MOLECULAR POPULATION GENETICS OF THE ALCOHOL DEHYDROGENASE GENE REGION OF DROSOPHILA MELANOGASTER

CHARLES F. AQUADRO,* SUSAN F. DESSE,* MOLLY M. BLAND,*,+ CHARLES H. LANGLEY* AND CATHY C. LAURIE-AHLBERG†

* Laboratory of Genetics, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, and † Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695

Manuscript received January 12, 1985
Revised copy accepted September 4, 1986

ABSTRACT

Variation in the DNA restriction map of a 13-kb region of chromosome II including the alcohol dehydrogenase structural gene (Adh) was examined in Drosophila melanogaster from natural populations. Detailed analysis of 48 D. melanogaster lines representing four eastern United States populations revealed extensive DNA sequence variation due to base substitutions, insertions and deletions. Cloning of this region from several lines allowed characterization of length variation as due to unique sequence insertions or deletions [nine sizes; 21–200 base pairs (bp)] or transposable element insertions (several sizes, 340 bp to 10.2 kb, representing four different elements). Despite this extensive variation in sequences flanking the Adh gene, only one length polymorphism is clearly associated with altered Adh expression (a copia element approximately 250 bp 5' to the distal transcript start site). Nonetheless, the frequency spectra of transposable elements within and between Drosophila species suggests they are slightly deleterious. Strong nonrandom associations are observed among Adh region sequence variants, ADH allozyme (Fast vs. Slow), ADH enzyme activity and the chromosome inversion In(2L)t. Phylogenetic analysis of restriction map haplotypes suggest that the major twofold component of ADH activity variation (high vs. low, typical of Fast and Slow allozymes, respectively) is due to sequence variation tightly linked to and possibly distinct from that underlying the allozyme difference. The patterns of nucleotide and haplotype variation for Fast and Slow allozyme lines are consistent with the recent increase in frequency and spread of the Fast haplotype associated with high ADH activity. These data emphasize the important role of evolutionary history and strong nonrandom associations among tightly linked sequence variation as determinants of the patterns of variation observed in natural populations.

FOR the past two decades, population geneticists have focused on electrophoretically detectable protein polymorphism as an index of genetic variation. A great wealth of variation has been found in natural populations. However, unequivocal determination of the evolutionary significance of most

1 Present address: Section of Genetics and Development, Cornell University, Ithaca, New York 14853.
2 To whom correspondence should be addressed.

of this allozyme polymorphism has eluded even the most single-minded efforts (e.g., Kimura 1982; Koehn, Zera and Hall 1983; Nei and Graur 1984). This failure has contributed to renewed speculation that changes in the expression of structural genes might be of major, if not primary, evolutionary importance (e.g., Britten and Davidson 1969; Wilson 1976; MacIntyre 1982). These arguments have been fueled by the disparity between the total size of the eukaryotic genome and the portion estimated to be structural genes (see Lefevre and Watkins 1986; Lewin 1985). Is much of the genome simply spacer or dispensable DNA? How much and what kind of variation exists in these nontranscribed sequences, and of what relevance is it to the individual’s fitness?

Advances in molecular biology over the past several years have made it possible to survey DNA sequence variation within not only the transcribed structural genes but also the nontranscribed matrix in which these structural genes are embedded. In addition, application of these new molecular approaches has revealed a whole new source of variation, that due to transposable elements. In fact, a great many classical spontaneous “point” mutants at numerous loci in Drosophila melanogaster are due to the insertion or imprecise excision of transposable elements (Rubin 1983). With their potential for rapid movement around the genome, implication in altered expression of genes and association with interstrain sterility (e.g., the P-element), a great deal of excitement has been generated as to their role in, for example, reproductive isolation and speciation (Bingham, Kidwell and Rubin 1982; Rose and Doolittle 1983; Ginzburg, Bingham and Yoo 1984). Variation in transposable element position and their occurrence adjacent to various genes from natural populations has been recently documented (Montgomery and Langley 1983; Leigh Brown 1983). Nonetheless, the evolutionary significance of this newly accessible diversity remains largely unknown.

The successful examination of the quantity and significance of genetic variation in these newly accessible regions of the genome relies heavily on having a wealth of information on the molecular biology of the gene of interest and the biochemistry, ecological relevance and evolutionary biology of the gene product. One particularly suitable region is that around the structural gene encoding alcohol dehydrogenase (Adh) in D. melanogaster.

Alcohol dehydrogenase is a typical eukaryotic gene in most ways. The 255 amino acid sequence is encoded by three exons which are shared by two developmentally regulated transcripts (distal/adult and proximal/larval), differing only in their 5’ untranslated leader sequences (Benyajati et al. 1983; Savakis and Ashburner 1985). Promoter sequences (TATA boxes) are located approximately 30 base pairs (bp) 5’ to the start sites of both distal and proximal transcripts, the latter being contained within an intron of the distal transcript. Thus, the region shown to be involved in Adh transcription spans approximately 1.9 kb of genomic DNA.

Studies of D. melanogaster populations worldwide have revealed a pattern of Adh polymorphism characterized by two common electrophoretically detectable protein variants, Fast and Slow, which show similar frequency clines on several
continents (David 1982; Oakeshott et al. 1982; Singh, Hickey and David 1982). In addition, a third variant, distinguishable from Fast by its greater thermal stability, is apparently widespread in distribution, although generally in very low frequency (Gibson, Wilks and Chambers 1982; Chambers, Wilks and Gibson 1984; Sampsell and Steward 1983). The two predominant ADH electromorphs differ in a variety of enzymatic characteristics ($V_{\text{max}}, K_m$, thermal stability). There is also a strong correlation between total ADH activity, level of ADH protein and allozyme [Fast being generally twice as concentrated as Slow on a per fly basis; see Chambers (1981) for a review].

Both protein and DNA sequencing suggest that only two forms of the ADH protein exist at appreciable frequency in most natural populations, and that they differ by only a single lysine (Slow) to threonine (Fast) replacement (Thatcher 1980; Benyajati et al. 1981; Kreitman 1983). The distribution of this protein polymorphism in natural populations, together with the biochemical differences between electromorphs and the frequency changes observed in population cages in response to ethanol stress, have led to the proposition that the observed protein variation at the Adh locus is maintained by some form of natural selection (see Van Delden 1982).

We present here a detailed analysis of genomic DNA restriction map variation in a 13-kb region around the Adh gene isolated from several natural populations of D. melanogaster, documentation of substantial insertion variation in flanking sequences as due to transposable elements, and an examination of the possible effects of the sequence variation of the expression of Adh and on individual fitness. We also examine patterns of linkage disequilibrium among DNA sequence variants (as well as with a second chromosome inversion) for clues to the contribution of drift, selection and evolutionary history to the pattern of polymorphism observed in the Adh region.

MATERIALS AND METHODS

Samples and strategy: Our approach has been to examine restriction map variation in an approximately 13-kb region containing Adh in a set of 48 D. melanogaster lines, each homozygous for a second chromosome sampled from one of four natural populations in the eastern United States (Adh is located on the second chromosome). The X and third chromosomes are homozygous across lines for chromosomes from a line from Japan (included here as the 49th line). In this way, variation at modifier loci on the X and third chromosomes that affect ADH activity is controlled (McDonald and Ayala 1978; Laurie-Ahlberg et al. 1980). These lines are listed in Table 1 and have been the subject of extensive analysis for the sources of variation in enzyme activity for several proteins including ADH (e.g., Laurie-Ahlberg et al. 1980, 1982; Maroni et al. 1982). Details of the construction of these lines are given in Laurie-Ahlberg et al. (1980).

