Historical Selection, Amino Acid Polymorphism and Lineage-Specific Divergence at the *G6pd* Locus in *Drosophila melanogaster* and *D. simulans*

Walter F. Eanes, Michele Kirchner, Jeanne Yoon, Christiane H. Biermann, Ing-Nang Wang, Michael A. McCartney[1] and Brian C. Verrelli

Department of Ecology and Evolution, State University of New York, Stony Brook, New York 11794

Manuscript received June 22, 1995
Accepted for publication July 17, 1996

ABSTRACT

The nucleotide diversity across 1705 bp of the *G6pd* gene is studied in 50 *Drosophila melanogaster* and 12 *D. simulans* lines. Our earlier report contrasted intraspecific polymorphism and interspecific differences at silent and replacement sites in these species. This report expands the number of European and African lines and examines the pattern of polymorphism with respect to the common *A/B* allozymes. In *D. melanogaster* the silent nucleotide diversity varies 2.8-fold across localities. The *B* allele sequences are two- to fourfold more variable than the derived *A* allele, and differences between allozymes are twice as among *B* alleles. There is strong linkage disequilibrium across the *G6pd* region. In both species the level of silent polymorphism increases from the 5′ to 3′ ends, while there is no comparable pattern in level of silent site divergence or fixation. The neutral model is not rejected in either species. Using *D. yakuba* as an outgroup, the *D. melanogaster* lineage shows a twofold greater rate of silent fixation, but less than half the rate of amino acid replacement. Lineage-specific differences in mutation fixation are inconsistent with neutral expectations and suggest the interaction of species-specific population size differences with both weakly advantageous and deleterious selection.

UNDERSTANDING the nature of natural selection acting on genetic variation, in particular protein polymorphisms, is a fundamental problem in population genetics. Analysis of allele differences in both *in vitro* and *in vivo* function is an approach that has been used to reveal the *potential* for selection to act on different polymorphisms (see EANES 1987, GILLESPIE 1991, and WATT 1994 for discussion), but it discloses nothing of the historical consequences of selection on allele fate. As simplistic working hypotheses, one either might expect natural selection to move new alleles to polymorphic frequencies more rapidly than expected if they were neutral (as under recent directional selection), or conversely polymorphisms may be too old to be consistent with the life expectancy of a neutral polymorphism (if under balancing selection). The introduction of large-scale DNA sequencing at the population level has greatly enhanced our ability to resolve these patterns of historical selection at single genes. From DNA sequences, a sample of alleles may be projected onto a genealogical framework, and features of this framework have the potential to tell us much about the history of natural selection (HUDSON 1990; HUDSON *et al.* 1987, 1994). KREITMAN's definitive analysis of 11 copies of the *Adh* gene region in *Drosophila melanogaster* revealed the power of this type of analysis (KREITMAN 1983; HUDSON *et al.* 1987), and there is now a growing list of such studies, often finding patterns that are inconsistent with the null expectation supplied by a strict neutral theory (TAKANO *et al.* 1993; BEGUN and AQUADRO 1994; HUDSON *et al.* 1994; WALTHOUR and SCHAEFFER 1994).

Sequence-based studies also provide estimates of important population parameters such as population size and recombination. For example, sequence studies of the *Adh* region in *D. pseudoobscura* by SCHAEFFER and MILLER (1992, 1993) contrasted sharply with those of *D. melanogaster,* and these data are consistent with *D. pseudoobscura* having a much larger historical population size. Finally, levels of polymorphism, when partitioned, by locus or functional class, have been used with interspecific divergence and fixation to develop statistical tests for recent balancing or adaptive natural selection (HUDSON *et al.* 1987; MCDONALD and KREITMAN 1991; EANES *et al.* 1993). In this paper we apply this type of analysis to the study of the G6PD locus in *D. melanogaster* and *D. simulans.*

Much is known about the geographic variation and functional phenotypes of the G6PD electrophoretic polymorphism in *D. melanogaster* (see for example, YOUNG *et al.* 1964; CAVENER and CLEGG 1981; GANGULY *et al.* 1985; EANES *et al.* 1989; MIYASHITA 1990). The diallele allozyme polymorphism is cosmopolitan, showing reciprocal latitudinal clines in northern and southern hemispheres (OAKESHOTT *et al.* 1983). The *A* allele, whose protein migrates as a dimer under nondenaturing electrophoresis, predominates in Europe and Japan, while the tetrameric *B* variant predominates in

Corresponding author: Walter F. Eanes, Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794. E-mail: walter@life.bio.sunysb.edu

[1] *Present address:* Smithsonian Tropical Research Institute, Unit 0948, APO AA 34002.

sub-Saharan Africa. North American and Australian populations show clinal allele frequencies with the *A* allele increasing with latitude. We thus designate the *A* variant as the temperate allele. The *A* allele has a leucine at residue 384, and this change destabilizes the *in vitro* quaternary structure and results in the electrophoretic *A/B* polymorphism (EANES *et al.* 1993). This substitution also causes a catalytic change that produces an approximately 20% difference in pentose shunt flux between genotypes (LABATE and EANES 1992). Finally, our analyses with respect to the amino acid changes associated with human G6PD deficiency suggests that the common Pro/Leu polymorphism is in the NADP binding site (EANES *et al.* 1996).

Using 32 *D. melanogaster* and 12 *D. simulans* coding sequences of *G6pd* we (EANES *et al.* 1993) earlier contrasted intraspecific polymorphism and interspecific fixed differences for silent and replacement sites using the test and logic proposed by MCDONALD and KREITMAN (1991). We reported that G6PD is evolving under positive selection with respect to amino acid substitutions that became fixed between *D. melanogaster* and *D. simulans*. However, that contribution did not address a number of other issues.

This report expands the data set of that study geographically and, in particular, examines the intragenic pattern of nucleotide polymorphism with respect to the allozyme polymorphisms. In addition to five new European lines, the study includes 12 lines from Zimbabwe. BEGUN and AQUADRO (1993) have reported significantly higher levels of restriction site polymorphism at a number of loci including *G6pd* (BEGUN and AQUADRO 1993), and African populations are presumed to have given rise (relatively recently) to cosmopolitan populations as those in North America. They may thus represent populations closer to genetic equilibrium. For this reason, a collection of east African lines were considered important for comparison with the mostly cosmopolitan collection from EANES *et al.* (1993). Finally, motivated by the dramatic number of fixed amino acid differences between *D. melanogaster* and *D. simulans*, we have sequenced *D. yakuba* as an outgroup, thus allowing the assignment of each fixation event to a lineage. This will allow us to test for unequal rates of substitution at both replacement and silent sites. A recent report by AKASHI (1995) examined differential rates of silent substitution across a number genes that had been sequenced and found significant disparity in the number of unfavored codons fixed in the *D. melanogaster* lineage compared to the *D. simulans* lineage. The large number of amino acid substitutions for *G6pd* between these species allows us to examine this relationship as well.

From this analysis we are interested in assessing a number of questions central to the issue of intra- and interspecific patterns of selection at this locus. Is the level and distribution of polymorphism associated with the two allozymes indicative of recent arrival? What is the pattern of silent polymorphism in *G6pd* across North American, European and African populations? Is the level of polymorphism nonrandomly associated with the allozyme polymorphism in *D. melanogaster*, and is its level across the gene region inflated in the region of the allozyme polymorphism, as might be expected for a polymorphism under balancing selection (HUDSON *et al.* 1987)? Finally, is there parity between the *D. melanogaster* and *D. simulans* lineages in the long-term rates of fixation for silent and replacement mutation?

MATERIALS AND METHODS

Origin of wild lines: Thirty-two of the *D. melanogaster* lines are first described in the restriction map study of EANES *et al.* (1989). The twenty-one North American line sequences were first introduced in EANES *et al.* (1993). Lines designated as *CC* and *C* are from Watsonville, California and were collected in 1985. Lines designated as *DPF* and *D* are from Davis Peach Farm, Mt. Sinai, New York and were collected in 1985. A single line *F34* is from Orchid, Florida. Seven of the 13 lines of European origin were first described in EANES *et al.* (1993). These are designated as *F* (one line *F24.1*; Menetreal, France) and *G* (six lines; Tübingen, Germany), and were collected in 1986. The four lines designated as *OK* were collected in the Okavango Delta, Botswana, and were part of the EANES *et al.* (1993) data. Overall, 18 line sequences, as well as a single sequence from *D. yakuba* are added to the 32 first reported in EANES *et al.* (1993) to constitute the total data. These new sequences are prefixed *MT* (three lines from Montpellier, France, collected July 25, 1991), *F* (three lines from Menetreal, France, collected 1986) and *Z* (12 lines from Zimbabwe; see BEGUN and AQUADRO 1993). The 12 *D. simulans* lines were first introduced in EANES *et al.* (1993) and are from Davis Peach Farm, Mt. Sinai, New York (*DPF*), Montpellier, France (*MT*), and Vera Cruz, Mexico (*VC*). A single *D. yakuba* male from line number BG1016 from the Bowling Green stock center was sequenced.

PCR amplification and DNA sequencing: DNA sequence representing the 1673 nucleotide segment of the *G6pd* locus (545–2216 in FOUTS *et al.* 1988) in *D. melanogaster* was amplified via PCR from genomic DNA prepared as in MCGINNIS *et al.* (1983). Approximately 10 ng of genomic DNA was amplified in 10 mM Tris (pH 8.3), 50 mM KCl, 0.01% gelatin, 1 mM $MgCl_2$, 2 units of *AmpliTaq* polymerase (Perkin-Elmer), and 100 nM of each primer. The resulting amplified 1.67-kb fragment was then excised from 3% NuSieve agarose and used as template to amplify two smaller segments (545–1436 and 1408–2216). Single-strand template for sequencing was generated by the kinased primer/λ exonuclease digestion described by HIGUCHI and OCHMAN (1989). DNA template was separated from PCR primers using Millipore filters. Primers for Sanger dideoxy sequencing using Sequenase (U.S. Biochemical) were spaced about every 300 base pairs. Each sequencing reaction was run on both standard acrylamide gels with an electrolyte gradient (SHEEN and SEED 1988), and Long Ranger acrylamide gels (AT Biochem). Both strands were completely sequenced for each allele, with occasional gaps of 5–10 bp (less than 1–2% of the total sequence), where only one strand produced readable sequence. All polymorphisms were confirmed on both strands, and no errors were observed. The 1673 base sequence for the 50 *D. melanogaster* gene copies described here are stored under GenBank Accession numbers L13885–L13890, L13895–L13920, L13880, U42738–U42749, U43165–U43167, U44721, U45985; the 12 *D. simulans* copies

are L13876–L13879, L13891–L13894, and L13881–L13884. *D. yakuba* is stored under U42750.

