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 (Hyytia *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-

Corresponding author: Martha T. Hamblin, Department of Human Genetics, University of Chicago, 924 E. 57th St., Chicago, IL 60637. E-mail: mhamblin@genetics.uchicago.edu

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; Satta 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 Gloor *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 1

Population samples

Location	Date of collection	Collected by
Yaounde, Cameroon	December 1997	B. Riera
Nairobi, Kenya	September 1995	C. Wilson
Mt. Kilimanjaro, Tanzania	April 1996	D. Lachaise
Harare, Zimbabwe	February 1997	D. Lachaise
St. Martin, Lesser Antilles	March 1995	J. David
Sticiano, Italy	August 1996	C. Montchamp-Moreau

```
                                       position
          iiiiiiiiiiiiissssssssssiiiiiiiiiiiissssssiiiiiiiiiiiiissssssssssss
          111111111111111111111111111111111111111111111111111111111111111111
          222222222222222333333344444444444445556666666666666667777777888899
          001112223344599013445613455556667880361123334445666777225789778811
          451490562734825764695158114581566188226923675684368017477643473903
cons      accgccattcctcgggctcagctatgtggaagagacccgcgctcctaatctatctcccccgggcaa

Tanzania
ta5       ................t..g.....................................t........
ta12      ...................g..............t..t............................
ta4       ........a........................c...................t............
ta10      ....................cg................................a...........
ta13      ....................cgc............t..............a...a...........
ta11      ............t.......cgc.ca.......c............t............a......
ta8       .t.........c............ca..........................c.............
ta14      ........................ca.......c..t...........c.....gt.......a..
ta1       ..........a.............ca.......c.....ta..........c..a.....aaa...
ta2       ...................gcg..ca....ga.......ta........t.c..a...t.......
ta15      ........................ca....ga........at......................g.

Kenya
ke15      .t.........c............ca..............c.........................
ke2       ............t.......cgc.ca.......c............t............a......
ke10      ............t.......cgc.ca.......c............t............a......
ke11      ............t.......cgc.ca....ga.c............t............a......
ke3       .........t...t..........ca....ga......................a..........g
ke4       .t.............c........ca.......c......................t.........
ke5       ....tt..................ca.......c..t...........c.....gt..t....a..
ke12      ........................ca....ga.......tat......................g.
ke7       ........................ca.........t..tta.......................g.
ke8       ......................................tta.............a...........
ke1       ........a........................c...................t............
ke6       .......g.....................c.......t................a...........
ke9       ......................................tta.........................

Zimbabwe
zi15      ..............a.....cg......................tc........gt..........
zi6       .......................t.........c..t............................g
zi4       gta.................cg.....t....t...t.....aa.....................g
zi5       gta.................cg.....t....t...t.....aa.....................g
zi13      gta.................cg.....t....t...t.....aa.....................g
zi14      gta.................cg.....t....t...t....t......................g.
zi7       .........t.......c.gcg...a..........t....................t........
zi8       .........t.......c.gcg...a..........t....................t........
zi10      ........................ca....ga.......tat......................g.
zi11      .........................a...c........................a...........
```

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. [a] The U.S. alleles are from Begun and Aquadro (1995).