ADH activity assays: The ADH activities for the 49 lines represent the unweighted average of the means of two different experiments, one performed in 1980 and the other in 1982. In each case, the isogenic lines were sampled according to a randomized block design. In the 1980 experiment (described further in Laurie-Ahlberg et al. 1982), the design included six observations per line; the 1982 experiment included 24 observations per line. Assay conditions were as described (Stam and Laurie-Ahlberg 1982) with activity adjusted by linear regression on weight and general protein content of the flies in each homogenate.
Restriction map analysis and cloning: Genomic DNA from each line was prepared as described (Bingham, Levis and Rubin 1981), digested with single or pairs of restriction endonucleases (Maniatis, Fritsch and Sambrook 1982) and size-fractionated on agarose gels (0.8–2.0% agarose, depending on the fragment sizes expected; McDonnell, Simon and Studier 1977). DNA fragments were transferred to nitrocellulose (Southern 1975), which was then hybridized, washed and autoradiographed as described by Wahl, Stern and Stark (1979). Plasmid probes containing D. melanogaster DNA sequences between EcoRI sites at position −2.8 and +2.1 (Figure 1) and between SacI sites (not shown) at −6.8 and +4.6 (SacI and sAS1, respectively; Goldberg 1980) were nick-translated according to Rigby et al. (1977).

Insertions were cloned as follows. Libraries were made in EMBL-4 (Friscauf et al. 1983) from complete BglII digests of genomic DNA from each insert line. Clones containing the Adh region and adjacent insertions were isolated by plaque hybridization (Benton and Davis 1977) using SacI as a probe. The right half of insert “m” and adjacent Adh region sequence was cloned using an M13 clone of genomic sequence from the SalI site at +2.3 to EcoRI site +4.2 as a probe (clone a gift from W. Quattlebaum and G. Carmody). This was necessary due to the presence of a BglII site within insertion “m.”

Initial characterization of a deleted sequence or an insertion as repeated or unique was based on use of the cloned region to probe a Southern blot of EcoRI-digested genomic DNA. Those clones containing repeated sequences (as evidenced by the probing of a large number of bands) were then used to probe a nitrocellulose filter dotted with DNA from clones representing the majority of currently known and characterized transposable elements from D. melanogaster. This transposable element filter was prepared by dotting approximately 100 ng of DNA in 1–5 µl of buffer (10 mM Tris-HCl, 1 mM EDTA, pH 7.5) directly onto a dry filter, which was then allowed to dry at room temperature, denatured in 0.5 M NaOH, 1.5 M NaCl for 20 sec, neutralized in 0.5 M Tris, 3.0 M NaCl for 3 min, blotted dry and baked for 2 hr at 80°C in a vacuum. The filter was then probed as for a genomic Southern blot, as described above. After autoradiography, the transposable element filter was washed according to Thomas (1980; modified to contain 25% formamide) and then prehybridized and hybridized with a new insert clone. Homology to particular transposable elements of inserts in the Adh region identified by this method were subsequently confirmed by reciprocal probing of Southern blots of the Adh region clones and the transposable element clones (run, transferred and probed as described for genomic Southern blots, above). The following transposable-element containing clones were screened for homology to the Adh region inserts: cDm5002 (copia), cDm2042 (412), cDm4006 (297), pD75.3 and FB-4 (Foldback elements), pr25.1 (P element), B104A and B104B (B104 elements), cDm2173 (homologous to B104 element, unpublished data), cDm2161 (an F-like element, unpublished data), cDm2144, cDm2154, cDm2156, cDm2157, cDm2179, cDm2180, cDm2186, cDm2198, cDm2210, cDm2217, cDm2218, cDm2219, cDm2228. Plasmids for these clones were kindly provided by G. M. Rubin and K. O’Hare, S. Potter, V. Pirrotta and E. Strobel. Descriptions of clones are given in Rubin et al. (1981), Levis, Collins and Rubin (1982), Potter et al. (1980) and Scherer et al. (1982).

RESULTS

Restriction map variation

Restriction site polymorphism: A summary restriction map for the 48 second-chromosome lines of D. melanogaster from four eastern United States populations (and one line from Japan) is shown in Figure 1. The Adh structural gene is located in the approximate center of this 13-kb region. We have mapped and scored 30 restriction sites, eight of which are polymorphic. From
Figure 1.—Summary restriction site map of naturally occurring DNA sequence variation in the Adh region of D. melanogaster. The Adh gene is located approximately in the center of the region examined, as indicated by the boxed region (filled boxes indicate the three coding exons). Insertions and deletions, relative to the most common restriction map, are indicated approximately to scale by triangles pointing toward or away from the map, respectively. Each size insertion/deletion is denoted by a lower case letter. Insertion/deletions of unique sequence as identified by Southern blot analysis are indicated in solid black. The identity of transposable element insertions is indicated in the upper left of the figure. Variant and invariant restriction sites are indicated above and below the line, respectively, as follows: B (BamHI), Bg (BglII), Bn (BanII), E (EcoRI), H (HindIII), Hp (HpaI), S (Sall), and X (Xhol). Each of these sites was scored for all 49 lines, with the exception of the HpaI sites in the adult intron of Adh and in insert “i” which were examined only in the lines with insertions “i,” “j” or deletions “q” or “h” in order to refine the location of these variants. The scale, in kilobases (kb), is oriented with 0.0 at the BamHI site in the Adh gene. Insertion/deletion sizes were estimated to be as follows: a = 40 bp; b = 400 bp; c = 700 bp; d = 180 bp; f = 34 bp; g = 200 bp; h = 27 bp; i = 5.2 kb; j = 31 bp; k = 340 bp; m = 10.2 kb; n = 450 bp; o = 4.8 kb; p = 21 bp; q = 38 bp; r = 28 bp.

The frequency of site variation in these samples we estimate that one in 37 nucleotides is polymorphic (approximately 347 nucleotide sites over the 13-kb region examined; equation 7 of Hudson 1982), and that between any two randomly chosen chromosomes, one in 156 nucleotide sites will differ (83 over the 13-kb region; equation 11 of Engels 1981). These estimates may be slightly biased since our choice of restriction enzymes was not entirely random (BanII was chosen because it allowed us to score two nucleotide polymorphisms detected by Kreitman's (1983) sequencing). Nonetheless, this level of variation is similar to that previously reported for this region for 18 of the same 49 lines by Langley, Montgomery and Quattlebaum (1982) from restriction map variation, and to that reported by Kreitman (1983) for exons of the Adh gene as measured by direct sequence analysis. This level is, however, lower than that observed by Kreitman (1983) for introns and silent sites of the Adh gene, consistent with Kreitman and Aguadé's (1986) evidence that an excess of nucleotide polymorphism exists within Adh relative to surrounding sequences in D. melanogaster.

All lines have also been scored for their ADH allozyme (Fast vs. Slow) by
## Table 1

