Copyright © 2001 by the Genetics Society of America

Clinal Variation for Amino Acid Polymorphisms at the *Pgm* Locus in *Drosophila melanogaster*

Brian C. Verrelli and Walter F. Eanes

Department of Ecology and Evolution, State University of New York, Stony Brook, New York 11794-5245

Manuscript received September 18, 2000
Accepted for publication December 26, 2000

ABSTRACT

Clinal variation is common for enzymes in the glycolytic pathway for *Drosophila melanogaster* and is generally accepted as an adaptive response to different climates. Although the enzyme phosphoglucomutase (PGM) possesses several allozyme polymorphisms, it is unique in that it had been reported to show no clinal variation. Our recent DNA sequence investigation of *Pgm* found extensive cryptic amino acid polymorphism segregating with the allozyme alleles. In this study, we characterize the geographic variation of *Pgm* amino acid polymorphisms at the nucleotide level along a latitudinal cline in the eastern United States. A survey of 15 SNPs across the *Pgm* gene finds significant clinal differentiation for the allozyme polymorphisms as well as for many of the cryptic amino acid polymorphisms. A test of independence shows that pervasive linkage disequilibrium across this gene region can explain many of the amino acid clines. A single *Pgm* haplotype defined by two amino acid polymorphisms shows the strongest correlation with latitude and the steepest change in allele frequency across the cline. We propose that clinal selection at *Pgm* may in part explain the extensive amino acid polymorphism at this locus and is consistent with a multilocus response to selection in the glycolytic pathway.

CHARACTERIZING the explicit nature of amino acid variation and not just its level is necessary to understand the mechanisms behind protein evolution and its role in adaptation. Statistical approaches utilizing intra- and interspecific comparisons at the DNA level have revealed patterns of variation consistent with balancing selection acting on amino acid polymorphism (Kreitman and Hudson 1991; Takahata *et al.* 1992; Wayne *et al.* 1996; Katz and Harrison 1997; Terauchi *et al.* 1997). Likewise some genes, particularly those associated with mitochondrial variation, have shown an apparent excess of amino acid polymorphism (Ballard and Kreitman 1994; Nachman *et al.* 1996; Rand and Kann 1996; Hasegawa *et al.* 1998; Kennedy and Nachman 1998; Nielsen and Weinreich 1999). While consistent with diversifying selection, this pattern is also expected under a slightly deleterious model of molecular evolution that permits amino acid mutations to be polymorphic, but limits their fixation (Ohta 1992, 1996). A complementary approach that begins to focus more specifically on the amino acid polymorphisms *per se* is to compare the geographic variation of the locus under potential selection with that of neutral markers. If the different classes of polymorphisms exhibit different patterns of geographic variation, this is evidence for natural selection (Berry and Kreitman 1993; McDonald 1994; Pogson *et al.* 1995; McDonald *et al.* 1996; Salamon *et al.* 1999; Schmidt and Rand 1999). The most compelling evidence in favor of selection comes from studies of glycolytic enzymes in natural populations of *Drosophila melanogaster* (see review by Eanes 1999). Given that there is evidence for substantial gene flow for this species (David and Capy 1988; Hale and Singh 1991; Berry and Kreitman 1993; our unpublished data), the observation of reciprocal clines on several continents is generally accepted as adaptation to an environmental gradient (Oakeshott *et al.* 1981, 1982, 1983, 1984; David and Capy 1988). Identifying and characterizing latitudinal clines for enzymes in the glycolytic pathway can help determine how natural populations adapt to environmental diversity and further elucidate the role specific amino acid polymorphisms play in this adaptation in a more functional and physiological context.

Verrelli and Eanes (2000) recently reported an apparent excess of amino acid polymorphisms for the glycolytic enzyme phosphoglucomutase (*Pgm*) in *D. melanogaster*. In addition to the polymorphisms responsible for the three common allozyme alleles, their study discovered many electrophoretically cryptic, but common, amino acid polymorphisms. Although this might have been predicted from thermostability studies on the allozyme alleles (Trippa *et al.* 1976, 1978), it is unexpected given the absence of amino acid divergence for this gene when compared to *D. simulans*. For *D. melanogaster* loci, significant McDonald-Kreitman tests always show an excess of amino acid fixations (McDonald and Kreitman 1991; Eanes *et al.* 1993; Moriyama and Powell 1996), whereas the excess variation at the *Pgm* locus

Corresponding author: Brian C. Verrelli, Department of Biology, University of Maryland, College Park, MD 20742.
E-mail: verrelli@wam.umd.edu

could be associated with adaptive amino acid polymorphisms. Given these issues, an important contribution would be to study geographic patterns associated with this variation.

For *D. melanogaster*, geographic variation in allozyme polymorphisms is common for glycolytic enzymes (see Eanes 1999). However, early surveys of PGM allozyme variation found no geographic variation (Oakeshott *et al.* 1981). In retrospect, this is not surprising since our results show PGM allozyme mobility classes are extensive mixtures of amino acid polymorphisms (Verrelli and Eanes 2000). The question now emerges if amino acid variation is more explicitly resolved, whether the emerging *Pgm* alleles will exhibit clinal variation. The model for this approach is that of Berry and Kreitman (1993) where nucleotide variation associated with the *Adh-F/S* polymorphism in *D. melanogaster* was investigated along a latitudinal cline. Their study discovered an insertion/deletion polymorphism in the 5′ adult intron of *Adh* that is in linkage disequilibrium with *Adh-F* and shows a stronger cline than the *Adh-F/S* polymorphism. Their results suggest that epistatic selection on both the intron in/del and the *Adh-F/S* polymorphisms is producing the latitudinal cline at this locus.

In this study, we survey 10 populations of *D. melanogaster* along the Atlantic coast of the United States for the 12 *Pgm* amino acid polymorphisms reported by Verrelli and Eanes (2000). We intend to characterize any potential cline at the nucleotide level to address two issues. First, we are interested in determining whether any of the allozyme alleles exhibit clinal variation now that they are better resolved at the nucleotide level. If so, patterns of PGM allozyme variation are consistent with other *D. melanogaster* enzymes, a trend that argues for a multilocus response to selection along an environmental gradient for enzyme variation in the glycolytic pathway (Eanes 1999). Second, we were interested in whether any of the cryptic amino acid polymorphisms underlying the allozyme mobility alleles also show evidence for clinal variation. Diversifying selection, reflected in the form of geographic variation, could maintain intraspecific amino acid polymorphism at the *Pgm* locus, while limiting amino acid divergence. Therefore, the presence of clinal variation would support the results of Verrelli and Eanes (2000), that amino acid polymorphism at the *Pgm* locus is selectively favored and reflects adaptive protein evolution in another glycolytic enzyme in *D. melanogaster*.

MATERIALS AND METHODS

Origin of wild lines: In excess of 100 *D. melanogaster* isofemale lines were collected from each of 10 populations along the Atlantic coast of the United States in 1997. Population summaries are listed in Table 1. Flies were collected directly from rotting fruit at apple, peach, watermelon, and orange farms by sweep netting, and all isofemale lines were immediately established in the field.

Allozyme survey: Single flies from each isofemale line (~300 alleles per population) were assayed for their PGM allozyme genotype. The 12% starch gel (S-5651; Sigma, St. Louis) electrophoresis conditions were as follows: electrode and gel buffer: 41 mm Tris, 6 mm boric acid, pH 8.5. 14 × 8-inch gels were run for 5 hr (with ice) at 4° at 15 mA and 900 V. Staining procedures were as follows: 100 mg glucose-1-phosphate, disodium salt (Sigma G-1259), 1.0 ml 10 mg/ml $MgCl_2$, 18 mg β-NADP, ~5 mg each of MTT and PMS, 70 units glucose-6-phosphate dehydrogenase, and bring to total volume of 50 ml with 0.1 m Tris-HCl, pH 7.5. A total of 50 ml of a 2% agar solution was added for a 1% agar overlay. Gels were incubated in the dark in 37° ovens for 1–3 hr before visualization. This protocol follows Hjorth (1970), with adjustments to the buffer for a more precise separation of alleles. In addition to the allozyme data from the 10 populations collected in 1997, frequency data from a Mt. Sinai, New York, population (1995) was added to this study. Table 1 lists the frequencies for the three common PGM allozymes and the total number of alleles screened for each of the 11 populations.

Single nucleotide polymorphism analysis: Fifty isofemale lines from each of the 10 populations collected in 1997 were made homozygous for the third chromosome after three generations using the *TM3/TM6* balancer chromosome. All 500 extracted third chromosome lines were assayed for their PGM allele by starch gel electrophoresis as above. The entire 2354-bp *Pgm* gene was amplified from all 500 extracted third chromosome lines in 10-μl volumes in an Idaho Technologies (Idaho Falls, ID) Air-Thermo-Cycler by PCR from single-fly CTAB genomic preps (Winnepenninckx *et al.* 1993). PCR products were excised from 2% agarose gels and used as template for 50-μl reamplifications using internal primers. DNA fragments were purified from PCR products (Prep-A-Gene kit; Bio-Rad, Richmond, CA) and double-stranded DNA templates were manually sequenced using the Sequenase kit (United States Biochemical, Cleveland) and [^{35}S]dATP (Amersham, Arlington Heights, IL). A simple single nucleotide polymorphism (SNP) screen was performed as follows. Because the amino acid polymorphism at nucleotide site 25 is a G → A mutation, the single A nucleotide for each individual is sequenced for the region spanning the SNP at nucleotide site 25. These single nucleotide reactions were run side-by-side on standard acrylamide gels with an electrolyte gradient and electrophoresed for 2–3 hr. On one gel as many as 96 individuals can be screened for the presence (or absence) of this SNP and any other polymorphisms that involve the A nucleotide and occur along the region sequenced (sequencing reaction can span 300–400 bp). One individual, for which all four bases are sequenced for the region spanning the SNP being scored, is run alongside the single base pair reactions for sequence alignment. Using this approach, we first screened 50 alleles from each of two northern populations (VT97 and MA97) and two southern populations (MFL97 and HFL97) for the 12 amino acid polymorphisms previously reported by Verrelli and Eanes (2000).

