Contrasting Molecular Population Genetics of Four Hexokinases in *Drosophila melanogaster, D. simulans* and *D. yakuba*

David D. Duvernell and Walter F. Eanes

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

Manuscript received January 13, 2000

Accepted for publication July 14, 2000

ABSTRACT

As part of a larger study contrasting patterns of variation in regulatory and nonregulatory enzymes of the central metabolic pathways we have examined the molecular variation in four uncharacterized hexokinase genes unique to muscle, fat body, and testis in *Drosophila melanogaster, D. simulans*, and *D. yakuba*. Earlier isoenzyme studies had designated these genes as Hex-A, Hex-C, and Hex-t. There are two tightly linked testis-specific genes designated here as Hex-t1 and Hex-t2. Substantial and concordant differences across species are seen in levels of both amino acid and silent polymorphism. The flight muscle form Hex-A is the most conserved followed by the fat body hexokinase Hex-C and testis-specific hexokinases Hex-t1 and Hex-t2. While constraints acting at the amino acid level are expected, the silent polymorphisms follow this pattern as well. All genes are in regions of normal recombination, therefore hitchhiking and background selection are not likely causes of interlocus differences. In *D. melanogaster* latitudinal clines are seen for amino acid polymorphisms at the Hex-C and Hex-t2 loci. There is evidence for accelerated amino acid substitution in Hex-t1 that has lost residues known to be associated with glucose and glucose-6-phosphate binding. *D. simulans* shows substantial linkage phase structuring that suggests historical population subdivision.

The generation of complex phenotypes involves many steps, irrespective of whether the phenotype is a simple metabolic pool or a morphology whose expression involves a developmental pathway. Implicit in the understanding of genetic architecture is the expectation that all steps in a pathway may not be equal in their contribution to the phenotypic variation. A major challenge in understanding both local adaptation and long-term evolutionary divergence is to establish an understanding of which steps, because of their inherent properties, will be the most responsive to selection. In studies of metabolic regulation, both physiological biochemistry (Hochachka and Somero 1984) and metabolic control theory (Kacser and Burns 1973) have focused on the issue of pathway regulation. Physiological biochemistry has emphasized the role of key regulatory enzymes that display allosteric control and nonequilibrium reaction states. In contrast, metabolic control has emphasized the distribution of pathway regulation across many steps. In the 1970s the relationship between selective responses and pathway position first emerged in arguments about whether regulatory and nonregulatory enzymes should show more or less genetic polymorphism (Johnson 1971), then defined by allozymes.

Amid much early speculation these questions remained unanswered. However, the introduction of DNA sequencing adds enormously more power and insight into the study of the impact of selection across steps in a pathway.

The concentration of pathway control in a few critical steps requires specialization in those enzymes. The central metabolic pathway, represented by glycolysis and the Krebs cycle, is active in all tissues; yet, for a given pathway step, selective pressures may be expected to be tissue or stage specific. The duplication and functional divergence of enzymes to meet the differing metabolic demands and modulator environments of different tissues is a pervasive theme in metazoan evolution, especially in the vertebrates. Divergence of function is typically achieved through gene duplication or the appearance of isozymes, which may arise from post-translational modification or alternative splicing. It appears in *Drosophila melanogaster* that most of the enzymes of the central metabolism show single forms derived from single-copy genes, although several, such as *Gpdh* and *Ald*, show multiple forms derived from alternative splicing (Bewley et al. 1989; Kalm et al. 1989; Shaw-Lee et al. 1992). The notable exception is the hexokinases that are routinely encoded by multiple loci in *Drosophila* (Murray and Ball 1967; Beckenbach and Prakash 1977; Cavener 1980). The hexokinases represent the initial step in the oxidative phosphorylation of hexoses and are critically viewed as textbook “regulatory” or nonequilibrium enzymes that control flux into the pathway in response to several effectors, in particular the...
adenylate charge (Hochachka and Somero 1984). In *D. melanogaster* four different prominent hexokinase iso-
yzymes have been reported from studies (Murray and Ball 1967). The so-called HEX-A and HEX-B isozymes are
products of a single X-linked locus, *Hex-A*, while the *Hex-C* form is the product of a locus on the second
chromosome, *Hex-C*, and segregates a common allozyme polymorphism in nature populations (Mukai and
Voelker 1977). The HEX-A and HEX-B forms are largely expressed in adult thorax and larval muscle tis-
ues, respectively, while the *Hex-C* enzyme appears con-

ined to the adult and larval fat body, an organ similar
in function to the vertebrate liver. In addition, a fourth
form, HEX-T, which is expressed in the testis, has been
reported (Murray and Ball 1967; Cavener 1980). Given
that these three hexokinases may regulate glycolysis
in tissues with very different metabolic demands, we
might expect very different evolutionary constraints
acting on each gene. This report focuses on the molecu-
lar population genetics of the hexokinase loci of *D.
melanogaster* and its relatives, *D. simulans* and *D.
yakuba*. It extends our study of the population genetic and mo-
lecular evolution of genes involved in central metabo-

lism and raises the possibility that genes dedicated to
control of pathways that are operational in different
tissues can have very different patterns of molecular
variation and evolution that are perhaps predictable by

their metabolic biology (Eanes 1999).

