prepared from single male flies by the method of Gillor et al. (1993). An 809-bp product was amplified from the vermilion locus using primers corresponding to bases 602–622 (forward) and 1410–1390 (reverse) of GenBank accession no. U27204. A 769-bp product was amplified from the G6pd locus using primers corresponding to bases 917–939 (forward) and 1685–1664 (reverse) of GenBank accession no. L13876. PCR products were separated on 1.2% agarose gels and the desired bands were cut out and purified using the QiaexII kit (Qiagen, Valencia, CA). DNAs were sequenced manually using the Thermosequenase kit (Amersham, Arlington Heights, IL). The PCR primers were used as sequencing primers, with the exception of the reverse sequencing primer for vermilion, which was an internal primer corresponding to bases 1376–1357. DNAs were sequenced on one strand except for a small area of overlap in the middle of the fragment. GenBank accession numbers for the vermilion sequences are AF149122–149191; accession numbers for the G6pd sequences are AF148146–148207.

**Data analysis:** The program DnaSP (Rozas and Rozas 1997) was used to obtain summary statistics of sequence polymorphism within populations, Tajima’s D (Tajima 1989), and divergence between populations. The fixation index (\(F_{ST}\)) was calculated according to Hudson et al. (1992b) and the probability of panmixis was determined using the method of Hudson et al. (1992a).

Tests of haplotype number and haplotype diversity were conducted using the method of Depaulis and Veuille

<table>
<thead>
<tr>
<th>Location</th>
<th>Date of collection</th>
<th>Collected by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yaounde, Cameroon</td>
<td>December 1997</td>
<td>B. Riera</td>
</tr>
<tr>
<td>Nairobi, Kenya</td>
<td>September 1995</td>
<td>C. Wilson</td>
</tr>
<tr>
<td>Mt. Kilimanjaro, Tanzania</td>
<td>April 1996</td>
<td>D. Lachaise</td>
</tr>
<tr>
<td>Harare, Zimbabwe</td>
<td>February 1997</td>
<td>D. Lachaise</td>
</tr>
<tr>
<td>St. Martin, Lesser Antilles</td>
<td>March 1995</td>
<td>J. David</td>
</tr>
<tr>
<td>Sticiano, Italy</td>
<td>August 1996</td>
<td>C. Montchamp-Moreau</td>
</tr>
</tbody>
</table>
Population History in *D. simulans*

**Figure 1.**—Polymorphism at *vermilion* in *D. simulans*. The numbering system is the same as in Begun and Aquadro (1995), based on the sequence of *D. melanogaster* as reported in GenBank accession no. M34147. cons, the most common base at each segregating site; i, intron; s, synonymous site.

(1998). For G6pd, a recombination rate of $N_r/bp = 0.005$ was used; for *vermilion*, $N_r/bp = 0.01$. Note that for *vermilion*, estimates of $r$ based on genetic data from *D. melanogaster* are $2 \times 10^{-5}/bp$ (Searles et al. 1990) and that if $N = 10^{-5}$, $N_r = 2$. (Rates of recombination in this region of the X chromosome are very similar in *D. simulans* and *D. melanogaster*; Sturtevant 1929.) Limitations of computational power prevented us from running the simulations under higher values of $N_r$; our $P$ values therefore are conservative. Tests were one-tailed because the hypothesis predicted a departure in the direction of too few haplotypes.

**RESULTS**

**DNA sequence variation:** For estimation of historical patterns of gene flow between populations of *D. simulans*, we collected DNA sequence data from two unlinked loci on the X chromosome in samples from four continental African, one West Indian, and one European population. We determined the sequence of 728 bp of the *vermilion* gene in 10–13 individuals from each of the six populations. This 728-bp region spans three small introns, which total 186 bp in length. Haplotypes at *vermilion*, including the corresponding *vermilion* data for the United States (Begun and Aquadro 1995), are shown in Figure 1.