Linkage disequilibrium analysis: The initial survey of variation at the *Pgm* locus by Verrelli and Eanes (2000) detected significant linkage disequilibrium across the entire 2354-bp gene in 22 allele copies. We were interested in associations between SNPs within and among populations as a factor in generating parallel latitudinal clines. The populations compared here may not be discrete breeding units but rather simple "subpopulations" of a larger quasi-panmictic unit. We are interested in whether associations between SNPs within populations are similar to associations between SNPs among populations. The Rozas and Rozas (1999) DnaSP program was used to compute the correlations for all possible pairwise comparisons between SNPs. Significant associations were iden-

tified at the 5% level using chi-square tests with a Bonferroni correction for multiple comparisons. For each pairwise comparison, an estimate of linkage disequilibrium (R^2) was computed for each of the 10 subpopulations and then averaged over all 10 populations. This is compared to the estimate based on all 500 individuals pooled from the 10 subpopulations.

Analysis of the cline: We used two analyses to investigate patterns of geographic variation. First, we used Nei's G'_{ST} (1986) as a relative measure of genetic differentiation among subpopulations. This was computed for the allozyme variation, the SNPs, and the protein haplotypes and is generally an unbiased estimator of G_{ST} given sufficiently large and equal samples of alleles across subpopulations (McDonald 1994). Second, we used a linear regression analysis to measure the association of allele frequency with population sample latitude. Frequencies were arcsine transformed (Sokal and Rohlf 1995) and regressed on population latitude to determine statistical significance. If selection on *Pgm* variation is responding to an environmental gradient covarying with latitude, then we expect allele frequencies to show a latitudinal cline. We analyzed the SNP data on both a site-by-site and a haplotype-by-haplotype basis. A site-by-site analysis treats each nucleotide polymorphism as a separate locus; however, because of the association between sites, geographic variation at one nucleotide site can cause another site in close linkage to exhibit apparent geographic variation. We were interested in testing for "independent" clinal sites, where clinal variation for one SNP cannot be explained by its association with another clinal SNP in our sample.

Combining sites into haplotypes treats entire amino acid sequences as alleles. Haplotypes with stronger latitudinal clines than their composite single site polymorphisms implies the presence of epistatic selection or recent selection on an unscreened linked site. We were interested in determining whether protein haplotypes exhibit geographic variation and if they can better explain any geographic pattern found for individual SNPs. We analyzed the haplotype data with a linear regression analysis and a Monte Carlo sampling as described below.

Monte Carlo sampling: After performing the linear regression of transformed allele frequency on population latitude for each SNP, we used a Monte Carlo sampling to assess the probability that significant geographic variation at one site can be explained by linkage disequilibrium with another site in our sample. The basic design of this analysis is adopted from Berry and Kreitman (1993).

For an example, consider polymorphic nucleotide sites X and Y. We can hold the frequency of the A allele at site X in each population constant and recalculate the frequency of the B allele at site Y for each population based on its association with the A allele at site X in the pooled data set of 500 alleles. For example, the B allele at site Y is found 10% of the time when the A allele at site X is present and is found 20% of the time when the A allele at site X is absent, for an overall frequency of 30% in the pooled data set. If the A allele at site X is present in 30 individuals out of 50 in the k_1 population, then the expected frequency of the B allele at site Y in the k_1 population is $(0.1 \times 30) + (0.2 \times 20) = 7/50$. To determine whether sampling 50 alleles from each population provides sufficient power to detect a significant latitudinal cline, we performed a binomial sampling of 50 alleles around this expected frequency of 7/50 and obtained 1000 simulated data sets each of $n = 50$. This resulted in 1000 simulated frequencies for the B allele at site Y in the k_1 population based on the observed frequency of the A allele at site X in the k_1 population. This was continued for populations k_2 through k_{10}. These 1000 simulated frequencies for each of the 10 populations resulted in 1000 r^2 values of the linear regression of the frequency of the B allele at site Y on latitude based on the observed frequency of the A allele at site X in each of the 10 populations. An observed r^2 value for the regression of allele frequency on latitude was calculated for each SNP. If the observed r^2 value for the B allele at site Y falls within the 95% confidence interval of the 1000 simulated r^2 values, the observed geographic variation for the B allele at site Y can be explained simply by linkage with the A allele at site X, which exhibits significant geographic variation. This sampling was used to test the significance of clinal variation for all SNPs scored.

We used the same approach to investigate the distribution of amino acid haplotypes along the latitudinal cline. Rare haplotypes of overall low frequency (<2%) were excluded from the analysis, which did not result in the loss of any of the SNPs scored. We performed linear regressions of transformed haplotype frequencies on population latitude and the observed r^2 values were tested for significant association by a Monte Carlo sampling as described below. The mean frequency of each haplotype (p_{all}) was calculated from pooling the entire 500 alleles. The frequency of the haplotype in each population (p_k) was then simulated with binomial sampling (samples of $n = 50$) around the expected value (p_{all}) to obtain 1000 new frequencies for the haplotype for each population. Similar to the SNP analysis, 1000 r^2 values for the regression of the simulated frequencies on population latitude were obtained for each protein haplotype and the observed r^2 value was compared with the 95% confidence interval from the 1000 simulated r^2 values. This approach tests whether the observed haplotype clines can be explained by the variance associated with our sample size of 50 alleles from each subpopulation. The observed r^2 values are compared with the 95% confidence interval to determine statistical significance.

RESULTS

Allozyme survey: Table 1 reports the results of the PGM allozyme survey of the latitudinal cline. The previous allozyme survey by Oakeshott *et al.* (1981) had focused on the widespread *Medium* allozyme and combined the *Fast* and *Slow* allozyme mobility classes because of their overall low respective frequencies. Our survey revealed several *Fast* and *Slow* allozyme mobility alleles across the latitudinal cline, and Figure 1 shows the relationship between the *Medium*, *Fast*, and *Slow* allozyme frequencies and latitude. Our multiple-allele estimate of G_{ST} for PGM allozyme variation is consistent with very little geographic variation ($G'_{ST} = 0.010$). However, while there is no geographic variation for the common *Medium* allele, the less frequent *Fast* and *Slow* alleles each show significant clines with latitude. Several *Slow* and *Fast* mobility alleles were found in our sample. However, a single *Fast* allozyme mobility allele dominates this class and exhibits a significant latitudinal cline, decreasing in frequency with increasing latitude ($r^2 = 0.463$; $P < 0.05$). A single *Slow* allozyme mobility allele dominates this class as well and exhibits a significant latitudinal cline, increasing in frequency with increasing latitude ($r^2 = 0.598$; $P < 0.01$). Figure 1 shows that the *Medium* allozyme mobility class maintains a uniform frequency across all populations, with the *Fast* and *Slow* allozymes reciprocating each other with the change in latitude. Although the analysis of allozyme mobility

TABLE 1

PGM allozyme frequencies for 11 *D. melanogaster* populations

Samples			Allozyme frequencies			
Population	Abbreviation	Latitude (°N)	*Medium*	*Fast*	*Slow*	*n*
Whiting, VT	VT97	43.6	0.880	0.013	0.095	316
Concord, MA	MA97	42.0	0.850	0.068	0.069	392
Middlefield, CT	CT97	41.2	0.810	0.132	0.028	362
Mt. Sinai, NY	DPF95	40.8	0.860	0.050	0.070	300
Churchville, MD	MD97	39.3	0.860	0.072	0.065	336
Richmond, VA	VA97	37.3	0.830	0.109	0.053	266
Smithfield, NC	NC97	35.3	0.820	0.126	0.048	210
Eutawville, SC	SC97	33.2	0.850	0.086	0.047	338
Jacksonville, FL	JFL97	30.2	0.820	0.114	0.045	202
Merritt Island, FL	MFL97	28.3	0.810	0.148	0.018	332
Homestead, FL	HFL97	25.2	0.840	0.130	0.020	320

Frequencies do not add up to 1.00 due to rare electrophoretic alleles (see text). *n* refers to the number of alleles sampled per population.

classes shows apparent clinal variation, the SNP survey will further resolve the allozyme alleles by their specific amino acid mutations.

SNP survey: Table 2 lists the 12 amino acid replacements that were discovered in the initial characterization of nucleotide variation at the *Pgm* locus by Verrelli and Eanes (2000). That study showed that three independent substitutions result in *Slow* allozyme alleles. A single change (R240L, indicates Arg to Leu at amino acid residue 240) generates a *Fast* allozyme allele and this change also has a closely linked amino acid polymorphism at nucleotide site 1340 (E245D). The remaining 7 amino acid polymorphisms do not create electrophoretic alleles, but segregate within the *Medium* allozyme. We were especially interested in these amino acid polymorphisms because of the uniform frequency of the *Medium* allele across the cline. Given the high level of amino acid polymorphism found in just 22 *Pgm* sequences in our initial study (Verrelli and Eanes 2000), there are undoubtedly other rare amino acid polymorphisms segregating in our sample of 500 alleles. However, we were interested in whether these 12 specific amino acid polymorphisms exhibited geographic variation (Verrelli and Eanes 2000). Three amino acid polymorphisms at nucleotide sites 1194, 1308, and 1617 (E197K, E235K, and A338S, respectively) were each found at frequencies <2% in the initial survey of the two northern populations (VT97 and MA97) and the two southern populations (MFL97 and HFL97) and were subsequently not scored in the remaining six populations. These 3 polymorphisms were found as singletons by Verrelli and Eanes (2000) and represent rare polymorphisms based on their frequencies in this study.