MATERIALS AND METHODS

Origin of wild lines: More than 1500 isofemale lines of
*D. melanogaster* representing 10 local populations distributed
from Florida to Vermont were collected in the fall of 1997.
They are prefixed as HFL (Homestead, FL), MFL (Merritt
Island, FL), JFL (Jacksonville, FL), SC (Eutawville, SC), NC
(Smithfield, NC), VA (Richmond, VA), MD (Churchville,
MD), CT (Middlefield, CT), MA (Concord, MA), and VT
(Whiting, VT). Lethal-free second and third chromosomes
were genetically extracted to establish isochromosomal lines
using the *CyG*/*and TM6 balancers. A total of 14 *D. simulans*
lines collected in 1996 and 1997 from Davis Peach Farm (DPF96s),
Middleton, CT (CT96s), Homestead, FL (HFL97s), and Vir-
ginia (VA96) were subjected to 14 generations of full-sib
mating. Samples of flies from east Africa (ZimH and ZimS)
were kindly donated by the lab of Chung-I Wu. The line of
*D. yakuba* was obtained from the Bowling Green Stock Center
(BG1013).

Identification and isolation of hexokinase sequences: Short
353- and 640-bp segments of the *Hex-A* and *Hex-t2* genes,
respectively, were recovered from a third instar cDNA library
using degenerate PCR primers designed to amino acid motifs
conserved among hexokinase genes from yeast and mammals.
These degenerate PCR products were sequenced, confirmed to
have high similarity to hexokinases, and the 353-bp frag-
ment was subsequently used to isolate the complete transcript
sequence of *Hex-A* from the cDNA library by the method of
*Verrelli and Eanes* (2000). The complete genomic se-
quences for *Hex-C*, *Hex-t1*, and *Hex-t2* emerged from BLAST
searches of the Drosophila Genome Project database (acces-
sion nos. AA567642 and AC005813) using the degenerate
PCR-amplified sequences of both *Hex-A* and *Hex-t2* as the query
sequences.

Northern blot analysis: Probes for each hexokinase gene
included an ~1.2- to 1.4-kb fragment of coding sequence and
were prepared by direct incorporation of digoxigenin-
11-dUTP (Boehringer Mannheim, Mannheim, Germany) dur-
ing PCR. Total RNA was prepared according to standard proto-
cols (Sambrook et al. 1989) and 30 μg of each sample was
electrophoresed through a 1% agarose-formaldehyde gel.
RNA samples were transferred by capillary blotting onto posi-
tively charged nylon membranes (Boehringer Mannheim).
Hybridization and detection were performed using the nonra-
dioactive digoxigenin system (Boehringer Mannheim) follow-
ing the protocols of Engler-Blum et al. (1993).

PCR amplification and DNA sequencing: DNA sequences repres-
enting the complete coding regions for each of the hexokinase genes were amplified according to the following
conditions: 10 μg of genomic DNA was amplified in 10 mM
Tris (pH 8.3), 50 mM KCl, 0.01% gelatin, 1.5 mM MgCl2,
2 units of platinum Taq polymerase (Life Technologies, Rockville,
MD), and 100 mM of each primer. The amplification products
were purified (Prep-A-Gene DNA purification kit; Bio-Rad, Her-
cules, CA), and double-stranded templates (~500 ng per reac-
tion) were sequenced manually (Sequenase, v.2.0; United
States Biochemical, Cleveland) using primers spaced at ~500-
bp intervals. The primary sequences for each hexokinase gene
were determined for both strands. In the population level
studies, all ambiguous sites were verified on both strands.

Single nucleotide polymorphism screening: Geographic
data on allozyme polymorphisms were collected by screening a
single female from each isofemale line using standard staining
techniques (Jelnes 1971). For the two third chromosome loci,
553 lethal-free third chromosomes were extracted with the
TM3 balancer. Targeted single nucleotide polymorphisms
(SNPs) were screened by performing single-base sequencing
reactions in regions spanning the target polymorphisms.

Sampling strategy and statistical methods: Sample size var-
ied among loci because of different questions associated with
specific amino acid polymorphisms. For estimation of popula-
tion genetic parameters, the subsampled eastern population
was considered the sampling unit and gene copies were sam-
ples across the range from Florida to Vermont. A stratified
sampling strategy was used. For mutations of unique interest,
such as allozymes and chromosome inversions, sampling was
expanded to examine within-lineage variation. For statistical
tests and parameter estimation, a constructed random sample
(Hudson et al. 1994) was generated for each gene from the
stratified data sets by random subsampling conditioned on
independent population estimates of targeted allozyme and
inversion frequencies. Parameters and statistics were deter-
mined using DnaSP (Rozas and Rozas 1999) and MEGA
(Sudhir et al. 1993).

To assess whether balancing or directional selection was
operating on each gene, a series of tests were carried out to
evaluate departures in the data from expectations for neutral
variation evolving in a Wright-Fisher population. To evaluate
the expected correlation between intra- and interspecific poly-
morphisms a four-locus HKA test (Hudson et al. 1987) was
carried out. Failure to reject this model would imply that
differences among locus polymorphisms arise from differ-
ences in neutral mutation rates. The two-locus version of this
test is most commonly seen, but as originally proposed by
Hudson et al. (1987) the test can be extended to *L* loci with
2f/2 d.f. Tajima (1989) and Fu and Li (1993) tests, which
examine departures from the expected neutral allele fre-
quency spectrum, were also carried out to detect historical
balancing selection or, conversely, vestiges of hitchhiking
events (Charlesworth et al. 1995; Fu 1997). The sequencing
of *D. yakuba* as an out-group permitted a designation of all changes as derived or ancestral and allowed us to use the D-statistic of Fu and Li’s (1993) test (the F-statistic gives a similar result). Monte Carlo simulations were also carried out to evaluate the distribution of silent polymorphism with respect to specific mutation events that are defined by amino acid polymorphisms (Hudson 1990, 1993). These follow the logic and design described in Eanes et al. (1996). To assess if any of the hexokinase genes possess elevated rates of amino acid substitution between species (see Akashi 1993a,b) or, conversely, elevated levels of amino acid polymorphism, we applied McDonald-Kreitman (1991) tests to each of the four genes using polymorphism and fixed differences recovered from the *D. melanogaster* and *D. simulans* data sets.