Statistical methods: Because of our interest in the *G6pd* allozyme polymorphism we are often interested in the sequence diversity associated with a particular electromorph, which may be rare. This often requires prescreening large samples, with the inclusion of all recovered copies of a minority allele, and a random subsample of the majority allele. The statistical testing of such conditioned or stratified sets is complex (see HUDSON and KAPLAN 1986), and the tests of neutrality generally assume random sampling of alleles. To comply with this assumption, we apply the practice of assembling "constructed random samples" (HUDSON *et al.* 1994), where our construction is conditioned on *a priori* estimates of allele frequencies.

Two tests of the polymorphism frequency spectrum were applied to the data. Both TAJIMA (1989) and FU and LI (1993) proposed tests of the neutrality of silent variation. The TAJIMA test examines the allele frequency spectrum for all segregating sites in a sample of sequences. It contrasts the estimate of the parameter $\theta = 4N\mu$, based on the number of segregating polymorphisms, with an estimate of θ derived from the observed number of pairwise differences (π), the latter including information on polymorphic site frequencies. Under neutrality both values are expected to be the same. The FU and LI (1993) test contrasts the number of interior and exterior (singleton) branch mutations in the genealogy of a sample of sequences with numbers expected under a neutral model. Positive values of their D statistic indicate an excess of intermediate frequency polymorphisms, a negative value an excess of singletons as might be expected under deleterious mutation, or after a recent purge of molecular variation from a region. The interlocus contrast of intraspecific polymorphism and interspecific divergence proposed by HUDSON *et al.* (1987) was used as test of historical selection.

Finally, we are interested in assessing the potential of amino acid polymorphisms to distort the genealogical relationships in our sample of alleles. As proposed by HUDSON (1990, 1993), we have applied a Monte Carlo simulation of the Wright-Fisher population process to investigate the significance of specific features of the observed sequence data, and these features are defined with respect to the amino acid polymorphisms *per se*. The specific features defined here *a priori* are (1) the number of silent polymorphisms associated within the subset of copies bearing a particular derived mutation, and (2) the number of "fixed" mutations between those same subsets. Single-copy amino acid mutations in a sample may be examined with respect to the latter question only. For example, with respect to the latter question we may ask whether given the observed number of silent polymorphisms m in the total sample of size n, what is the probability that a subset of i alleles (or a singleton allele) will have j or more fixed unique mutations? These particular outcomes have biological relevance. An observed subset of alleles with fewer silent polymorphisms than generally observed under the Monte Carlo process would be indicative of a mutation recently favored by directional selection. Likewise, a subset with too many fixed differences would be an indication of an old balanced polymorphism.

The Monte Carlo simulations were done by generating a large number of replicate gene trees each with m polymorphic sites and determining what proportion of trees possesses the observed number of polymorphisms (or fewer) associated with the subset of i alleles, or in the case of fixed differences, the observed number (or more) associated with the defining subset of alleles. We have simulated the coalescent process using the logic and an adaptation of the computer algorithm presented by HUDSON (1990). The simulation uses only the observed number of polymorphisms as a parameter and makes no assumptions about θ (HUDSON 1993; HUDSON *et al.* 1994). These simulations also assume no recombination in the true sequences, recognizing that these will be, in the presence of recombination, conservative probabilities (HUDSON *et al.* 1994). Samples of 10,000 trees, each with m polymorphisms and conditioned on i copies per derived subset were generated in each simulation run.

Linkage disequilibrium was quantified and tested using the estimator of HILL (1974). To assign a sign to the direction of disequilibrium that was not arbitrary, nucleotide changes were identified as ancestral or derived in state relative to the *D. simulans* sequence. Thus the direction of disequilibrium indicates associations in these states. In addition the structure of associations was also examined where state was defined with respect to preferred and unpreferred codon usage as defined by AKASHI (1995). The method of HUDSON (1987) was used to generate the estimate of population level recombination.

RESULTS

Levels of intraspecific polymorphism: The basic sequence of the 1943 bp *G6pd* region and corrections to the *G6pd* sequence reported by FOUTS *et al.* (1988) has been reported in EANES *et al.* (1993). Polymorphisms are distributed across three exons of 264, 216 and 1078 bp, as well as two small introns of 62 and 85 bp. The length of intron 3 varies depending on the presence of a small insertion/deletion polymorphism; however, 85 bp are common to all lines. Based on the codon composition of the coding region spanning 519.3 codons, there are estimated to be 354 silent site equivalents (KREITMAN 1983). We have ignored the small first exon, which encodes only six amino acids and is separated from our sequence by a highly variable 2.7–12-kb intron 1 (EANES *et al.* 1989). The EANES *et al.* (1993) report did not discuss intron variation because the focus of that report was contrasting levels of synonymous and replacement polymorphism and divergence only. This current report includes polymorphisms in the introns as part of the general pattern of polymorphism and divergence across the gene.

For *D. melanogaster*, 50 copies of the *G6pd* region have been sequenced from North America (*DPF, D, CC,* and *C*), Europe (*G, MT, F*) and east Africa (*Z, OK*). This includes 18 copies designated *A*, 28 as *B*, and four as the *AF1* allozyme. In total there are 49 single base sites that are variable. These polymorphic sites are listed by line in Table 1 and their distribution is plotted in Figure 1. They add one replacement and 12 silent polymorphisms to the list presented in EANES *et al.* (1993). Other than single base changes, there are two insertion/deletion polymorphisms. First there is a 7 (Δ_1) or 24 bp (Δ_2) insertion/deletion polymorphism in intron 3 at position 1109–1110. The Δ_2 sequence possesses two additional single base polymorphisms at positions 1111 and 1122. In four lines there is an additional GATCAA copy (designated Δ_3) of the normal three repeats of this six base sequence seen in all other lines.

We have sequenced this same region in 12 lines of

TABLE 1

List of 51 polymorphic sites identified across 1705 nucleotide sites of *G6pd* in 50 lines of *Drosophila melanogaster*

	Nucleotide position
	1111111111111111111111111111111122222222
	6667788889901111122223444556677788888889999000111 2
	1476812332690123803486136370201913367892345178124 64
Line	9224946393855127863875041355643773938430576992512 94
SIM	GCGCTGGCCGG ACATAGCCCTCCCCGCCCCCGCCCCCCGGCCATGCTTC
Z16	.......T...32...C.....CTT...T.....T.........CCCTAC.
Z27	.....A.T.T. 1 .CC.T..C.G...T.....T.T.T....TC.CGAC.
Z5	.....A.T... 2C...C....CTG.T.T.....T..TT.A.....CGA..
Z55	.......T... 1 ..C....CTT...T.....T..TT.....C.TGAC.
Z42	.....A.T...32...C.....CTT...T..T..T.........CCCTAC.
Z3	.......T..A 2C...C....C.G...T.....T...TA....CCTGAC.
Z21	.....A.T... 2C...C....CTG.T.T.....T...T.A.T...CGAC.
Z64	.......T...32...C...T...G..A.T....T...T.....CCTGAC.
Z74	...A.A.T... 2C...C....C.T.T.T.....T.....A...CCTG.C.
Z11	.......T.T. 1 .CC.T..C.G...T.......T.TA...TCCTGAC.
OK16	.....A..A.. 2C...C.....TT...T.T...T...T.....C.CGAC.
OK5	.....A.T... 2C..C.....CTT...T.....T.........CCCGAC.
OK93	......AT... 2CA..C....C.T......T..T...........CTA..
OK70	......AT... 1 ..C....CTT...T...T.TA..T.....CCCTAC.
Z62	......AT... 1 .CC....C.G...T...T.TA..TA....C.CTAC.
Z41	......AT... 1 .CC...TC.T...T...T.TA..T...T.CCCTAC.
G4.2	......AT... 1 .CC....C.G...T...T.TA..T.....CCCTAC.
G11.1	......AT... 1 .CC....C.G...T...T.TA..TA....C.CTAC.
G2.2	......AT... 1 .CC....C.G...T...T.TA..T.....CCCTAC.
G22.2	......AT... 1 .CC....C.G...T...T.TA..T...T.CCCTAC.
G40.2	......AT... 1 .CC....C.G...T...TATA..T...T.CCCTAC.
G13.2	......AT... 1 .CC....C.G...T...T.TA..T...T.CCCTAC.
F22.3	T.....AT... 1 .CC....C.G...T...T.TA..TA..T.CCCTAC.
F24.1	T.....AT... 1 TCC....C.G...T...T.TA..T...T.CCCTAC.
MT68	T.....AT... 1 .CC....C.G...T...T.TA..T...T.CCCTAC.
MT32	T.....AT... 1 .CC....C.G...T...T.TA..T...T.CCCTAC.
F23.3	......AT... 2CA..C....C.G...T..T..T...T.....C.CGACT
F9.2	..A..A.T... 2...CC....C.T...T..T..T...T.....C.CGAC.
MT41	....CA.TA..32C..CCT....TGTT.T.....T...T....TCCTGAC.