Figure 2 summarizes the frequencies of 15 SNPs that were scored from each of 50 *Pgm* alleles sampled from each of the 10 populations. In addition to the nine amino acid replacements, other polymorphisms were

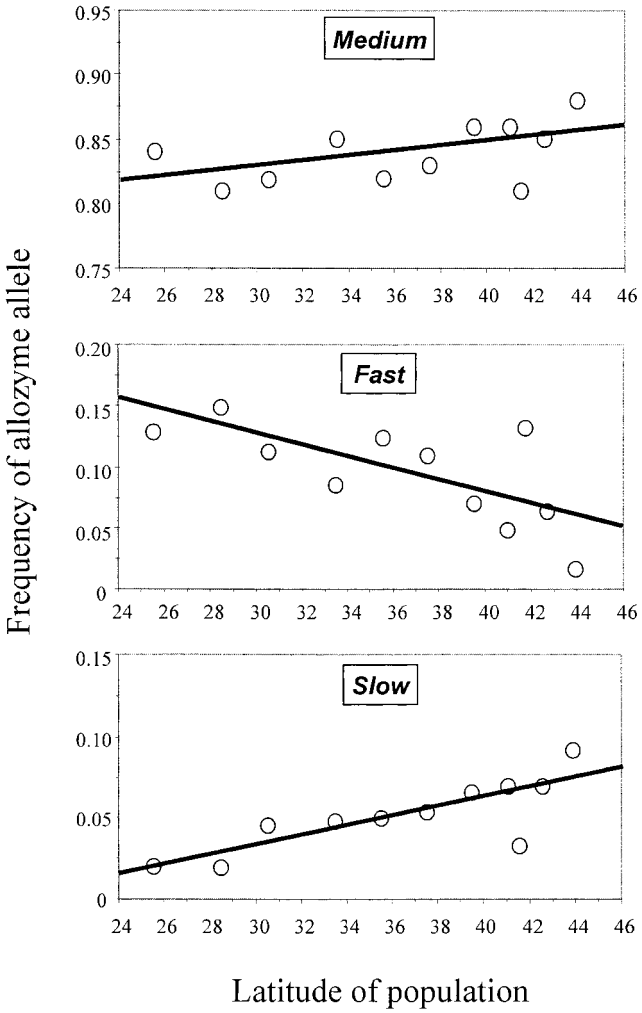


Figure 1.—Plot of the relationship of nontransformed allozyme allele frequencies with latitude. All regressions (r^2) and slopes (m) are computed from transformed data: *Medium* allele $r^2 = 0.249$, $m = 0.005$; *Fast* allele $r^2 = 0.463$, $m = -0.016$; *Slow* allele $r^2 = 0.598$, $m = 0.013$.

Site number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Nuc. position	**25**	**178**	200	**226**	**1324**	**1340**	**1626**	**1642**	**1998**	**2055**	*2156*	*2178*	*2195*	**2259**	2327
SNP	A	T	T	C	T	T	A	A	T	T	A	C	A	A	C
Population															
VT97	0.16	0.04	0.00	0.86	0.00	0.00	0.14	0.00	0.16	0.06	0.00	0.84	0.90	0.00	0.00
MA97	0.06	0.02	0.16	0.84	0.02	0.02	0.10	0.00	0.10	0.30	0.16	0.60	0.90	0.04	0.16
CT97	0.04	0.02	0.24	0.86	0.14	0.14	0.02	0.00	0.18	0.26	0.22	0.56	0.82	0.04	0.22
MD97	0.04	0.10	0.10	0.78	0.04	0.04	0.02	0.02	0.26	0.24	0.08	0.60	0.84	0.00	0.08
VA97	0.04	0.12	0.26	0.74	0.14	0.14	0.02	0.00	0.12	0.48	0.22	0.42	0.90	0.00	0.22
NC97	0.02	0.06	0.28	0.76	0.14	0.14	0.00	0.02	0.06	0.40	0.28	0.46	0.84	0.04	0.28
SC97	0.00	0.00	0.26	0.88	0.12	0.12	0.02	0.02	0.16	0.32	0.24	0.48	0.80	0.00	0.24
JFL97	0.02	0.02	0.32	0.80	0.16	0.16	0.02	0.00	0.14	0.42	0.32	0.44	0.84	0.06	0.32
MFL97	0.00	0.02	0.32	0.74	0.20	0.20	0.02	0.02	0.08	0.52	0.36	0.30	0.82	0.08	0.36
HFL97	0.00	0.00	0.26	0.66	0.12	0.12	0.00	0.08	0.10	0.46	0.38	0.26	0.72	0.06	0.38
G'_{ST}	0.062	0.044	0.061	0.029	0.044	0.044	0.067	0.038	0.029	0.084	0.082	0.109	0.023	0.029	0.082
m	0.030	0.013	- 0.029	0.018	- 0.024	- 0.024	0.026	- 0.022	0.011	- 0.036	- 0.041	0.047	0.018	- 0.017	- 0.041
r^2	0.752	0.249	0.447	0.428	0.485	0.485	0.471	0.471	0.177	0.572	0.629	0.774	0.567	0.275	0.629
	***		*	*	*	*	*	*		**	**	***	**		**

FIGURE 2.—Frequencies of the 15 SNPs across the 10 populations. Nucleotide positions in boldface are amino acid polymorphisms and in italics are intron polymorphisms; all others are silent polymorphisms. The SNP row displays the derived state of the nucleotide polymorphism. r^2 refers to the regression of the transformed allele frequencies for each SNP on latitude (*$P <$ 0.05; **$P < 0.01$; ***$P < 0.001$). m refers to the slope of each linear regression.

inadvertently scored because they resulted in the same base change as a polymorphism that was already being scored in that region. For example, nucleotide site 200 has a T → C mutation that was scored because this region was sequenced for the C → T mutation at nucleotide site 226. This resulted in data for five silent SNPs and one additional amino acid SNP at nucleotide site 178 (T36M). Although other amino acid SNPs were discovered at nucleotide sites 49 (K17Q, *Fast* allozyme), 155 (K28N, *Fast* allozyme), and 1656 (E351K, *Slow* allozyme), none were higher than 1% in frequency in the entire sample.

TABLE 2

List of 12 amino acid variants from 22 *D. melanogaster Pgm* alleles

Nucleotide position	Amino acid residue	Polymorphism	Electrophoretic change
25	9	Ala/Thr	Slow[a]
226	52	Val/Ala	—
1194	197	Glu/Lys	Slow
1308	235	Glu/Lys	Slow
1324	240	Arg/Leu	Fast
1340	245	Glu/Asp	—
1617	338	Ala/Ser	—
1626	341	Val/Met	—
1642	346	Arg/Lys	—
1998	465	Thr/Ser	—
2055	484	Val/Leu	—
2259	530	Ala/Thr	—

Polymorphism refers to the ancestral to derived amino acid change.

[a] Refers to a *Slow* allozyme allele that is not the result of a charge change.

All 500 lines used for the SNP survey were also surveyed for their PGM allozyme allele. This allowed us to determine how often *Fast* and *Slow* alleles were generated by novel amino acid changes. The polymorphism at nucleotide site 25 (A9T) is responsible for over 90% of all *Pgm* alleles that exhibit a *Slow* allozyme phenotype, and it exhibits significant clinal variation (observed $r^2 = 0.752$; $P < 0.001$). The remaining *Slow* alleles are due to Glu to Lys substitutions that all converge on a *Slow* mobility (E197K, E235K, and E351K). The polymorphism at site 1324 (R240L) exhibits significant clinal variation (observed $r^2 = 0.485$; $P < 0.05$) and is responsible for over 96% of all *Fast* allozyme alleles in our sample (rare substitutions K17Q and K28N account for the other *Fast* alleles). The E245D polymorphism was found exclusively with the R240L polymorphism, which is consistent with the *Fast* allele sequence reported by VERRELLI and EANES (2000). As is the case with the *Slow* allozyme mobility class, rare *Fast* alleles will potentially obscure the true pattern of clinal variation exhibited by the single *Fast* allozyme allele (R240L).

On a site-by-site basis, intron site 2178 exhibits the greatest level of geographic variation ($G'_{ST} = 0.109$), the

strongest association with latitude (observed $r^2 = 0.774$; $P < 0.001$), and the steepest cline for any SNP in our sample (slope = 0.047). Other nucleotide sites (silent, replacement, and intron) also exhibit strong associations with latitude. For example, the substitutions at nucleotide sites 226 (V52A) and 2055 (V484L) are the most common among all replacement polymorphisms and demonstrate changes of 20 and 40%, respectively, across the latitudinal cline. Although most *Pgm* SNPs show latitudinal clines, the likely explanation for this observation is the linkage disequilibrium between sites across the *Pgm* gene.

Linkage disequilibrium analysis: To better understand the clines at many SNPs, we are interested in describing the general pattern of association among all sites across all 10 population samples. To describe this general correlation structure or pattern of associations among sites, standardized estimates of linkage disequilibrium (R^2) were generated from the full collection of individual alleles pooled among all populations. Correlations were similar in magnitude and direction across samples. Whether computed from pooling all individuals or from averaged R^2 values across population samples, both estimates were similar in magnitude and sign. This indicated that the covariance within samples constituted most of the overall covariance between SNP alleles in the pooled collection (data not shown). Figure 3a displays the relationship between the strength of linkage disequilibrium and the distance between the 15 SNPs across the *Pgm* gene. In all, 93 out of a possible 105 pairwise comparisons are significant at the 5% level with a chi-square test, with a total of 76 comparisons significant with a Bonferroni correction ($P < 0.0005$). Figure 3b displays a table of the correlations between all SNPs. This diagram also shows that 19 out of a possible 45 pairwise comparisons between only amino acid polymorphisms are significant by a chi-square test with a Bonferroni correction ($P < 0.0005$). This demonstrates that although many silent sites are highly correlated with amino acid polymorphisms, many amino acid polymorphisms also exhibit associations with each other.