### RESULTS

**Identification and characterization of paralogous hexokinase genes:** We isolated the complete coding sequences and intervening introns for four paralogous hexokinase genes in *D. melanogaster*, *D. simulans*, and *D. yakuba*. The chromosomal locations and tissue-specific expression patterns of these genes are consistent with the three previously described *D. melanogaster* hexokinase genes inferred from isoenzyme studies (Murray and Ball 1967; Cavener 1980). Although not available at the time of recovery of these genes, a recent search of the complete *D. melanogaster* genome sequence (http://www.fruitfly.org/annot) has failed to find any other genes with hexokinase homology.

The putative *Hex-A* locus exhibits an uninterrupted coding region of 1347 bp and has subsequently been localized to polytene bands 8E10 in the Drosophila Genome Project (Gadfly gene CG3001). Its cytogenetic position had been mapped previously to band 8D using deletion mapping (Lindsay and Zimm 1992). A Northern blot analysis confirmed that *Hex-A* is strongly expressed in the thorax of male and female adult flies, with no detectable transcript in adult abdomens (results not shown). This gene was reported to be expressed primarily in adult flight muscle (Moser et al. 1980). The *Hex-C* locus exhibits an uninterrupted coding region of 1365 bp and was localized to band 51E9 by the Drosophila Genome Project (Gadfly gene CG8094). The *Hex-C* locus had been previously cytologically mapped to bands 51B–52E (Burkhart et al. 1984). A Northern blot analysis demonstrated that *Hex-C* is expressed only in the abdomens of male and female adult flies, consistent with its reported localization to fat body tissues (Moser et al. 1980). Finally, the identity of the *Hex-C* locus was unequivocally confirmed by the identification, at the nucleotide level, of two inferred charge-changing amino acid polymorphisms that are correlated with known electrophoretic allozyme polymorphisms at this locus (described below).

The third hexokinase gene recovered in our PCR using degenerate primers was localized to band 97B2 of the third chromosome by Drosophila Genome Project (Gadfly gene CG5443). The 3086-bp region revealed two tandemly repeated hexokinase genes separated by a short variable-length intergenic sequence of 185–230 bp. The first gene exhibits an uninterrupted coding sequence of 1392 bp while the second gene contains a coding sequence of 1359 bp that is interrupted by a single variable-length intron of 88 bp in *D. melanogaster* and 71 bp in *D. simulans*. This intron is missing in *D. yakuba*. The Northern blot hybridization patterns for both genes were identical and localized to the abdominal segment in adult male flies. Although the specific function for these latter hexokinase genes is not known, the Northern hybridization patterns are consistent with the expression pattern reported previously for *Hex-t*, which was shown by isoenzyme study (Murray and Ball 1967; Cavener 1980) to be male specific and localized to the testes. We have designated the two loci as *Hex-t1* and *Hex-t2*. The predicted isoelectric points of the four proteins are also consistent with the relative electrophoretic mobilities of *Hex-A*, *Hex-C*, and *Hex-T*.

Inferred amino acid sequences show high similarity to other hexokinases in regions of established conservation (Figure 1). The amino acid sequences from the four hexokinase genes indicate that *Hex-t1* and *Hex-t2* are most recently derived and because of their close linkage perhaps have originated from a tandem duplication event. Levels of amino acid divergence among the four genes are quite high, ranging from 38 to 58% (ignoring gaps). This level of divergence suggests that the duplication events that gave rise to this gene family are quite ancient, allowing ample time for divergent expression patterns and different functional constraints to have developed.

**Intraspecific polymorphism and interspecific divergence:** Tables of the distribution of polymorphic sites across lines for each of the four genes for *D. melanogaster* and *D. simulans* samples can be viewed on the Eanes lab web site (http://www.life2.sunysb.edu/ee/eaneslab/hexdata) and all 130 sequences (including *D. yakuba*) are deposited in GenBank (accession nos. AF257522–AF257652). Table 1 provides the associated summary statistics and parameter estimates for the silent sites, as well as the *Hex-t2* intron and *Hex-t1-Hex-t2* intergenic region. *Hex-A* did not exhibit any amino acid polymorphisms in either species. At the *Hex-C* locus, four amino acid replacement polymorphisms were detected in 50 gene copies of *D. melanogaster* while none were observed in 14 copies from *D. simulans*. Two of the replacement polymorphisms (nucleotides 113 and 1277) in *D. melanogaster* appear as singletons in our sample, while the other two (1049 and 1231) occur at intermediate frequencies. The polymorphism at position 1231 results in a Lys-to-Glu replacement at residue 411 and is responsible for the *Hex-C* allele (*Hex-C*, Lindsay and Zimm 1992), a widely distributed allozyme polymorphism at the *Hex-C* locus (Jelnes 1971; Mukai and Voelker 1977). The polymorphism at position 1277 results in an Ala-to-Glu replacement at residue 426 and the combined charge
Figure 1.—Amino acid sequence comparisons of the four candidate D. melanogaster hexokinases, Schistosoma mansoni hexokinase (Q26609), Saccharomyces cerevisiae hexokinase A (6321168), and human glucokinase (P35557). Glucose- and glucose-6-phosphate-binding residues are shown with dark backgrounds.
TABLE 1
Summary statistics for the silent site variation (including intron and intergenic region)