	Nucleotide position
	1111111111111111111111111111111122222222
	6667788889901111122223444556677788888889999000111 2
	1476812332690123803486136370201913367892345178124 64
Line	9224946393855127863875041355643773938430576992512 94
SIM	GCGCTGGCCGG ACATAGCCCTCCCCGCCCCCGCCCCCCGGCCATGCTTC
CC10	......AT... 1 .CC....C.G...T...T.TA..T...T.CCCTAC.
CC19	......AT... 1 .CC....C.G...T...T.TA..T...T.CCCTAC.
CC38	.G....AT... 1 .CC....C.G...T...T.TA..T...T.CCCTAC.
DPF1A	......AT... 1 .CC....C.G...T...T.TA..T...T.CCCTAC.
D64A	......AT... 1 .CC....C.G...T...T.TA..T...T.CCCTAC.
DPF2A	......AT... 1 .CC....C.G...T...T.TA..T...T.CCCTAC.
CC22	......AT... 1 .CC....C.G...T...T.TA..T.....CCCTAC.
CC2	......AT... 1 .CC....C.G...T...T.TA..T.....C.CGACT
CC34	......AT... 1 .CC....C.G...T...T.TA..T.....C.CGAC.
D11A	......AT... 1 .CC....C.G...T...T.TA..T.....CCCTAC.
DPF1B	......AT... 2CA..C....C.G...T.....T......A....CGAC.
DPF11	......AT... 2CA..C....C.G...T.....T......A....CGAC.
DPF14	......AT... 2CA..C....C.G.........T......A....CGAC.
DPF21	......AT... 2CA.CC....C.G.........T......A....CGAC.
DPF5B	......AT... 2CA..C....C.G...T..T..T...T.......CGAC.

TABLE 1

Continued

Line	Nucleotide position
	1111111111111111111111111111111122222222
	66677888899011111222234445566777888888899990001111 2
	14768123326901238034861363702019133678923451781246 4
	92249463938551278638750413556437739384305769925129 4
F34	......AT... 2CA..C....C.G...T..T..T...T.......CGAC.
DPF3B	......AT..A 1 ..C....C.T.........T......A....CGAC.
DPF7B	......AT... 2CA..C....C.G.........T......A....CGAC.
CC28	......AT... 1 .CC....C.T...T.....T......A....CGAC.
CC37	......AT... 2CA..C....C.G...T..T..T..........CGAC.
CC39	......AT..A 2CA..C....C.G...T..T..T...T...T...CGAC.

Characters 1, 2, and 3 designate the Δ_1, Δ_2, and Δ_3 insertion/deletion polymorphisms. SIM refers to *D. simulans* line DPF88S. Nucleotide polymorphisms shown in bold are amino acid polymorphisms.

D. simulans from Europe (*MT*), New York (*DPF*) and Mexico (*VC*). The levels of polymorphisms were first summarized in EANES *et al.* (1993) and the data are presented for the first time in Table 2. There are 18 polymorphisms including a small insertion/deletion polymorphism in intron 3 after nucleotide 1119. This polymorphism is similar to the small insertion/deletion polymorphism seen in *D. melanogaster.*

Levels of polymorphism in *D. melanogaster* were examined separately for samples from three geographic regions. We will treat the two North American localities of *D. melanogaster* (including the single Florida sequence) as part of an effectively panmictic unit because of the general sharing of haplotypes seen in other studies (see KREITMAN and AGUADÉ 1986; EANES *et al.* 1989; MIYASHITA 1990) and pool the data. This results in data for 21 lines of which 10 and 11 carry the *A* and *B* allozyme polymorphisms, respectively. There are a total of 14 polymorphisms in this sample. At silent positions the average number of differences between North American *G6pd* copies is 5.18 (4.94 in California, 5.16 in NY, and 5.34 between), and the nucleotide diversity (NEI 1987) is $\pi = 0.0146$, using our estimate of 354 effectively silent sites. By electromorph (differentiated by the Leu/Pro polymorphism), the average number of differences within *A* and *B* allele classes are 1.64 and 3.13, respectively, and on the average *A* and *B* alleles differ at 7.85 silent sites.

Thirteen lines were sampled from Europe. Six lines bear the *A* allozyme, 3 the *B* and 4 the *AF1* electrophoretic allele. Using in each case all the *A* and *AF1* bearing lines, and a single *B* different allele, we have three constructed random samples (*sensu* HUDSON *et al.* 1994). This constitution is based on our *a priori* knowledge of the allozyme frequencies in populations north of the Sahara where *B* has a frequency of about 5–10%. The nucleotide diversity (the average of three samples of $n = 11$, each with a different *B* allele thus reflecting natural allozyme frequencies) in European localities is estimated as $\pi = 0.0079$.

In contrast to the two temperate populations, 38 silent polymorphisms are found in the 16 east African lines. One line, *Z74,* shows an additional amino acid change of Thr to Asn at position 764. From the east African sites in Zimbabwe and the Okavango Delta, we constructed three samples, each with all the 13 *B* alleles, and one of the three *A* alleles. From these constructed random samples we estimate the nucleotide diversity to be $\pi = 0.0219$. The average number of differences within African *A* and *B* allele classes are 3.11 and 7.90, respectively, and the average between *A* and *B* allele classes is 8.15. Table 3 summarizes the information on levels of polymorphism among populations and allele classes. We have pooled the line data for *D. simulans* since studies of other genes have failed to uncover any regional structure and most *G6pd* haplotypes are com-

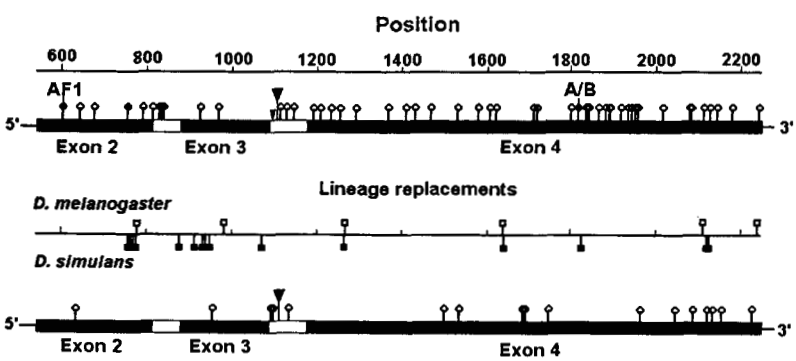

FIGURE 1.—The distribution of polymorphic sites across the *G6pd* gene region in *Drosophila melanogaster* (top) and *D. simulans* (bottom). Circles designate single base mutations, and filled circles represent amino acid changes. Triangles refer to insertion/deletion polymorphisms. The positions of the *A/B* and *AF1* allozyme polymorphisms are indicated. The position of the 21 amino acid replacements that occur with respect to each species lineage are shown along the scale between the two species.

TABLE 2

List of 18 polymorphic sites identified across 1705 nucleotide sites of the *G6pd* locus in 12 lines of *Drosophila simulans*

Line	Nucleotide position																	
	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2		
	6	9	0	0	1	1	5	5	6	6	7	9	0	0	1	1	1	2
	4	6	9	9	1	5	0	3	8	8	5	6	4	8	1	2	5	2
	8	2	3	4	9	0	0	9	3	6	2	5	6	5	5	1	1	0
MEL	T	T	A	G	1	T	C	T	G	C	C	C	C	G	C	G	T	C
VC902	.	.	T	A	0	.	.	A	A	.	.	.	.	A	A	C	.	T
VC908	.	.	T	A	0	.	.	A	A	.	.	.	.	A	A	C	.	T
VC913	.	.	.	A	0	.	.	A	.	T	.	.	.	A	A	C	.	.
DPF87S	.	.	T	A	0	.	G	A	.	.	T	.	.	A	A	C	.	.
DPF92S	.	.	T	A	0	.	G	A	.	.	T	.	.	A	A	C	.	.
MT11	.	.	T	A	0	.	G	A	.	.	T	.	.	A	A	C	.	.
MT6	C	G	.	.	.	C	G	A	.	.	T	.	T	.	A	C	.	.
DPF88S	C	G	.	.	.	C	.	A	.	.	T	.	T	.	A	C	.	.
DPF101S	C	.	.	A	.	C	.	C	.	.	.	T	.	.	G	.	C	.
VC903	C	.	.	A	.	C	.	C	.	.	.	T	.	A	G	.	C	.
MT5	C	.	.	A	.	C	.	C	.	.	.	T	.	A	G	.	C	.
MT12	C	.	.	A	.	C	.	C	.	.	.	T	.	A	G	.	C	.

MEL refers to line OK93 of *D. melanogaster*. 0 and 1 refer to the Δ_0 and Δ_1 insertion/deletion polymorphisms at position 1119.

mon to several sites. The nucleotide diversity for *D. simulans* is 0.0127 and there are no replacement polymorphisms. Earlier studies of *D. simulans* failed to find any allozyme variation (EANES 1983).

Figure 2 shows the neighbor-joining (SAITOU and NEI 1987) analysis of all 50 sequences of *D. melanogaster* and 12 sequences of *D. simulans*. A UPGMA clustering yielded virtually identical groupings with respect to electrophoretic allele. It is based on the 52 synonymous site polymorphisms, nine polymorphisms within the second intron, and 24 fixed silent differences between species. No replacement variation is included because we wish to examine the clustering independent of allozyme type and because between species replacement differences are not informative. The additional polymorphisms which involve variable sites within the seven (Δ_1) and 24 bp (Δ_2) insertion/deletion polymorphism in intron 3 are also not used. This clustering illustrates the basic relationships between sequences, but is not meant to explicitly reflect the genealogical process, since it will be obscured by recombination. It would appear that recombination has not erased much information, since the analysis clearly partitions the set of alleles into three clusters related to allozyme, and cosmopolitan or African origin. The first cluster of 23 sequences (*G40.1* to *CC34*) groups all alleles that possess the Pro → Leu amino acid change responsible for the derived *A* allozyme. As well this group includes the four copies of the *AF1* allozyme polymorphism, which is created by a Gly → Cys substitution in nucleotide 619. This cluster includes *A* alleles from all three continents. The second group (*F23.3* to *DPF14B*) clusters only *B* alleles. One subgroup clusters the 11 *B* alleles from North America, and the second group of *B* alleles represent those sampled from east African sources, although a European *B* copy (*F9.2*) also falls in this group. It should also be noted that the common *X*-linked inversion, *In(1)A* (associated with lines *Z16, Z27,* and *Z21*), which is unique to east African populations (EANES *et al.* 1992), shows no particular association with variation at the *G6pd* locus even though its proximal breakpoint (18D1-10) is relatively near the *G6pd* locus (at 18D12-13).