We also determined the sequence of 700 bp of the third exon of the G6pd gene in 10–12 individuals from each of the four African populations. Due to problems with amplification in the Italian lines, we have data from only five sequences from Italy, which we combined with the four sequences from France published by Eanes.
et al. (1996) to create a European sample. The four haplotypes from Italy are very similar to the three haplotypes from France, suggesting that the sample has not been biased by the amplification problems and that it is appropriate to combine the samples. Haplotypes at G6pd are shown in Figure 2.

Summary statistics for both loci are presented in Table 2. Note that these data are not appropriate for comparing variation between loci nor are they appropriate for making inferences about absolute effective population size, because the regions sequenced were not chosen at random; the goal was to understand population structure, not to estimate $4N_e\mu$. Nevertheless, these statistics are appropriate for use in comparing variation between populations at a particular locus. At vermilion, the samples from African populations, particularly Tanzania and Kenya, are the most variable. This pattern is consistent with the hypothesis of Lachaise et al. (1988) that
Genetic differentiation between populations: We tested the null hypothesis that our population samples were drawn from a single panmictic population using the method of Hudson et al. (1992b). Estimates of $F_{ST}$ were calculated according to Hudson et al. (1992a). In cases where the sample size differed between populations, the average pairwise difference was weighted by the sample size. The significance of $F_{ST}$ was determined by resampling the data 1000 times. Populations were tested in all pairwise combinations, and data from the two loci were analyzed separately (Table 3). Because multiple nonindependent tests were performed, we have not chosen a significance threshold for this analysis. Instead, we make qualitative comparisons of $P$ values.

At the vermilion locus, $F_{ST}$ is essentially zero between the samples from Tanzania and Kenya, which were collected only 200 miles apart. The other distances, and most of the other $F_{ST}$'s, are much larger: 13 of the 15 comparisons involving Antilles, Zimbabwe, and Cameroon have $P$ values $\leq 0.01$. $P$ values for $F_{ST}$'s involving Tanzania, the United States, and Italy are generally low (0.01 $< P < 0.1$) but probably not significant. Thus, our samples seem to fall into four genetically distinct groups: Zimbabwe, Cameroon, Lesser Antilles, and Tanzania/Kenya/Italy/United States. The Antilles sample, dominated by seven copies of a unique haplotype, is quite different from all other populations. Again, the relationships between populations at the $G_{6pd}$ locus are different from those at vermilion: Cameroon appears
to be differentiated ($P < 0.005$) from all other populations except Tanzania, but most other comparisons have much larger $P$ values.

**Haplotype tests:** Haplotypic diversity is quite high; at vermilion, there are 41 different haplotypes, of which 27 occur only once in the total of 82 sequences. At G6pd, there are 28 haplotypes, 16 of which are unique, in a total of 66 sequences. However, inspection of the data in Figures 1 and 2 reveals that all the samples that do not come from Africa contain multiple copies of a few highly divergent haplotypes: Italy/Europe and Antilles for both loci and the U.S. sample of vermilion. One prediction of a hypothesis of recent population admixture is that the number of haplotypes will be smaller than expected in a population at mutation-drift equilibrium, given the observed number of segregating sites, similar to the pattern produced by an old balanced polymorphism. We tested the observed number of haplotypes in each of our samples against the expectation of a neutral, equilibrium model, conditioned on the number of segregating sites ($S$) and the number of sequences surveyed ($n$) (Depaulis and Veuille 1998). Recombination was included in the model (see materials and methods), because both loci in our study experience significant amounts of recombination. All the samples from Italy/Europe, Antilles, and the United States show a significant departure from the expectation for at least one of the tests (Table 4), while none of the tests of the samples from Africa is significant. The $P$ values for the vermilion sample from Zimbabwe, however, are much lower than those for the three other African populations.

