Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*

Martin Kreitman

Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138, USA

The sequencing of eleven cloned Drosophila melanogaster *alcohol dehydrogenase* (Adh) *genes from five natural populations has revealed a large number of previously hidden polymorphisms. Only one of the 43 polymorphisms results in an amino acid change, the one responsible for the two electrophoretic variants (fast, Adh-f, and slow, Adh-s) found in nearly all natural populations. The implication is that most amino acid changes in Adh would be selectively deleterious.*

MANY issues in population genetics require, for proper evaluation, measurements of genetic variation in natural populations. In recent years, this measurement has been made primarily by gel electrophoresis[1,2], which detects differences in amino acid sequence of proteins. Such studies have uncovered a remarkable amount of genetic polymorphism at loci coding for soluble enzymes[3]. However, because bands on electrophoretic gels are phenotypes, not genotypes, the total extent of genetic variation at structural loci remains unknown.

Recent advances in recombinant DNA technology, including rapid sequencing techniques[4,5], now allows allelic differences between individual genes to be identified directly at the nucleotide level. The complete resolution of genetic variation afforded by DNA sequence comparison not only solves the problem of detecting all amino acid substitutions between proteins, it also identifies nucleotide variation that is not translated into protein differences. Sequence variation in noncoding regions such as introns, or between synonymous codons, which is presumably not subject to natural selection acting through the phenotype of the encoded protein, is the best available source for estimating the frequency of polymorphism due to neutral gene substitution. This estimate, which can only be made from direct DNA sequence comparison, is crucial for understanding the role of natural selection and genetic drift in the maintenance of genetic variation.

I report here a DNA sequence comparison of 11 independently cloned *Adh* genes from five geographically distinct populations of the fruit fly, *Drosophila melanogaster.* Virtually all natural populations in this species are polymorphic for two electrophoretically distinguishable alleles[6,7], *Adh-s* and *Adh-f*, which differ by a single amino acid (Thr versus Lys at codon 192)[8,9]. However, several other variants have been identified, some of which are electrophoretically indistinguishable from the *Adh-f* or *Adh-s* electromorphs. The most common one, a heat-resistant variant of *Adh-f*[10], has been reported at frequencies as high as 10% in some populations[11,12]. Therefore, I was particularly interested in whether a DNA sequence analysis could disclose genetic heterogeneity among proteins with identical electrophoretic mobilities. This comparative DNA sequence study of five *Adh-f* and six *Adh-s* alleles, shows the presence of many silent polymorphisms in exons and introns but no amino acid replacement polymorphism within the two electromorphs. This distribution of nucleotide polymorphism is a strong indication that virtually all amino acid replacement mutations have been selectively deleterious. This study also reports extensive sequence conservation in the 3′ flanking region of the *Adh* gene.

Physical organization

Figure 1 shows the structure of the *Adh* gene, the protein-encoding sequence of which is split into three sections (of lengths 96 base pairs (bp), 405 bp and 264 bp) separated by two small introns (65 bp and 70 bp). Two transcripts are produced, a larval and an adult form, differing only in their 5′ noncoding leader sequences. The larval mRNA leader sequence, 70 nucleotides in length, is contiguous with the first coding sequence in exon 2 and therefore is unspliced in the mature message. The adult mRNA contains an 87 nucleotide leader sequence that is transcribed from a segment located 690 bp upstream from the first coding sequence. This leader sequence (positions 1–87) is spliced to a site 36 nucleotides 5′ to the AUG initiation codon in the mature mRNA (position 741/742). Therefore, in addition to the two small introns dividing the coding regions, the adult primary transcript contains an additional intron of 654 bp at the 5′ end of the gene. A larval consensus promoter sequence 'TATAATA' as well as part of the larval leader sequence is included in the 3′ end of the adult intron.