Adh region variation among *D. melanogaster* lines

<table>
<thead>
<tr>
<th>Line</th>
<th><strong>In(2L)B</strong></th>
<th>BamHI</th>
<th>EcoRI</th>
<th>HindIII</th>
<th>BanII</th>
<th>h, q</th>
<th>HindIII</th>
<th>BanII</th>
<th>Allo.</th>
<th>BanII</th>
<th>XhoI</th>
<th>k, b, c, m, n, o, a</th>
<th>ADH activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>RI42</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>-</td>
<td>+</td>
<td>1</td>
<td>2.14</td>
</tr>
<tr>
<td>R131</td>
<td><strong>In</strong></td>
<td>-</td>
<td>p</td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td>-</td>
<td>S</td>
<td>+</td>
<td>2</td>
<td>2.49</td>
</tr>
<tr>
<td>NC35</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>f</td>
<td>+</td>
<td></td>
<td>+</td>
<td>-</td>
<td>S</td>
<td>+</td>
<td>3</td>
<td>2.51</td>
</tr>
<tr>
<td>RI16</td>
<td><strong>In</strong></td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>+</td>
<td></td>
<td>+</td>
<td>-</td>
<td>S</td>
<td>+</td>
<td>4</td>
<td>2.66</td>
</tr>
<tr>
<td>KA33</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>+</td>
<td>-</td>
<td>5</td>
<td>2.72</td>
</tr>
<tr>
<td>NC14</td>
<td>+</td>
<td>+</td>
<td>r</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>S</td>
<td>+</td>
<td>6</td>
<td>2.82</td>
</tr>
<tr>
<td>NC11</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>f</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>+</td>
<td>-</td>
<td>7</td>
<td>2.84</td>
</tr>
<tr>
<td>RI25</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-f</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>+</td>
<td>-</td>
<td>8</td>
<td>2.94</td>
</tr>
<tr>
<td>W107</td>
<td>+</td>
<td>p</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>-</td>
<td>+</td>
<td>9</td>
<td>3.05</td>
</tr>
<tr>
<td>KA27</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td>q</td>
<td></td>
<td>-</td>
<td>S</td>
<td>-</td>
<td>+</td>
<td>10</td>
<td>3.13</td>
</tr>
<tr>
<td>RI41</td>
<td><strong>In</strong></td>
<td>-</td>
<td>p</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td>-</td>
<td>S</td>
<td>-</td>
<td>b</td>
<td>11</td>
</tr>
<tr>
<td>KA06</td>
<td>+</td>
<td>+</td>
<td>r</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>-</td>
<td>+</td>
<td>d</td>
<td>12</td>
</tr>
<tr>
<td>RI36</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td></td>
<td>S</td>
<td>-</td>
<td>+</td>
<td>d</td>
<td>13</td>
<td>3.29</td>
</tr>
<tr>
<td>W109</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td></td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>14</td>
<td>3.30</td>
</tr>
<tr>
<td>NC25</td>
<td>+</td>
<td>+</td>
<td>r</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td>S</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>15</td>
<td>3.33</td>
</tr>
<tr>
<td>NC23</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>f</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>+</td>
<td>-</td>
<td>16</td>
<td>3.35</td>
</tr>
<tr>
<td>RI37</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td></td>
<td>S</td>
<td>-</td>
<td>+</td>
<td>d</td>
<td>17</td>
<td>3.42</td>
</tr>
<tr>
<td>RI45</td>
<td><strong>In</strong></td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>-</td>
<td>+</td>
<td>18</td>
<td>3.47</td>
</tr>
<tr>
<td>KA16</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>-</td>
<td></td>
<td>S</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>19</td>
<td>3.55</td>
</tr>
<tr>
<td>RI105</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>-</td>
<td></td>
<td>S</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>20</td>
<td>3.60</td>
</tr>
<tr>
<td>RI08</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>f</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>-</td>
<td>+</td>
<td>d</td>
<td>21</td>
</tr>
<tr>
<td>RI03</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td></td>
<td>S</td>
<td>-</td>
<td>+</td>
<td>m</td>
<td>a</td>
<td>22</td>
</tr>
<tr>
<td>KA14</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>f</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>+</td>
<td>+</td>
<td>23</td>
<td>3.86</td>
</tr>
<tr>
<td>RI28</td>
<td><strong>In</strong></td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>-</td>
<td>+</td>
<td>24</td>
<td>3.96</td>
</tr>
<tr>
<td>RI02</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>-</td>
<td></td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>4.12</td>
</tr>
<tr>
<td>KA12</td>
<td>+</td>
<td>+</td>
<td>p</td>
<td></td>
<td>-</td>
<td>+</td>
<td></td>
<td>+</td>
<td>-</td>
<td>F</td>
<td>-</td>
<td>26</td>
<td>4.13</td>
</tr>
<tr>
<td>W103</td>
<td>+</td>
<td>p</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>-</td>
<td>+</td>
<td>27</td>
<td>4.18</td>
</tr>
<tr>
<td>RI06</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>S</td>
<td>+</td>
<td>-</td>
<td>28</td>
<td>4.21</td>
</tr>
<tr>
<td>Line</td>
<td>( In(2L)k )</td>
<td>( BanH1 )</td>
<td>( EcoRI )</td>
<td>( g, r )</td>
<td>( HindIII )</td>
<td>( BglII )</td>
<td>( q )</td>
<td>( HindIII )</td>
<td>( j )</td>
<td>( BglII )</td>
<td>( Allo. )</td>
<td>( BanH1 )</td>
<td>( XhoI )</td>
</tr>
<tr>
<td>------</td>
<td>--------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>R115</td>
<td>+</td>
<td>p</td>
<td>+</td>
<td>f</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>S</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>c</td>
<td>29</td>
</tr>
<tr>
<td>R132</td>
<td>+</td>
<td>p</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>h</td>
<td>+</td>
<td>-</td>
<td>F</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>30</td>
</tr>
<tr>
<td>R109</td>
<td>+</td>
<td>p</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>b</td>
<td>31</td>
</tr>
<tr>
<td>KA01</td>
<td>+</td>
<td>p</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>b</td>
<td>32</td>
</tr>
<tr>
<td>KA15</td>
<td>+</td>
<td>p</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>b</td>
<td>33</td>
</tr>
<tr>
<td>RI01</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>j</td>
<td>F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>b</td>
<td>34</td>
</tr>
<tr>
<td>KA25</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>j</td>
<td>F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>b</td>
<td>35</td>
</tr>
<tr>
<td>RI43</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>j</td>
<td>F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>36</td>
</tr>
<tr>
<td>RI34</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>j</td>
<td>F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>37</td>
</tr>
<tr>
<td>NC16</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>j</td>
<td>F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>38</td>
</tr>
<tr>
<td>KA10</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>F</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>39</td>
</tr>
<tr>
<td>WI12</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>F</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>40</td>
</tr>
<tr>
<td>RI46</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>F</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>41</td>
</tr>
<tr>
<td>WI05</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>j</td>
<td>F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>42</td>
</tr>
<tr>
<td>HO-R</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>g</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>43</td>
</tr>
<tr>
<td>KA03</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>44</td>
</tr>
<tr>
<td>WI08</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>j</td>
<td>F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>45</td>
</tr>
<tr>
<td>KA13</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>j</td>
<td>F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>46</td>
</tr>
<tr>
<td>RI07</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>j</td>
<td>F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>47</td>
</tr>
<tr>
<td>R123</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>j</td>
<td>F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>48</td>
</tr>
<tr>
<td>KA21</td>
<td>-</td>
<td>p</td>
<td>+</td>
<td>g</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>k</td>
<td>49</td>
</tr>
</tbody>
</table>

\( Adh \) region variation for 48 eastern United States populations and a line from Japan (HO-R). The lines are ranked according to \( Adh \) activity (from Figure 3), from lowest to highest activity. Four U.S. populations are represented: NC, North Carolina; RI, Rhode Island; WI, Wisconsin; and KA, Kansas (Laurie-Ahlberg et al. 1980). Polymorphic restriction sites, insertions, deletions and \( Adh \) allozyme are given in the 5' to 3' order along the \( Adh \) region restriction map and are identified in Figure 1. \( In(2L)k \) is a cosmopolitan inversion with breakpoints distal to \( Adh \) at polytene bands 22D8-E1 and 34A8-9 (Lindsley and Grell 1968); \( Adh \) is at 35A-B (Woodruff and Ashburner 1979). The presence of a restriction site is indicated by a "+" and absence of the site by a "-". \( Adh \) allozyme (Allo.) is indicated as "F" for fast or "S" for slow as determined by starch gel electrophoresis (Laurie-Ahlberg et al. 1980). Lines with \( In(2L)k \) or insertions/deletions are indicated by the symbol "In" or the appropriate letter, respectively. Other lines are standard gene arrangement on the left arm of the second chromosome.
starch gel electrophoresis (Laurie-Ahlberg et al. 1980). From Kreitman's (1983) sequencing we infer that the difference in electrophoretic mobility, due to a Thr vs. a Lys at amino acid position 192, reflects a single nucleotide substitution and, thus, provides an additional marker at both the protein and DNA level.

**Sequence length variation:** Eighty percent of the lines were found to have at least one insertion or deletion relative to the consensus restriction map shown in Figure 1. These insertions/deletions range in size from approximately 21 bp to 10.2 kb in length. This length variation is distributed over most of the 13-kb region, with some notable exceptions. Clustering of length variation is apparent approximately 1.5 kb 3' to Adh. Four sizes of inserts, ranging from 340 bp to 10.2 kb, were observed in eight of the 49 lines (insert “b” occurs in five different lines). No insertions/deletions were observed in Adh coding sequence or in the 3' untranscribed region found to be so conservative on the basis of direct sequence by Kreitman (1983). Three small (27–38 bp) insertions/deletions were, however, located in sequence corresponding to the 5' untranslated portion of the distal Adh transcript.