Linkage disequilibrium: There is notable linkage disequilibrium across the 1.7-kb *G6pd* region, and the *A/B* polymorphism is involved in the pattern of association. Inspection of the matrix of *D* values and associated squared correlations r^2 for the 43 polymorphisms in 50 lines shows that 108 of 903 pairwise correlations (12%) are statistically significant ($P < 0.05$), without correcting for multiple comparisons. Much of the covariance among silent polymorphisms is contributed by the geographic structuring, since the east African lines possess many unique polymorphisms. A more informative analysis is one limited to North America where there are 21 chromosomes and the *A/B* polymorphism is observed in more intermediate frequencies than seen in Europe and east Africa. Table 4 is the matrix of associations for the 16 polymorphic sites (including the Δ_1 and Δ_2 insertion/deletion polymorphism). Out of 112 possible values, 53 are statistically significant (at $P < 0.05$), and these are largely in association with the *A/B* polymorphism (11 of 15 are significant). Given that

TABLE 3

Summary population parameter estimates for θ and π for three samples of *Drosophila melanogaster* and *D. simulans*

Locality	*n*	$\hat{\theta}_A$	$\hat{\pi}_A$	$\hat{\theta}_B$	$\hat{\pi}_B$	$\hat{\theta}_{total}$	$\hat{\pi}_{total}$
D. melanogaster							
N. America	21	0.0048 (10)	0.0046	0.0083 (11)	0.0088	0.0109	0.0146
Africa	16	0.0108 (3)	0.0088	0.0293 (13)	0.0223	0.0271[a]	0.0219[a]
Europe	13	0.0048 (10)	0.0039	0.0200 (3)	0.0163	0.0092[b]	0.0079[b]
D. simulans	12	—	—	—	—	0.0127	0.0160

Data are partitioned into the *A* and *B* electrophoretic alleles and the total. Estimates are based on 354 silent site equivalents.
[a] Based on the average of three "constructed random" samples of 13 *B* and 1 *A* allele.
[b] Based on the average of three "constructed random" samples of 6 *A*, 4 *AF1*, and 1 *B* allele.

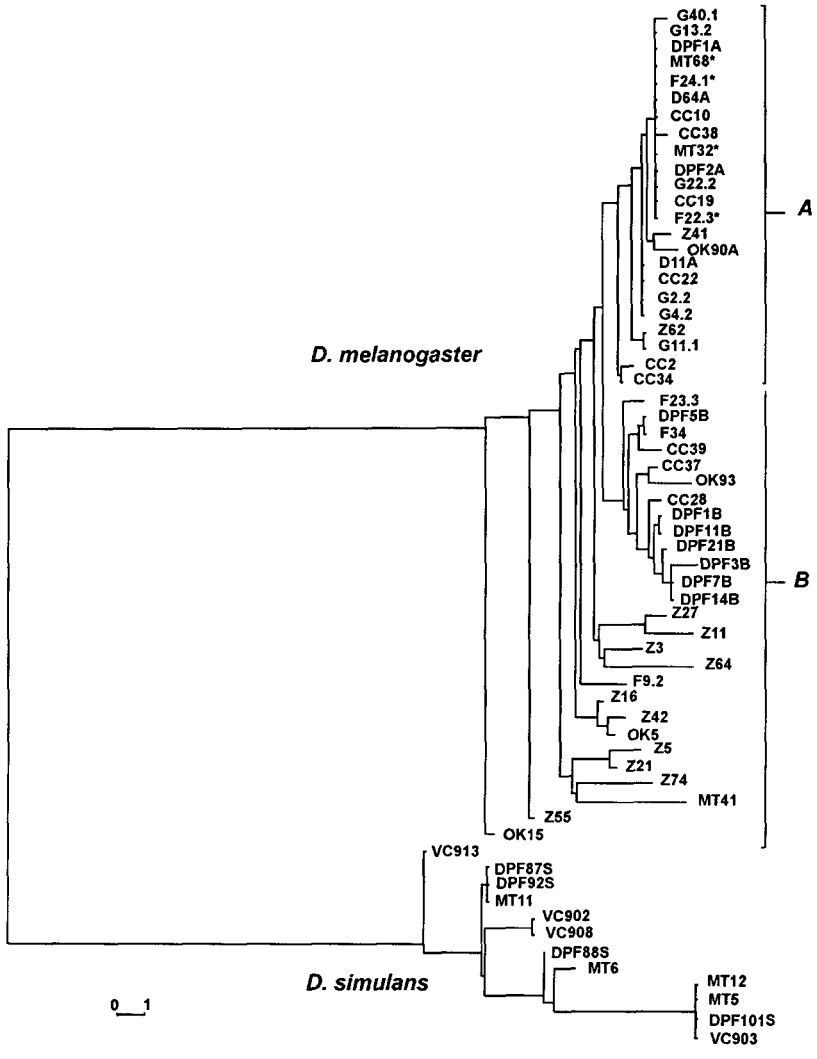

FIGURE 2.—A Neighbor-joining clustering based on silent substitutions among and between the 50 *Drosophila melanogaster* and 12 *D. simulans G6pd* sequences. The analysis is based on the number of silent differences, and the scale indicates one difference. The asterisks designate the *AF1* electrophoretic allele.

the sample size is only 21 chromosomes, statistical significance also implies a high correlation, since no values of r^2 below 0.182 will be significant at $P < 0.05$. Polymorphisms at positions 1863 and 2079 are 46 and 262 base pairs from position 1817 and show perfect associations with the allozyme polymorphism. Even though there are several individual sequences representing likely recombination products, overall the entire region is associated as a block. Finally, there is also no systematic direction with respect to the sign of the disequilibrium when state is defined as either derived or ancestral (Table 4), or preferred or unpreferred with respect to codon bias. With respect to the former definition, 44 of 88 informative disequilibria (leaving out sites 1105 and 1817) are positive and the latter 48 of 88 informative values are positive.

Tests of the polymorphism frequency spectrum: Both TAJIMA (1989) and FU and LI (1993) tests were applied to the data. For *D. melanogaster* we tested the data by partitioning into "constructed random samples" (HUDSON *et al.* 1994) of African (n = 14; all 13 *B* alleles and 1 *A* allele, sampled at random), North American (n = 21; entire data set), and European (n = 11; all 6*A* alleles, all 4 *AF1* alleles, and 1 *B* allele sampled at random) samples. The total data of 12 alleles for *D. simulans* are pooled. The results are nonsignificant for all tests.

Tests contrasting intraspecific polymorphism and interspecific divergence: Two issues arise when contrasting inter "locus" levels of polymorphism and divergence, as in HKA tests. These are defining (1) the choice of a "neutral locus" (see HUDSON *et al.* 1994), and (2) the limits of the span of sequence whose genealogy is assumed to be correlated with a site under natural selection, as when one has identified an allozyme polymorphism. One nonarbitrary approach would be to define the selected region as the entire gene in question. This might be appropriate if no *a priori* interest exists for a specific site or region in the gene. However, both directional and balancing selection may impact only a small span of sequence, perhaps hundreds of base pairs or less (KAPLAN *et al.* 1988; HUDSON and KAPLAN 1988); therefore special interest may require further partitioning of the region in question.

In this paper we have attempted to specify *a priori* some expectations about this region, recognizing that the Pro/Leu polymorphism is in the second half of the

TABLE 4

Matrix of disequilibra (D) and squared correlations (r^2 below diagonal) across the 16 polymorphic sites in the *G6pd* region observed in 21 North American lines

Position	*642*	*968*	*1095*[a]	*1188*	*1461*	*1626*	*1797*	*1817*	*1863*	*1893*	*1947*	*1956*	*2079*	*2082*	*2121*	*2244*
642		−0.01	0.02	0.02	−0.01	−0.01	−0.01	0.03	0.03	0.02	−0.02	0.03	−0.03	0.03	0.03	−0.00
968	0.01		−0.01	−0.05	0.04	0.03	0.03	−0.05	−0.05	−0.01	0.02	0.02	0.05	−0.04	−0.04	−0.01
1095	0.04	0.02		0.20	0.04	−0.06	−0.11	0.20	0.20	0.12	−0.10	0.10	−0.20	0.16	0.16	0.02
1188	0.04	0.14	**0.65**		−0.01	−0.06	−0.11	0.20	0.20	0.12	−0.10	0.10	−0.20	0.16	0.16	0.02
1461	0.01	**0.20**	0.08	0.00		0.03	−0.02	−0.05	−0.05	−0.06	0.06	−0.03	0.05	−0.04	−0.04	−0.01
1626	0.01	0.07	0.10	0.10	0.07		−0.04	−0.10	−0.10	−0.12	0.13	−0.06	0.09	−0.07	−0.07	−0.01
1797	0.00	0.07	**0.31**	**0.31**	0.03	0.06		−0.10	−0.10	0.03	−0.06	−0.02	0.09	−0.07	−0.07	−0.01
1817	0.06	0.10	**0.68**	**0.68**	0.01	**0.21**	**0.21**		0.25	0.18	−0.16	0.13	−0.25	0.20	0.20	0.03
1863	0.06	0.10	**0.68**	**0.68**	0.01	**0.21**	**0.21**	**1.00**		0.18	−0.16	0.13	−0.25	0.20	0.20	0.03
1893	0.03	0.01	**0.26**	**0.26**	0.17	**0.38**	0.02	**0.56**	**0.56**		−0.21	0.13	−0.18	0.15	0.15	0.02
1947	0.03	0.01	0.17	0.17	**0.21**	**0.47**	0.12	**0.46**	**0.46**	**0.81**		−0.11	0.16	−0.13	−0.13	−0.02
1956	0.10	0.01	0.17	0.17	0.05	0.12	0.01	**0.29**	**0.29**	**0.32**	**0.25**		−0.13	0.16	0.16	−0.02
2079	0.06	0.10	**0.68**	**0.68**	0.01	**0.21**	**0.21**	**1.00**	**1.00**	**0.56**	**0.46**	**0.29**		−0.20	−0.20	−0.03
2082	0.08	0.07	**0.46**	**0.46**	0.07	0.15	0.15	**0.68**	**0.68**	**0.38**	**0.31**	**0.48**	**0.68**		0.24	−0.02
2121	0.08	0.07	**0.46**	**0.46**	0.07	0.15	0.15	**0.68**	**0.68**	**0.38**	**0.31**	**0.48**	**0.68**	**1.00**		−0.02
2244	0.00	0.01	0.04	0.04	0.01	0.01	0.01	0.06	0.01	0.03	0.03	0.03	0.01	0.03	0.03	
H	*0.09*	*0.17*	*0.49*	*0.49*	*0.17*	*0.31*	*0.31*	*0.50*	*0.50*	*0.47*	*0.44*	*0.44*	*0.50*	*0.47*	*0.47*	*0.09*

See text for designation of sign of D. *H* is the heterozygosity at each site, where a value of 0.090 indicates a single change. Bold faced values are statistically significant at $P < 0.05$.