Sequence polymorphism

A consensus DNA sequence based on the six *Adh-s* genes is shown in Fig. 2. It contains the complete structural gene, and also flanking 5′ (63 bp) and 3′ (800 bp) sequence. Table 1 shows a summary of DNA sequence variation among the 11 genes over this 2,721 bp region. Forty-three nucleotide positions are polymorphic overall, 14 in the three coding regions (765 bp), 11 in the adult intron (654 bp), 7 in introns 2 and 3 (135 bp), 3 in the 5′ and 3′ nontranslated sequences (332 bp) and 8 in the 5′ and 3′ flanking regions (863 bp). Thirteen of the 43 sites

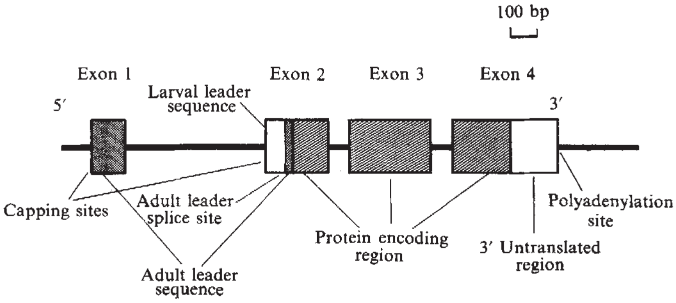

Fig. 1 Structure of the *Adh* gene of *D. melanogaster.* Two different transcripts have been identified in adult and larval tissues[26] which differ only in their 5′ noncoding sequences. The larval 5′ leader is contiguous with the first coding region whereas 87 bp of the adult 5′ leader is transcribed from a region located 690 bp further upstream and is spliced to a site located 36 bp upstream from the translation initiation. The two leader sequences overlap, therefore, by 36 bp.

0028–0836/83/310412—06$01.00

Table 1

Reference sequence	5′ Flanking sequence	Adult leader (exon 1)	Intron 1 (Adult intron, larval non-coding)	Larval leader	Translated region of exon 2	Intron 2	Exon 3	Intron 3	Translated region of exon 4	3′-Untranslated region	3′ Flanking sequence
	C C G		C A A T A T G G G ∇1 C ∇2 G	C	T	A C	C C C C	G G A A T	C T C C A* C T A G	A ∇3 C	A G C ∇4 C ∇5 T Δ6
Strain											
Wa-S	. . .		. . . . . . A T	.	.	. .	T T . A	C A . T A	A C	. . .	. . . . Δ
Fl-1S	. . C		. . . .	.	.	. .	T T . A	C A . T A	A C	. . .	. . . . Δ
Af-S	. . .		. . . .	.	.	. .	. . . .	. . . . .	. . . . A	. . .	. . T ∇ . 1 A .
Fr-S	. . .		. . . .	.	.	G T	. . . .	. . . . .	. . . . A	.-1 .	T A . . . .
Fl-2s	. . .		A G . . . A . T C	A	G	G T	. . . .	. . . . .	. . . .	C 3 .	. . . .
Ja-S	. . C		. . . .	.	G	. .	. . . .	. . . . .	. . . T . T . C A	C 4 .	. . . . . T . . .
Fl-F	. . C		. . . .	.	G	. .	. . . .	. . . . .	. . G T C T C C .	C 4 .	. . . .
Fr-F	T G C		A G . . . A . T C ∇ G ∇ .	.	G	. .	. . . .	. . . . .	. . G T C T C C .	C 4 G	. . . .
Wa-F	T G C		A G . . . A . T C ∇ G ∇ .	.	G	. .	. . . .	. . . . .	. . G T C T C C .	C 4 G	. . . .
Af-F	T G C		A G . . . A . T C ∇ G ∇ .	.	G	. .	. . . .	. . . . .	. . G T C T C C .	C 5 G	. . . .
Ja-F	T G C		A G G G G A . . . ∇ . . T	.	G	. .	. . A .	. . G . .	. . G T C T C C .	C 4 .	. . . . . .-1 . .
No. of polymorphic sites	3	0	11	1	1	2	4	5	9	2	5
Average no. of Nucleotides compared	63	87	620	70	99	65	405	70	264	178	767
% Sites polymorphic	4.7	0	1.8	1.4	1.0	3.1	1.0	7.1	3.5	1.1	0.6