<table>
<thead>
<tr>
<th></th>
<th>Hex-A</th>
<th>Hex-C</th>
<th>Hex-t1 (exons)</th>
<th>Hex-t2 (intron)</th>
<th>Hex-t1,2 (intergenic)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Polymorphism</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>D. melanogaster</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total bases</td>
<td>1347</td>
<td>1365</td>
<td>1395</td>
<td>1359</td>
<td>88</td>
</tr>
<tr>
<td>Silent sites</td>
<td>320.3</td>
<td>334.3</td>
<td>321.4</td>
<td>319.4</td>
<td>88.0</td>
</tr>
<tr>
<td>n</td>
<td>14</td>
<td>26</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>θ</td>
<td>0.0020</td>
<td>0.0094</td>
<td>0.0258</td>
<td>0.0231</td>
<td>0.0315</td>
</tr>
<tr>
<td>θ&lt;sub&gt;lower 95%&lt;/sub&gt;</td>
<td>0.0006</td>
<td>0.0045</td>
<td>0.0142</td>
<td>0.0126</td>
<td>—</td>
</tr>
<tr>
<td>θ&lt;sub&gt;upper 95%&lt;/sub&gt;</td>
<td>0.0127</td>
<td>0.0231</td>
<td>0.0720</td>
<td>0.0660</td>
<td>—</td>
</tr>
<tr>
<td>π</td>
<td>0.0016</td>
<td>0.0152</td>
<td>0.0291</td>
<td>0.0236</td>
<td>0.0217</td>
</tr>
<tr>
<td><strong>D. simulans</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total bases</td>
<td>1347</td>
<td>1365</td>
<td>1395</td>
<td>1359</td>
<td>71</td>
</tr>
<tr>
<td>Silent sites</td>
<td>322.2</td>
<td>331.1</td>
<td>323.6</td>
<td>319.9</td>
<td>71.0</td>
</tr>
<tr>
<td>n</td>
<td>14</td>
<td>14</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>θ</td>
<td>0.0000</td>
<td>0.0095</td>
<td>0.0270</td>
<td>0.0151</td>
<td>0.0091</td>
</tr>
<tr>
<td>θ&lt;sub&gt;lower 95%&lt;/sub&gt;</td>
<td>0.0000</td>
<td>0.0030</td>
<td>0.0148</td>
<td>0.0078</td>
<td>—</td>
</tr>
<tr>
<td>θ&lt;sub&gt;upper 95%&lt;/sub&gt;</td>
<td>0.0053</td>
<td>0.0300</td>
<td>0.0790</td>
<td>0.0440</td>
<td>—</td>
</tr>
<tr>
<td>π</td>
<td>0.0000</td>
<td>0.0141</td>
<td>0.0194</td>
<td>0.0162</td>
<td>0.0087</td>
</tr>
<tr>
<td><strong>Divergence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mel/Sim</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D&lt;sub&gt;c&lt;/sub&gt;</td>
<td>0.0632</td>
<td>0.1076</td>
<td>0.1247</td>
<td>0.1600</td>
<td>0.0540</td>
</tr>
<tr>
<td>Proportion fixed</td>
<td>0.0624</td>
<td>0.0808</td>
<td>0.0653</td>
<td>0.1284</td>
<td>0.0428</td>
</tr>
<tr>
<td>Mel/Yak D&lt;sub&gt;c&lt;/sub&gt;</td>
<td>0.1444</td>
<td>0.2030</td>
<td>0.3367</td>
<td>0.3305</td>
<td>0.0412</td>
</tr>
<tr>
<td>Sim/Yak D&lt;sub&gt;c&lt;/sub&gt;</td>
<td>0.1492</td>
<td>0.1724</td>
<td>0.3109</td>
<td>0.2822</td>
<td>—</td>
</tr>
</tbody>
</table>

* 171 bp total.
* 213 bp total.
* 70 bp total.
* 170 bp total.

change with residue 411 results in the widespread but rarer Hex-C* (Hex-C*, LINDSEY and ZIMM 1992) allozyme allele. The second common amino acid polymorphism (Lys to Arg at residue 350) is seen at position 1049 and results in no charge change. We also sequenced 12 copies found on In(2R)NS-bearing chromosomes. This inversion lies distally to the Hex-C locus, and since both standard and inversion-associated copies of Hex-C show the same haplotypes and harbor both common amino acid polymorphisms, recombination between arrangements in this region of the chromosome is apparently common.

Both D. melanogaster and D. simulans possess intraspecific amino acid polymorphisms in Hex-t1 and Hex-t2. In D. melanogaster three of the amino acid polymorphisms (positions 97 in Hex-t1, 2117 and 2764 in Hex-t2) appear as singletons and the other three (positions 728 in Hex-t1, 3076 and 3081 in Hex-t2) are at intermediate frequencies. In D. simulans one of the sites (position 4 in Hex-t1) appears as a singleton, while three other sites, all in Hex-t2 (positions 1895, 2149, and 2964), are at intermediate frequencies.

While numbers of amino acid polymorphisms vary across genes, there are also marked differences in levels of silent polymorphism, varying nearly 13-fold across loci in D. melanogaster (θ ranges from 0.0020 to 0.0258).