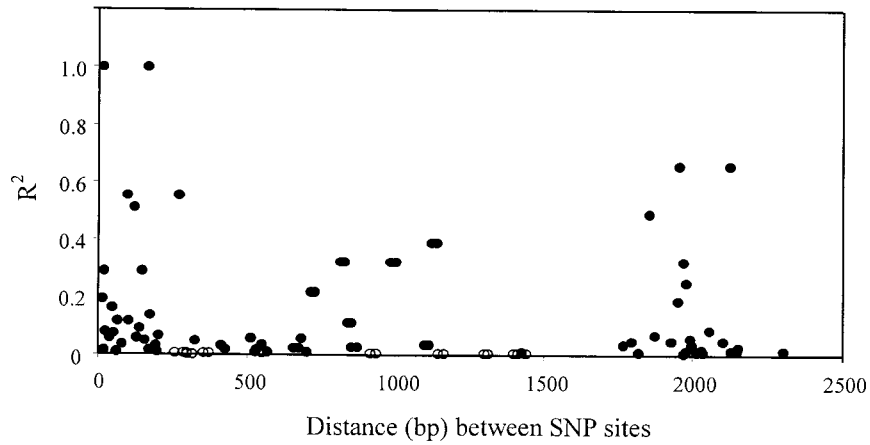
As was seen in the initial *Pgm* study by Verrelli and Eanes (2000), Figure 3a shows variable sites as far as 2 kb apart are highly correlated. Figure 3, a and b, indicates that many of these sites are not independent of each other in the regression analyses for each variable site on latitude. Therefore, we investigated the overall pattern of clinal variation at this locus by both site-by-site and haplotype-by-haplotype models.

Site-by-site analyses: We used the approach by Berry and Kreitman (1993) to investigate the extent to which the observed clinal variation at single sites can be explained by the linkage disequilibrium between sites. Figure 4 shows a few examples from the Monte Carlo sampling performed with each SNP. Replacement site 1998 can explain the observed geographic pattern at only a few variable SNPs. This is expected for a polymorphism that does not exhibit clinal variation and therefore cannot explain the clinal variation for other polymorphisms. In contrast, replacement site 2055 (V484L), which exhibits one of the strongest associations with latitude, can explain the observed clinal variation at all other variable sites, except replacement site 25 (A9T) and intron site 2178. Finally, if we choose intron site 2178 for the same comparison, this SNP can explain the clinal variation at all other SNPs except replacement site 25 (A9T). This stepwise approach was performed to determine if there is a single variable site (or sites) that can explain most of the geographic variation. The A9T polymorphism demonstrates a strong association with latitude that is not explained by covariation with any other SNP; however, alone it cannot explain the clinal variation at other SNPs. Figure 5 shows that although intron site 2178 exhibits the strongest correlation with latitude, geographic variation at this site can be effectively explained by other SNPs as well. Because of the strong association between many of the SNPs, it is unclear from this analysis alone whether there is a single identifiable source of the *Pgm* clines.

Haplotype-by-haplotype analyses: Although the site-by-site analysis fails to identify the cause of the strong clinal variation for *Pgm*, a second approach incorporates the strong linkage disequilibrium between sites in an analysis of haplotype structure. This analysis collapses the structure associated with the SNPs and, as expected from the degree of linkage disequilibrium, there are relatively few haplotypes. The 15 SNPs in this study were preferentially scored to look specifically for clinal variation and were not surveyed with respect to a random sampling (*i.e.*, several sites that were relatively rare were omitted earlier in the analysis). Therefore, it is invalid to subject our data to a typical test of the haplotype frequency distribution (Berry and Kreitman 1993; Kirby and Stephan 1995; Andolfatto *et al.* 1999). From the 15 SNP sites we find 30 haplotypes, of which 18 are individually <2% in frequency in our sample. The fact that 12 haplotypes account for 90% of the total number of haplotypes reflects the strong association between all SNPs. The site-by-site analysis confirmed that many SNPs are effectively linked and even a few of these polymorphisms can effectively explain the overall pattern and structure of variation for *Pgm*. If we examine only amino acid SNPs, there are a total of 16 protein haplotypes, of which 6 are individually <2% in frequency. As seen in Figure 3b, silent SNPs are highly correlated with replacement SNPs, which is demonstrated by the total number of haplotypes only decreasing from 12 to 10 when silent sites are omitted. Because it is apparent that many of the silent SNPs may simply be hitchhiking with other high frequency clinal replacement polymorphisms, we were initially interested in the replacement SNPs for the haplotype analysis of the *Pgm* cline.

All protein haplotypes are listed in Figure 6. Only

a



b

* 25	* 178	200	* 226	* 1324	* 1340	* 1626	* 1642	* 1998	* 2055	2156	2178	2195	* 2259	2327
0.002														
0.011	**0.012**													
0.010	**0.158**	***0.074***												
0.005	0.005	***0.389***	***0.032***											
0.005	0.005	***0.389***	***0.032***	***1.000***										
0.001	0.001	**0.010**	0.005	0.004	0.004									
0.001	0.001	0.005	0.004	0.002	0.002	0.001								
0.006	**0.007**	**0.045**	***0.036***	**0.019**	**0.019**	0.005	0.003							
0.015	***0.067***	***0.490***	0.001	***0.216***	***0.216***	**0.019**	***0.031***	**0.070**						
0.011	**0.012**	***0.656***	***0.048***	***0.321***	***0.321***	**0.010**	***0.056***	**0.046**	***0.552***					
0.025	**0.033**	**0.248**	***0.185***	**0.110**	**0.110**	***0.036***	**0.016**	***0.133***	**0.508**	**0.287**				
0.001	***0.008***	***0.055***	***0.318***	***0.023***	***0.023***	***0.007***	0.003	***0.030***	***0.088***	***0.057***	***0.190***			
0.001	0.001	***0.082***	***0.009***	0.004	0.004	0.001	0.001	0.005	***0.063***	***0.113***	**0.032**	***0.006***		
0.011	**0.012**	***0.656***	***0.048***	***0.321***	***0.321***	**0.010**	***0.056***	**0.046**	***0.552***	***1.000***	**0.287**	***0.057***	***0.113***	

FIGURE 3.—(a) Plot of the relationship of the measure of linkage disequilibrium R^2 with the distance between SNPs across the *Pgm* gene. Solid circles refer to the 76 of 105 possible pairwise comparisons significant by a chi-square test with a Bonferroni correction ($P < 0.0005$), and open circles refer to all remaining pairwise comparisons. (b) Correlation matrix for pairwise comparisons between all SNPs expressed as R^2. Asterisks designate replacement SNPs, values in boldface are the 76 significant pairwise comparisons in 3a, and boldface italics refer to significant pairwise comparisons where derived alleles are positively associated.

those haplotypes that have frequencies in excess of 2% over the entire cline (except for haplotype 11, which is discussed below) were analyzed for geographic variation. While our G_{ST} analysis shows very little evidence for geographic differentiation (overall $G'_{ST} = 0.038$), five haplotypes show significant clinal variation by regression of their transformed frequencies on latitude ($P < 0.05$). Haplotype 5 is the *Fast* allozyme defined by the SNP at replacement site 1324 (R240L), and it shows strong clinal variation. As mentioned above, this haplotype also contains the replacement site 1340 (E245D), which is in complete linkage disequilibrium with R240L. Replacement site 25 (A9T), which exhibits strong clinal variation in the site-by-site analysis and is the dominant *Slow* allozyme, is found on haplotype 8 all but once in the entire data set and shows the steepest change across the cline for any single haplotype in Figure 6. Haplotype 1 is the most common protein haplotype and shows strong clinal variation. However, other haplotypes with the same allele at replacement site 226 do not show the same geographic pattern and, in fact, show clines in the opposite direction. This implies that there may be some underlying cause independent of polymorphic site 226 that generates the observed clinal variation. While other replacement SNPs are found almost exclusively on single haplotypes, the replacement SNPs 226 and 2055 alone constitute four of the haplotypes in Figure 6. Therefore, because even many replacement SNPs are strongly linked to these two common replacement SNPs, we conducted a haplotype "equivalence" test to determine if haplotypes defined by these two replacement polymorphisms can explain the significant haplotype clines found in Figure 6 (BERRY and KREITMAN 1993).

To determine if there is significant clinal variation within the 226T/C or the 2055G/T haplotype classes, all haplotypes in Figure 6 were analyzed as a function of (1) the frequency of the 226T/C site in each population and (2) the frequency of the 2055G/T site in each population. We then used a Monte Carlo sampling (as described above) to determine if significant haplotype clines could be explained simply by their nesting within the 226/2055 haplotypes. For example, haplotype 5 (*Fast* allele) is a 226C haplotype that is found six times in the HFL97 population, where the total number of all 226C haplotypes is 33. Therefore, the frequency for haplotype 5 is 18% in HFL97 as a function of the 226C

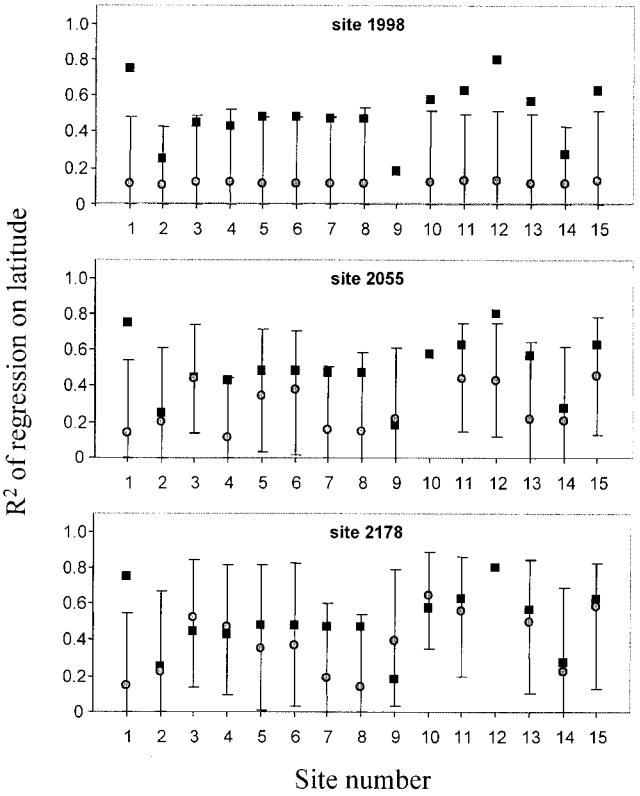


FIGURE 4.—Example plots of the Monte Carlo sampling results for three SNPs. The *x*-axis refers to the SNP numbers in Figure 2. Squares refer to the observed r^2 values for the regression of SNP transformed allele frequencies on latitude. Circles refer to the predicted r^2 values for each SNP based on the linkage disequilibrium with the SNP appearing at the top of each plot box (see text). Bars around the predicted r^2 values refer to 95% confidence intervals.