[a] 1095 is the Δ_1/Δ_2 polymorphism.

gene (see Figure 1). Our first approach is to simply treat the entire 1705-bp *G6pd* region as the selected locus, and contrast it with the 5′ *Adh* region (HUDSON *et al.* 1987). However, HUDSON and KAPLAN's (1988) analysis of *Adh* showed a best fit where only a region of about 500 bp shows an "excess" of polymorphism. Therefore, our second approach has been to partition the *G6pd* gene into two regions of equal numbers of potentially neutral sites, although the 5′ region will also include the second and third introns. In total there are 467 silent site equivalents. Region 1 spans from nucleotide 543 to 1222, and region 2 from 1223 to 2247.

Table 5 lists the number of polymorphisms, number of fixed differences, and divergence for regions 1 and 2, and the total *G6pd* region. The number of synonymous polymorphisms increases across the gene in both species. There are a larger number of polymorphisms in the 3′ half of the *G6pd* coding region (ignoring intron sequences and dividing into 177 silent site equivalents each) in both species but this is only significant in the *D. simulans* (13:25 for *D. melanogaster*, $X^2 = 3.79$, NS and 2:12 for *D. simulans*, $X^2 = 7.14$, $P < 0.01$) and combined species set ($X^2 = 9.31$, $P < 0.005$). Given the inherent statistical weakness of the above partitioning and to further test if there is what appears to be excessive clustering around the *A/B* polymorphism site in *D. melanogaster*, we carried out a randomization of the silent polymorphism positions, each time examining the distribution of average distance in base pairs between site 1817 and each of the 38 silent polymorphisms. Each replicate sample involved distributing 38 polymorphic sites at random across the coding gene and asking how often is the average distance of the randomized set equal to, or less than, the observed average distance in the true data. Only 9 of 10,000 randomized sets possessed smaller than the observed value. Thus, as is suggested by Figure 1, the distribution of silent polymorphisms is not random, but is significantly clustered with respect to the *A/B* polymorphism.

Under neutral theory, one explanation for this pattern in polymorphism could be an increase in the neutral mutation rate from the 5′ to 3′ end. The overall level of divergence between *D. melanogaster* and *D. simulans*, using reference lines *OK93* and *DPF88S*, shows a typical overall level of silent site divergence of 11% (see

TABLE 5

Summary data on numbers of polymorphisms in *Drosophila melanogaster*, diverged sites between reference lines *OK93* and *DPF88S* and fixed differences with *D. simulans* for the *G6pd* region

Classification	5′ Region 1	3′ Region 2	Total *G6pd*
Total synonymous ($n = 50$)	12	32	44
North American ($n = 21$)	3	11	14
East African ($n = 16$)	8	27	35
Diverged silent (*OK93 vs. DPF88S*)	20	23	43
Fixed silent	11	13	24
D. melanogaster lineage	9	8	17
D. simulans lineage	2	5	7
Fixed replacement	14	7	21

The data are partitioned into regions corresponding to nucleotides 542 to 1222 (region 1) and nucleotides 1223 to 2242 (region 2). Region 1 includes 61 and 51 bases from introns 2 and 3 as well as 121 "silent site equivalents."

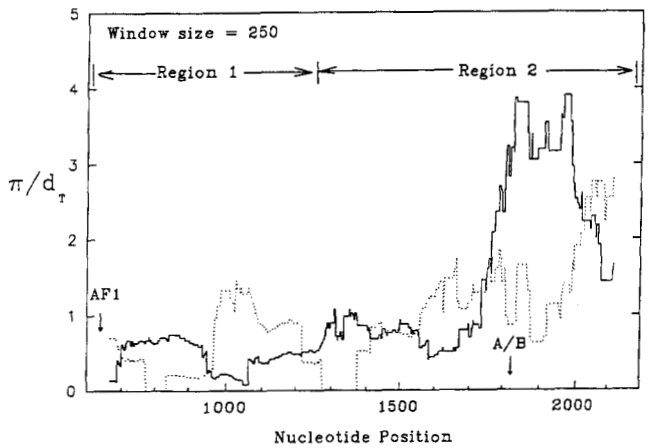

FIGURE 3.—The distribution in *Drosophila melanogaster* (solid line) and *D. simulans* (dotted line) of the ratio of the percent pairwise differences π over scaled interspecific divergence d_T. The plot is for sliding windows of 250 bases. *D. yakuba* is used for the estimate of divergence. The positions of the *A/B* and *AF1* allozyme polymorphisms are indicated.

HUDSON *et al.* 1994); however, it also appears that divergence slightly increases from the 5′ to 3′ ends (16:22), but this is not statistically significant ($X^2 = 0.95$, NS) using an expectation of equal silent divergence for equal numbers of effectively silent sites. Likewise the number of "fixed" silent sites is not different between 5′ and 3′ regions (12:12, $X^2 = 0.0$, NS). These results are consistent with a normal level of neutral mutation for the *G6pd* region, as well as equal rates across the *G6pd* region.

Figure 3 plots for *D. melanogaster* and *D. simulans* the ratio π/d_T (for sliding windows of 250 bp) across the *G6pd* region, where π is the nucleotide diversity (average heterozygosity per silent site; NEI 1987) and d_T is a scaled percent interspecific divergence across the same sequence. This scaling provides an "expected relative heterozygosity" generated from the observed percent divergence of each species with *D. yakuba* (assuming $T = 11.9$; see HUDSON 1990). Therefore, for windows with value 1.0, the observed heterozygosity is concordant with an expected heterozygosity derived from the scaled divergence in the window, assuming the infinite sites neutral model (KIMURA 1969). Values less than 1.0 suggest regions that have relatively low heterozygosity, as might be expected from recent directional selection events, while values greater than 1.0 might be caused by historical balancing selection. It should be emphasized that the sliding window analysis is for the purpose of data and pattern exploration, and not statistical testing. We used the divergence with *D. yakuba*, rather than between *D. melanogaster* and *D. simulans* because of the much larger number of changes and the larger contribution of fixed differences rather than polymorphisms to divergence. Multiple hits are corrected for by the Jukes-Cantor distance (JUKES and CANTOR 1969), but are expected to affect very few sites given this level of divergence.

For *D. melanogaster* this figure shows a large excess of polymorphism at the 3′ end in the region of the *A/B* polymorphism at nucleotide 1817, every much reminiscent of that observed for *Adh* around the Thr/Lys polymorphism (HUDSON and KAPLAN 1988). It should be noted that no such pattern is observed for the *AF1* site, the second allozyme polymorphism endemic to regions north of the Sahara. *D. simulans* shows a relatively low level of polymorphism across much of the gene, with a sharp rise at the 3′ end.

We have carried out HKA tests contrasting the 5′ and 3′ halves of *G6pd* with each other, as well each half and their total against the flanking 5′ *Adh* region (KREITMAN and HUDSON 1991). This noncoding region is 5′ to the *Adh* coding region and has been used in a number of studies as a neutral contrast locus. This partitioning results in nine tests: the entire *G6pd* regions for all the data, North American samples only and African samples only contrasted with the *Adh* 5′ region (three tests), the 3′ *G6pd* region contrasted with the *Adh* 5′ region (three tests) and the respective 5′ *G6pd* regions (three tests). The expected numbers of polymorphisms in each case are determined assuming that the effective population size of the *X*-linked *G6pd* region is three quarters that of the autosomal *Adh* region. Because the original sampling was stratified to emphasize sampling of rarer variants (*A* alleles in Africa and *B* alleles in Europe), "constructed random samples" (see HUDSON *et al.* 1994) were generated for each test. In the cases of contrasting the 3′ *G6pd* region with the flanking 5′ *Adh* region, the HKA tests are statistically significant for the total lines ($X^2 = 3.86$, $P < 0.05$) or the African lines ($X^2 = 4.36$, $P < 0.05$) separately. The remaining partitionings and contrasts are not significant. The total gene data for *D. simulans* was contrasted against data for the *D. simulans* *Est-6* (KAROTAM *et al.* 1995), *Adh-dup* (AKASHI 1995) and *Adh* (MCDONALD and KREITMAN 1991) regions. None were significant.

Monte Carlo simulations: Our simulations of the coalescence process were designed to test two null hypotheses. These are associated with the distribution of polymorphisms within or between two subsamples defined by association with an amino acid polymorphism. Both hypotheses are designated *a priori*, and in each case the amino acid polymorphism is the Pro → Leu mutation responsible for the allozyme polymorphism. The first null hypothesis is that the number of polymorphisms associated with the derived (*A*) allele is not significantly less than expected given a Wright-Fisher population model conditioned on the total number of observed polymorphisms and the frequency of the derived allele. The second is that the number of fixed differences between the allele sets (*A* and *B*) is not significantly greater than that expected from the same Wright-Fisher model and parameters.