(1998). For *G6pd*, a recombination rate of $Nr/\text{bp} = 0.005$ was used; for *vermilion*, $Nr/\text{bp} = 0.01$. Note that for *vermilion*, estimates of r based on genetic data from *D. melanogaster* are $2 \times 10^{-6}/\text{bp}$ (Searles *et al.* 1990) and that if $N = 10^{-6}$, $Nr = 2$. (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 Nr; 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

```
Cameroon
ca3     ......................ca....ga.......tat.....................g.
ca2     ..............a...t..g.....................tc........gt.......a..
ca5     ..............a...t..g.....................t.........gt.......a..
ca10    .......g..............c............t..............a...a...........
ca11    ....................cgc............t..............a...a...........
ca4     ....................cgc..............t............a...a...........
ca8     gta..................cg.....t....t...t.....aa.....................g
ca6     ..............................c.......t................a...........
ca7     ..............................c.......t................a...........
ca12    .......g......................c.......t................a...........
ca1     .......g......................c.......t..............................
ca9     .......g......................c.......t..............................

Italy
it8     ........................a.............tat.....................g.
it12    ......................ca....ga.......tat.....................g.
it1     .t.........c............ca..........................c.............
it6     .t.........c............ca..........................c.............
it7     .t.........c............ca..........................c.............
it9     .t.........c............ca..........................c.............
it10    .t.........c............ca..........................c.............
it11    .t.........c............ca..........................c.............
it2     gta..................cg.....t....t...t....taa.....................g
it3     ...................gcg....c.....................g......a...........
it4     ...................gcg....c.....................g......a...........
it5     ...................gcg....c.....................g......a...........

Lesser Antilles
an3     ....................................tat.....................g.
an4     ....................................tat.....................g.
an5     ....................................tat.....................g.
an6     ....................................tat.....................g.
an7     ....................................tat.....................g.
an10    ....................................tat.....................g.
an11    ....................................tat.....................g.
an2     ......................ca....ga.......tat.....................g.
an9     .t.........c............ca..........................c.............
an1     ...................gcg....c.....................g......a...........
an8     .......................t.........c..t.............................g
an13    .......................t.........c..t.............................g

United States[a]
us18    ...................gcg....c.....................g......a...........
us27    ...................gcg....c.....................g......a...........
us11    ...................gcg....c.....................g......a...........
us3     ...................gcg....c.....................g......a...........
us5     ...................gcg....c.....................g......a...........
us28    ...................gcg....c.....................g......a...........
us10    .......................t............t.............................g
us14    ...t..c..................ca..a.......t.....aa.....................g
us15    ...t..c..................ca..a.......t.....aa.....................g
us29    ......................ca....ga.......ta.....................g.
us63    .t.........c............ca....ga...................c.............
us7     .t.........c............ca..........................c.............
```