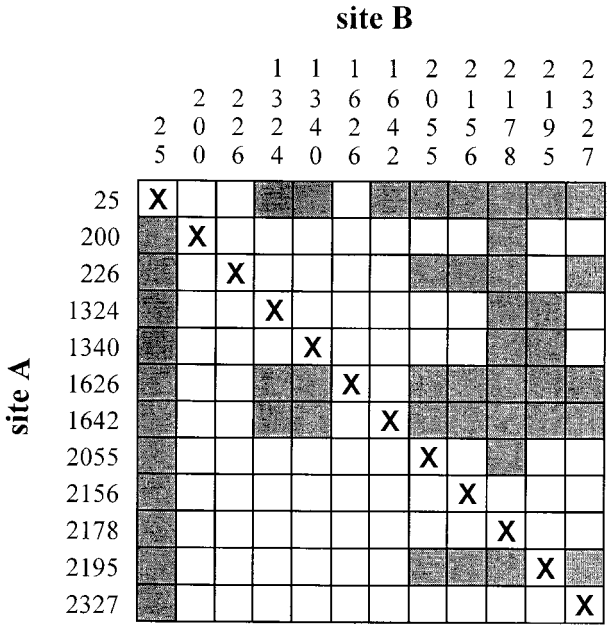


FIGURE 5.—Summary of the Monte Carlo sampling for the site-by-site analysis. Only those SNPs are shown that exhibited a significant regression with latitude (see Figure 2). Shaded boxes refer to significant clinal variation at SNP site B that cannot be explained by geographic variation at SNP site A as tested by the Monte Carlo sampling.

haplotype that it is nested within. This is done for all 10 populations, an r^2 value for the regression of the transformed frequencies on latitude is computed, and the Monte Carlo sampling tests the strength of each of the correlations with latitude. This examines whether the cline for haplotype 5 is a result of clinal variation in the 226C haplotype class (which it is nested within) or if the cline in haplotype 5 exists once the 226C haplotype cline is effectively removed. This test of haplotype equivalence finds significant clinal variation within the 226C haplotype class (haplotypes 1, 5, and 8) and within the 2055G haplotype class (haplotypes 3 and 8), but not within the 226T or 2055T haplotype classes.

Although this test finds that haplotypes 1, 3, and 5 have significant regressions with latitude, a Monte Carlo sampling indicates these correlations can be effectively explained by the geographic variation of 226C/T and 2055G/T haplotypes. Of interest is haplotype 8, which shows significant clinal variation within both the 226C and 2055G haplotype classes (observed $r^2 > 0.830$ for both classes; $P < 0.01$ from Monte Carlo sampling). Therefore, haplotype 8, as the dominant *Slow* allele (A9T), shows the same trend as the 226C/2055G haplotype class; both increase in frequency with increasing latitude. This implies there may be something interesting about all 226C/2055G haplotypes, and not just haplotype 8. When haplotype 8 is removed from the analysis, the remaining sample of 226C/2055G haplotypes still exhibits significant clinal variation ($n = 243$, observed $r^2 = 0.707$; $P < 0.01$ from Monte Carlo sampling). Therefore, while haplotype 8 exhibits clinal variation, it cannot explain the clinal variation for the 226C/2055G protein haplotype class that it is nested within.

DISCUSSION

In contrast to the earlier study by OAKESHOTT *et al.* (1981), we find extensive geographic variation of protein polymorphisms at the *Pgm* locus in *D. melanogaster*. Not only do allozyme alleles show latitudinal clines, but so do many of the amino acid polymorphisms screened as SNPs. Our examination of 15 *Pgm* SNPs from both a site-by-site and a haplotype-by-haplotype model finds substantial linkage disequilibrium, which can explain many of the covarying amino acid clines. Given this, we were interested in whether a single protein haplotype can explain both the overall pattern of clinal variation and the extensive amino acid variation at this locus.

Allozymes *vs.* SNPs: At the allozyme level, PGM is one of the most polymorphic enzymes surveyed in *D. melanogaster* population studies (TRIPPA *et al.* 1977; SINGH and RHOMBERG 1987). While electrophoretic variation is often accompanied by geographic variation

	Nucleotide position													
	25[s]	178	226	1324[f]	1340	1626	1642	1998	2055	2259				
Haplotype	G	C	T	G	A	G	G	A	G	G	*n*	G'_{ST}	*m*	r^2
1	-	-	C	-	-	-	-	-	-	-	161	0.039	0.025	0.623**
2	-	-	C	-	-	-	-	T	-	-	65	0.025	0.011	0.183
3	-	-	-	-	-	-	-	-	-	-	64	0.040	-0.028	0.708**
4	-	-	C	-	-	-	-	-	T	-	55	0.023	-0.015	0.412
5	-	-	C	T	T	-	-	-	T	-	53	0.044	-0.025	0.529*
6	-	T	-	-	-	-	-	-	T	-	19	0.047	0.011	0.176
7	-	-	-	-	-	-	-	-	T	-	19	0.035	-0.005	0.064
8	A	-	C	-	-	-	-	-	-	-	18	0.068	0.032	0.824***
9	-	-	C	-	-	A	-	-	-	-	16	0.058	0.024	0.361
10	-	-	C	-	-	-	-	-	T	A	16	0.029	-0.017	0.270
11	-	-	C	-	-	-	A	-	T	-	8	0.038	-0.022	0.473*
12	-	-	C	-	-	-	-	T	T	-	2			
13	-	T	-	-	-	A	-	-	-	-	1			
14	-	-	C	T	T	-	-	-	-	-	1			
15	-	-	-	-	-	-	-	T	-	-	1			
16	A	-	C	-	-	-	-	-	T	-	1			

FIGURE 6.—*m* refers to the slope and r^2 refers to the regression of the transformed haplotype frequencies on latitude (*$P <$ 0.05; **$P < 0.01$; ***$P < 0.001$). Superscripts s and f refer to nucleotide positions responsible for the common *Slow* and *Fast* allozyme alleles, respectively.

for glycolytic enzymes in this species (see review by EANES 1999), PGM appeared to be an exception (OAKESHOTT *et al.* 1981). The lack of clinal variation was not surprising given that cryptic amino acid variation was inferred from thermostability studies (TRIPPA *et al.* 1976, 1978) and confirmed in the DNA sequence-based study (VERRELLI and EANES 2000). The OAKESHOTT *et al.* (1981) study of PGM allozyme variation had combined all *Fast* allozyme alleles together and all *Slow* allozyme alleles together, where this study examined all electrophoretic groups separately for latitudinal clines. Although we find the *Slow* allozyme class exhibits clinal variation, our survey of nucleotide variation reveals that a single phenotypically *Slow* allele (caused by A9T) dominates in frequency and exhibits a stronger cline. Because the *Slow* allozyme class is composed of several different amino acid polymorphisms and exhibits clinal variation, it is possible that mutations that confer an increase in positive charge are favored. Because the A9T polymorphism is the most common of the *Slow* allozyme alleles, the correlation with latitude actually drops when all *Slow* allozyme alleles are examined. Thus, the significant cline for the entire *Slow* allozyme class can be explained by the strong A9T cline. In addition, the A9T polymorphism does not confer a charge change by traditional criteria (see Table 2). It is possible that this polymorphism is associated with a conformational change in the protein that reveals a buried charged residue. Not only does A9T have the strongest correlation with latitude for any amino acid polymorphism, but also the randomizations show it cannot be explained by geographic variation at any other SNP surveyed here (Figure 5). Conversely, the A9T change also cannot explain the significant clinal variation for many other sites.

While the *Fast* allozyme class is clinal, the dominant *Fast* allozyme allele (R240L, as revealed by the SNP survey) does not show a stronger cline. This may be explained by less allozyme allele heterogeneity within the *Fast* allozyme class than within the *Slow* allozyme class. While there is a weak latitudinal cline for the *Fast* allozyme class and for R240L, several other silent and replacement polymorphisms show stronger correlations with latitude. Unlike the A9T polymorphism, covariance with many other variable sites can potentially explain the R240L cline.

Single sites *vs.* haplotypes: Based on a full sequence study of 22 alleles, VERRELLI and EANES (2000) reported pervasive linkage disequilibrium across the entire 2354-bp gene. With a much larger sample of 500 alleles, this SNP study shows the same strong association between sites. In all, 12 of the 15 SNPs show clinal variation with especially strong clines at amino acid polymorphisms A9T (nucleotide site 25) and V484L (nucleotide site 2055). The Monte Carlo sampling demonstrates that, as a result of the linkage between many sites, clines in many of the SNPs can be effectively explained by other SNPs that exhibit stronger clines. Therefore, because of the strong linkage disequilibrium, the site-by-site analysis does not unambiguously reveal the source of the clinal variation at this locus.