In an effort to determine the origins of and mechanisms resulting in the sequence length variation observed, we have cloned and characterized the 11.8-kb BglII genomic DNA fragment containing the Adh gene from each of the 13 lines with inserts (Figure 1). Lines containing deletions were not cloned since we could determine the nature of the sequence (unique vs. repetitive) from the original clones for the region (sAS1 and sAC1 of Goldberg 1980). As indicated in Figure 1, the length variation falls into two classes: unique sequence events (DNA involved not present elsewhere in the genome) or repeated sequence events due to the presence of transposable elements (determined by Southern blots analysis, restriction mapping and direct cross hybridization). Unique sequence events refer here only to the sequence inserted or deleted as not being repeated elsewhere in the genome and is not meant to imply that they may not represent the imprecise excision of a transposable element (although we have no data to support, or refute, the latter mechanism of their origin).

Unique sequence length variation was of two types. Thirty percent of the lines contain one or both of two insertions, estimated to be 31 and 34 bp in length, respectively. One of these (“j” in Figure 1) probably corresponds to insert 2 of Kreitman (1983) judging from its size, location and relatively high frequency among Fast sequences. Sixty-one percent of the lines contain one or more of seven deletions ranging in size from 21 to 200 bp. Their locations are indicated in Figure 1 and their distributions among the 49 lines is noted in Table 1.

Seven sizes of insert, observed in 11 lines, were found to be repeated throughout the genome. By hybridization to a nitrocellulose filter dotted with DNAs representing a variety of transposable elements from D. melanogaster, and by subsequent reciprocal Southern blot probing and restriction mapping, we determined that all seven sizes had sequence homology to four different transposable elements. A 5.2-kb insert, occurring approximately 250 bp 5' to
the beginning of the distal Adh transcript in one line (RI42), appears to represent a complete copia element (FINNEGAN et al. 1978; POTTER et al. 1979), whereas a 4.8-kb insert 3' to Adh in another line appears to be homologous to a complete F element [101 of DAWID et al. (1981) and DiNocera, DIGAN and DAWID (1983); "giminy" of BENDER, SPIERER and HOGNESS (1983)].

Interestingly, all 11 inserts (four sizes) in the approximately 230-bp EcoRI/SalI fragment 1.5 kb 3' to Adh share sequence homology with a relatively uncharacterized genomic clone containing repeated sequences that was isolated and termed cDm2161 by RUBIN et al. (1981). We thus refer to this repetitive element as 2161. Our own sequencing of the Adh region 2161 inserts indicate they are similar in structure to F elements in lacking terminal repeats and having a polyadenylate terminus (C. AQUADRO, W. QUATTLEBAUM, D. BILLINGS and C. LANGLEY, unpublished results). These elements do not, however, show significant sequence homology to previously characterized F elements. The three smaller insert sizes at this site (340, 400 and 700 bp) have homology to 2161 alone, and represent abbreviated and/or defective copies of a complete element in much the manner as seen with P elements (O'HARE and RUBIN 1983). However, the fourth insert size in this EcoRI/SalI fragment (an insert of 10.2 kb, "m" in Figure 1) shows homology to 2161 at both ends, whereas the central 9 kb represents an apparently complete copy of the element B104 (SCHERER et al. 1982; MEYEROWITZ and HOGNESS 1982). Whether the insertion of B104 into the 2161 element occurred at this site or at some other location with the double element subsequently transposing into the Adh region is unknown. However B104 and 2161 do not appear to share any sequence homology based on reciprocal Southern blot analysis. The remaining repetitive insert, "n", appears to represent a 400-bp piece of a B104 element. Whether it represents imprecise excision of a B104 element or is itself still transposable is unknown.

Haplotype diversity: Few lines are identical at the restriction map level. Twenty-nine restriction map haplotypes are observed among the 49 chromosomes. Most are observed only once or a few times, with the most frequent haplotypes being found only five times (Table 1). The breakdown of this haplotype diversity by type of variation or by ADH allozyme is informative (Table 2). In viewing the relative diversity of restriction site haplotypes with those based on sequence length variation, it must be remembered that all length variation (over probably 30 bp on average) is detected, whereas only a fraction of the nucleotide differences are observed, because our seven restriction enzymes assay only a portion of the total sequence over the 13-kb region. It is quite likely that no two sequences examined are identical over the 13 kb. Haplotype diversity assayed by restriction site polymorphism does, however, provide a useful relative measure of diversity.

A large amount of haplotype diversity clearly exists even when considering either sequence length or restriction site variation alone. Considering all variation in the 13-kb region, Fast allozyme chromosomes are similar to Slow allozyme chromosomes in the level of haplotype diversity. This comparison, however, masks a significant difference between Fast and Slow allozyme chro-
TABLE 2

Adh restriction map haplotype diversity

<table>
<thead>
<tr>
<th>Variation considered</th>
<th>Fast and Slow (49)</th>
<th>Slow (32)</th>
<th>Fast (17)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restriction sites</td>
<td>0.856 (13)</td>
<td>0.819 (9)</td>
<td>0.419 (4)</td>
</tr>
<tr>
<td>Sequence length variation</td>
<td>0.917 (19)</td>
<td>0.863 (11)</td>
<td>0.890 (9)</td>
</tr>
<tr>
<td>Unique sequence insertions and deletions</td>
<td>0.853 (12)</td>
<td>0.744 (8)</td>
<td>0.794 (5)</td>
</tr>
<tr>
<td>Transposable elements</td>
<td>0.222 (8)</td>
<td>0.339 (5)</td>
<td>0.507 (3)</td>
</tr>
<tr>
<td>All variation</td>
<td>0.972 (29)</td>
<td>0.946 (17)</td>
<td>0.956 (12)</td>
</tr>
</tbody>
</table>

Haplotype diversity in the Adh region is tabulated for the complete data set (49 chromosomes) and for the 32 Slow and 17 Fast allozyme chromosomes separately. Haplotypes are based on all characters scored in the 13-kb region (from Table 1) as well as several subsets of these characters: Restriction sites (including allozyme variation) and sequence length variation due to unique sequence insertions and deletions and transposable element insertions. Shown is a measure of haplotype diversity that is influenced not only by the number of haplotypes observed but also by their frequencies in the sample (ranges from 0 to 1.0, where 1.0 indicates the highest diversity, with each chromosome a unique haplotype). This measure is calculated as

\[ h = \frac{n(1 - \sum x_i^2)}{(n - 1)} \]

where \( n \) is the number of chromosomes sampled and \( x_i \) is the frequency of the \( i \)th haplotype. Also given in parentheses are the numbers of different haplotypes for each category.

Haplotype diversity of the Adh region is revealed by examination of haplotype diversity for each allozyme considering restriction sites separately from unique sequence insertion/deletion and transposable element insertion variation. A substantial portion of the haplotype diversity of Fast chromosomes is due not to extensive restriction site variation but, rather, to a high frequency of sequence length variation among Fast chromosomes (Table 1). Length variation haplotype diversity for Fast chromosomes is comparable to, even slightly higher than, that for Slow chromosomes. For transposable element haplotypes, Fast chromosomes are significantly more variable (\( P < 0.01 \)) than Slow chromosomes using haplotype diversity variances of \( 1.5 \times 10^{-3} \) and \( 3.9 \times 10^{-4} \) calculated by the jackknife procedure (see Reynolds, Weir and Cockerham 1983). In marked contrast, diversity among restriction site haplotypes for Fast chromosomes is only one-half that of Slow chromosomes (significant at \( P < 0.001 \) level using jackknife variances of \( 1.5 \times 10^{-3} \) and \( 5.8 \times 10^{-5} \), respectively). This lower level of restriction site diversity among Fast relative to Slow allozyme chromosomes is also reflected in lower per nucleotide heterozygosity (equation 11 of Engels 1981), which for Fast chromosomes is 3.6 times lower than that of Slow chromosomes (0.0018 vs. 0.0064, respectively; this difference is significant (\( P < 0.001 \)) using the sampling variance of heterozygosity (4.15 \( \times 10^{-7} \) and 3.18 \( \times 10^{-7} \), respectively, from equation 21 of Engels 1981). This difference in nucleotide diversity between Slow and Fast haplotypes in the Adh region is consistent with Kreitman’s (1983) DNA sequence results for a 2.7-kb region containing the Adh gene.