Our tests were carried out separately on the three geographic collections. In each case the same con-

structed random samples were used as in the estimation of diversity parameters (π and θ). Neither the North American collection of 21 chromosomes nor the three east African constructed random samples of single *A* and 13 *B* alleles rejected the null expectations (only the question of number of fixed differences could be tested in the African samples), and probabilities were between $P = 0.1835$ and 0.8631. In contrast, all three European constructed random samples of 10 *A* (also includes *AF1* which are derived from *A*) alleles and a single *B* allele yielded probabilities of $P = 0.002$, 0.0004 and <0.0001 for 5, 8 and 13 "fixed" differences, respectively. It would appear that the European sample does not reflect an equilibrium population, but rather one where the derived *A* allele has become unexpectedly common.

Partitioning the rates of silent and replacement fixation between lineages: Using *D. yakuba* as an outgroup we have assigned silent and replacement fixation events to either the *D. melanogaster* or *D. simulans* lineages. The null expectation under neutral theory is that there will be parity between branches in the number of fixed differences. The first observation is that there is a two-fold difference in the rate of fixation at silent sites, with the *D. melanogaster* lineage possessing more than twice the number of silent fixations as the *D. simulans* lineage (17:7, $X^2 = 4.16$, $P < 0.05$; Table 5). The second observation is the remarkable difference in the number of amino acid fixations by lineage. There is a greater than two-fold higher rate of amino acid fixation in the *simulans* lineage (6:15, $X^2 = 3.86$, $P < 0.05$; Figure 1).

DISCUSSION

The major goal of this report is to describe within and between population levels of polymorphism, their associations, and to test specific hypotheses concerning historical selection on the contemporary amino acid polymorphisms. The collection of *D. simulans* sequences is effectively random. However, with respect to *D. melanogaster*, we are examining this locus with prior knowledge of, and special interest in, the segregating amino acid polymorphisms responsible for the *A*, *B* and *AF1* allozyme polymorphisms. In the absence of sequence information, a number of hypotheses could be proposed for the history of the *A*/*B* polymorphism. One simple hypothesis might propose that the *A* allele is actually a collection of dimer-forming mobility variants of heterogeneous molecular origin, as is the case for human *G6PD* deficiency polymorphisms (see review by VULLIAMY *et al.* 1992). This was rejected by EANES *et al.* (1993) because all the *A* allozyme sequences possess the Leu replacement. Also it might be proposed that given *D. melanogaster*'s Afrotropical origin, and relatively recent invasion of temperate regions, the temperate-associated *A* allele would be derived, and possibly possess lower sequence polymorphism. The *A* allele is clearly derived and possesses two- to fourfold lower sequence polymorphism than the *B* allele. Earlier restriction map studies also supported this scenario (EANES *et al.* 1989; MIYASHITA 1990).

The neighbor-joining (and UPGMA) algorithm groups all *A* alleles into a monophyletic class irrespective of geographic origin. African *A* alleles, which are quite rare, still cluster in this group, but also possess sequence polymorphisms consistent with their being recombinants with *B* sequences. The lower level of polymorphism associated with *A* sequences would suggest the Pro → Leu mutation is relatively recent, although it is also consistent with it simply possessing an historically smaller population number. If its appearance were recent, it would possess no more differences from *B* alleles than two *B* alleles sampled at random. However, the average number of differences between *A* and *B* sequences supports the interpretation that, while it has been the minority allele, it has persisted for sufficient time to collect fixed or nearly fixed differences. The *AF1* allele, which is found only north of the Sahara, appears to be a very recent derivative of the *A* allele, yet it reaches frequencies as high as 30% in Tunisia.

The allelic diversity among these three geographic samples portray very different histories. East African populations of *D. melanogaster* are clearly unique, and studies of restriction site (BEGUN and AQUADRO 1993), inversion (EANES *et al.* 1992), and mating isolation (WU *et al.* 1995) suggest that populations in the east African region have been isolated from cosmopolitan ones for some time. They may also represent a population at genetic equilibrium. The cosmopolitan and east African lines still share allozyme and inversion polymorphisms, but the *B* alleles are further diverged supporting the hypothesis that this is a unique lineage and that cosmopolitan flies, as typified by the North American samples, are derived from other African sources, possibly west African populations. The higher level of polymorphism is also consistent with east African populations possessing a historical population size that is several times larger than the populations from which the cosmopolitan alleles are derived and that have been subsequently sampled in North America. In contrast, European lineages possess low polymorphism because the predominant allele is *A* and it possesses low sequence diversity. Nevertheless, two of three European *B* sequences are distinctly different from those of North America and possess many of the polymorphisms seen in the east African lines. Our Monte Carlo simulations support the hypothesis that this is a nonequilibrium population where the *A* allele has risen, perhaps under selection (clinal data favors the *A* allele in temperate regions), to become the majority allele. The *A*/*B* polymorphism is the most prevalent in North America and that collection of 21 alleles does not appear to be simply a sample of the other two potentially originating populations. While the three *B* alleles in the European collection

segregate for 17 polymorphisms, the North American collection of 11 *B* alleles segregate for only 8 polymorphisms. It would appear that either the North American sample reflects a population that has undergone a significant bottleneck during founding, or the originating population is neither of the other two sampled here.

Although separated by 722 bp, there are no unequivocal examples of recombination between Δ_2 and the *A*/*B* polymorphism. All *A* alleles posses the Δ_1 arrangement and this insertion sequence was likely present in the sequence in which the Pro → Leu substitution first appeared, since this arrangement is polymorphic among *B* alleles sampled from the isolated east African lines. *In vivo* activity differences are associated with the *G6PD* allozyme polymorphism (EANES *et al.* 1990; LABATE and EANES 1992), and since there is an analogous insertion/deletion polymorphism correlated with ADH activity differences in the *Adh* region (LAURIE and STAM 1994), this site might be considered a candidate for the source of that allozyme activity difference. However, all evidence from this laboratory points to the activity variation being associated with different K_M's for glucose-6-phosphate, and no differences between *A* and *B* lines were observed in *G6PD* protein level (EANES *et al.* 1990). Therefore, there is no evidence that the Δ_1/Δ_2 arrangement has any functional significance. In *D. simulans* there is an analogous polymorphism at precisely the same site, although the derived insertion sequence is different. This is likely to be simply an inherently unstable region subject to replication slippage (LEVINSON and GUTMAN 1987).

There is substantial linkage disequilibrium among polymorphisms within the 1.7-kb *G6pd* region, and in particular there are strong associations with the allozyme polymorphism. This is apparent from both the pattern of correlation as well as the neighbor-joining clustering, which clearly separates clusters by allozyme type, a result that would not have occurred in the absence of strong disequilibrium. This disequilibrium is not simply due to the mixing of New York and California samples since each individual sample shows the same strong disequilibria, except that some are not statistically significant because of the smaller sample size associated with the partitioning.

LEWONTIN (1995) has recently raised the issue of reduced power to detect disequilibria when allele frequencies are skewed and the pervasive presence of singletons in some data sets makes this an issue. However, the issue here is not an inability to detect significant equilibria; the observation is the pervasive presence of disequilibria between most sites. The results are even more dramatic when the two singleton sites (642 and 2244) are excluded, as these sites contribute no significant associations to the total observation. Most of this disequilibrium is likely to have a mutational origin; the associations generated as a consequence of the original mutation capturing a single sequence, and for which there has been insufficient time for the resulting association to decay. This is in contrast to disequilibrium resulting strictly from a steady-state balance between the stochastic sampling of finite numbers of gametes and recombination. How to resolve these two processes in such data is not apparent.

While there is pervasive linkage disequilibrium, there is no reason to propose that recombination is unusual for the *G6pd* region. Coefficients of exchange are normal if not higher than average in the *G6pd* region (LINDSLEY and SANDLER 1977). Using the estimated level of recombination between *car* (18D1-2) and *sw* (19B3) as 2.23 centiMorgans (EANES 1983), and recognizing there are 30 polytene bands in this interval (where the estimated amount of DNA per band is about 25 kb), we directly estimate a level of recombination of about 3×10^{-8} per base pair. This may be compared to an indirect estimate obtained from assuming a steady-state neutral population model that estimates the population parameter $C = 4Nc$ (where N is the effective population size and c is the per base recombination rate) from the variance in the number of pairwise differences (HUDSON 1987). Restricting this analysis to the North American sample we estimate the locus-specific value of $4Nc$ as 35.75 and the per base estimate of $4Nc$ as 0.022. Assuming that the *D. melanogaster* effective population size is on the order of 10^6, and adjusting for the *X* linkage of *G6pd* and its recombination bias in females (two thirds of the chromosomes are in females), we estimate c as 1.1×10^{-8} per base, which is smaller than our direct estimate. Of course this may simply result from an overestimated N. Another relative measure, less sensitive to population size estimation, is the ratio of C/θ (ultimately c/μ), which is 0.022/0.013 = 1.7. This ratio for *G6pd* is comparable to the estimates for the *Adh* (c/μ = 1.6, HUDSON 1987) and *Sod* (c/μ = 0.8, HUDSON *et al.* 1994) regions, but an order of magnitude lower that estimated for the *Mlc1* (c/μ = 13.4, LEICHT *et al.* 1995) region. One explanation for the relatively high value for *Mlc1* region might be due to an intrinsically lower level for μ because most of the polymorphisms in *Mlc1* involve intron sequences that typically have lower levels of divergence. These estimates assume a neutral Wright-Fisher population at steady-state equilibrium, but historical selection as already implicated for all three loci will confound estimates in $4Nc$. Finally, we can also estimate $C = 4Nc$ for our *D. simulans* data for 12 sequences of *G6pd*, and this is estimated as C = 118.55. This three- to fourfold higher estimate is consistent with both the proposed larger effective population size of *D. simulans* and the higher rate of recombination (see TRUE *et al.* 1996), although these may be confounded if background selection is a significant factor in determining the levels of polymorphism and linkage disequilibrium in natural populations (CHARLESWORTH *et al.* 1995; HUDSON and KAPLAN 1995).