In D. simulans the contrast is even more striking: the Hex-A locus shows no variation at silent sites (Tables 1 and 2). The levels of intraspecific polymorphism are also correlated between species. Using TAVARE’s (1984) equation 9.5 we have put approximate upper and lower confidence limits on our silent site-based θ estimates for all four loci. The 95% confidence limits of θ for Hex-A and the Hex-t1 and Hex-t2 loci do not overlap in either species (Table 1).

In contrast to the large disparities in levels of intraspecific polymorphism across genes, interspecific divergence between D. melanogaster and D. simulans is more modest (pairwise differences vary from 0.0151 to 0.0494, Table 1). The number of fixed amino acid differences per replacement site differs substantially across loci (Table 2). There were no amino acid replacements found at Hex-A, 2 fixed replacement differences were found at Hex-C, and 8 and 13 fixed amino acid replacements were observed in Hex-t1 and Hex-t2, respectively. The inclusion of sequence data for D. yakuba further extends this pattern (Table 2). Notably, there are still no amino acid differences at Hex-A.

A four-locus HKA test (HUDSON et al. 1987) was carried out to test the null hypothesis of concordance of
silent polymorphism levels with silent site divergence in D. melanogaster and D. simulans. We derived joint parameter estimates of \( \hat{\theta}_1 = 6.8 \times 10^{-3} \), \( \hat{\theta}_2 = 12.2 \times 10^{-3} \), \( \hat{\theta}_3 = 21.5 \times 10^{-3} \), \( \hat{\theta}_4 = 21.0 \times 10^{-3} \), \( \hat{T} = 6.77 \) and used these to generate expectations and variances for levels of polymorphism and divergence. For the combined species the result was not statistically significant (\( X^2 = 9.33 \), d.f. = 6), and we cannot reject the explanation that the varying silent polymorphism levels are due to among-locus variation in the neutral mutation rate. When the HKA test is partitioned into separate species tests, the D. simulans test is statistically significant (\( X^2 = 8.512 \), d.f. = 3, \( P < 0.05 \)), and the largest deviation is associated with the absence of polymorphism associated at the Hex-A locus. Although not significant, the same deviation is apparent in the D. melanogaster data, suggesting that there is possibly more to the lower polymorphism levels at Hex-A than simply decreased mutation. To increase the power the four-locus HKA analysis was also carried out separately for D. melanogaster and D. simulans using D. yakuba to estimate divergence. Deviations from neutral expectations were nonsignificant and smaller than in the joint melano-gaster-simulans contrast.

There is clear evidence for decreasing constraint in amino acid substitution from the flight- to the sperm-specific enzymes. A G-test of the homogeneity across genes for ratios of numbers of fixed synonymous and replacement mutations between D. melanogaster and D. simulans (Table 2) is highly significant (\( G = 13.78 \), d.f. = 3, \( P < 0.003 \)). As a measure of relative constraint (assuming synonymous sites are neutral), \( K_s/K_a \) ratios for all three interspecific comparisons range from 0 for Hex-A to 0.128 in Hex-t2 (for D. simulans/D. yakuba) with a rank order of Hex-A < Hex-C < Hex-t1 < Hex-t2 (Table 2). The results for the four McDonald-Kreitman (1991) tests were statistically significant (\( P < 0.006 \)) for only the Hex-t1 gene, and it appears to deviate from the expectation either because there are too many amino acid fixations or too few silent fixations.

Samples of amino acid polymorphisms were surveyed for large-scale patterns of geographic variation. The Hex-C\( ^2 \), Hex-C\( ^6 \), and Hex-C\( ^z \) alleles were screened in isofemale lines in the 10 D. melanogaster populations (\( n = 150 \)). The Hex-C\( ^6 \) allele is universally rare, while the ancestral Hex-C\( ^2 \) allele shows statistically significant frequency variation (\( F_{ST} = 0.006, P < 0.05; F_{ST} \) adjusted for sampling variance), decreasing in more northern populations and forming a subtle but statistically significant cline against latitude (\( r = 0.74, P < 0.01 \)). We also examined the pattern of variation for the Hex-C\( ^6 \) and Hex-C\( ^z \) alleles combined, which represents the change in frequency of the derived Glu mutation. That cline is also highly significant (\( r = 0.69, P < 0.02 \)) and is shown in Figure 2. The Hex-t1\( ^{772A} \) (Tyr/Phe polymorphism) allele was surveyed in 533 third chromosomes sampled across the same populations using an SNP screen. There was no significant allele frequency variation and the mean frequencies are not clinal (Figure 2). In contrast, although located downstream only 2348 bases from the Hex-t1 polymorphism, the derived Hex-t2\( ^{AAA} \) allele (the Asn-Ile haplotype) shows significant geo-

### Table 2

Summary data and test statistics for polymorphism and divergence of silent and replacement changes in the four hexokinase genes in D. melanogaster, D. simulans, and D. yakuba

<table>
<thead>
<tr>
<th>Gene</th>
<th>Hex-A</th>
<th>Hex-C</th>
<th>Hex-t1</th>
<th>Hex-t2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silent polymorphisms</td>
<td>2</td>
<td>12</td>
<td>27</td>
<td>24</td>
</tr>
<tr>
<td>Replacement polymorphisms</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Tajima’s ( D )</td>
<td>−0.532</td>
<td>1.402</td>
<td>0.509</td>
<td>0.073</td>
</tr>
<tr>
<td>Fu and Li’s ( D )</td>
<td>−0.540</td>
<td>0.897</td>
<td>0.838</td>
<td>−0.306</td>
</tr>
</tbody>
</table>