The significance of the haplotype tests for the vermilion data may be somewhat overestimated, because we chose to survey the 5' half of the gene because of its higher level of linkage disequilibrium. The entire region surveyed by Begun and Aquadro (1995) does not contain significantly too few haplotypes, although the haplotype diversity is still significantly lower than expected (Table 3). However, it should be noted that the rate of

**TABLE 3**

<table>
<thead>
<tr>
<th></th>
<th>Antilles</th>
<th>Kenya</th>
<th>Tanzania</th>
<th>Zimbabwe</th>
<th>Cameroon</th>
<th>Italy/Europe</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antilles</td>
<td>—</td>
<td>0.134</td>
<td>0.163</td>
<td>0.286</td>
<td>0.231</td>
<td>0.199</td>
<td>0.238</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.002)</td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.010)</td>
<td>(0.000)</td>
<td></td>
</tr>
<tr>
<td>Kenya</td>
<td>0.1857</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.107</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td></td>
<td>(0.854)</td>
<td>(0.001)</td>
<td>(0.006)</td>
<td>(0.050)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>Tanzania</td>
<td>0.1516</td>
<td>0</td>
<td>—</td>
<td>0.033</td>
<td>0.178</td>
<td>0.093</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td>(0.057)</td>
<td>(0.518)</td>
<td>(0.007)</td>
<td>(0.001)</td>
<td>(0.014)</td>
<td>(0.110)</td>
<td>(0.072)</td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>0.037</td>
<td>0.021</td>
<td>0.041</td>
<td>—</td>
<td>0.159</td>
<td>0.162</td>
<td>0.130</td>
</tr>
<tr>
<td></td>
<td>(0.220)</td>
<td>(0.297)</td>
<td>(0.167)</td>
<td>(0.014)</td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>Cameroon</td>
<td>0.204</td>
<td>0.182</td>
<td>0.015</td>
<td>0.179</td>
<td>—</td>
<td>0.178</td>
<td>0.139</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.001)</td>
<td>(0.343)</td>
<td>(0.004)</td>
<td>(0.001)</td>
<td>(0.003)</td>
<td></td>
</tr>
<tr>
<td>Italy/Europe</td>
<td>0.077</td>
<td>0.269</td>
<td>0.175</td>
<td>0.127</td>
<td>0.188</td>
<td>—</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>(0.147)</td>
<td>(0.003)</td>
<td>(0.025)</td>
<td>(0.064)</td>
<td>(0.002)</td>
<td>(0.181)</td>
<td></td>
</tr>
</tbody>
</table>

$F_{ST}$ statistics and $P$ values (in parentheses) calculated according to Hudson et al. (1992a,b). Statistics for vermilion are above the diagonal; statistics for G6pd are below the diagonal.
TABLE 4
Tests of haplotype number and haplotype diversity

<table>
<thead>
<tr>
<th>Locus/ population</th>
<th>n</th>
<th>S</th>
<th>K</th>
<th>P value</th>
<th>H</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>vermilion/ Tanzania</td>
<td>11</td>
<td>39</td>
<td>11</td>
<td>1.000</td>
<td>0.909</td>
<td>1.000</td>
</tr>
<tr>
<td>vermilion/ Kenya</td>
<td>13</td>
<td>37</td>
<td>12</td>
<td>0.906</td>
<td>0.911</td>
<td>0.906</td>
</tr>
<tr>
<td>vermilion/ Zimbabw</td>
<td>10</td>
<td>31</td>
<td>7</td>
<td>0.118</td>
<td>0.820</td>
<td>0.089</td>
</tr>
<tr>
<td>vermilion/ Cameroon</td>
<td>12</td>
<td>32</td>
<td>10</td>
<td>0.566</td>
<td>0.889</td>
<td>0.566</td>
</tr>
<tr>
<td>vermilion/ Italy</td>
<td>12</td>
<td>25</td>
<td>5</td>
<td>&lt;0.001</td>
<td>0.667</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>vermilion/ Antilles</td>
<td>12</td>
<td>21</td>
<td>5</td>
<td>&lt;0.001</td>
<td>0.611</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>vermilion/ U.S.(^a)</td>
<td>12</td>
<td>40</td>
<td>9</td>
<td>0.195</td>
<td>0.833</td>
<td>0.037</td>
</tr>
<tr>
<td>G6pd/ Tanzania</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>0.910</td>
<td>0.840</td>
<td>0.933</td>
</tr>
<tr>
<td>G6pd/ Kenya</td>
<td>12</td>
<td>10</td>
<td>9</td>
<td>0.955</td>
<td>0.898</td>
<td>0.861</td>
</tr>
<tr>
<td>G6pd/ Zimbabwe</td>
<td>11</td>
<td>13</td>
<td>9</td>
<td>0.927</td>
<td>0.876</td>
<td>0.927</td>
</tr>
<tr>
<td>G6pd/ Cameroon</td>
<td>12</td>
<td>10</td>
<td>10</td>
<td>0.993</td>
<td>0.889</td>
<td>0.993</td>
</tr>
<tr>
<td>G6pd/ Europe</td>
<td>9</td>
<td>10</td>
<td>4</td>
<td>0.044</td>
<td>0.741</td>
<td>0.213</td>
</tr>
<tr>
<td>G6pd/ Antilles</td>
<td>12</td>
<td>10</td>
<td>3</td>
<td>&lt;0.001</td>
<td>0.542</td>
<td>0.008</td>
</tr>
<tr>
<td>Pgd/ U.S.(^b)</td>
<td>19</td>
<td>11</td>
<td>3</td>
<td>0.016</td>
<td>0.526</td>
<td>0.042</td>
</tr>
<tr>
<td>gld/ U.S.(^c)</td>
<td>11</td>
<td>26</td>
<td>10</td>
<td>0.982</td>
<td>0.893</td>
<td>0.982</td>
</tr>
</tbody>
</table>