While several of the haplotypes in Figure 6 show significant clinal variation, the Monte Carlo sampling indicates that some of these clines can be explained by their nesting within major haplotypes. The major outcome

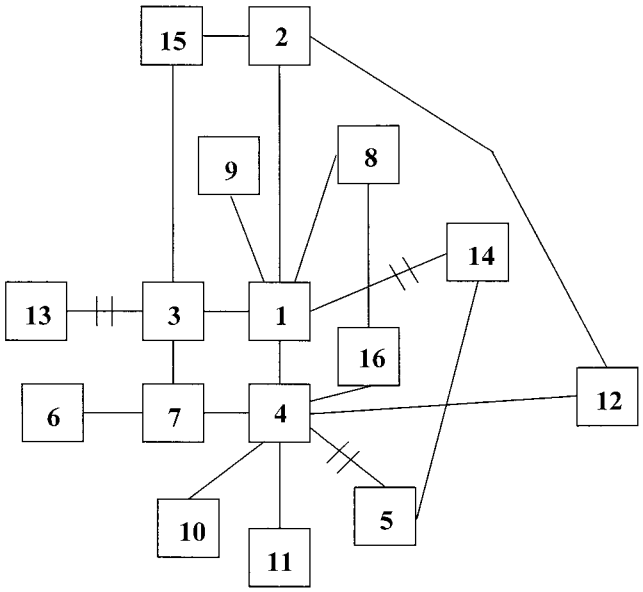


FIGURE 7.—Diagram of the haplotype network for the 16 *Pgm* protein haplotypes listed in Figure 6. Network is constructed from only amino acid polymorphisms. Haplotypes connected by a single line differ by one amino acid polymorphism. Haplotypes connected by a line with two hash marks differ by two amino acid polymorphisms.

is that the four haplotypes defined by the replacement polymorphisms at nucleotide sites 226 (V52A) and 2055 (V484L) can explain most of the *Pgm* haplotype clines. Furthermore, the combination of these two SNPs shows a common haplotype (226C/2055G) that explains the clines for each of these two SNPs independently. While the *Slow* allozyme allele (A9T-bearing haplotype 8), lies within this major 226C/2055G haplotype, the strong cline for A9T cannot explain the 226C/2055G haplotype cline, and vice versa. As is the case for many of the silent sites, many of the less common replacement polymorphisms appear to be hitchhiking along with the clinal 226C/2055G protein haplotype.

As previously mentioned, we cannot examine the *Pgm* haplotype frequency distribution with a typical statistical test because our SNPs were not randomly sampled. However, we were interested in describing the relatedness and the haplotype diversity among individuals in our sample with respect to amino acid polymorphisms. A haplotype network (Figure 7) was constructed from the differences between all amino acid haplotypes listed in Figure 6. Although a phylogenetic framework is useful for probing the structure of variation, the relatedness among alleles is obscured by recombination. This network can be used to identify some of the recombination events and clustering of haplotypes in our sample of 500 alleles. Most of the haplotypes are closely related and the diversity is likely the result of a stepwise mutational process that alters the common 226/2055 haplotypes (haplotypes 1, 3, 4, and 7). Haplotype 5 (*Fast* allele) is the most derived allele compared to the ancestral haplotype 3 (226T/2055G), owing to four amino acid point mutations. The extremely rare haplotype 13 is the most distantly related to the network, and it differs by as many as five amino acid mutations from some haplotypes. Because the true relationship of this haplotype to other haplotypes has been obscured by recombination, it is simply connected to the most closely related haplotype in the network.

We were interested in determining the effect of recombination relative to the mutational process in producing new protein haplotypes. Haplotypes that are connected in an enclosed box in Figure 7 (*i.e.*, haplotypes 1, 3, 4, and 7 or 1, 8, 16, and 4) represent a probable recombination event between these haplotypes. Our estimate of the recombination parameter $C = 4Nc$ from HUDSON (1987) for the 15 SNPs in this sample of 500 alleles ($C = 12.0$) is consistent with that obtained from 42 sites in the full *Pgm* sequence analysis ($C = 11.2$; VERRELLI and EANES 2000). Because many polymorphic sites are highly correlated, it is obvious that only a few SNPs are needed to explain the haplotype structure at this locus. Using the criteria from HUDSON and KAPLAN (1985), at least three recombination events are apparent from our sample. However, while recombination has apparently generated protein diversity, Figure 6 shows that >97% of the *Pgm* alleles result in a total of 10 protein haplotypes from 10 replacement polymorphisms. This suggests that *Pgm* haplotype variation is predominantly explained by reoccurring mutational input.

Most haplotypes are one or two steps from haplotypes 1 and 4 (226C/2055G and 226C/2055T, respectively). Of these two, haplotype 4 is the more derived haplotype compared to the ancestral haplotype 3. Although haplotypes 1 and 4 seem to exhibit similar levels of protein diversity, they show different patterns of variation. Figure 3b shows strong correlations between several derived replacement polymorphisms, which are largely associated with haplotype 4. This pattern of variation may suggest that this haplotype is not relatively new and was historically higher in frequency. Because there is ample evidence for recombination in our sample, it is possible that the strong association among derived replacement alleles is the result of epistatic selection for protein haplotypes. The excess of rare alleles associated with haplotype 1 (VERRELLI and EANES 2000) suggests this haplotype had been historically lower in frequency. If this 226C/2055G haplotype was historically lower in frequency and has recently become very common, this would have provided less opportunity for recombination with the 226C/2055T haplotype and could explain the strong correlation between derived alleles associated with the 226C/2055T haplotype.

What causes the *Pgm* cline? *D. melanogaster* likely colonized North America in the last 200–300 years from African and European populations (DAVID and CAPY 1988). It is possible that two geographic regions that differed in their respective *Pgm* haplotypes may have

initially colonized the extreme points of the latitudinal cline. Therefore, the *Pgm* cline might be the result of migration of individuals outward from their initial points of colonization, with limited gene flow. However, several observations are inconsistent with the hypothesis that historical population structure generates the *Pgm* cline. If *D. melanogaster* populations are, or were at some time, subdivided along the Atlantic coast of North America, we might expect the same clinal pattern at all loci for effectively neutral variation. However, while many loci show allozyme clines (see Eanes 1999 for review), several studies show no evidence of population structure for effectively neutral markers (Hale and Singh 1991; Berry and Kreitman 1993; our unpublished data). These conflicting patterns imply that the allozyme clines are maintained by natural selection in the face of gene flow. In addition, we find several SNPs at the *Pgm* locus that show no clinal variation. The effective number of migrants per subpopulation, or *Nm*, can be estimated from the population parameter F'_{ST} (Hudson *et al.* 1992), which is equal to the estimator G'_{ST} under a two-allele model (Nei 1986). Although many sites exhibit clinal variation, our locus-specific estimate of 7.43 is comparable with other loci in *D. melanogaster* (Hale and Singh 1991; Berry and Kreitman 1993; our unpublished data).

The cosmopolitan inversion *In(3L)P* is clinal in North America, increasing in frequency with decreasing latitude (Mettler *et al.* 1977; Knibb *et al.* 1981). Although the *Pgm* locus (chromosome 3L; 72D7) is ~180 kb inside the proximal breakpoint (at 73E3), Verrelli and Eanes (2000) found shared variation between arrangements that indicates events of gene conversion. We found significant clinal variation for the 15 copies of the inversion recovered from a screen of our 500 lines ($r^2 = 0.710$; $P < 0.001$). Consistent with our previous results, we found a heterogeneous sampling of *Pgm* haplotypes associated with these 15 inverted copies, although one replacement polymorphism in our sample, site 1642 (G346K), is completely associated with the inversion and represents haplotype 11 in Figure 6. Interestingly, the Verrelli and Eanes (2000) full *Pgm* sequence study had sampled the site 1642 polymorphism eight times, three of which are associated with standard arrangements, and two of these three are from Zimbabwe. Because the inversion *In(3L)P* is not typically found in East Africa (Eanes *et al.* 1992), the different association of site 1642 with the inversion in the two populations suggests this polymorphism originated on a standard allele and has recently exchanged onto an inverted copy. Because the 1642 site polymorphism is found almost exclusively with the inversion, it is also likely that this polymorphism was captured by the original inversion event, and variation found on all other inverted copies represents rare events of gene conversion or recombination. Nonetheless, with the exception of the site 1642 polymorphism, this inversion cannot explain the amino acid polymorphism clines at the *Pgm* locus.

The best explanation for the pattern of clinal variation at the *Pgm* locus is one that includes selection. The most compelling evidence comes from the analysis of the 226C/2055G protein haplotype. While there are observed individual clines at both replacement sites 226 and 2055, the combination of these two sites into a protein haplotype shows a very strong correlation with latitude ($r^2 = 0.776$) and the steepest cline for any single site or haplotype in our sample (slope = 0.052). Although many SNPs are strongly correlated, Figure 3b shows that there is very little association between sites 226 and 2055, suggesting that linkage disequilibrium alone cannot explain the strong cline for this protein haplotype. If the 226C/2055G haplotype cline was the result of one of these nucleotide sites, then either the 226C allele or the 2055G allele should exhibit comparable or stronger clines than the combined pair, but this is not the case. Figure 8 shows the geographic variation for the four major haplotypes defined by the 226 and 2055 replacement polymorphisms. Although the 226T and 2055T alleles both decrease in frequency with increasing latitude, there is nothing noteworthy about the 226T/2055T haplotype; it is found only 38 times in the entire data set and exhibits no clinal variation. The remaining two of the four haplotype classes, 226T/2055G and 226C/2055T, could explain the clinal variation of the 226T and 2055T alleles.

Verrelli and Eanes (2000) observed that although the 226C/2055G protein haplotype is relatively high in frequency, it possesses a significant excess of low frequency silent site polymorphisms. This is possible evidence for a recent increase in this haplotype by directional selection (Hudson *et al.* 1994; Braverman *et al.* 1995). This haplotype is also the most frequent protein haplotype in our sample, ranging from 84% in the north to 28% in the south. The 226C allele dominates our Zimbabwe sample (12 of 13 lines), suggesting that this derived allele was historically high in frequency (Verrelli and Eanes 2000). In this same sample, the 2055T allele is also high in frequency (10 of 13 lines), altogether accounting for nine 226C/2055T haplotypes out of 13 lines, with only three 226C/2055G haplotypes. Given that the colonization of North America is relatively recent, this implies a rapid increase of this haplotype along the latitudinal cline and may explain both the high linkage disequilibrium in the overall sample and the significant excess of low frequency variants associated with the 226C/2055G haplotype. It is unclear whether selection is acting on the 226C/2055G protein haplotype or a site in strong linkage with this haplotype. Whatever the cause, it is likely that this protein haplotype has increased in frequency very recently.