### Interspecific divergence

<table>
<thead>
<tr>
<th>Gene</th>
<th>Hex-A</th>
<th>Hex-C</th>
<th>Hex-t1</th>
<th>Hex-t2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silent fixed differences</td>
<td>20</td>
<td>27</td>
<td>21</td>
<td>41</td>
</tr>
<tr>
<td>Replacement fixed</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>McDonald-Kreitman test (( P ))</td>
<td>0.406(^a)</td>
<td>0.006(^a)</td>
<td>0.322(^a)</td>
<td></td>
</tr>
<tr>
<td>( K_s/K_a )</td>
<td>0.000</td>
<td>0.021</td>
<td>0.065</td>
<td>0.077</td>
</tr>
<tr>
<td>D. melanogaster/D. yakuba ( K_s/K_a )</td>
<td>0.000</td>
<td>0.084</td>
<td>0.110</td>
<td></td>
</tr>
<tr>
<td>D. simulans/D. yakuba ( K_s/K_a )</td>
<td>0.000</td>
<td>0.104</td>
<td>0.128</td>
<td></td>
</tr>
</tbody>
</table>

\( ^a P < 0.05 \).  
\( ^a \)Probability associated with G-test.
Drosophila Hexokinase Population Genetics

1197

ing selection (Hudson 1990; Hudson et al. 1994; Eanes et al. 1996). It should be emphasized, however, that such tests are conservative since recombination will decouple individual site genealogies and make inferences about specific sites under selection more difficult. There are a number of examples where all four gametes segregate in two-site combinations and this provides ample evidence for recombination. Using the method of Hudson and Kaplan (1985) there are estimated to be a minimum number of four recombination events in the sample of D. melanogaster Hex-C sequences from North America, while the Hex-t1 and Hex-t2 sequences show eight and seven recombination events, respectively, in the same sample.

DISCUSSION

Most enzymes of the glycolytic pathway in Drosophila are present as single forms, irrespective of the cellular environment in which they are expressed, but a few notable exceptions occur. Among these are the hexokinases. The hexokinases are believed to tightly regulate the entry of glucose into the glycolytic pathway in response to changes in reducing equivalents or metabolic charge. Consequently, features of the regulatory control function of hexokinases will be expected to vary along with the diverse requirements of various tissue environments. The four hexokinase genes that occur in the Drosophila genome show dramatically distinctive patterns of polymorphism and divergence, and this may reflect the intensities and patterns of selection imposed in order to regulate glycolytic flux in each of the unique tissue-specific environments where these genes are expressed. We begin with a discussion of general patterns of variation across all four hexokinase loci and then follow with a discussion of the specific patterns of polymorphism present at each of the individual loci.

The most notable differences exhibited among the four hexokinase paralogs were in the levels of both

Figure 2.—Plot of derived (using D. simulans as an outgroup) allele frequency vs. latitude for 10 D. melanogaster population samples from the eastern United States. The Hex-C polymorphism, the SNP sites at the Hex-t1 (Tyr/Phe polymorphism at nucleotide 728), and Hex-t2 (sites at nucleotides 3076, 3080, and 3081). Two additional points for Hex-C were added (Raleigh, NC, and Stony Brook, NY).

graphic variation ($F_{st} = 0.061, P < 0.0001$) and a highly significant latitudinal cline (Figure 2, $r = 0.80, P < 0.01$).

Although there is evidence for selection from the latitudinal clines for Hex-C and Hex-t2, there is no evidence from the Monte Carlo simulations for significant genealogical departures from an expected Wright-Fisher model. Significant departures might be expected if there were adaptive polymorphisms that had recently increased in frequency or were under persistent balanc-
lent and replacement sites suggests that the constraints operating on amino acid changes may transfer to synonymous site variation.

The substantial differences observed in silent polymorphism among the four hexokinases could be explained by extrinsic genomic factors if the recombinational landscape differed among loci. In D. melanogaster, low variation as a consequence of hitchhiking is correlated with regionally low levels of recombination, as seen in telomeric and centromeric regions (Begun and Aquadro 1992; Hudson and Kaplan 1995), or with chromosomes lacking recombination entirely (Berry et al. 1991; Zurovcova and Eanes 1999). However, since all four genes occur in regions of normal recombination, it is unlikely that hitchhiking due to selective sweeps or background selection has substantially reduced levels of variation as seen, for example, with Hex-A.

Another explanation for the concordance of polymorphism and divergence at silent and replacement sites could be if intensity of selection for codon bias varied among the four genes according to levels of functional constraint. However, levels of codon bias were unexpectedly low for all four hexokinase loci, Hex-A and Hex-C in particular. Most enzymes of central metabolism possess very high codon bias. For example, enolase (Eno), aldolase (Ald), and triose phosphate isomerase (Tpi), all farther down the glycolytic corridor, rank at or near the top of 428 genes surveyed and most others rank in the top 10% (Figure 3, R. Kliman, personal communication). This index is typically correlated with genes that undergo high rates of translation and are typically abundant (see Moriyama and Powell 1997). An explanation for the relatively low bias might be that, unlike the other steps in the pathway, hexokinases operate as nonequilibrium enzymes that have not evolved to an upper neutral limit (see Hartl et al. 1985). Enzyme levels for such regulatory steps must be fine tuned to effectors that may occur at low concentrations. As a result, extremely low interspecific divergence and a near absence of polymorphism without high codon bias, such as seen for Hex-A, could be the consequence of stabilizing selection for tightly regulated protein levels. In contrast, many equilibrium enzymes, which are not duplicated and must function in many different metabolic contexts, have been selected to operate well below their catalytic capacity and do so by being abundant in many tissues with regulation occurring by high protein turnover. This will lead to these enzymes possessing high codon bias because of the need for high translational efficiency. It will be interesting to examine the pattern of polymorphism and interspecific divergence in phosphofructokinase (Pfk), which is also a textbook regulatory enzyme and has only average codon bias (Figure 3) even though it is flanked in the glycolytic pathway by high-bias genes such as Pgi and Ald. Whether patterns of codon bias can be used to make this type of inference is an interesting question.