Tests were done according to Depaulis and Veuille (1998). \(K\), haplotype number; \(H\), haplotype diversity. \(P\) values for both the \(K\)-test and \(H\)-test are one tailed (see Materials and Methods).

\(^a\) Based on sites 1204–1993.
\(^b\) Based on sites 1204–2587 (Begun and Aquadro 1995).
\(^c\) From Begun and Aquadro (1994).

recombination used in the simulations was considerably lower than the rate of recombination estimated for the vermilion region (by ~100-fold if we accept the estimate of 2 \(\times\) 10\(^{-2}\)/bp from Searles et al. 1990) due to limited computational ability. A more realistic recombination rate would have decreased the \(P\) value by an unknown amount. At G6pd, it is unlikely that including the entire gene would have changed the outcome of the tests, because there is little variation in the remaining sequence, and linkage disequilibrium was high throughout the gene (Eanes et al. 1996). Note also that the significant result for the European sample is not due to the combining of alleles from Italy and France; the divergent haplotypes are not associated with different locations.

Because the hypothesis of population admixture predicts a similar pattern throughout the genome, we conducted the same tests of haplotype number and diversity for all the other surveys of DNA polymorphism that we could find in the literature for population samples of D. simulans; unfortunately there are very few (Table 4). A four-cutter restriction site survey of the phosphogluconate dehydrogenase region (Pgd) from North Carolina (Begun and Aquadro 1994) gave very significant results; a sequence survey of the glucose dehydrogenase (Gld) region from the same population did not, although this sample did have a significantly positive Fu and Li’s \(D\) statistic due to a complete lack of singletons (Hamblin and Aquadro 1996). Surveys of the yellow-achete region from Europe and the United States (Martín-Campos et al. 1992; Begun and Aquadro 1991) as well as the suppressor of forked region (Langley et al. 1993) had too little variation to be informative.

Phylogeny of G6pd alleles: Because the \(F_{ST}\) analysis pools all alleles in a sample to produce one statistic of distance, it is a poor reflection of the complexity of these particular samples, and the relationships of these populations are problematic (see discussion). We therefore wanted to estimate phylogenies of alleles using parsimony to see where these divergent alleles arose with respect to the potentially ancestral African populations. Because both loci are from recombining regions of the nuclear genome, simple bifurcating phylogenies cannot be reconstructed.