One *Adh-f*, and either one or two *Adh-s* electrophoretic alleles were randomly chosen from isochromosomal lines derived from each of five population samples. S, *Adh-s* alleles; F, *Adh-f* alleles. Collection sites and year collected: Fl, West Palm Beach, Florida, 1979; Wa, Seattle, Washington, 1979; Af, Burundi, Africa, 1977; Fr, Bully, France, 1977; Is, Ishigaki, Japan, 1978. The reference nucleotide sequence is the most common *Adh-s* nucleotide at each of the polymorphic sites. Differences are shown in the body of the table. ∇/Δ: insertion/deletion polymorphisms. The numbers in columns ∇3 and ∇5 are the differences in homopolynucleotide run lengths compared with the consensus sequence. *: Thr–Lys amino acid replacement polymorphism. All other polymorphisms are either silent or noncoding.

Table 2 Summary of silent polymorphism frequencies

	Intron 1	Introns 2+3	Translated region of exon 2	Exon 3	Translated region of exon 4	Complete translated sequence	Non-translated	3′ Non-transcribed
Average no. nucleotides compared	654	135	99	405	264	765	335	767
Effective no. silent sites	654	135	25.5	103.6	63.1	192.2	335	767
No. silent polymorphisms	11	7	1	4	8	13	3	5
Per cent silent polymorphism	1.7	5.2	3.9	3.9	14.3	6.7	0.9	0.6

The proportion of nucleotide changes that would be silent among all nine possible substitutions at each codon is averaged over all codons for an *Adh-s* allele, excluding the initiation and stop codon, and excluding any substitution that changes a codon to a stop codon. The effective number of silent sites is calculated for each exon as the product of the total number of nucleotides and the average proportion of all possible substitutions that are silent.

have nucleotide differences that are represented only once in the 11 strains, 11 of which occur in one strain, *Ja-f*. The adult intron and the 3′ flanking regions each contain five unique polymorphisms. Of the 14 polymorphic sites in the three coding regions, only one is a polymorphism leading to an amino acid replacement, and it accounts for the two electromorphs. The remaining 13 exon polymorphisms represent silent changes.

Adjacent polymorphisms at the first three non-transcribed nucleotide positions immediately 5′ to the adult mRNA cap site (positions −1, −2 and −3 in Fig. 1) represent the only example of consecutively varying nucleotide sites. Polymorphism at these positions strongly suggests that their conservation is not necessary for proper initiation of transcription as all of the fly strains used in this study produce *Adh* (all show *Adh* bands on native acrylamide gels). Although this pattern of polymorphisms is suggestive of a mutational 'hotspot', these same positions are conserved between the commonest sequence in *D. melanogaster* and two sibling species, *Drosophila simulans* and *Drosophila mauritiana*[13]. In addition, there are no analogous polymorphisms in the corresponding larval 5′ non-coding region.

Length polymorphism

In addition to nucleotide site variation, the adult intron and the 3′ flanking region have six sites containing length polymorphisms, ranging in size from 1 to 37 bp (see Fig. 2). Five of the six polymorphisms involve direct repeats. Two length polymorphisms are structurally similar, involving expansions or contractions of homonucleotide repeats (runs), at position 1,698–1,708 and at position 2,303–2,311. Five different 'run' lengths are associated with the former site, ranging from 10 to 16 bp in length, and three with the latter site, ranging from 9 to 11 bp. The fact that the relationship of lengths among the 11 genes conforms well to the overall pattern of relationship based on shared nucleotide polymorphisms (Table 1), strongly suggests that this length variation is not a cloning artefact. Thus, these homonucleotide runs appear to be mutational hotspots.