Nonrandom associations along the sequence

Within the Adh region: Significant nonrandom associations (linkage disequilibrium) are observed among most of the restriction site and insertion/deletion sites.

deletion variants in the Adh region (Table 3). Notably, there is strong linkage disequilibrium between restriction sites and allozyme. For example, the absence of the BamHI -7.1 site is strongly associated with the Fast allozyme, presence of the site with the Slow allozyme. Interestingly, an examination of the 17 Fast allozyme chromosomes separately reveals complete linkage disequilibrium between all pairs of polymorphic sites within the 13 kb. This means that in no two-site comparisons are all four gametic types observed and also implies that few recombinants (or multiple substitutions; see below) among Fast haplotypes are present or have occurred. In addition, 75% of the Fast chromosomes are of the same restriction site haplotype (ignoring for the moment insertion/deletions and transposable elements). Slow chromosomes, in contrast, can be divided into several reasonably divergent restriction site haplotypes and show significant, although often much lower, levels of linkage disequilibrium among sequence variants (Table 3).

Since recombination has not randomized the sequences in the Adh region, we have constructed a phylogenetic tree for the 49 sequences that attempts to relate sequences by the minimum number of mutational and recombinational events (Figure 2). While we have included the possibility of recombination, we have conservatively weighted substitutions more favorably than recombination events in the network shown; that is, for the recombination event to be included it must result in a reduction of more than one substitution. The evidence that recombination has occurred among these sequences in the 13-kb region is of two types. First, there is a positive association between distance between any two sites in kilobases and the occurrence of all four haplotypes (Figure 3). Such a trend would be produced only by recombination, not by recurrent gain or loss of restriction map variation. Second, several lines, discussed below, show a distribution of restriction map variation that would require several adjacent recurrent events to have occurred nearly simultaneously.

Given the correlation of nonrandom associations with distance, it is reasonable to assume that the occurrence of these apparently recurrent substitutions is best explained by a recombination or gene conversion event between divergent sequences. The two parts of the sequence consequently have two different evolutionary histories. Ignoring this will lead to erroneous inferences about the phylogenetic relationships among the sequences. We are, of course, also faced with the fact that, with relatively few markers over this 13-kb region, we probably miss many past recombination events. For this reason in particular, the tree presented in Figure 2 should be viewed primarily as a visual summary of the data and as only a rough approximation to the true phylogenetic relationships among the sequences. This caution applies particularly to Slow haplotypes among which linkage disequilibrium is reduced (although still significant) and because other parsimony trees involve shifts in branching within the Slow haplotypes. Nonetheless, phylogenies estimated by UPG, modified UPG and maximum likelihood algorithms (GOLDING, AQUADRO and LANGLEY 1986 and unpublished results) show the same general relationships as illustrated in Figure 2.

Thus, with the exception of two apparent recombinants between Fast and
<table>
<thead>
<tr>
<th>Site</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In</td>
<td>-7.1</td>
<td>p</td>
<td>-2.7</td>
<td>f</td>
<td>j</td>
<td>Allo.</td>
<td>+0.26</td>
<td>+1.9</td>
<td>XhoI</td>
<td>b</td>
<td>d</td>
</tr>
<tr>
<td>1</td>
<td>-0.60*</td>
<td>-1.00***</td>
<td>0.05</td>
<td>1.00*</td>
<td>-1.00***</td>
<td>-0.90***</td>
<td>0.79**</td>
<td>-1.00**</td>
<td>-0.68*</td>
<td>1.00</td>
<td>-0.68*</td>
<td>0.33</td>
</tr>
<tr>
<td>2</td>
<td>0.29</td>
<td>-1.00*</td>
<td>-0.18</td>
<td>-1.00*</td>
<td>1.00***</td>
<td>-0.82**</td>
<td>1.00***</td>
<td>0.64</td>
<td>-1.00</td>
<td>0.64</td>
<td>-1.00</td>
<td>0.44</td>
</tr>
<tr>
<td>3</td>
<td>0.09</td>
<td>-0.49</td>
<td>0.09</td>
<td>-1.00</td>
<td>-1.00*</td>
<td>-0.36</td>
<td>-0.47</td>
<td>-0.02</td>
<td>0.44</td>
<td>-1.00</td>
<td>-0.22</td>
<td>-1.00</td>
</tr>
<tr>
<td>4</td>
<td>-1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-1.00*</td>
<td>0.80***</td>
<td>-0.61**</td>
<td>-1.00</td>
<td>0.22</td>
<td>-1.00</td>
<td>0.70*</td>
</tr>
<tr>
<td>5</td>
<td>-1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-1.00*</td>
<td>1.00***</td>
<td>-1.00</td>
<td>1.00</td>
<td>0.28</td>
<td>-1.00</td>
<td>0.70*</td>
</tr>
<tr>
<td>6</td>
<td>-1.00*</td>
<td>0.18</td>
<td>-0.51</td>
<td>-0.55</td>
<td>0.76*</td>
<td>-1.00*</td>
<td>-1.00**</td>
<td>1.00**</td>
<td>0.10</td>
<td>-1.00</td>
<td>-1.00</td>
<td>0.70*</td>
</tr>
<tr>
<td>7</td>
<td>-1.00*</td>
<td>-1.00</td>
<td>1.00*</td>
<td>0.10</td>
<td>-0.52</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-1.00*</td>
<td>-0.48*</td>
<td>0.26</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>0.60*</td>
<td>-0.63</td>
<td>0.60*</td>
<td>1.00*</td>
<td>-1.00</td>
<td>-1.00</td>
<td>0.18</td>
<td>-1.00</td>
<td>0.18</td>
<td>-1.00</td>
<td>-1.00</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-1.00</td>
<td>1.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>0.15</td>
<td>-1.00</td>
<td>1.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$D'_{ij}$ is a relative measure of linkage disequilibrium and is calculated as $D_{ij} = P_{ab} - P_A P_B$ relative to its theoretical maximum (min ($p_A(1 - p_B)$, $(1 - p_A)p_B$) when $D_{ij} < 0$ or min ($p_A p_B$, $(1 - p_A)(1 - p_B)$) when $D_{ij} > 0$; Lewontin 1964). Its value ranges from -1 to 1, with 0 indicating random association of variants at the two "loci." Statistical significance is indicated as follows: * ($P < 0.05$), ** ($P < 0.01$), *** ($P < 0.001$). Calculations are based on the 48 lines representing four eastern United States populations for all chromosomes (above diagonal) and for Slow chromosomes alone (below diagonal). Similar calculations for Fast chromosomes alone give $D'_{ij} = -1.0$ or +1.0 for all comparisons and thus are not shown. Similar results are obtained when populations are examined individually, indicating that the nonrandom associations are not due to pooling divergent samples (data not shown). Variants "j" and "a" were not observed among Slow chromosomes and thus are not included in the table.
**Figure 2.**—Phylogenetic tree for the Adh region haplotypes. This particular tree represents one of several most parsimonious networks constructed to minimize the total number of mutational and recombination events required to relate all 29 restriction map haplotypes observed among the 49 lines. Our approach to tree construction followed that outlined by Fitch (1977) with the addition of recombination (see text). Each haplotype is composed of restriction map variants shown 5' to 3' (see key in lower right of figure). Further explanation of these variants can be found in Figure 1. In order to save space, variants unique to a single haplotype are shown as superscripts in the haplotype “balloon.” Solid haplotype “balloons” are observed haplotypes, and dashed “balloons” represent hypothetical intermediate haplotypes. The ADH activities for lines with a given haplotype are shown below the appropriate “balloon” (e.g., 7.21, 7.32; see Table 1 to convert this to line number and location). Haplotypes are connected by solid lines to designate mutation or dotted lines to indicate hypothetical recombination/gene conversion events. For mutation events, the variant changing along each branch of the tree is indicated, with those apparently showing parallel or convergent (multiple) changes indicated by an asterisk. Putative recombinant haplotypes have solid and hatched bars above and below the haplotype to indicate the two segments that have been joined by recombination; dotted lines indicate the approximate region where the recombination event occurred. Dashed lines and arrows indicate the possible ancestral haplotype for each segment of the recombinant haplotype.