Figure 1.—*Continued.*

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

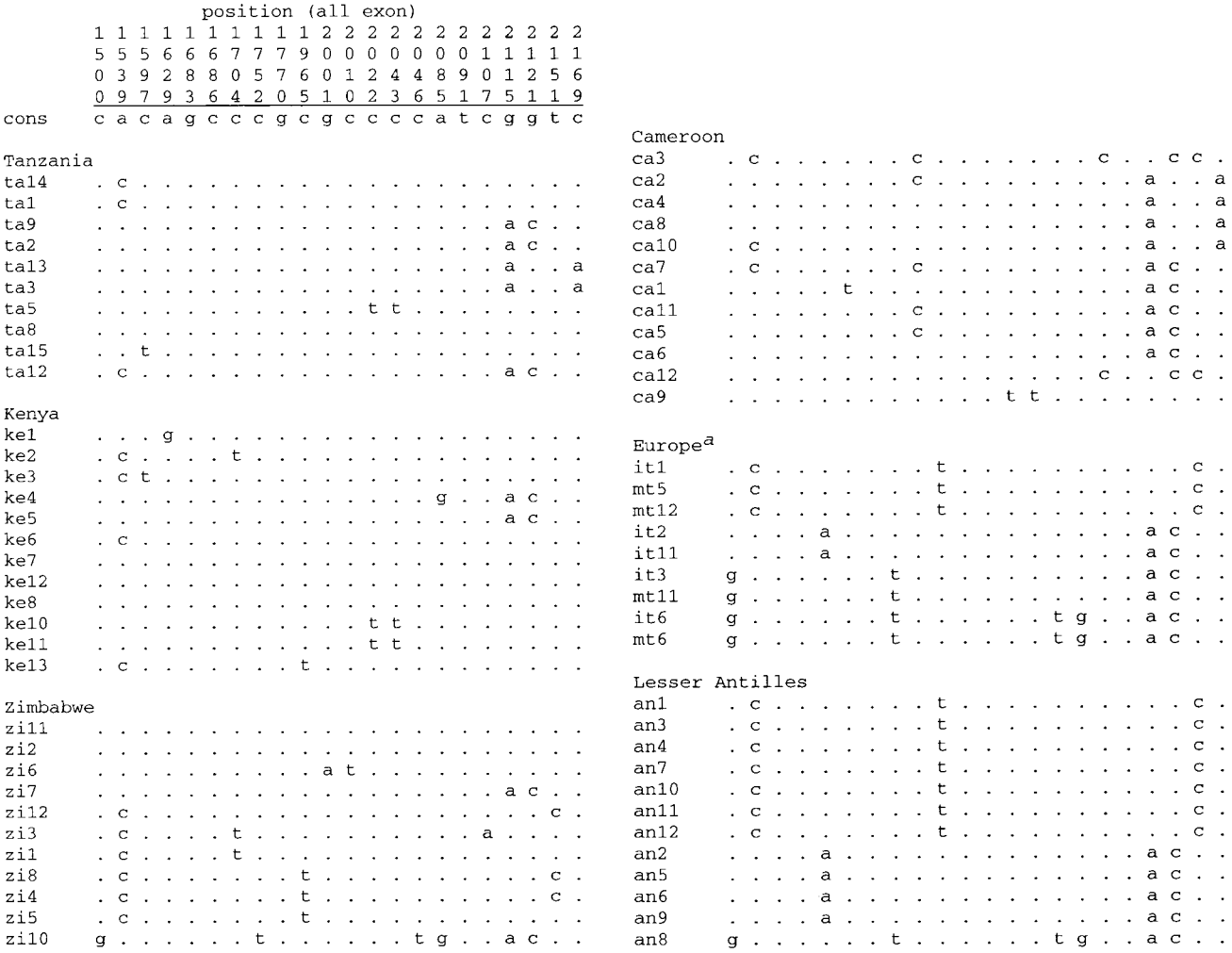

Figure 2.—Polymorphism at *G6pd* in *D. simulans*. The numbering system is the same as that in Eanes *et al.* (1996). cons, the most common base at each segregating site. All sites are synonymous coding sites. [a] The alleles labeled "it" are from Italy. The alleles labeled "mt" are from France (Eanes *et al.* 1996).

the ancestral population that became *D. simulans* inhabited coastal eastern Africa and/or the islands in the Indian Ocean. At *G6pd*, however, this pattern is not observed: the populations differ very little in the number of segregating sites, and Tanzania has the fewest. The ability to detect differences between populations is weaker at *G6pd* because there are about three times fewer polymorphic sites in this region, but other analyses suggest that the evolutionary history of the *G6pd* region has indeed been different from that of *vermilion* (see below).

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 ≤ 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 *G6pd* locus are different from those at *vermilion*: Cameroon appears

TABLE 2

Summary statistics of population variation at *vermilion* and *G6pd*

	vermilion					*G6pd*				
Population	*n*	*S*	θ	*k*	*D*	*n*	*S*	θ	*k*	*D*
Tanzania	11	40	13.66	10.47	−1.09	10	7	2.47	2.44	−0.051
Kenya	13	39	12.57	10.09	−0.88	12	10	3.31	2.53	−0.970
Zimbabwe	10	32	11.31	10.98	−0.14	11	13	4.44	3.67	−0.755
Cameroon	12	33	10.93	8.80	−0.83	12	10	3.31	3.47	0.197
Italy (Europe)[a]	12	25	8.28	8.23	−0.03	9	10	3.68	4.78	1.390
Lesser Antilles	12	21	6.95	5.59	−0.86	12	10	3.31	3.80	0.611
United States[b]	12	24	7.95	8.85	0.50					

n is the number of alleles surveyed; *S* is the number of mutations; θ is S/a from Watterson (1975); *k* is the average number of pairwise differences; *D* is Tajima's *D* (1989).