One apparent feature is the uneven distribution of silent polymorphisms across the *G6pd* coding region in both species. In *D. melanogaster* this might be explained by balancing selection on the Leu/Pro polymorphism and its effect on linked variation in the 3′ region. However, aside from the African data, we have been unable to reject specific statistical models based on neutrality. Furthermore, the east African HKA test involving the *Adh* 5′ flanking region as a contrast is not valid. East African populations are exceptionally polymorphic (BEGUN and AQUADRO 1993), and no comparable east African data exists for the *Adh* 5′ flanking region. Intrinsic to the HKA test is the assumption that samples for both contrasting loci are (in the absence of selection) drawn from the same potential genealogical histories. The absence of east African lines from the *Adh* 5′ set violate this assumption, and the significant HKA test is suspect as evidence for balancing selection.

Using *D. yakuba* as an outgroup to determine the polarity of the differences between the *D. melanogaster* and *D. simulans* lineages, we have observed a significant lineage-specific difference in the rate of silent substitution, where *D. melanogaster* shows the highest silent rate. If silent mutations are strictly neutral, observed lineage-dependent differences in silent fixation are difficult to explain unless there is a substantial generation-time difference between species, which seems unlikely. However, if many silent mutations are slightly deleterious (as codon usage bias indicates) then the species with the smaller lineage-specific population size would possess the higher rate of substitution. This is because the probability of fixation of a deleterious mutation increases with decreasing population size. Based on restriction map data it has been suggested that *D. melanogaster* possesses a three- to sixfold smaller effective population size than *D. simulans* (AQUADRO 1992), so the difference in long-term fixation rates is consistent with the prediction based on contemporary levels of polymorphism. Furthermore, AKASHI (1995) recently reported a substantially elevated silent fixation rate in the *D. melanogaster* lineage, for five other genes where polarity of change could be established. To examine this further, he also classified each derived silent mutation as "preferred" or "unpreferred," depending on the pattern of overall codon usage in Drosophila (SHIELDS *et al.* 1988). It is expected that for a population of historically stable size the ratio of changes for these two classes will be unity, while in a population of historically diminished size it would show an excess of unpreferred (deleterious) fixations until a new equilibrium is reached (codon usage becomes less biased). For *G6pd, D. simulans* shows a unpreferred:preferred ratio of 4:3, while the ratio is 14:3 in *D. melanogaster*. These data when pooled with the those reported by AKASHI (1995) on the numbers of unpreferred to preferred codon fixations (by lineage) give total unpreferred:preferred codon ratios of 9:10 for *D. simulans* and 46:8 for *D. melanogaster*. Therefore, *D. melanogaster* mostly shows fixation of unpreferred, or slightly deleterious synonymous mutations. As suggested by AKASHI (1995), this departure from expected unity for *D. melanogaster* is consistent with the proposed smaller population size of *D. melanogaster* being a long-term feature of its history since separation from the common ancestor with *D. simulans*. This *apparent* reduction in population size could also be generated by increased background selection (see CHARLESWORTH *et al.* 1995; HUDSON and KAPLAN 1995) assuming the *D. melanogaster* lineage has had a long term reduction in its genome-wide level of recombination, as proposed by the studies of TRUE *et al.* (1996) in the *D. simulans* complex.

The observation with respect to amino acid changes contrasts with that seen for silent changes. The major observation in EANES *et al.* (1993) was an excess of fixed amino acid replacement fixations relative to numbers of silent fixations, when contrasted with the same ratio for replacement and silent polymorphisms (MCDONALD and KREITMAN 1991). This is consistent with mechanistic processes involving the fixation of adaptive replacement substitutions. An obvious question emerging from that study was whether these replacement fixations were concentrated in either lineage, and this required a *G6pd* sequence from an outgroup. Our partitioning by lineage shows that 15 of 21 replacement fixations are in the *D. simulans* lineage. If, as proposed by OHTA (1972, 1992), most amino acid fixations involve deleterious mutations, then *D. melanogaster*, which is presumed to have the historically smaller populations size, should show the greater rate of fixation. Since the rate is substantially greater in the *D. simulans* lineage, this is inconsistent with any model where most amino acid substitutions are deleterious; it is consistent with a higher rate of advantageous substitution for the lineage with the larger population size. However, it would appear that *G6pd* may be the exception in this regard. If one examines the polarity of the replacement substitutions between *D. melanogaster* and *D. simulans* for the *Adh* (MCDONALD and KREITMAN 1991), *Adhr* (*Adh-dup* in KREITMAN and HUDSON 1991; JEFFS *et al.* 1994), *per* (THACKERAY and KYRIACOU 1990; HEY and KLIMAN 1993) *Pgi* (J. MCDONALD, personal communication), and *boss* (AYALA and HARTL 1993) loci, there are a total of only 13 fixed amino acid differences, and the substitution pattern is in favor of replacement fixations in the *Drosophila melanogaster* lineage (11:2, $X^2 = 6.23$, $P < 0.025$). This is significantly different from the ratio in *G6pd* ($G = 10.84$, $P < 0.005$), and would argue that such amino acid substitutions are deleterious in these genes. Therefore, *G6pd* may well have undergone an episodic burst of adaptive mutation (*sensu* GILLESPIE 1991) in the early history of the *D. simulans* lineage (see below), and this burst runs counter to the typical pattern of amino acid substitution between the species, where most amino acid substitutions are in the *D. melanogaster* lineage.

Both of these lineage-specific observations direct attention to the nearly neutral theory (OHTA 1972, 1992) and its variations (*e.g.*, TAKAHATA 1987; KIMURA 1979, 1981; TACHIDA 1991) as an explanation for the evolution of both silent and replacement changes. OHTA (1972) introduced her theory to explain the narrow range of allozyme heterozygosities across taxa relative to the assumed large range of taxa population sizes, but it has also been proposed as an explanation for codon bias, where mutations are presumed to possess selection coefficients on the order of $1/N_e$. With respect to amino acid substitutions, and their general overdispersion in many analyses (LANGLEY and FITCH 1974; GILLESPIE and LANGLEY 1979; GILLESPIE 1988), the possibility of evolution via strictly deleterious mutation has been strongly criticized by GILLESPIE (1991, 1994) on the basis of a number of issues. Among others, these include the lack of the biological realism associated with the assumed distributions of selection coefficients affiliated with new mutations, and the narrow range of population sizes (assuming the selection intensities are comparable) over which the observed variation in amino acid substitution rates are compatible, relative to likely ranges of population size for humans, *Drosophila,* and *Escherichia coli.* The burst of replacements early in the *D. simulans* lineage is not consistent with fixation of weakly deleterious mutation. However, with respect to silent variation a nearly neutral model may be appropriate. In Drosophila the estimated parameter Ns is within the range where several-fold differences in N will affect codon bias and the rate of fixation of preferred and unpreferred codons, although as pointed out by AKASHI (1995), this may become a problem when broader ranges of N are considered. As suggested by AKASHI it would also appear that *D. simulans* is near equilibrium where unpreferred fixations are compensated for by preferred mutations. In contrast, it appears that *D. melanogaster* is not at equilibrium, either because of relaxed selection or reduced population size.

Finally, the level of silent *G6pd* polymorphism in *D. simulans* ($\hat{\theta} = 0.013$) is substantially lower than seen for eight of nine (mean of all nine is $\hat{\theta} = 0.037$) other genes in this species found in regions of normal levels of recombination (see MORIYAMA and POWELL 1996). While the polymorphism predominates in the 3′ end of the gene, the 15 amino acid substitutions in *D. simulans* predominate in the 5′ region and are clustered (Figures 1 and 3). One hypothesis would be that some of these substitutions have depressed the polymorphism level in the 5′ region in *D. simulans.* Using *D. mauritiana* and *D. sechellia* sequences as outgroups (W. F. EANES, unpublished results) it can be shown that only one substitution (Gln → Arg at position 967) occurred since those species diverged from *D. simulans.* Therefore, most of the substitutions must have occurred earlier in the period following divergence and prior to the splitting off of the *D. mauritiana* and *D. sechellia* lineages (barring they were not still polymorphic at that time). In order for selective hitchhiking to have a depressing effect on contemporary levels of neutral polymorphism, linked adaptive fixations must occur within the last $4N$ generations, since that is the time to the common ancestor of all $2N$ copies. Fixations prior to that time will have no impact on linked neutral polymorphism. In *D. melanogaster* with restriction map based average heterozygosity of 0.004 ($4Nu$) and *mel-sim* divergence of 0.0296 (AQUADRO 1992), only fixations within the last 25% of time since separation could in theory impact linked neutral polymorphism. However, in *D. simulans* with an average heterozygosity several-fold higher (0.014), neutral genealogies will in principle coalesce near the time of divergence between the species, and a majority of the observed 15 replacement fixation events could potentially contribute to reducing neutral polymorphism below the level expected in its absence. Under hitchhiking the ability to reduce variation depends on the strength of selection associated with each fixed mutation, the level of recombination in the region, and population size (KAPLAN *et al.* 1989). For example, assuming a population size of $2N = 10^8$, a recombination rate of 10^{-8} per base, a single fixation event would substantially reduce the polymorphism from a span of 200 to 2000 bp if associated selection coefficients are on the order of 10^{-4} to 10^{-3} (see KAPLAN *et al.* 1989). The 15 amino acid replacements specific to this lineage may have reduced the polymorphism in this region. The fact that *tra* shows very little synonymous polymorphism in *D. melanogaster* (WALTHOUR and SCHAEFFER 1994), yet is one of the most diverged genes at the amino acid level yet described (O'NEIL and BELOTE 1992) may be causally connected.

We thank MARTY KREITMAN for his advice and interest in this project during its early development. Thanks are also extended to CHIP AQUADRO, DAVE BEGUN, ROLLIN RICHMOND, and RICK ROUSH for supplying numerous fly lies. CEDRIC WESLEY helped in the development of a number of new molecular methods in the lab. DAN DYKHUIZEN kindly commented on an early version of the manuscript, and two anonymous reviewers added valuable criticism in revision. This research was supported by U.S. Public Health Service grant GM-45247 to W.F.E. This is contribution number 908 from the Graduate Program in Ecology and Evolution, State University of New York at Stony Brook.

LITERATURE CITED

AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076.

AQUADRO, C. F., 1992 Why is the genome variable? Insights from Drosophila. Trends Genet. **8:** 355–362.