Slow haplotypes (Figure 2; discussed below), there appears to be a defined split in the phylogeny between Slow and Fast haplotypes over the 13-kb region. This dichotomy is consistent with Kreitman’s (1983) DNA sequences from the region immediately surrounding the Adh transcribed region and suggests a monophyletic origin of the Fast allozyme.

Although many of the smaller unique sequence insertions/deletions are in strong linkage disequilibrium with adjacent restriction site variation and are consistent with the haplotype phylogeny in Figure 2, the transposable element insert size “b” shows reduced association with several flanking sites. The occurrence of “b” in several very different haplotypes scattered over the tree suggests the possibility of independent origin of at least three of the “b”-size
Discordant comparisons are thus those where $|D'|$ is less than 1.0 (Table 2). The points plotted represent all pairwise comparisons (for all 49 chromosomes) among the 11 polymorphic restriction map variants (those at which two or more of the less frequent "allele" were observed). The mean distances are indicated above the arrows. The distance along the sequence (in kilobases) was estimated from Figure 1. Since restriction mapping often does not allow precise location of insertions or deletions, their location was assumed to be in the middle of the fragment to which they were localized.

inserts seen in the five lines. Sequences of the breakpoints of these inserts (C. AQUADRO, W. QUATTLEBAUM, D. BILLINGS and C. LANGLEY, unpublished results) support this hypothesis.

The pattern of linkage disequilibrium and the hypothesized phylogeny also suggest the repeated (parallel or convergent) gain of the HindIII $-2.7$ site in both Fast and Slow haplotypes. Sequence data in this region (KREITMAN and AGUADÉ 1986) show the presence of a sequence located 11 bp from the $-2.7$ site that differs by only one nucleotide from the recognition sequence for HindIII. Thus, it is possible that what we have scored as the presence of a single site at $-2.7$ may be two polymorphic HindIII sites not distinguishable by routine restriction mapping.

Adh region variants and inversions: Significant nonrandom associations have been reported between ADH allozymes and second chromosome inversions from several D. melanogaster populations (e.g., KOJIMA, GILLESPIE and TOBARI 1970; MUKAI, METTLER and CHIGUSA 1971; LANGLEY, TOBARI and KOJIMA 1974; MUKAI, WATANABE and YAMAGUCHI 1974). For example, the inversion $In(2L)t$ has only been found associated with the ADH Slow allozyme. This rearrangement has breakpoints close to, but not including, the Adh gene (bands 22D3-E1 and 34A8-9, LINDSLEY and GRELL 1968, with Adh at 35A–B, WOODRUFF and ASHBURNER 1979). Our results are consistent with this pattern for ADH allozyme; the chromosomes from five lines carry $In(2L)t$ and are all the Slow ADH allozyme (Table 1). In addition to allozymes, several restriction map variants from the Adh region are in strong (and, in many cases, complete) linkage disequilibrium with this inversion (Table 3).

The history of Adh on $In(2L)t$ chromosomes appears rather complicated. Three of the five $In(2L)t$ lines share one haplotype, whereas the other two show a second haplotype differing from the first at five sites. For two of these divergent sites, the HindIII site at $-2.7$ and transposable insert "b," other evidence summarized above indicates independent gain or loss. The other
REGION VARIATION AND EVOLUTION

FIGURE 4.—ADH activity of 49 second-chromosome lines of *D. melanogaster* plotted in ascending rank order as listed in Table 1. Each point is the average of the line means from two different experiments, one performed in 1980 and the other in 1982. In both experiments, allozymes and lines within allozymes were highly significant sources of variation. The minimum difference between two lines that is significant at the 95% confidence level (Tukey’s multiple comparison procedure) varies among allozymes and experiments as follows: for Slow lines it is 1.2 and 1.3 units/fly in the 1980 and 1982 experiments, respectively, and for Fast lines the corresponding values are 1.6 and 1.5. The correlation over line means between the two experiments is 0.97 (*P* < 0.0001). The ADH activity units here are nanomoles NAD reduced/min/fly with ethanol substrate, and they have been adjusted by linear regression on weight and general protein content of the flies in each homogenate.

three divergent sites are the three 5’-most sites constituting the haplotype. As indicated by Figure 2, the two *In(2L)t* lines R131 and R141 (activities 2.49 and 3.16) have a haplotype explained most easily by recombination between a Fast [never *In(2L)t*] and a Slow sequence that was *In(2L)t*. However, since both inversion breakpoints are 5’ (centromere distal) to *Adh* (inferred from Chia et al. 1985), it appears the putative event was a gene conversion or double crossover. Alternatively, the R131 and R141 inversions may not be *In(2L)t*, or else *In(2L)t* arose more than once. The breakpoints in these five lines are, however, cytologically indistinguishable, and the complete association between ADH Slow and *In(2L)t* in many population samples argues against multiple origins of the inversion.

Significant linkage disequilibrium between ADH allozymes and an inversion, *In(2R)NS*, on the opposite arm of the second chromosome from *Adh* has also been reported (e.g., Mukai, Watanabe and Yamaguchi 1974). Unfortunately, only two lines in our study, KA21 (a Fast) and KA14 (a Slow), were of the *In(2R)NS* arrangement, precluding any formal analysis of nonrandomness.

**ADH enzyme activities**

Mean ADH activity per fly (adjusted for body weight and general protein content) for each of the 49 lines is presented in Figure 4. (Recall that effects on activity due to modifier loci on the X and third chromosomes have been eliminated since all lines are homozygous for the same X and third chromosomes.) It is worth noting the high repeatability of ADH activities reported
for these lines. Despite the often small differences between adjacent lines, the
correlation of activity for these 49 lines was very high (0.97, \( P < 0.0001 \))
between experiments done 2 years apart (see Figure 4 legend).

While ADH activities for the 49 lines are spread over a five fold range, two
clusters of lines are apparent; termed here high and low activity. Of particular
interest is the nearly complete association between the Fast allozyme and high
activity, and the Slow and low activity. A persistent question since this corre-
lation was noted several years ago (MCDONALD and AYALA 1978; LAURIE-
AHLBerg et al. 1980; MCDONald, ANDERSON and SANTOS 1980) has been
whether the high/low difference is solely due to the Thr-Lys replacement that
underlies the allozyme difference. Our data suggest that one or more addi-
tional (or alternative) sites are involved.

There are three obvious activity outliers among the 49 lines we have studied;
that is, lines that appear to have the “wrong” activities for their allozyme
(Figure 4). Heat and urea denaturation studies and comparison of activity
ratios for different alcohol substrates reveal no deviation from typical Slow or
Fast properties, providing no evidence for new amino acid substitutions in any
of the three lines (MARONI et al. 1982, and our unpublished data). These lines
may simply represent overlap between the tails of two frequency distributions
of activity, variation within and between allozymes being simply due to many
variants, each of small effect. However, the restriction map of each of the
three lines is distinctive, with two of the lines appearing to represent recom-
binants 5’ to the allozyme between Fast/high and Slow/low activity haplotypes.

One line (RI32; a Fast allozyme and low activity of 4.34) has an approxi-
mately 27-bp deletion very close to, or at the start of, the distal Adh transcript
that may be related to its lower activity. The restriction maps for the other
lines with the “wrong” activity for their allozyme (NC16, Slow but high activity;
and KA12, Fast but low activity) appear to be recombinant in the region
between the allozyme and BamHI –7.1 site. For example, the restriction site
haplotype of NC16 has the allozyme and 3’ region of a typical Slow, low
activity line but has a 5’ region typical of a high-activity Fast haplotype. The
high activity of line NC16 is tightly linked to the allozyme polymorphism, and
the effect is cis-acting (MARONI and LAURIE-AHLBERG 1983). (In fact, no re-
combinants between activity and allozyme have been recovered in experiments
carried out to date.) In addition, the high activity of NC16 is associated in
both larvae and adults with a high level of ADH protein (MARONI et al. 1982).
The third line (KA12; Fast) appears to represent an ancestral intermediate in
the evolution of Fast from Slow, although it, too, may be the product of a 5’
recombination event (Figure 2). Notably, lines that might be considered recom-
binants 3’ to the allozyme site or possible recombinants in the extreme 5’ end
of the 13-kb region do not show shifts in ADH activity.