[a] For *g6pd*, includes four alleles from France from Eanes *et al.* (1996).

[b] From Begun and Aquadro (1995).

to be differentiated ($P \leq 0.005$) 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

Estimates of F_{ST} in *D. simulans* populations

	Antilles	Kenya	Tanzania	Zimbabwe	Cameroon	Italy/Europe	United States
Antilles	—	0.134	0.163	0.286	0.231	0.199	0.238
		(0.004)	(0.002)	(0.001)	(0.001)	(0.010)	(0.000)
Kenya	0.1857	—	−0.033	0.178	0.107	0.093	0.118
	(0.011)		(0.854)	(0.001)	(0.006)	(0.050)	(0.010)
Tanzania	0.1516	0	—	0.122	0.077	0.052	0.053
	(0.057)	(0.518)		(0.007)	(0.014)	(0.110)	(0.078)
Zimbabwe	0.037	0.021	0.041	—	0.159	0.162	0.130
	(0.220)	(0.297)	(0.167)		(0.014)	(0.012)	(0.009)
Cameroon	0.204	0.182	0.015	0.179	—	0.178	0.139
	(0.005)	(0.001)	(0.343)	(0.004)		(0.001)	(0.003)
Italy/Europe	0.077	0.269	0.175	0.127	0.188	—	0.040
	(0.147)	(0.003)	(0.025)	(0.064)	(0.002)		(0.181)

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

Locus/population	*n*	*S*	*K*	*P* value	*H*	*P* value
vermilion/Tanzania	11	39	11	1.000	0.909	1.000
vermilion/Kenya	13	37	12	0.906	0.911	0.906
vermilion/Zimbabwe	10	31	7	0.118	0.820	0.089
vermilion/Cameroon	12	32	10	0.566	0.889	0.566
vermilion/Italy	12	25	5	<0.001	0.667	<0.001
vermilion/Antilles	12	21	5	<0.001	0.611	<0.001
vermilion/U.S.[a]	12	24	6	0.007	0.694	0.003
vermilion/U.S.[b]	12	40	9	0.195	0.833	0.037
G6pd/Tanzania	10	7	7	0.910	0.840	0.933
G6pd/Kenya	12	10	9	0.955	0.898	0.861
G6pd/Zimbabwe	11	13	9	0.927	0.876,	0.927
G6pd/Cameroon	12	10	10	0.993	0.889	0.993
G6pd/Europe	9	10	4	0.044	0.741	0.213
G6pd/Antilles	12	10	3	<0.001	0.542	0.008
Pgd/U.S.[c]	19	11	3	0.016	0.526	0.042
gld/U.S.[d]	11	26	10	0.982	0.893	0.982

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).

[d] From Hamblin and Aquadro (1996).

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 2×10^{-6}/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-achaete* region from Europe and the United States (Martin-Campos *et al.* 1992; Begun and Aquadro 1991) as well as the *supressor 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.