The HEX-A isoenzyme, which shows very little or no intra- and interspecific variation and is specialized to the adult flight muscle, exhibits the highest affinity for glucose (Moser et al. 1980), which, along with trehalose (which is converted to glucose), is used as the primary source of energy for flight in insects (Sacktor and Wormser-Shavit 1966). The importance of hexokinase as a regulator in high-metabolic-demand tissues, such as in muscle and brain, is evidenced by its being bound to the porins of the outer membrane of the mitochondria, where it can be the most sensitive to the immediate ATP demands of the cell (Wilson 1995). In contrast to other glycolytic corridor enzymes [e.g., phosphoglucose isomerase (PGI)], it is shown that in the high-flux flight muscle, hexokinase is operating at its maximum catalytic capacity or near saturation (Suarez et al. 1996, 1997). It would appear that these stringent demands have imposed very tight functional constraints. Furthermore, the Hex-A gene is highly conserved not only at the amino acid level, but also at synonymous positions and possesses the lowest $K_a/K_s$ ratios (zero) in all interspecific comparisons. The lack of silent and replacement polymorphism in both D. melanogaster and D. simulans is probably not the result of a recent hitchhiking event since it would require that such an event occur independently in each species. Moreover, for Hex-A the nearest linked coding regions are over 9 kb upstream and 12 kb downstream. An argument against internal hitchhiking events is also sup-
ported by the complete conservation of the amino acid sequence among all three species surveyed in this study.

Two isozymes are associated with this locus (Moser et al. 1980), but it is not known if these arise from alternate splicing or post-translational modification. We are not able to address this issue since our screen recovered only one cDNA, but an exhaustive search for potential exons in the 5′ and 3′ sequences flanking the gene (GenBank accession no. AC012813) failed to detect any potential exons.

The HEX-C enzyme is expressed in the adult and larval fat body and should be involved in the partitioning of energy resources and perhaps gluconeogenesis. While it exhibits activity on glucose and mannose substrates, HEX-C has its highest affinity for fructose and is likely primarily involved in metabolism of the latter (Moser et al. 1980). There are multiple potential fates for fructose. It can be kinased to fructose-6-phosphate and enter the glycolytic pathway at phosphoglucose isomerase, or to fructose-1-phosphate and then be lysed into dihydroacetone phosphate and glyceraldehyde. In this sense HEX-C serves as a branch point enzyme, not unlike G6PD and phosphoglucomutase (PGM), and is correspondingly less conserved than Hex-A at both nonsynonymous and synonymous sites. Furthermore, consistent with amino acid polymorphisms at other branch point enzymes in glycolysis (Gpdh, Oakeshott et al. 1982; G6pd, Oakeshott et al. 1983; Pgm, B. C. Verrelli and W. F. Eanes, unpublished results) we see clines in allele frequencies with the derived Hex-C′ allozyme systematically increasing with latitude in our samples (Figure 2). The Lys-to-Glu amino acid polymorphism at residue 411, responsible for this common allozyme in D. melanogaster, is in a functionally important structural position. The three-dimensional structure of hexokinase from Schistosoma reveals that the position homologous to HEX-C site 411 (also a lysine residue) is located at a bend in an a-helix that forms part of the ATP-binding domain (see Hex-C residue 411, Figure 1; Mulichak et al. 1998). While the Lys is not strictly conserved, it is never occupied by a negatively charged amino acid and is flanked by a number of conserved residues in nearly all other hexokinases.

The pair of testis-specific genes represent an old duplication event in which new roles appear to have emerged. Because the HEX-T1 enzyme does not retain crucial hexose-binding residues (discussed below), we predict that the HEX-T2 enzyme is the testis-specific form identified by Murray and Ball (1967). At Hex-t2 there are 13 amino acid replacements fixed between D. melanogaster and D. simulans (Table 2), and in D. melanogaster we see a pair of amino acid polymorphisms at the carboxy terminus of the protein. These two amino acid polymorphisms may be under latitudinal variable selection since the Asn-Ile haplotype varies over 45% in frequency in a steep cline between Vermont and Florida. This is comparable in magnitude to the well-characterized clines in Adh and Gpdh (Oakeshott et al. 1982; Berry and Kreitman 1993). There is evidence that selection is targeted to this site and not a larger chromosome region since a silent polymorphism at nucleotide site 3041, only 35 nucleotides away and also detected in our SNP screen, is only weakly associated with this pair of sites and shows no significant geographic variation. Furthermore, the common Try-Phe polymorphism in Hex-t1, which is 2395 bp away, is only in weak disequilibrium with these polymorphisms (n = 533, R² = 0.017, P < 0.0016) and also shows no cline. How selection might operate on these residues is unknown, although a look at the three-dimensional structure of hexokinase in Schistosoma reveals that the carboxy terminus, while not highly conserved, lies in very close proximity to the glucose-binding domain.