For G6pd, it was possible to construct a network (Figure 3) using all 22 segregating sites and all but two haplotypes (Cameroon 3 and 12). We also included the alleles from the United States and Mexico surveyed by Eanes et al. (1996). There are two haplotypes (the consensus or “dot” and the Tanzania 9-type), separated by two steps, that are found in all three eastern African populations. Many of the Cameroon alleles are located between these two haplotypes in the middle of the network. Most of the Kenyan alleles are more closely related to the dot haplotype. There are two clusters of alleles from the derived populations at opposite ends of the network, separated by between five and nine steps. Both clusters are associated with alleles from Zimbabwe, so the network provides no evidence that they have different geographic origins.
DISCUSSION

This study, the first to compare African populations of D. simulans at the DNA sequence level, has revealed substantial genetic differentiation among those populations at the vermilion locus. These findings contrast with patterns of allozyme and mitochondrial DNA variation, which show little geographic structure. Haplotype structure in the non-African population samples at both vermilion and G6pd departs significantly from the expectation under a neutral, equilibrium model, suggesting a history of bottleneck (i.e., founder effect) and possible admixture in these recently established populations. At the vermilion locus, the three non-African population samples also have lower genetic diversity (average \( \theta = 7.73 \pm 0.692 \)) than the four African samples (average \( \theta = 12.03 \pm 2.656 \)), consistent with the inference of a founder effect.

Variation at the G6pd locus is not lower in the non-African samples, and there is much less evidence for population structure among all the samples. If both loci were neutral indicators of population sizes and structure, we would expect relative distances and diversity among populations at the two loci to be the same. While such a discrepancy might be due to chance, we believe that it is due primarily to past episode(s) of positive selection at G6pd in D. simulans as discussed in Eanes et al. (1996). In the following two sections, we present our case that the vermilion data are likely to reflect much more closely the population structure of this species in Africa.

Reduced coalescence time at G6pd: Of the 21 amino acid differences fixed between D. melanogaster and D. simulans at G6pd, 15 have become fixed in the D. simulans population.
lineage, although only 1 has been fixed since the divergence of D. simulans and D. sechellia (Eanes et al. 1996). A significant McDonald-Kreitman test supports the interpretation that many of those amino acid differences have been fixed due to positive selection (Eanes et al. 1993). The coalescence time of D. simulans is very long and probably goes back before the simulans-mauritiana-sechellia split because D. simulans and D. mauritiana share polymorphisms (Kliman and Hey 1993). An episode of positive selection at a locus is expected to reduce the coalescence time and eliminate evidence of prior population history for that region. Using the method of Hudson et al. (1987), we found that the levels of polymorphism and divergence at G6pd in our samples were concordant with those at vermilion. One possible interpretation of this result is that the episodes of positive selection at G6pd happened sufficiently long ago that they have no detectable effect on present-day levels of polymorphism. However, the lack of a departure from neutrality at G6pd makes it very unlikely that those amino acid differences could also be due to the nonrandom nature of our sample: we deliberately collected sequence data from the more variable 3’ region of G6pd. The unsurveyed 5’ end of the gene was much less variable in the alleles surveyed by Eanes et al. (1996). While it therefore remains unproven that there is a significant hitchhiking effect at G6pd in D. simulans, this is clearly a strong possibility that must be kept in mind in the interpretation of these data. If the migration of D. simulans out of Africa has been very recent, however (i.e., more recent than selection at G6pd), our G6pd data are still useful for inferring the relationship between African and non-African populations (see below).

Variation at vermilion reflects population history: In contrast to G6pd, the vermilion locus appears to have been evolving under purifying selection since the D. melanogaster-D. simulans split. There is only one amino acid difference (a serine to threonine change) between the genes in the two species (Begun and Aquadro 1995). Rates of recombination in the vermilion region are high (2.2 × 10^{-6}/bp/generation in D. melanogaster), so evolution at vermilion is substantially decoupled from evolution at linked loci. The D. simulans population samples surveyed by Begun and Aquadro (1995) showed no departure from neutrality in the original analyses or in a further analysis of these data using several heterogeneity tests (McDonald 1998). In our own data, Tajima’s D’s for several of the vermilion samples, although not significant, are large and negative. This is most likely due to recent population expansion, as the high levels of variation observed in most of these samples argue against other explanations such as a bottleneck or selective sweep. It therefore seems reasonable to use our vermilion data to make inferences about the history of African populations of D. simulans, keeping in mind that these inferences need to be confirmed by data from other, independent loci.