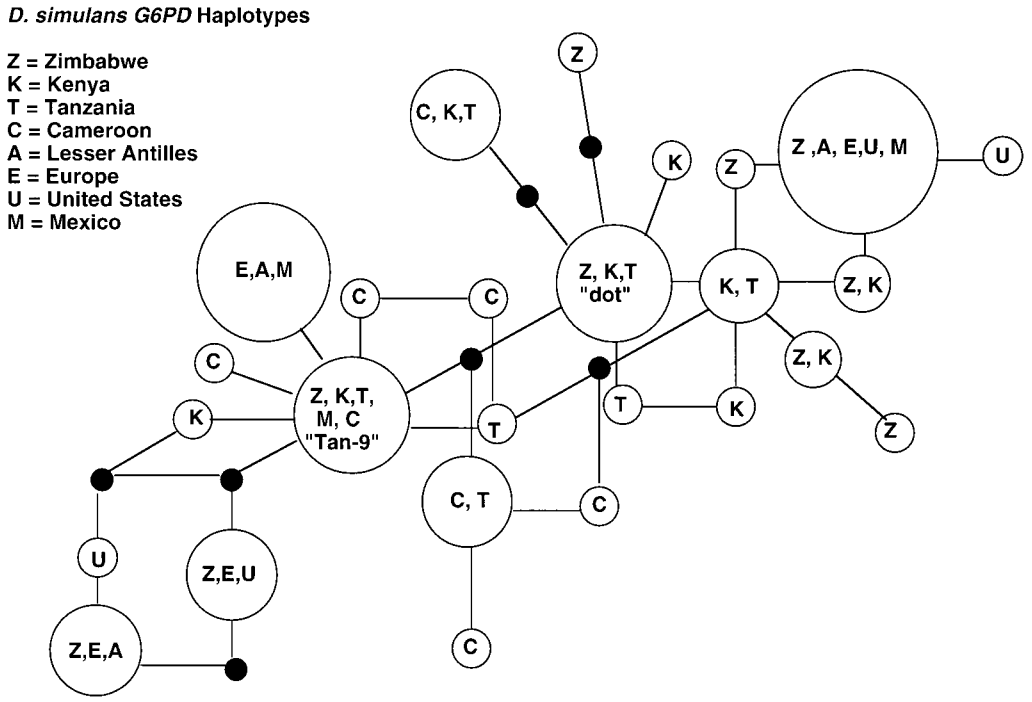

Figure 3.—Haplotype network of alleles at *G6pd*. Small black circles represent hypothetical intermediates not observed in our samples.

Informative site analysis: Because of the large amount of recombination in the history of the *vermilion* alleles, we tried to construct phylogenies using short regions (Templeton *et al.* 1992). This approach was not successful, however; the regions for which it was possible to construct a network with little recombination had few informative sites, while even slightly larger regions produced extremely reticulate phylogenies (data not shown).

Instead, we did a site-by-site analysis of what we call "informative segregating sites," *i.e.*, those sites that are segregating in at least two populations, but not in all of them. There were a total of 35 informative sites. For each pair of populations, the number of times they shared one of these informative sites was counted. The resulting counts were then divided by the average total number of segregating sites in the two populations, giving the fraction of the polymorphism shared by the two populations that is not present in all populations. The results are presented in Table 5. There are six comparisons in which >30% of the polymorphism is shared and informative. One of these, not surprisingly, is Kenya/Tanzania. What is not obvious from the F_{ST} data is that Zimbabwe and Cameroon have the highest proportion of shared informative polymorphisms and that the other most similar pairs all involve comparisons among Italy, Antilles, the United States, and Zimbabwe.

At *G6pd*, there is also a high proportion of shared informative polymorphisms among Zimbabwe, Antilles, and Europe, none of which shares any informative sites with Tanzania. These associations can be seen in the network in Figure 3, where several derived haplotypes are shared among these populations.

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*

TABLE 5

Shared informative sites

	Antilles	Kenya	Tanzania	Zimbabwe	Cameroon	Italy/Europe	United States
Antilles	—	4	5	5	0	7	7
		0.13	0.16	0.19	0	0.30	0.31
Kenya	2	—	15	6	9	3	2
	0.20		0.38	0.17	0.25	0.09	0.06
Tanzania	0	3	—	5	7	3	4
	0	0.35		0.14	0.19	0.09	0.13
Zimbabwe	6	3	0	—	13	9	5
	0.52	0.26	0		0.41	0.32	0.18
Cameroon	1	2	3	1	—	8	3
	0.10	0.20	0.27	0.09		0.28	0.11
Italy/Europe	7	2	0	6	1	—	8
	0.70	0.20	0	0.52	0.10		0.33

[a] Number and proportion (see text) for *vermilion* are above the diagonal and for *G6pd* are below the diagonal.