The DNA sequences of the two length polymorphisms in the adult intron are shown in Fig. 2. ∇1, located between positions 447/448 and 476/477, contains a 6 bp direct repeat which is tandemly reiterated four times in the insertion. An identical 6 bp sequence is also located adjacent (3′) to the point

```
                         *         *         *         *         *         *      TGC
                  -63 GTCGACTGCACTCGCCCCCACGAGAGAACAGTATTTAAGGAGCTGCGAAGGTCCAAGTCACCG

              *         *         *         *         *         *         *         *
   1 ATTATTGTCTCAGTGCAGTTGTCAGTTGCAGTTCAGCAGACGGGCTAACGAGTACTTGCATCTCTTCAAATTTACTTAAT

              *         *      A  *  G      *         *         *  G      *         *
  81 TGATCAAGTAAGTAGCAAAAGGGCACCCAATTAAAGGAAATTCTTGTTTAATTGAATTTATTATGCAAGTGCGGAAATAA

             G*  G A    *         *         *         *         *         *         *
 161 AATGACAGTATTAATTAGTAAATATTTTGTAAAATCATATATAATCAAATTTATTCAATCAGAACTAATTCAAGCTGTCA

              *         *         *         *      T  *  T      *   C     *         *
 241 CAAGTAGTGCGAACTCAATTAATTGGCATCGAATTAAAATTTGGAGGCCTGTGCCGCATATTCGTCTTGGAAAATCACCT

              *         *         *         *         *         *         *         *
 321 GTTAGTTAACTTCTAAAAATAGGAATTTTAACATAACTCGTCCCTGTTAATCGGCGCCGTGCCTTCGTTAGCTATCTCAA

              *         *         *         *      ⇓ ←*         *∇1       *   → ⇓   *
 401 AAGCGAGCGCGTGCAGACGAGCAGTAATTTTCCAAGCATCAGGCATAGTTGGGCATAAATTATAAACATACAAACCGAAT
                                                  TAATATACTAATACTAATACTAATACTAATATAA
              *         *         *     G   *         *         *         *⇓   ∇2   *
 481 ACTAATATAGAAAAAGCTTTGCCGGTACAAAATCCCAAACAAAAACAAACCGTGTGTGCCGAAAAATAAAAATAAACCAT
                                                                            AATAAACCAT
              *         *     T   *         *         *         *         *         *
 561 AAACTAGGCAGCGCTGCCGTCGCCGGCTGAGCAGCCTGCGTACATAGCCGAGATCGCGTAACGGTAGATAATGAAAAGCT
     AAACTAGGCTGCTCAGCCGGCGACGGC
              *         *         *         *         *         *         *  A      *
 641 CTACGTAACCGAAGCTTCTGCTGTACGGATCTTCCTATAAATACGGGGCCGACACGAACTGGAAACCAACAACTAACGGA

              *         *         *         *         *         *         *         *
 721 GCCCTCTTCCAATTGAAACAGATCGAAAGAGCCTGCTAAAGCAAAAAAGAAGTCACCATGTCGTTTACTTTGACCAACAA
                                                              MetSerPheThrLeuThrAsnLy

              *     G   *         *         *         *         *         *         *
 801 GAACGTGATTTTCGTTGCCGGTCTGGGAGGCATTGGTCTGGACACCAGCAAGGAGCTGCTCAAGCGCGATCTGAAGGTAA
     sAsnValIlePheValAlaGlyLeuGlyGlyIleGlyLeuAspThrSerLysGluLeuLeuLysArgAspLeuLys

              *     G   *         *         *    T    *         *         *         *
 881 CTATGCGATGCCCACAGGCTCCATGCAGCGATGGAGGTTAATCTCGTGTATTCAATCCTAGAACCTGGTGATCCTCGACC
                                                                  AsnLeuValIleLeuAspA

              *         *         *         *         *         *         *         *
 961 GCATTGAGAACCCGGCTGCCATTGCCGAGCTGAAGGCAATCAATCCAAAGGTGACCGTCACCTTCTACCCCTATGATGTG
     rgIleGluAsnProAlaAlaIleAlaGluLeuLysAlaIleAsnProLysValThrValThrPheTyrProTyrAspVal

              *         *       T *         *         *         *         *         *
1041 ACCGTGCCCATTGCCGAGACCACCAAGCTGCTGAAGACCATCTTCGCCCAGCTGAAGACCGTCGATGTCCTGATCAACGG
     ThrValProIleAlaGluThrThrLysLeuLeuLysThrIlePheAlaGlnLeuLysThrValAspValLeuIleAsnGl

              *         *         *         *         *         *         *         *
1121 AGCTGGTATCCTGGACGATCACCAGATCGAGCGCACCATTGCCGTCAACTACACTGGCCTGGTCAACACCACGACGGCCA
     yAlaGlyIleLeuAspAspHisGlnIleGluArgThrIleAlaValAsnTyrThrGlyLeuValAsnThrThrThrAlaI