We assume that the silent and intron sites in our sample are effectively neutral and are simply hitchhiking along with the amino acid variation. Figure 3b shows

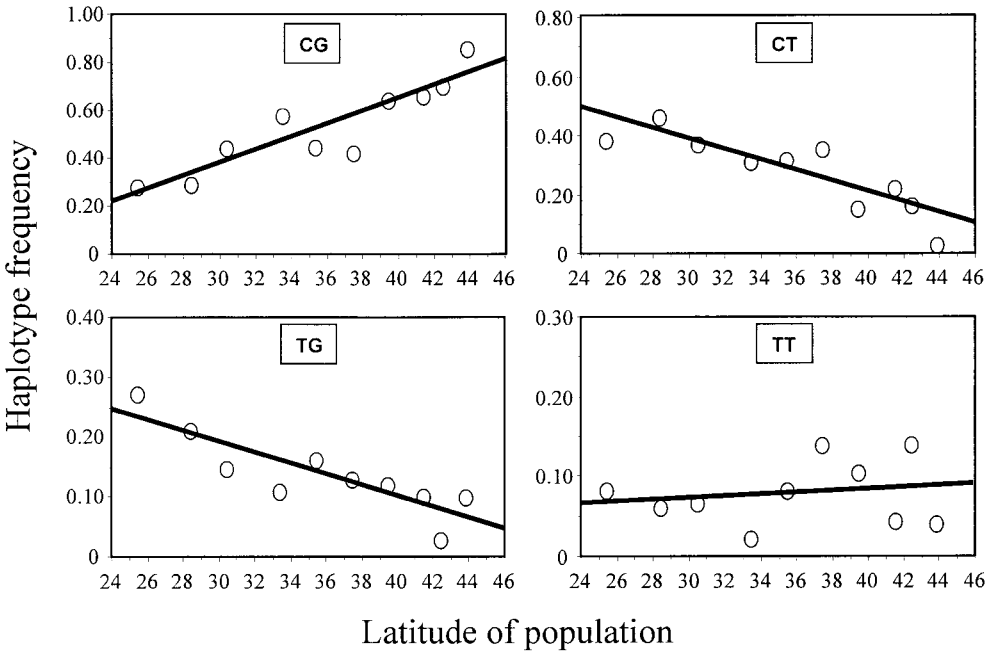


FIGURE 8.—Plot of the relationship of nontransformed *Pgm* haplotype frequency with latitude for the four major 226/2055 protein haplotypes. All regressions (r^2) and slopes (m) are computed from transformed data: CG $r^2 = 0.776$, $m = 0.052$; CT $r^2 = 0.687$, $m = -0.041$; TG $r^2 = 0.631$, $m = -0.026$; TT $r^2 = 0.018$, $m = 0.004$.

that there is strong linkage disequilibrium between nucleotide sites 226 and 2178 and between nucleotide sites 2055 and 2178. Although the 2178T allele segregates evenly with all four 226/2055 protein haplotypes, the derived 2178C allele is strongly associated with the 226C/2055G haplotype (239 of 248 times that the 2178C allele is found). Because of this strong linkage disequilibrium, inclusion of the 2178C allele in the analysis of the 226C/2055G haplotype cline has no added effect (CG $r^2 = 0.773$ with 2178C, CG $r^2 = 0.776$, without 2178C). It is possible this intron SNP may possess an adaptive regulatory role like that suggested for the clinal intron polymorphism at *Adh* (BERRY and KREITMAN 1993), and this may explain the strong cline 2178C exhibits with the 226C/2055G haplotype. Our sample of 13 Zimbabwe *Pgm* alleles finds the 2178C allele as a singleton, and segregates with one of the three 226C/2055G haplotypes (VERRELLI and EANES 2000). This also suggests that this polymorphism was initially rare and has recently increased in frequency with the 226C/2055G haplotype in temperate North America. As of yet, we have no evidence that suggests the intron SNP 2178 is selectively favored and generates the *Pgm* cline; however, a functional analysis shows that the 226C/2055G protein haplotype confers significantly greater enzyme activity and glycogen content compared to all other *Pgm* protein haplotypes (our unpublished data). Given the strong association between all SNPs, the fact that sites 226 and 2055 are weakly associated, yet show the strongest cline when combined, makes a strong case for clinal selection acting on these two sites in unison.

Linkage disequilibrium is expected to decrease with increasing distance between variable sites, but our analysis shows strong disequilibrium between variable sites more than 2 kb apart. It is possible that a polymorphic site not identified at the *Pgm* locus, and that is in strong disequilibrium with the clinal sites, can explain all clinal variation at *Pgm*. However, this hypothetical variant must be very low in frequency to not be found in the 44 full *Pgm* sequences in VERRELLI and EANES (2000) and, therefore, cannot explain the cline for haplotype 226C/2055G, which is as high as 84% in our sample of subpopulations. It is also possible that a variable site that lies outside of the *Pgm* locus is driving the latitudinal clines. If the clinal variation at *Pgm* is the result of hitchhiking with a variable site outside this gene region, this variable site must be under strong selection for latitudinal variation and either has recently and rapidly risen in frequency or is in very close proximity to the *Pgm* locus. Because several polymorphic sites that are distantly separated show strong associations (Figure 3), it would be interesting to investigate the extent to which this strong linkage disequilibrium extends outside the *Pgm* gene region.

Glycolytic enzymes and latitudinal clines: Surveyed to date, all glycolytic enzymes in *D. melanogaster* possess much less replacement polymorphism than found for *Pgm* (VERRELLI and EANES 2000; our unpublished data). While allozyme clines are common, only BERRY and KREITMAN (1993) had examined latitudinal clines in nucleotide variation to resolve targets of natural selection. Albeit complex, the extensive amino acid variation at the *Pgm* locus offers the opportunity to investigate a large number of protein haplotypes and their distribution along a latitudinal cline. Our study finds that clines in allozyme alleles may simply be hitchhiking with other

Pgm amino acid polymorphisms that exhibit stronger latitudinal clines. In fact, a *Pgm* haplotype network shows that these allozyme alleles represent only a small fraction of the amino acid polymorphism that has accumulated on the common 226/2055 protein haplotypes. The strong latitudinal cline for the 226C/2055G protein haplotype suggests that *Pgm* amino acid variation is under selection in natural populations.

The question remains whether latitudinal variation for *Pgm* protein haplotypes can explain the extensive amino acid polymorphism at the *Pgm* locus (Verrelli and Eanes 2000). As a form of diversifying selection, spatially varying selection may favor different protein haplotypes along an environmental gradient. *Pgm* is similar to *Tpi* (Hasson *et al.* 1998), *Sod* (Hudson *et al.* 1994), and *Pgi* (John H. McDonald, personal communication), in that they have low levels of amino acid fixation. While these other enzymes have a simple two-allele amino acid polymorphism, *Pgm* would have at least 20 amino acid polymorphisms in a comparable sample. In addition, a similar level of silent site polymorphism is not seen for *Pgm*. Therefore, while diversifying selection may result in an accumulation of linked silent site polymorphism, this may not be the case if adaptive polymorphism is relatively recent or short lived (Hudson *et al.* 1994; Gillespie 1994, 1997). If diversifying or epistatic selection favors different combinations of amino acid polymorphisms in varying environments (*i.e.*, combinations of the 226 and 2055 polymorphisms in temperate and tropical regions), amino acid polymorphism may be limited from reaching appreciable frequencies and contributing to fixation. While epistatic selection can explain several amino acid polymorphisms, this does not imply that all *Pgm* amino acid polymorphism is adaptive. In fact, our analysis shows that many are simply hitchhiking along with the 226/2055 protein haplotype classes. The Monte Carlo sampling emphasizes that selection may favor both the 226C/2055G protein haplotype and a specific *Slow* allele in higher latitudes. While the *Fast* allele (R240L) cline can in effect be explained by the geographic variation for the 226/2055 protein haplotype, it still may be favored in lower latitudes. While we have no other evidence for these scenarios, an examination of the three-dimensional structural and functional differences between the vast numbers of different *Pgm* protein haplotypes suggests that these amino acid polymorphisms have adaptive value (our unpublished data).

Finally, most glycolytic enzymes that possess latitudinal clines show the derived allele is higher in frequency in temperate regions (Eanes 1999). This is also the case for *Pgm*. The 226C/2055G protein haplotype is a derived allele that has apparently undergone a recent and rapid increase in frequency in northern latitudes since *D. melanogaster* colonized North America from Afrotropical regions. It is possible that this *Pgm* allele plays a role in this species' adaptation to temperate regions. This recent colonization and adaptation may explain latitudinal clines for a suite of life history characters including body size and development (Coyne and Beecham 1987; James *et al.* 1995), ethanol and acetic acid tolerance (David and Bocquet 1975; Cohan and Graf 1985; Chakir *et al.* 1996), and ovariole number and egg size (Capy *et al.* 1993; Azevedo *et al.* 1996; our unpublished data). Because many glycolytic enzymes have a derived allele associated with northern latitudes, the burden associated with overwintering in temperate regions may require a multilocus response to selection for a different genetic architecture (Eanes 1999). This enzyme variation could involve a change in enzyme activity or thermostability to adapt to different environmental conditions (Argos *et al.* 1979; McDonald *et al.* 1999). Nonetheless, differences in enzyme structure and function must be manifested in a change in metabolic flux and have fitness consequences (Hartl *et al.* 1985; Dykhuizen *et al.* 1987; Labate and Eanes 1992; Eanes 1999). PGM resides at a branch point that partitions flux into glycogen metabolism, the pentose shunt, and the main glycolytic pathway, and Clark and Keith (1988) also show a strong association between PGM enzyme activity and glycogen content. Therefore, our examination of the physiological effects of *Pgm* amino acid polymorphisms on glycogen storage will further explore the relationship between enzyme activity variation and metabolic flux.