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 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 F_{ST}'s in comparisons with the non-African populations. The other two African populations have lower levels of polymorphism and are genetically differentiated from Tanzania/Kenya and from each other. The accumulation of genetic differences between distant populations suggests that these populations have not been very recently established. These observations are consistent with the biogeographic hypothesis of Lachaise *et al.* (1988), which proposes that early populations of *D. simulans* were restricted to coastal eastern Africa and islands in the Indian Ocean until the end of the Pleistocene (Lachaise *et al.* 1988), at which time they expanded into central Africa. The time of the putative expansion corresponds to $\sim 0.1\ N_e$ generations ago (assuming $N_e \geq 10^6$ and 10 generations/year), far less than the time required to reach mutation-drift equilibrium.

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 $N_e \geq 10^6$). 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). Wakely (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 θ 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 ~14% of their informative sites, while Zimbabwe and Cameroon share ~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 F_{ST} 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, 1225, 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-Aïssa 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 siII 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-

cally differentiated lineages in population(s) ancestral to European and American populations.

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 subdivision 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

Akashi, H., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics **144:** 1297–1307.

Aquadro, C. F., K. M. Lado and W. A. Noon, 1988 The *rosy* region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. Genetics **199:** 875–888.

Baba-Aïssa, F., and M. Solignac, 1984 La plupart des populations de *Drosophila simulans* ont probablement pour ancêtre une femelle unique dans un passé récent. C. R. Acad. Sci. Paris **299:** 289–292.

Ballard, J. W. O., J. Hatzidakis, T. L. Karr and M. Kreitman, 1996 Reduced variation in *Drosophila simulans* mitochondrial DNA. Genetics **144:** 1519–1528.

Begun, D. J., and C. F. Aquadro, 1991 Molecular population genetics of the distal portion of the X chromosome in Drosophila: evidence for genetic hitchhiking of the yellow-achaete region. Genetics **129:** 1147–1158.

Begun, D. J., and C. F. Aquadro, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA sequence level. Nature **365:** 548–550.

Begun, D. J., and C. F. Aquadro, 1994 Evolutionary inferences from DNA variation at the *6-phosphogluconate dehydrogenase* locus in natural populations of Drosophila: selection and geographic differentiation. Genetics **136:** 155–171.

Begun, D. J., and C. F. Aquadro, 1995 Molecular variation at the *vermilion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*. Genetics **149:** 1019–1032.

Benassi, V., and M. Veuille, 1995 Comparative population structuring of molecular and allozyme variation of *Drosophila melanogaster Adh* between Europe, West Africa and East Africa. Genet. Res. **65:** 95–103.

Choudhary, M., and R. S. Singh, 1987 A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. III. Variations in genetic structure and their causes between *D. melanogaster* and its sibling species *D. simulans*. Genetics **117:** 697–710.

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite site model. Mol. Biol. Evol. **15:** 1788–1790.

Eanes, W. F., M. Kirchner and J. Yoon, 1993 Evidence for adaptive evolution of the *G6pd* gene in *Drosophila melanogaster* and *Drosophila simulans* lineages. Proc. Natl. Acad. Sci. USA **90:** 7475–7479.

Eanes, W. F., M. Kirchner, J. Yoon, C. H. Biermann, I-N. Wang *et al.*, 1996 Historical selection, amino acid polymorphism and lineage-specific divergence at the *G6pd* locus in *Drosophila melanogaster* and *D. simulans*. Genetics **144:** 1027–1041.

Gloor, G. B., C. R. Preston, D. M. Johnson-Schlitz, N. A. Nassif, R. W. Phillis *et al.*, 1993 Type I repressors of P element mobility. Genetics **135:** 81–95.

Hamblin, M. T., and C. F. Aquadro, 1996 High nucleotide variation in a region of low recombination in *Drosophila simulans* is consistent with the background selection model. Mol. Biol. Evol. **13:** 1133–1140.