              *         *        T*    A    *         *         *         *         *
1201 TTCTGGACTTCTGGGACAAGCGCAAGGGCGGTCCCGGTGGTATCATCTGCAACATTGGATCCGTCACTGGATTCAATGCC
     leLeuAspPheTrpAspLysArgLysGlyGlyProGlyGlyIleIleCysAsnIleGlySerValThrGlyPheAsnAla

       A      *         *         *         *         *         *         *   C     *
1281 ATCTACCAGGTGCCCGTCTACTCCGGCACCAAGGCCGCCGTGGTCAACTTCACCAGCTCCCTGGCGGTAAGTTGATCAAA
     IleTyrGlnValProValTyrSerGlyThrLysAlaAlaValValAsnPheThrSerSerLeuAla

      A       *         *       G *         T    A    *         *    A    *C        *
1361 GGAAACGCAAAGTTTTCAAGAAAAAACAAAACTAATTTGATTTATAACACCTTTAGAAACTGGCCCCCATTACCGGCGTG
                                                             LysLeuAlaProIleThrGlyVal

       G      * T       *         *         *         C         *         *       T *
1441 ACCGCTTACACCGTGAACCCCGGCATCACCCGCACCACCCTGGTGCACAAGTTCAACTCCTGGTTGGATGTTGAGCCCCA
     ThrAlaTyrThrValAsnProGlyIleThrArgThrThrLeuValHisLysPheAsnSerTrpLeuAspValGluProGl

           C  *         *         *      C  *         *         *         *     A   *
1521 GGTTGCTGAGAAGCTCCTGGCTCATCCCACCCAGCCATCGTTGGCCTGCGCCGAGAACTTCGTCAAGGCTATCGAGCTGA
     nValAlaGluLysLeuLeuAlaHisProThrGlnProSerLeuAlaCysAlaGluAsnPheValLysAlaIleGluLeuA

              *         *         *         *         *         *         *         *
1601 ACCAGAACGGAGCCATCTGGAAACTGGACTTGGGCACCCTGGAGGCCATCCAGTGGACCAAGCACTGGGACTCCGGCATC
     snGlnAsnGlyAlaIleTrpLysLeuAspLeuGlyThrLeuGluAlaIleGlnTrpThrLysHisTrpAspSerGlyIle

              *  C      *  ∇3     *         *         *         G         *         *
1681 TAAGAAGTGATAATCCCAAAAAAAAAAACATAACATTAGTTCATAGGGTTCGCGAACCACAAGATATTCACGCAAGGCAA
                            A

              *         *         *         *         *         *         *         *
1761 TTAAGGCTGATTCGATGCACACTCACATTCTTCTCCTAATACGATAATAAAACTTTCCATGAAAAATATGGAAAAATATA

              *         *         *         *         *         *       T *         *
1841 TGAAAATTGAGAAATCCAAAAAACTGATAAACGCTCTACTTAATTAAAATAGATAAATGGGAGCGGCAGGAATGGCGGAG

         A    *      T  *         *         *         *         *         *         *
1921 CATGGCCAAGTTCCTCCGCCAATCAGTCGTAAAACAGAAGTCGTGGAAAGCGGATAGAAAGAATGTTCGATTTGACGGGC

              *         *         *         *         *         *         *         *
2001 AAGCATGTCTGCTATGTGGCGGATTGCGGAGGAATTGCACTGGAGACCAGCAAGGTTCTCATGACCAAGAATATAGCGGT

     ⇓ ∇4     *         *         *         *         T         *         *         *
2081 GAGTGAGCGGGAAGCTCGGTTTCTGTCCAGATCGAACTCAAAACTAGTCCAGCCAGTCGCTGTCGAAACTAATTAAGTTA
     GT
              *         *         *         *         *         *         *         *
2161 ATGAGTTTTTCATGTTAGTTTCGCGCTGAGCAACAATTAAGTTTATGTTTCAGTTCGGCTTAGATTTCGCTGAAGGACTT

              *         *         *         *         *         *      ∇5 *         *
2241 GCCACTTTCAATCAATACTTTAGAACAAAATCAAAACTCATTCTAATAGCTTGGTGTTCATCTTTTTTTTTAATGATAAG
                                                                       T
              *         *      A  *         *         *        ⇓*←        * Δ6      *
2321 CATTTTGTCGTTTATACTTTTTATATTTCGATATTAAACCACCTATGAAGTTCATTTTAATCGCCAGATAAGCAATATAT

       → ⇓    *         *         *         *         *         *         *         *
2401 TGTGTAAATATTTGTATTCTTTATCAGGAAATTCAGGGAGACGGGGAAGTTACTATCTACTAAAAGCCAAACAATTTCTT

              *         *         *         *         *         *         *         *
2481 ACAGTTTTACTCTCTCTACTCTAGAAACTGGCCATTTTACAGAGTACGGAAAATCCCCAGGCCATCGCTCAGTTGCAGTC

              *         *         *         *         *         *         *         *
2561 GATAAAGCCGAGTACCCAAATATTTTTCTGGACCTACGACGTGACCATGGCAAGGGAAGATATGAAGAAGTACTTCGATG

              *
2641 AGGTGATGGTCCAAATGG
```