Relationships among African populations: At vermilion, Tanzania and Kenya have the highest levels of polymorphism and haplotype diversity and are not significantly differentiated from each other. Collectively, they contain 18 polymorphisms that are not observed in any other population and have the lowest 

<table>
<thead>
<tr>
<th>Antilles</th>
<th>Kenya</th>
<th>Tanzania</th>
<th>Zimbabwe</th>
<th>Cameroon</th>
<th>Italy/Europe</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Antilles</td>
<td>---</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Kenya</td>
<td>2</td>
<td>---</td>
<td>15</td>
<td>6</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Tanzania</td>
<td>0.20</td>
<td>---</td>
<td>0.38</td>
<td>0.17</td>
<td>0.25</td>
<td>0.09</td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>0</td>
<td>0.35</td>
<td>---</td>
<td>0.14</td>
<td>0.19</td>
<td>0.09</td>
</tr>
<tr>
<td>Cameroon</td>
<td>0.52</td>
<td>0.26</td>
<td>---</td>
<td>0.41</td>
<td>0.32</td>
<td>---</td>
</tr>
<tr>
<td>Italy/Europe</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>---</td>
<td>8</td>
</tr>
<tr>
<td>United States</td>
<td>0.10</td>
<td>0.20</td>
<td>0.27</td>
<td>0.09</td>
<td>---</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Number and proportion (see text) for vermilion are above the diagonal and for G6pd are below the diagonal.
The higher level of genetic variation in the Tanzania/Kenya samples suggests that the population in this region may be older and/or larger than the others. This interpretation, however, assumes that the observed differences in levels of polymorphism reflect real differences in long-term effective population size. Several issues need to be considered here; first is the variance associated with estimates of $4N_e\mu$. While the stochastic variance of independent samples of a single, nonrecombining locus is very large, note that substantial recombination at vermilion results in estimates of $4N_e\mu$ that represent several partially independent evolutionary trajectories, substantially reducing the variance (Pluzhnikov and Donnelly 1996). Even if we assume free recombination, however, none of our estimates of $4N_e\mu$ is significantly different from any other.

Perhaps more importantly, the populations surveyed in this study share a recent common ancestry in a species with a very long coalescence time ($4N_e$ generations, where $N_e \approx 10^6$). Many alleles within these populations are likely to share a most recent common ancestor with an allele from another population, and much of the genealogy of these alleles probably took place prior to the divergences of these populations from the ancestral population(s). Wakeley (1996) has shown that the variance of pairwise differences within and between samples decreases with the decrease in time of population splitting ($T$). At very small $T$, the variance converges on the value expected in a single, randomly mating population. In other words, the variance of $\theta$ among recently separated populations approaches the sampling variance of a single population.

Another issue is that, in populations recently descended from a common ancestral population, estimates of $4N_e\mu$ will predominately reflect ancestral population size and contain little information about the sizes of the descendant populations. This is particularly true if these populations have experienced rapid expansion, which slows down genetic drift (Nichols and Beaumont 1996). Instead, the observed differences in levels of polymorphism among these population samples may reflect differences in the amount of genetic variation sampled at foundation, which may in turn reflect proximity to an ancestral population. Interestingly, the African samples show a significant relationship between levels of genetic diversity and distance from Tanzania ($r^2 = 0.948, P = 0.026$). This strong pattern in the data suggests that the observed differences in variation are not simply random fluctuations about a common value of $4N_e\mu$. When the distance from either Zimbabwe or Cameroon is used as the independent variable, the $P$ values are 0.910 and 0.199, respectively. When the non-African samples were included in the analysis, their much lower variation and long distances result in a significant regression regardless of which African location was the focus, showing only that variation in Africa is higher than variation out of Africa.