We hypothesize that the Hex-t1 gene has acquired a new sperm- or testis-related function that has co-opted its hexokinase properties. The overall amino acid alignment shows Hex-t1 is unmistakably related to other hexokinases, yet it shows changes in 6 of the 13 residues considered important for glucose and glucose-6-phosphate binding (see Figure 1; Wilson 1995; Mulichak et al. 1998), and these residues are all conserved in the HEX-A, HEX-C, and HEX-T2 proteins, with a single exception in HEX-T2. The importance of these critical residues is well established and mutations at these sites are associated with maturity onset type diabetes, in the human 50-kD pancreatic glucokinase homologue (see Gidhjain et al. 1993; Velho et al. 1997).

Nonetheless, even with this apparent loss of hexose-binding ability, Hex-t1 is still clearly under selective constraint. Most of the variable sites in Hex-t1 are synonymous and the average Ks/Ka ratio (Table 2) is smaller than that for Hex-t2, implying greater constraint. When the physiochemical differences (see Grantham 1974) for each amino acid substitution among all three species are examined, the mean differences for Hex-t1 and Hex-t2 are not statistically different when compared by Wilcoxon two-sample test (Sokal and Rohlf 1981). This indicates that on the average replacement substitutions at Hex-t1 are no less physiochemically conservative than changes at Hex-t2. Most importantly, by the criteria of the McDonald-Kreitman test Hex-t1 shows a significant excess of interspecific replacement substitutions and this argues that it has been subjected to directional selection.

If Hex-t1 and Hex-t2, which are relatively weakly expressed only in the whole abdomens of males, are functioning exclusively in the sperm (or are more broadly expressed in the testes) then they could be influenced by sexual selection. In mammalian sperm there are multiple isoforms of hexokinase-1, and a germ line isoform, HK-sc, is believed to act in both energy generation and as a sperm surface receptor for a ligand in the zona pellucida (Travis et al. 1998). This suggests that the glucose-binding ability of hexokinases can be co-opted.
for carbohydrate binding in the extracellular matrix of the egg. With respect to the Drosophila hexokinases this is only speculation, but the studies in mammal sperm emphasize the multiple roles that hexokinases can play. Given the recurring observation of new function in genes associated with testis-specific specialization (Nurminsky et al. 1998) and the associated observation that male-specific genes are under accelerated evolution (Eanes and Wyckoff et al. 2000), the pattern at Hex-t1 further extends this class of observations to an enzyme whose function was probably limited to glycolytic energy production in sperm before duplication.

Patterns of polymorphism are more complex in D. simulans at Hex-C, Hex-t1, and Hex-t2. The Hex-C locus has no amino acid polymorphisms, and there is very strong linkage disequilibrium (resulting in a significant positive Fu and Li D-statistic), where only three haplotype types emerge from 10 segregating polymorphisms. Likewise, there is very strong linkage disequilibrium in the Hex-t1, -t2 region. The Hex-t1 locus shows only a singleton amino acid variant, while Hex-t2 shows a complex set of replacement polymorphisms: a common ancestral amino acid sequence segregating most of the silent-site polymorphisms at the locus, and two derived sequences, one bearing two amino acid differences from the ancestral sequence and a second bearing a single difference from the ancestral sequence. Both derived sequences show no within-lineage variability and this extends across the entire Hex-t1, -t2 region. Such complex haplotype structure is a recurring theme in D. simulans (Hasson et al. 1998; Hamblin and Veukille 1999; Labate et al. 1999) and could be the result of historical population subdivision. D. simulans simply may not be a good candidate for a species in genetic equilibrium, and this makes contrasts of polymorphism with D. melanogaster less informative.

Finally, the polymorphism difference between species also differs with other studies. D. simulans possesses lower amino acid polymorphism than D. melanogaster, consistent with other studies, but it also possesses lower silent polymorphism (f = 0.86 from the HKA test), unlike the general trend seen for the loci in Moriyama and Powell’s (1996) survey, where silent polymorphism is on average two-to threefold higher. Reduced levels of replacement polymorphism in D. simulans are generally assumed to reflect greater efficacy of purifying selection in a species with larger effective population size (Aquadro et al. 1988; Akashi 1995). Therefore, reduced levels of silent polymorphism in D. simulans may provide an additional indication that silent sites are subject to purifying selection at hexokinase loci. Finally, while in D. melanogaster amino acid polymorphisms in several metabolic genes show evidence of age and geographic patterning favoring selection (Eanes 1999), D. simulans continues to show no such patterns.

We thank Brian Verrelli, Luciano Matzkin, Efe Sezgin, and Yihao Duan for their contributions to numerous genetic constructions and other demands of this project. This project was supported in part by U.S. Public Health Service grant GM-15247 and National Science Foundation grant DEB-9318831 to W. F. Eanes. This is contribution number 1057 from the Graduate Program in Ecology and Evolution at Stony Brook.

LITERATURE CITED


Akashi, H., 1999a Inference of the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151: 221–238.


Hamblin, M. T., and M. Veukille, 1999 Population structure among African and derived populations of Drosophila simulans: evidence...


Sudder, K., K. Tanuma and M. Nei, 1993 MEGA: Molecular Evolutionary Genetics Analysis, version 1.0. The Pennsylvania State University, University Park, PA.


Communicating editor: C.-I Wu