Fig. 2 Consensus sequence of the *D. melanogaster Adh* gene and flanking regions. The sequence shown is the consensus for the six *Adh-s* alleles. The sites of all the polymorphisms are indicated above the consensus sequences, including insertions (∇) and deletions (Δ). The sequences of the insertions are given below the consensus sequence, and ⇓ indicates the precise point of insertion. The horizontal arrows above the consensus sequence at sites ∇1 and Δ6 delineate the regions deleted in these two polymorphisms. The A ← → C polymorphism at position 1,490 is responsible for the Lys ← → Thr difference between the fast and slow *Adh* alleles. The 5′ untranslated leader sequence of the adult mRNA is made up of residues 1–87 and 742–777, and that of the larval mRNA of residues 708–777. The 3′ end of both mRNAs is at position 1,858. Strain *Af-s* has not been sequenced 3′ to position 2,347.

Methods: Cloning and sequencing strategy: An 11.8 kb *Bgl*II restriction fragment containing the *Adh* gene was isolated by plaque hybridization[27] from a λ1059 recombinant *Bgl*II genomic library[28–30], using as a probe sAC-1, a 4.5 kb *Eco*RI *Adh* genomic subclone in pBR322[31]. A 2.7 kb *Cla*I–*Sal*I fragment was subsequently subcloned[32] into pBR322. Overlapping nested deletions were constructed by ligating *Msp*I partial digestion products into the *Cla*I site of pBR322. Sets of 5–7 overlapping deletions were selected[33] and sequenced from the common pBR322 *Eco*RI site using the method of Maxam and Gilbert[34]. Additional restriction sites were used to confirm sequences and to determine flanking region sequences.

of insertion and therefore may represent a template for the tandem repeat. Because 29 bp are deleted in the genes carrying the insertion, the net increase in the size of the intron due to the insertion is only 5 bp.

∇2 is a 37 bp point insertion at position 550/551. The 5′ half of this insertion is a 19 bp tandem repeat of a sequence directly adjacent (3′) to the point of insertion. The 3′ half of the insertion is a 23 bp inverted complementary repeat of a sequence located further downstream from the point of insertion (positions 576–597) and so can form a large, extremely stable secondary structure in either RNA ($\Delta G = -66.9$)[14] or possibly in DNA as shown in Fig. 3. This secondary structure is related to a less stable one ($\Delta G = -28.9$), also shown in Fig. 3, that can form at the same position in genes lacking the insertion. This coincidence argues for the existence of one of the two secondary structures *in vivo*, as a means of generating the length polymorphism.

Relative rates of nucleotide substitution

Transitions and transversions. There is no evidence for an unequal distribution of nucleotide substitutions among the six possible classes of transitions and transversions for the 43 observed polymorphisms (homogeneity $G = 2.6$, $P > 0.5$). The observed ratio of transitions : transversions is 18 : 25, which does not differ significantly from expectation assuming equal mutation rates and no differential selection ($G = 1.36$, $P > 0.5$). Therefore, transitions do not appear to be accumulating at a significantly faster rate than transversions at the *Adh* locus.