**Inferences of gene flow based on shared informative sites:** As discussed above, biogeographical analysis suggests that continental African populations of *D. simulans* may have undergone an expansion on the order of $0.1N_e$ generations ago. Nichols and Beaumont (1996) have shown that exponential growth following foundation has the effect of reducing genetic drift: distributions of genetic variation become "frozen in place" and reflect the foundation event and early migration events, when the population is still small, much more than they reflect subsequent gene flow. This means that our estimates of $F_{ST}$ also reflect early events and may be misleading when used to make inferences about recent gene flow.

To try to separate foundation events from subsequent gene flow, we identified a subset of polymorphic sites that are segregating in more than one sample but not in all samples as informative segregating sites. We have not yet explored the expected properties of informative sites, so this analysis is only qualitative and suggestive, but it reveals some interesting properties of the dataset that are not revealed in any of the other analyses and may reflect underlying processes different than the $F_{ST}$ analysis.

The $F_{ST}$ statistics indicate rather similar levels of differentiation among all the African populations except between Kenya and Tanzania, presumably because all these populations share an ancestral set of polymorphisms that have been sorting for similar amounts of time. For example, $F_{ST}$ at vermilion between Zimbabwe and Tanzania is 0.122 ($P = 0.007$) and $F_{ST}$ at vermilion between Zimbabwe and Cameroon is 0.159 ($P = 0.014$). Yet Zimbabwe and Tanzania share only $\sim 14\%$ of their informative sites, while Zimbabwe and Cameroon share $\sim 40\%$, the same proportion as Kenya and Tanzania (Table 5). The high proportion of informative sites shared by Zimbabwe and Cameroon suggests that gene flow between these populations has been significant. Similarly, $F_{ST}$ is very high between Antilles and the United States (0.238; $P < 0.000$), but shared informative sites are high (30%).

Conversely, $F_{ST}$'s at vermilion are much smaller in comparisons between Tanzania, Italy, and the United States than in comparisons between Zimbabwe, Italy, and the United States. However, there are more derived polymorphisms shared among Italy, the United States, and Zimbabwe than among Italy, the United States, and Tanzania (Table 5). This suggests that Zimbabwe is as likely as Tanzania to be ancestral to these recently established populations or that alleles from another population not sampled in this study have migrated differentially.

As we would expect, at G6pd, where much of the shared ancestral polymorphism appears to have been eliminated during episodes of directional selection, the $F_{ST}$ and shared informative sites analyses are much more concordant. Both indicate a closer relationship among Zimbabwe, Europe, and Antilles than among Tanzania,
Europe, and Antilles, in agreement with the informative sites analysis at vermilion.

Interaction of selection, migration, and recombination at G6pd: If we did not have independent evidence of positive selection at G6pd (Eanes et al. 1996), the discordance between geographic patterns of variation at G6pd and vermilion would make interpretation of our data for both loci problematic. In contrast to the distribution of variation at vermilion, the G6pd alleles from Tanzania have the least variation of all five samples (Table 2). At G6pd, we see three private polymorphisms each in Cameroon and Zimbabwe, one in Kenya, and none in Tanzania. At vermilion, where there is no evidence of selection, Tanzania and Kenya each have five private polymorphisms, while Zimbabwe and Cameroon each have only one.

Positive selection at the G6pd locus has most likely reduced the coalescence time for this region such that it no longer reflects the early population history of D. simulans in Africa. In this case, the pattern observed at G6pd reflects the impact of directional selection on a recombining locus in a geographically structured species. A selectively favored mutation will go to fixation most quickly within the population in which it arises, in a process that may approximate a simple selective sweep model. Migration to distant populations will be slower and may allow opportunities for recombination during the fixation process, such that more variation may be preserved in those distant populations. Many different patterns of variation across populations could result from such a process, depending on the strength of selection, rate of recombination, rate of migration, and effective population size. The fact that variation at G6pd is lowest in Tanzania and is not reduced in the derived populations is one such outcome and is not inconsistent with the vermilion data, given that there is strong independent evidence for directional selection at this locus.