AYALA, F. J., and D. L. HARTL, 1993 Molecular drift of the *bride-of-sevenless* (*boss*) gene in *Drosophila.* Mol. Biol. Evol. **10:** 1030–1040.

BEGUN, D. J., and C. F. AQUADRO, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA 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.

CAVENER, D. R., and M. T. CLEGG, 1981 Evidence for biochemical and physiological differences between genotypes in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **78:** 4444–4447.

CHARLESWORTH, D., B. CHARLESWORTH and M. T. MORGAN, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141:** 1619–1632.

EANES, W. F., 1983 Genetic localization and sequential electrophoresis of *G6pd* in *Drosophila melanogaster*. Biochem. Genet. **21:** 703–711.

EANES, W. F., 1987 Allozymes and fitness: evolution of a problem. Trends Ecol. Evol. **2:** 44–48.

EANES, W. F., J. W. AJIOKA, J. HEY and C. WESLEY, 1989 Restriction map variation associated with the *G6PD* polymorphism in natural populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6:** 384–397.

EANES, W. F., L. KATONA and M. LONGTINE, 1990 Comparison of *in vitro* and *in vivo* activities associated with the *G6PD* allozyme polymorphism in *Drosophila melanogaster*. Genetics **125:** 845–853.

EANES, W. F., C. WESLEY and B. CHARLESWORTH, 1992 Accumulation of P elements in minority inversions in natural populations of *Drosophila melanogaster*. Genet. Res. **59:** 1–9.

EANES, W. F., M. KIRCHNER and J. YOON, 1993 Evidence for adaptive evolution of the *G6PD* gene in the *Drosophila melanogaster* and *D. simulans* lineages. Proc. Natl. Acad. Sci. USA **90:** 7475–7479.

EANES, W. F., M. KIRCHNER, D. R. TAUB, J. YOON and J. CHEN, 1996 Amino acid polymorphism and rare electrophoretic variants of *G6PD* from natural populations of *Drosophila melanogaster*. Genetics **143:** 401–406.

FOUTS, D., R. GANGULY, A. G. GUTIERREZ, J. C. LUCCHESI and J. E. MANNING, 1988 Nucleotide sequence of the *Drosophila* glucose-6-phosphate dehydrogenase gene and comparison with the homologous human gene. Gene **63:** 261–275.

FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

GANGULY, R., N. GANGULY and J. E. MANNING, 1985 Isolation and characterization of the glucose-6-phosphate dehydrogenase gene in *Drosophila melanogaster*. Gene **35:** 91–101.

GILLESPIE, J. H., 1988 More on the overdispersed molecular clock. Genetics **118:** 385–386.

GILLESPIE, J. H., 1994 Substitution processes in molecular evolution. III. Deleterious alleles. Genetics **138:** 943–952.

GILLESPIE, J. H., 1991 *The Causes of Molecular Evolution*. Oxford University Press, New York.

GILLESPIE, J. H., and C. H. LANGLEY, 1979 Are evolutionary rates really variable? J. Mol. Evol. **13:** 27–34.

HEY, J., and R. M. KLIMAN, 1993 Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Mol. Biol. Evol. **10:** 804–822.

HIGUCHI, R. G., and H. OCHMAN, 1989 Production of single-stranded DNA templates by exonuclease digestion following the polymerase chain reaction. Nucleic Acids Res. **17:** 5865.

HILL, W. G., 1974 Estimation of linkage disequilibrium in randomly mating populations. Heredity **33:** 229–239.

HUDSON, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Series in Evolutionary Biology*, Vol. 7, edited by D. FUTUYMA and J. ANTONOVICS. Oxford University Press, New York.

HUDSON, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Japan Scientific Press, Tokyo, and Sinauer Associates, Inc., Sunderlund, MA.

HUDSON, R. R., and N. L. KAPLAN, 1986 On the divergence of alleles in nested subsamples from finite populations. Genetics **113:** 1057–1076.

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

HUDSON, R. R., and N. L. KAPLAN, 1995 Deleterious background selection with recombination. Genetics **141:** 1605–1617.

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

HUDSON, R. R., K. BAILEY, D. SKARECKY, J. KWIATOWSKI and F. J. AYALA, 1994 Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster. Genetics **136:** 1329–1340.

JEFFS, P. S., E. C. HOLMES and M. ASHBURNER, 1994 The molecular evolution of the alcohol dehydrogenase and alcohol dehydrogenase-related genes in the *Drosophila melanogaster* species subgroup. Mol. Biol. Evol. **11:** 287–304.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism* III, edited by H. N. MUNRO. Academic Press, New York.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

KAPLAN, N. L., T. DARDEN and R. R. HUDSON, 1988 The coalesence process in models with selection. Genetics **120:** 819–829.

KAROTAM, J., T. M. BOYCE and J. G. OAKESHOTT, 1995 Nucleotide variation at the hyperactive esterase 6 isozyme locus of *Drosophila simulans*. Mol. Biol. Evol. **12:** 113–122.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a population due to steady flux of mutations. Genetics **61:** 893–903.

KIMURA, M., 1979 Model of effectively neutral mutatiopns in which selective constraint is incorporated. Proc. Natl. Acad. Sci. USA **76:** 3440–3444.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

KREITMAN, M., and M. AGUADÉ, 1986 Excess polymorphism at the *Adh* locus in *Drosophila melanogaster*. Genetics **114:** 93–110.

KREITMAN, M., and R. R. HUDSON, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

LABATE, J., and W. F. EANES, 1992 Direct measurement of *in vivo* flux differences between electrophoretic variants of *G6PD* from *Drosophila melanogaster*. Genetics **132:** 783–787.

LANGLEY, C. H., and W. M. FITCH, 1974 An estimation of the constancy of the rate of molecular evolution. J. Mol. Evol. **3:** 161–177.

LAURIE, C. C., and L. F. STAM, 1994 The effect of an intronic polymorphism on alcohol dehydrogenase expression in *Drosophila melanogaster*. Genetics **138:** 379–385.

LEICHT, B. G., S. V. MUSE, M. HANCZYC and A. G. CLARK, 1995 Constraints on intron evolution in the gene encoding the myosin alkali light chain in Drosophila. Genetics **139:** 299–308.

LEVINSON, G., and G. A. GUTMAN, 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

LEWONTIN, R. C., 1995 The detection of linkage disequilibrium in molecular sequence data. Genetics **140:** 377–388.

LINDSLEY, D. L., and L. SANDLER, 1977 The genetic analysis of meiosis in female *Drosophila melanogaster*. Phil. Trans. Soc. Lond. B. **277:** 295–312.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive evolution at the Adh locus in Drosophila. Nature **351:** 652–654.

MCGINNIS, W., A. W. SHERMOEN and S. K. BECKENDORF, 1983 A transposable element inserted just 5′ to a Drosophila glue gene alters gene expression and chromatin structure. Cell **34:** 75–84.

MIYASHITA, N. T., 1990 Molecular and phenotypic variation of the *Zw* locus region in *Drosophila melanogaster*. Genetics **125:** 407–419.

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

NEI, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

OAKESHOTT, J. G., G. K. CHAMBERS, J. B. GIBSON, W. F. EANES and D. A. WILLCOCKS, 1983 Geographic variation in *G6pd* and Pgd allele frequencies in *Drosophila melanogaster*. Heredity **50:** 67–72.

OHTA, T., 1972 Population size and the rate of evolution. J. Mol. Evol. **1:** 305–314.

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

O'NEIL, M. T., and J. M. BELOTE, 1992 Interspecific comparison of the transformer gene of Drosophila reveals an unusually high degree of evolutionary divergence. Genetics **131:** 113–128.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new

method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

SCHAEFFER, S. W., and E. L. MILLER, 1992 Estimates of gene flow in *Drosophila pseudoobscura* determined from nucleotide sequence analysis of the alcohol dehydrogenase region. Genetics **132:** 471–480.

SCHAEFFER, S. W., and E. L. MILLER, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura.* Genetics **135:** 541–552.

SHEEN, J., and B. SEED, 1988 Electrolyte gradient gels for DNA sequencing. Bio Techniques **6:** 942–944.

SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS and F. WRIGHT, 1988 "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

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

TAKANO, T. S., S. KUSAKABE and T. MUKAI, 1993 DNA polymorphisms and the origin of protein polymorphisms at the Gpdh locus of *Drosophila melanogaster,* pp. 179–190 in *Mechanisms of Molecular Evolution,* edited by N. TAKAHATA and A. G. CLARK. Sinauer Associates, Sunderlund, MA.

THACKERAY, J. R., and C. P. KYRIACOU, 1990 Molecular evolution in the *Drosophila yakuba period* locus. J. Mol. Evol. **31:** 389–401.

TACHIDA, H., 1991 A study of the nearly neutral muttaion model in finite populations. Genetics **128:** 183–192.

TAKAHATA, N., 1987 On the overdispersed molecular clock. Genetics **116:** 169–179.

TRUE, J. R., J. M. MERCER and C. C. LAURIE, 1996 Differences in crossover frequency and distribution among three sibling species of Drosophila. Genetics **142:** 507–523.

VUILLIAMY, T. J., P. J. MASON and L. LUZZATTO, 1992 The molecular basis of glucose-6-phosphate dehydrogenase deficiency. Trends Genet. **8:** 138–143.

WALTHOUR, C. S., and S. W. SCHAEFFER, 1994 Molecular population genetics of sex determination genes—the transformer gene of *Drosophila melanogaster.* Genetics **136:** 1367–1372.

WATT, W. B., 1994 Allozymes in evolutionary genetics—self-imposed burden or extraordinary tool? Genetics **136:** 11–16.

WU, C.-I., H. HOLLOCHER, D. J., BEGUN, C. F., AQUADRO, Y. XU *et al.,* 1995 Sexual isolation in Drosophila melanogaster: a possible case of incipient speciation. Proc. Natl. Acad. Sci. USA **92:** 2519–2523.

YOUNG, W. J., J. E. PORTER and B. CHILDS, 1964 Glucose-6-phosphate dehydrogenase in Drosophilia: X-linked electrophoretic variants. Science **143:** 140–141.

Communicating editor: A. G. CLARK