Introns and exons. To test whether silent polymorphisms in the three exons occur at the same frequency as polymorphisms in introns, I calculated the number of 'effectively silent sites'[15] in the coding regions, as explained in Table 2. Overall, 25% (192 out of 765) of the nucleotide positions in the exons are effectively silent. Therefore the per cent polymorphism in the three coding regions, calculated as the observed number of silent polymorphisms × 100 divided by the effective number silent sites, is $13/192 \times 100 = 6.7\%$ (no significant difference between the three exons: homogeneity $G = 4.472$, $P > 0.1$). Expressed in terms of heterozygosity, two genes chosen at random from the sample differ, on average, at 2.3% of their silent sites.

There is a significant difference in the per cent silent polymorphism between introns (2.4%) and coding regions (6.7%) ($G = 7.36$, $P < 0.01$). However, this difference is accounted for by the significantly lower value in the adult intron (1.7%) than in the two small introns (5.2%) ($G = 5.17$, $P < 0.025$). This suggests that the larval mRNA leader sequence and adjacent noncoding sequence contained in the adult intron are functionally constrained. In contrast, the per cent polymorphism in the two small introns is not significantly different from that in the exons ($G = 0.33$, $P > 0.5$). Therefore, if there are selective constraints on the nucleotide sequence in the two small introns and at effectively silent sites in exons, they appear to be of roughly the same magnitude.

Only five nucleotide sites are polymorphic in the 3′ noncoding region out of an average of 767 sites under comparison. Furthermore, each of the five sites are substituted in only a single strain. Even by the conservative comparison of per cent polymorphism, this region (0.65%) has 10-fold lower silent polymorphism than the coding regions (6.7%) and the two small introns (5.2%). It is also significantly lower than the overall level of polymorphism in the exons (1.8%) which includes both amino acid replacement and non-replacement sites ($G = 4.4$, $P < 0.05$). A similar result has also been reported for a 5′ flanking region of the β-globin gene[16], which, in interspecific comparisons, is more highly conserved than the introns. Such conservation in flanking regions of structural genes must imply either strong functional constraints or low mutation rates.

Rather than being limited only to the flanking regions of structural loci, this conservation may be representative of the genome in general. Langley *et al.*[17] have obtained an estimate of 0.004 for the heterozygosity per nucleotide site over a 12 kb region around the *Adh* locus in *D. melanogaster* based on restriction map variation in 18 genomes from four populations. Although this large region may contain other structural loci in addition to *Adh*, their estimate is approximately the same as the average heterozygosity per nucleotide site in exons in this study (0.006). Therefore, large regions of the genome appear to be evolutionarily constrained.

Natural selection against replacement substitutions. The most striking result of the sequence analysis is the extent to which silent polymorphisms (13) outnumber amino acid replacement polymorphisms (1) in the coding region of the *Adh* gene (homogeneity $G = 29.4$, $P < 0.001$). Although it is possible that their underlying rates of mutation are different, it seems more reasonable to assume that the observed numbers reflect differences in the intensity of natural selection acting on the two classes of mutations. Under this assumption, the expected number of replacement polymorphisms, taking the observed number of silent polymorphisms as an estimate of the expected level, is $(765 - 192) \times 13/192 = 39$ amino acid replacements. The large discrepancy between the expected (39) and observed (1) number of replacement polymorphisms clearly indicates that the overwhelming majority of amino acids in the *Adh* polypeptide sequence are constrained by natural selection.

Even if replacement substitutions have been uniformly selected against, this does not imply that the magnitude of selective disadvantage of such mutations has been large. The effectiveness of natural selection at removing deleterious mutations depends on the effective population size, N_e, as well as the magnitude of selection against a mutant allele, s. For example, Kimura defines as 'effectively' neutral[18] only those alleles having selection coefficients, $s < 1/(2N_e)$. When population sizes are large, the fate of even very slightly deleterious mutations will be governed by natural selection.