As with the between-allozyme component of activity variation, there is a
tendency within an allozyme for similarity in ADH activity to be correlated
with propinquity of descent of the lines based on the Adh region. For example,
Slow allozyme lines RI02, RI09, KA01 and KA15 cluster together in the
phylogeny in Figure 2 and have activities of 4.12, 4.54, 5.20 and 5.25, re-
spectively. Lines NC14, KA06 and NC25 form another lineage of Slow sequences with activities of 2.82, 3.22 and 3.33 (Figure 2). Although this pattern does not hold for all lines, it suggests that some of the subtle differences in activity among lines within allozymic classes are due to sequence variation closely linked to Adh.

Only one chromosome appears to have a transposable element or large insertion/deletion in the Adh region that is clearly associated with altered ADH activity. The line (RI42) has a 5.2-kb copia element insertion about 250 bp 5' to the distal cap site of Adh. It is also characterized by an unusual level and pattern of Adh expression. The low activity effect in this line is cis-dominant, maps very close to the allozyme site and is associated with a low level of ADH protein (Maroni et al. 1982). Furthermore, it shows a stage-specific effect in that larvae have much lower activity than adults (Maroni and Laurie-Ahlberg 1983). This result appears paradoxical on the surface, since the distal (adult) start site is 708 bp upstream of the proximal (larval) start site, and thus lies in closer proximity to the copia insertion. However, P-element transformation results of Posakony, Fisher and Maniatis (1985) also suggest that the proximal promoter may be more sensitive to upstream sequences than the distal promoter. These observations further strengthen the suggestion that the copia element causes the altered ADH expression in RI42.

The remaining (majority of) transposable element insertions and large insertions/deletions do not appear to be simply correlated with alterations of ADH activity. For example, transposable elements are present in the majority of the lines having outlier activities. However, other lines with “normal” activities for their allozyme also have transposable elements, often of similar size and in the same location (e.g., “b”), suggesting that the altered activity does not result from the insertion. Their presence 1.5 kb 3' to Adh makes these results not particularly surprising.

DISCUSSION

Several important points emerge from our survey of naturally occurring restriction map variation in the Adh region of D. melanogaster. DNA sequence polymorphism is extensive, with both transposable element insertions and unique sequence insertions/deletions contributing to enormous variation among lines in the sequences flanking the Adh coding region. Eighty percent of the lines differ in sequence length from the most common restriction map due to insertions or deletions ranging from 21 bp to 10.2 kb in size. All insertions over 200 bp are one of four transposable elements (copia, B104, giminy, and a new F-like element we have called 2161). Strong nonrandom associations exist among variants in the Adh region, including the allozyme polymorphism, ADH activity and many of the flanking variants (Table 3). As a result, restriction maps for the 49 lines form only 29 distinct haplotypes out of $2^{25}$ or $3.4 \times 10^7$ possible combinations (Tables 1 and 2). This haplotype structure has allowed us to construct a phylogeny for the alleles segregating within this species, and has provided some revealing insights into the influence of evolutionary history on current patterns of DNA sequence polymorphism.
Sources of ADH activity variation: Despite the enormous nucleotide and sequence length polymorphism in the Adh region, most of the variation detected is not clearly associated with altered levels of ADH activity. ADH activities for our 49 lines are spread over a fivefold range (Figure 4). Only one of the chromosomes studied has a transposable element insertion clearly correlated with an unusual level and pattern of Adh expression. Despite this lack of association of activity with observed restriction map variants, activities are distributed in a pattern generally consistent with a phylogeny for the Adh restriction map haplotypes and with the distribution of the Fast/Slow polymorphism (Figure 2). This consistency suggests that many of the subtle differences in activity between lines are due to sequence variation close enough to the Adh gene to be in linkage disequilibrium with the allozymes and flanking restriction site variation.

Previous studies have indicated that differences in both the efficiency and amount of ADH enzyme appear to contribute to activity variation between Fast and Slow allozymes. Winberg, Hovick and McKinley-McKee (1985) have reported a higher catalytic-center activity for the Fast allozyme than that in the Slow allozyme. However, numerous studies have also shown that the higher activity of Fast lines is correlated with higher levels of ADH protein (Gibson 1972; Day, Hillier and Clarke 1974; Lewis and Gibson 1978; McDonald, Anderson and Santos 1980; Birley, Couch and Marson 1981; Maroni et al. 1982; Anderson and McDonald 1983). While the catalytic differences must be due to the Thr-Lys replacement, the molecular basis of the variation in amount of ADH is unclear.

The bimodality of ADH activity observed and association with restriction map haplotype indicate the possibility of a simple polymorphism that gives rise to different levels of ADH protein (high vs. low). Two lines with the "wrong" activity for their allozyme (and that possess haplotypes that appear to be recombinant) suggest that, if such a polymorphism exists, it is tightly linked to, but distinct from, the substitution underlying the Thr-Lys replacement that distinguished the allozymes. This hypothesis is directly testable by DNA sequencing of these putative recombinant lines coupled with in vitro recombinant between Fast-high and Slow-low lines for different segments of the Adh region and subsequent P-element mediated transformation.

Fitness effects of flanking variation: An examination of the frequency spectra of the enormous diversity of sequence variation in the regions flanking Adh suggests that sequence length variation has different evolutionary dynamics than do most base substitutions. Transposable elements in particular appear to be slightly deleterious. In Figure 5 we have plotted the frequency spectra for the sequence variation, broken into four natural classes: Protein variation (Fast vs. Slow), restriction site variation, unique insertions/deletions and transposable element variation. There tend to be too many rare unique-sequence insertions and deletions; this suggests that they are slightly deleterious. Moreover, transposable elements show a clear and significant departure from selective neutrality predictions. Watterson's (1978) homozygosity test indicates that each individual element (vs. nontransposable element haplotypes) is too
Adh REGION VARIATION AND EVOLUTION

FIGURE 5.—Frequency spectrum for naturally occurring DNA sequence variants in the 13-kb Adh region of D. melanogaster. Shown is the less frequent character at each variable site.

rare ($F = 0.6076, k = 8, n = 48, P < 0.025$). The departure from neutrality predictions for transposable elements is actually more dramatic, because sequence data from the breakpoints of three of these insertions indicate that the five “b” inserts represent at least two independent insertion events and are thus individually less frequent (C. Aquadro, W. Quattlebaum, D. Billings and C. Langley, unpublished results).

This departure from selective neutrality could be accounted for by a slight deleterious effect of insertion. Alternatively, a high excision rate could result in the same distribution (the elements insert and excise rapidly and never have a chance to accumulate in frequency at any one site). Theoretical results (Golding, Aquadro and Langley 1986) favor the former interpretation, although selection coefficients are estimated to be only several times the rate of insertion (perhaps in the range of $1 \times 10^{-5}$ to $1 \times 10^{-6}$). In addition, unlike base substitutions, very little sequence length variation is observed to accumulate through evolutionary time, as would be expected if length variation were selectively neutral (Barrie, Jefferys and Scott 1981; Langley, Montgomery and Quattlebaum 1982; Leigh Brown and Ish-Horowicz 1981). Note, for example, that unique sequence deletions in the Adh region average 2.3 times larger and were observed 3.5 times more often than unique sequence insertions. Unchecked, such a trend would lead to a progressive decrease in the size of the genome, a pattern that is not apparent in this region (Langley, Montgomery and Quattlebaum 1982; Zwiebel et al. 1982).

**Origin and implications of linkage disequilibrium:** Numerous mechanisms can generate correlations among variants along a sequence, particularly considering the close proximity of the DNA sequence markers on the chromosome and the consequential low rate of recombination among sites. Mechanisms not invoking selection include drift due to small population size, population subdivision, migration and inbreeding. Hitchhiking among linked genes, one or more of which are selected but which show no epistasis, can also lead to linkage
disequilibrium, as can selection for favored combinations of alleles at two loci due to epistatic interaction (i.e., coadapted gene complexes).