The authors thank Andrew Berry for his advice on sampling locations and John H. McDonald and Paul Schmidt for insightful discussion regarding analyses of geographic data. Jody Hey and two anonymous reviewers provided helpful criticism in revision. This research was supported by National Science Foundation dissertation improvement grant DEB9902327 to B.C.V. and U.S. Public Health Service grant GM-45247 to W.F.E. This is contribution number 1084 from the Graduate Program in Ecology and Evolution, State University of New York at Stony Brook.

LITERATURE CITED

Andolfatto, P., J. D. Wall and M. Kreitman, 1999 Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. Genetics **153:** 1297–1311.

Argos, P., M. G. Rossmann, U. M. Grau, H. Zuber, G. Frank *et al.*, 1979 Thermal stability and protein structure. Biochemistry **18:** 5698–5703.

Azevedo, R. B. R., V. French and L. Partridge, 1996 Thermal evolution of egg size in *Drosophila melanogaster*. Evolution **50:** 2338–2345.

Ballard, J. W. O., and M. Kreitman, 1994 Unraveling selection in the mitochondrial genome of Drosophila. Genetics **138:** 757–772.

Berry, A. J., and M. Kreitman, 1993 Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the east coast of North America. Genetics **134:** 869–893.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

Capy, P., E. Pla and J. R. David, 1993 Phenotypic and genetic variability of morphometrical traits in natural populations of *Drosophila melanogaster* and *D. simulans*. I. Geographic variations. Genet. Sel. Evol. **25:** 517–536.

Chakir, M., P. Capy, J. Genermont, E. Pla and J. R. David, 1996 Adaptation to fermenting resources in *Drosophila melanogaster*:

ethanol and acetic acid tolerances share a common genetic basis. Evolution **50:** 767–776.

Clark, A. G., and L. E. Keith, 1988 Variation among extracted lines of *Drosophila melanogaster* in triacylglycerol and carbohydrate storage. Genetics **119:** 595–607.

Cohan, F. M., and J. D. Graf, 1985 Latitudinal cline in *Drosophila melanogaster* for knockdown resistance to ethanol fumes and for rates of response to selection for further resistance. Evolution **39:** 278–293.

Coyne, J. A., and E. Beecham, 1987 Heritability of two morphological characters within and among natural populations of *Drosophila melanogaster*. Genetics **117:** 727–737.

David, J. R., and C. Bocquet, 1975 Similarities and differences in latitudinal adaptation of two Drosophila sibling species. Nature **257:** 588–590.

David, J. R., and P. Capy, 1988 Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. **4:** 106–111.

Dykhuizen, D. E., A. M. Dean and D. L. Hartl, 1987 Metabolic flux and fitness. Genetics **115:** 25–31.

Eanes, W. F., 1999 Analysis of selection on enzyme polymorphisms. Annu. Rev. Ecol. Syst. **30:** 301–326.

Eanes, W. F., C. S. 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.

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

Gillespie, J. H., 1997 Junk ain't what junk does: neutral alleles in a selected context. Gene **205:** 291–299.

Hale, L. R., and R. S. Singh, 1991 Contrasting patterns of genetic structure and evolutionary history as revealed by mitochondrial DNA and nuclear gene-enzyme variation. J. Genet. **70:** 79–89.

Hartl, D. L., D. E. Dykhuizen and A. M. Dean, 1985 Limits of adaptation: the evolution of selective neutrality. Genetics **111:** 655–674.

Hasegawa, M., Y. Cao and Z. Yang, 1998 Preponderance of slightly deleterious polymorphism in mitochondrial DNA: nonsynonymous/synonymous rate ratio is much higher within species than between species. Mol. Biol. Evol. **15:** 1499–1505.

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.

Hjorth, J. P., 1970 A phosphoglucomutase locus in *Drosophila melanogaster*. Hereditas **64:** 146–148.

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

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

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

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.

James, A. C., R. B. R. Azevedo and L. Partridge, 1995 Cellular basis and developmental timing in a size cline of *Drosophila melanogaster*. Genetics **140:** 659–666.

Katz, L. A., and R. G. Harrison, 1997 Balancing selection on electrophoretic variation of phosphoglucose isomerase in two species of field cricket: *Gryllus veletis* and *G. pennsylvanicus*. Genetics **147:** 609–621.

Kennedy, P., and M. W. Nachman, 1998 Deleterious mutations at the mitochondrial *ND3* gene in South American marsh rats (*Holochilus*). Genetics **150:** 359–368.

Kirby, D. A., and W. Stephan, 1995 Haplotype test reveals departure from neutrality in a segment of the *white* gene of *Drosophila melanogaster*. Genetics **141:** 1483–1490.

Knibb, W. R., J. G. Oakeshott and J. B. Gibson, 1981 Chromosome inversion polymorphisms in *Drosophila melanogaster*. I. Latitudinal clines and associations between inversions in Australasian populations. Genetics **98:** 833–847.

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.

McDonald, J. H., 1994 Detecting natural selection by comparing geographic variation in protein and DNA polymorphisms, pp. 88–100 in *Non-neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

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

McDonald, J. H., B. C. Verrelli and L. B. Geyer, 1996 Lack of geographic variation in anonymous nuclear polymorphisms in the American oyster, *Crassostrea virginica*. Mol. Biol. Evol. **13:** 1114–1118.

McDonald, J. H., A. M. Grasso and L. K. Rejto, 1999 Patterns of temperature adaptation in proteins from Methanococcus and Bacillus. Mol. Biol. Evol. **16:** 1785–1790.

Mettler, L. E., R. A. Voelker and T. Mukai, 1977 Inversion clines in populations of *Drosophila melanogaster*. Genetics **87:** 169–176.

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

Nachman, M. W., W. M. Brown, M. Stoneking and C. F. Aquadro, 1996 Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics **142:** 953–963.

Nei, M., 1986 Definition and estimation of fixation indices. Evolution **40:** 643–645.

Nielsen, R., and D. M. Weinreich, 1999 The age of nonsynonymous and synonymous mutations in animal mtDNA and implications for the mildly deleterious theory. Genetics **153:** 497–506.

Oakeshott, J. G., G. K. Chambers, J. B. Gibson and D. A. Willcocks, 1981 Latitudinal relationships of esterase-6 and phosphoglucomutase gene frequencies in *Drosophila melanogaster*. Heredity **47:** 385–396.

Oakeshott, J. G., J. B. Gibson, P. R. Anderson, W. R. Knibb, D. G. Anderson *et al.*, 1982 Alcohol dehydrogenase and glycerol-3-phosphate dehydrogenase clines in *Drosophila melanogaster* on different continents. Evolution **36:** 86–96.

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.

Oakeshott, J. G., S. W. McKechnie and G. K. Chambers, 1984 Population genetics of the metabolically related *Adh*, *Gpdh*, and *Tpi* polymorphisms in *Drosophila melanogaster*. I. Geographic variation in *Gpdh* and *Tpi* allele frequencies in different continents. Genetica **63:** 21–29.

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

Ohta, T., 1996 The current significance and standing of neutral and nearly neutral theories. Bioessays **18:** 673–677.

Pogson, G. H., K. A. Mesa and R. G. Boutilier, 1995 Genetic population structure and gene flow in the Atlantic cod *Gadus morhua*: a comparison of allozyme and nuclear RFLP loci. Genetics **139:** 375–385.

Rand, D. M., and L. M. Kann, 1996 Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Mol. Biol. Evol. **13:** 735–748.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Salamon, H., W. Klitz, S. Easteal, X. Gao, H. A. Erlich *et al.*, 1999 Evolution of HLA class II molecules: allelic and amino acid site variability across populations. Genetics **152:** 393–400.

Schmidt, P. S., and D. M. Rand, 1999 Intertidal microhabitat and selection at MPI: interlocus contrasts in the northern acorn barnacle, *Semibalanus balanoides*. Evolution **53:** 135–146.

Singh, R. S., and L. R. Rhomberg, 1987 A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. II. Estimates of heterozygosity and patterns of geographic differentiation. Genetics **117:** 255–271.

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry*, Ed. 3. W. H. Freeman, San Francisco.

Takahata, N., Y. Satta and J. Klein, 1992 Polymorphism and balancing selection at major histocompatibility complex loci. Genetics **130:** 925–938.

Terauchi, R., T. Terauchi and N. T. Miyashita, 1997 DNA polymorphism at the *Pgi* locus of a wild yam, *Dioscorea tokoro*. Genetics **147:** 1899–1914.

Trippa, G., A. Loverre and A. Catamo, 1976 Thermostability studies for investigating non-electrophoretic polymorphic alleles in *Drosophila melanogaster*. Nature **260:** 42–44.

Trippa, G., G. A. Danieli, R. Costa and R. Scozarri, 1977 A new allele at the PGM locus in *Drosophila melanogaster*. Dros. Inf. Serv. **52:** 74.

Trippa, G., A. Catamo, A. Lombardozzi and R. Cicchetti, 1978 A simple approach for discovering common nonelectrophoretic enzyme variability: a heat denaturation study in *Drosophila melanogaster*. Biochem. Genet. **16:** 299–305.

Verrelli, B. C., and W. F. Eanes, 2000 Extensive amino acid polymorphism at the *Pgm* locus is consistent with adaptive protein evolution in *Drosophila melanogaster*. Genetics **156:** 1737–1752.

Wayne, M. L., D. Contamine and M. Kreitman, 1996 Molecular population genetics of *ref(2)P*, a locus which confers viral resistance in Drosophila. Mol. Biol. Evol. **13:** 191–199.

Winnepenninckx, B., T. Backeljau and R. de Wachter, 1993 Extraction of high molecular weight DNA from molluscs. Trends Genet. **9:** 407.

Communicating editor: J. Hey