The hypothesis of population admixture: Our data indicate that populations of D. simulans in Europe and America are young and far from equilibrium. At vermilion, Italy, Antilles, and the United States have about half as much variation as Tanzania and Kenya. The five samples of D. simulans from non-African populations all show a deficiency in haplotype number and/or diversity. None of the eight samples from Africa shows such a departure (although the P values for the Zimbabwe samples are low). At vermilion, the exact significance values of these haplotype tests are somewhat compromised by the fact that the region surveyed was not chosen at random (see results), but the difference in results between African and non-African populations is nevertheless real. Unlike the FST analysis, both vermilion and G6pd give the same results in these tests, suggesting that the phenomenon responsible for these unusual haplotype structures is more recent than any episode of selection at G6pd. The fact that we observe the same significant haplotype structure at three out of four unlinked loci (Table 4) also suggests that this is a population-level phenomenon rather than multiple instances of diversifying selection.

The unusual haplotype structure and reduced variation provide strong evidence for a founder event in non-African populations as was inferred from the microsatellite survey of Irvin et al. (1998). The additional question of admixture is not resolved by our data. However, there are haplotypes observed at high frequencies in the non-African samples that are not present in any of the 46 alleles surveyed from the African populations. For example, the Italy 3 haplotype is present in 10 out of 36 non-African alleles, while the Antilles 3 haplotype is present seven times. These haplotypes, along with the U.S. 14 type, harbor five polymorphisms (at positions 1214, 1454, 1461, and 1654) that are not segregating in the 46 African alleles. These "non-African" variants may have come from some divergent African population not included in our survey. The most likely sources of divergent alleles are the subpopulations associated with the mitochondrial races siI, siII, and siIII (Baba-Aissa and Solignac 1984; Solignac and Monnerot 1986), which have distinct lineages over 1 million years old (Satta and Takahata 1990). In worldwide continental populations of D. simulans, there is no evidence for admixture based on mtDNA haplotypes, which are uniformly of the siI type; this may be a consequence of natural selection against the siI and siIII mtDNAs in these populations (Rand et al. 1994; Ballard et al. 1996). Although mtDNAs show evidence of admixture in only a few isolated populations (e.g., Madagascar, the Seychelles), worldwide patterns of neutral nuclear DNA variation may reveal the contributions from these ancient lineages in a way that allozyme variation does not.

Evidence of admixture should be reflected in patterns of haplotype structure throughout the genome, although this evidence will decay at different rates at different loci due to differences in recombination. Selection in a new environment could also eliminate evidence of admixture. At the two loci we have studied, as well as at Pgd, haplotype number and/or diversity are inconsistent with the equilibrium expectation; data from the Gld locus do not show such a departure. Clearly, more data from other loci are needed to test this hypothesis.

Conclusions and implications for population genetic studies:

1. Populations of D. simulans in continental Africa are genetically differentiated and therefore not very young, although they are probably still not at equilibrium.
2. Non-African populations of D. simulans are very young, far from equilibrium, and have experienced a bottleneck during their foundation.
3. Compelling, although not conclusive, evidence suggests that there has been admixture among geneti-
Estimates of nucleotide diversity in non-African populations do not appear to be inflated but are nonetheless unlikely to reflect long-term effective population size due to the combined effects of bottleneck and possible admixture. Ratios of silent to replacement variation might also be affected by a bottleneck if frequency distributions of those variants are different (i.e., if one or both are not neutral). For these reasons, non-African population samples of D. simulans are unsuitable for testing models that assume mutation-drift equilibrium. African populations of both D. melanogaster and D. simulans are probably more suitable for testing population genetic models but may also violate assumptions of equilibrium. In addition, comparisons between the species will need to be put in the context of their evolutionary histories, which may be quite different.

We thank Frantz Depaulis for providing the computer programs for the population substation test and the haplotype tests and for helpful discussions; Chip Aquadro for providing office space during preparation of the manuscript; the Aquadro lab group and two reviewers for comments on the manuscript; and Walt Eanes for editorial assistance. M.T.H. was supported by a Chateaubriand Fellowship from the French Embassy to the United States. M.V.'s research is supported by CNRS Unité Mixte de Recherche 7625.

LITERATURE CITED


Rozas, J., and R. Rozas, 1997 DnaSP version 2.0: a novel software


Communicating editor: W. F. Eanes