An examination of the patterns and relative levels of nucleotide haplotype diversity and linkage disequilibrium among Fast vs. Slow chromosomes indicates that the strong nonrandom associations we observe in the Adh region are the product of a complex evolutionary history probably influenced by hitchhiking with some selectively favored variant. Fast-allozyme chromosomes are two to three times less variable than Slow-allozyme chromosomes, on the basis of both heterozygosity per nucleotide (0.002 vs. 0.006) and restriction site haplotype diversity (0.419 vs. 0.819). In addition, complete linkage disequilibrium is observed in our sample of Fast-allozyme chromosomes over the 13-kb Adh region, whereas reduced, although significant, linkage disequilibrium is seen in Slow-allozyme chromosomes. Evidence for large effective population size for D. melanogaster (e.g., Mukai and Yamaguchi 1974; Henderson and Lambert 1982; Kreitman 1983) and the observation that the same qualitative patterns of linkage disequilibrium are observed for each of the four eastern United States populations we sampled (calculations not shown) further support the view that stochastic explanations alone cannot explain the evolution of the Adh region of these populations.

Kreitman (1983) suggested that the Fast allozyme was derived from the Slow allozyme and that Slow was the ancestral allozyme state for D. melanogaster. This interpretation has been supported by additional data and analyses (Bodmer and Ashburner 1984; Cohn, Thompson and Moore 1984; Ashburner, Bodmer and Lemeunier 1984; Stephens and Nei 1985). Our phylogeny for the Adh region restriction map haplotypes (Figure 2) is also consistent with this hypothesis and suggests that Fast/high activity is a new haplotype relative to the Slow/low activity haplotype.

While the Fast allozyme sequence has been estimated to have arisen 1- to 3-million-yr ago (Ashburner, Bodmer and Lemeunier 1984; Stephens and Nei 1985), our data suggest that the extant Fast sequences we sampled arose from a single progenitor more recently, recent enough for recombination not to have broken up the nonrandom associations among sequence variants that occurred initially by chance. Consider the rate of decay of linkage disequilibrium over the 7.3-kb region between the BamHI site at -7.1 and the allozyme site (at approximately +0.2 on our map; Figure 1). The rate of recombination over this region in natural populations is estimated to be approximately $3.6 \times 10^{-5}$ [c = (7.3 kb) (0.5) (1 $\times$ 10$^{-3}$ events/kilobase/generation)]. (Rate is estimated from the size of the Adh coding sequence and the genetic results of Ashburner et al. 1979; the factor of 0.5 is included because there is generally no recombination in males.) Linkage disequilibrium (complete initially) would be expected to be reduced by one-half by recombination in approximately 19,000 generations ($t = (\ln 0.5)/\ln (1-c)$). Nonetheless, we still observed complete linkage disequilibrium among the Fast-allozyme haplotypes we analyzed. The current level of linkage disequilibrium in our U.S. samples between the BamHI -7.1 site and the allozyme is -0.1875 (not given as a ratio to its current theoretical maximum). Under selective neutrality it would take approximately
7882 generations, or slightly less than 2000 yr (at four generations per year, probably an underestimate), for \( D_i \) to decay to the this current level from its theoretical maximum of \(-0.25\). Under strong selection, the rate of decay can be faster (AsmusSEN and Clegg 1981).

Given the evidence that natural selection can act on the Adh polymorphism (Van Delden 1982), the origin of a selectively favored substitution on the Fast allozyme chromosome (perhaps that resulting in higher ADH activity) may have contributed to a rapid increase and spread of this haplotype throughout the species. It is possible that the spread of the Fast/high activity haplotype was related to the movement of man into the temperate regions of the world within the last 5,000–10,000 yr. Currently, the Fast allozyme predominates in frequency in temperate regions of the four continents that have been examined to date, whereas Slow predominates in equatorial regions including Africa, which is thought to be the place of origin of the species (David 1982; OakESHOTT et al. 1982; Singh, Hickey and David 1982).

Nucleotide polymorphism also appears not to have reached equilibrium in the Fast-allozyme lineage relative to that in the Slow lineage (assuming neutrality and roughly similar effective population sizes for the two allozymes). Significantly, haplotype diversity due to both unique sequence insertion/deletion and transposable elements does appear to have reached roughly similar levels in both lineages, consistent with the proposition that both transposable element and unique sequence length variations occur at a higher rate than base substitutions.

Irrespective of the precise mechanisms generating these associations among variants in linked genes and regions, their existence as blocks of sequence variants must be considered when attempting to analyze allelic variation at any one gene or region within that block. Consider that for the 13-kb Adh region probably 65 or more nucleotide differ, on average, along the sequence between any two randomly chosen sequences. Some may have positive effects, some negative and others none with respect to fitness. However, they do not evolve independently, but rather as blocks with a shared evolutionary history. The magnitude of these blocks can be even more significant in regions of reduced recombination and in species with effective population sizes smaller than Drosophila (e.g., man, in whom significant linkage disequilibrium has been observed over as much as 100 kb; Orkin et al. 1982; Bech-Hansen et al. 1983; Murray et al. 1984).

We thank our colleagues, particularly the following people, for their help, discussion, constructive criticism and/or permission to cite unpublished work: M. Ashburner, D. Billings, G. R. Carmody, B. Charlesworth, G. B. Golding, R. R. Hudson, R. Jennings, M. Kreitman, A. J. Leigh Brown, E. A. Montgomery, M. Nei, J. Posakony, W. Quattlebaum, L. F. Stam, J. C. Stephens and B. S. Weir. This work was supported in part by Public Health Service grant NIH-GM11546. This is paper no. 9779 of the journal series of The North Carolina Agricultural Research Service.

LITERATURE CITED

ASHBURNER, M., M. BODMER and F. LEMEUNIER, 1984 On the evolutionary relationships of

ASHBURNER, M., R. CAMFIELD, B. CLARKE, D. THATCHER and R. WOODRUFF, 1979 A genetic
analysis of the locus coding for alcohol dehydrogenase, and its adjacent chromosome region,

ASMUSSEN, M. A. and M. T. CLEGG, 1981 Dynamics of the linkage disequilibrium function under


BENTON, W. D. and R. W. DAVIS, 1977 Screening lambda-gt recombinant clones by hybridization

BENYAJATI, C., N. SPOEREL, H. HAYMERLE and M. ASHBURNER, 1983 The messenger RNA for
alcohol dehydrogenase in *Drosophila melanogaster* differs in its 5' end in different develop-

BINGHAM, P. M., M. G. KIDWELL and G. M. RUBIN, 1982 The molecular basis of P-M hybrid

BINGHAM, P. M., R. LEVIS and G. M. RUBIN, 1981 Cloning of DNA sequences from the white

BIRLEY, A. J., P. A. COUCH and A. MARSON, 1981 Genetical variation for enzyme activity in a
population of *Drosophila melanogaster*. VI. Molecular variation in the control of alcohol de-

BODMER, M. and M. ASHBURNER, 1984 Conservation and change in the DNA sequences coding

165: 349–357.

93. In: *Genetic Studies of Drosophila Populations, Proceedings of the Kioloa Conference*, Edited by

CHAMBERS, G. K., A. V. WILKS and J. B. GIBSON, 1984 Variation in the biochemical properties

CHIA, W., R. KARP, S. MCGILL and M. ASHBURNER, 1985 Molecular analysis of the Adh region

COHN, V. H., M. A. THOMPSON and G. P. MOORE, 1984 Nucleotide sequence comparison of the

DAVID, J. R., 1982 Latitudinal variability of *Drosophila melanogaster*: allozyme frequencies diver-


DAY, T. H., P. C. HILLIER and B. CLARKE, 1974 The relative quantities and catalytic activities
of enzymes produced by alleles at the alcohol dehydrogenase locus in *Drosophila melanogaster*.


GIBSON, J. B., 1972 Differences in the number of molecules produced by two allelic electrophoretic enzyme variants in *D. melanogaster*. Experientia 28: 975–976.


LEWIS, R., M. COLLINS and G. RUBIN, 1982 FB elements are the common basis for the instability of the w^{D2} and w Drosophila mutations. Cell 30: 551–565.


Adh REGION VARIATION AND EVOLUTION


Communicating editor: W. W. ANDERSON