Hasson, E., I. N. Wang, L. W. Zeng, M. Kreitman and W. F. Eanes, 1998 Nucleotide variation in the triosephosphate isomerase (*Tpi*) locus of *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **15:** 756–769.

Hudson, R. R., M. Kreitman and M. Aguade, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Hudson, R. R., D. D. Boos and N. L. Kaplan, 1992a A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

Hudson, R. R., M. Slatkin and W. P. Maddison, 1992b Estimation of levels of gene flow from DNA sequence data. Genetics **132:** 583–589.

Hyytia, P., P. Capy, J. R. David and R. S. Singh, 1985 Enzymatic and quantitative variation in European and African populations of *Drosophila simulans*. Heredity **54:** 209–217.

Irvin, S. D., K. A. Wetterstrand, C. M. Hutter and C. F. Aquadro, 1998 Genetic variation and differentiation at microsatellite loci in *Drosophila simulans*: evidence for founder effects in New World populations. Genetics **150:** 777–790.

Kliman, R. M., and J. Hey, 1993 DNA sequence variation at the *period* locus within and among species of the *Drosophila melanogaster* complex. Genetics **133:** 375–387.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. **22:** 159–255.

Langley, C. H., J. McDonald, N. Miyashita and M. Aguade, 1993 Lack of correlation between interspecific divergence and intraspecific polymorphism at the *supressor of forked* region in *Drosophila melanogaster* and *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **90:** 1800–1803.

Martin-Campos, J. M., J. M. Comeron, N. Miyashita and M. Aguade, 1992 Intraspecific and interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster*. Genetics **130:** 805–816.

McDonald, J. H., 1998 Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **15:** 377–384.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

Nichols, R. A., and M. A. Beaumont, 1996 Is it ancient or modern history that we can read in the genes? pp. 69–87 in *Aspects of the Genesis and Maintenance of Biological Diversity*, edited by M. Hochberg, J. Clobert and R. Barbault. Oxford University Press, Oxford.

Ohta, T., 1992 The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. **23:** 263–286.

Parsons, P. A., 1975 The comparative evolutionary biology of the sibling species, *Drosophila melanogaster* and *D. simulans*. Quart. Rev. Biol. **50:** 151–169.

Pluzhnikov, A., and P. Donnelly, 1996 Optimal sequencing strategies for surveying molecular genetic diversity. Genetics **144:** 1247–1262.

Rand, D. M., M. Dorfsman and L. M. Kann, 1994 Neutral and nonneutral evolution of Drosophila mitochondrial DNA. Genetics **138:** 741–756.

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

package for extensive molecular population genetics analysis. Comput. Appl. Biosci. **13:** 307–311.

Satta, Y., and N. Takahata, 1990 Evolution of *Drosophila* mitochondrial DNA and the history of the *melanogaster* subgroup. Proc. Natl. Acad. Sci. USA **87:** 9558–9562.

Searles, L. L., R. S. Ruth, A.-M. Pret, R. A. Fridell and A. J. Ali, 1990 Structure and transcription of the *Drosophila melanogaster vermilion* gene and several mutant alleles. Mol. Cell. Biol. **10:** 1423–1431.

Solignac, M., and M. Monnerot, 1986 Race formation, speciation, and introgression within *Drosophila simulans, D. mauritiana*, and *D. sechellia* inferred from mitochondrial DNA analysis. Evolution **40:** 531–539.

Sturtevant, A. H., 1929 The genetics of *Drosophila simulans*. Carnegie Inst. Wash. Publ. No. **339:** 1–62.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Templeton, A. R., K. A. Crandall and C. F. Sing, 1992 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics **132:** 619–633.

Wakeley, J., 1996 Distinguishing migration from isolation using the variance of pairwise differences. Theor. Popul. Biol. **49:** 369–386.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: W. F. Eanes