That effective population sizes are large is suggested by an estimate of the parameter $\theta = 4N_e\mu$, where μ is the amino acid replacement rate per gene per generation. If the 11 *Adh* genes in this study were a random sample drawn from one population, at evolutionary equilibrium the number of silent polymorphisms in the exons would provide a minimum estimate of this parameter. The effect of using a species-wide sample is to overestimate this parameter to the extent that there is population subdivision within the species. However, most of the variation observed in this study is between *Adh-f* and *Adh-s* alleles within, rather than between, populations, implying that most of the species-wide polymorphisms are also present within

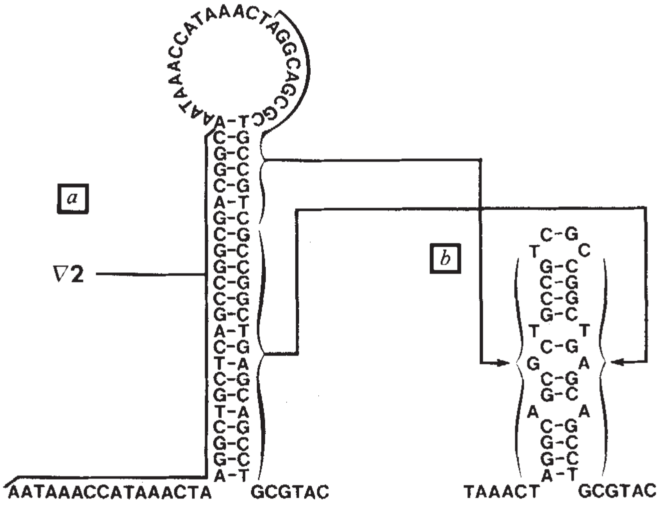

Fig. 3 *a*, Secondary structure predicted from the base sequence involving insertion ∇2. The 16 bp sequence at the 5′ end of the insertion that is not involved in the secondary structure is part of a 19 bp tandem repeat of the first 19 bp in the loop. *b*, A less stable secondary structure can also be formed between the two bracketed regions which is 3′ to the point of insertion. The 3′ locations of the two secondary structures are identical.

populations. Furthermore, in the one population in which the same electromorph is sampled twice (*Adh-s* in Southern Florida), the two genes differ in 6 of 13 possible silent polymorphisms. This suggests that the bias in the estimate of θ resulting from a species-wide sample will not be large. Of course, if silent polymorphisms are themselves selectively constrained, then θ may actually be underestimated.

Silent polymorphisms occur at approximately the same frequency in the coding region and the two small introns (see previous section): I assume this to be, as a first approximation, the equilibrium frequency of neutral polymorphisms. According to the infinite allele model of Kimura[19], the equilibrium number of segregating nucleotide sites at a locus, k, depends only on the effective population size, N_e, and the mutation rate to neutral variants per gene per generation, μ. Then for a sample size of r genes, an unbiased estimator of the parameter $\theta = 4N_e\mu$, is[20]

$$\hat{\theta} = k/S, \quad \text{where } S = (1/1)+(1/2)+\cdots+(1/(r-1))$$

Setting $k = 13$, the number of segregating silent sites observed in the sample of $r = 11$, then $4N_e\mu_s = 4.44$. In this case μ_s is the mutation rate at the 192 silent sites out of the 765 sites in the coding region. The mutation rate for amino acid replacements $\mu_r = 3\mu_s$, and therefore $4N_e\mu_r = 13.32$. Assuming a replacement mutation rate at the *Adh* locus of approximately 10^{-6} per generation, as suggested in a study on the mutation rate of several soluble enzyme loci (including *Adh*) to new electromorphs in *D. melanogaster*[21,22], the evolutionarily effective species size is then $N_e = 3.3 \times 10^6$. This estimate is consistent with an effective population size of much greater than 10^4 reported by Mukai and Yamaguchi[23] based on allelism of lethals.

If effective population sizes in *D. melanogaster* are reasonably close to the species estimate, the evolutionary fate of even slightly deleterious mutations at the *Adh* locus will be governed primarily by purifying selection rather than genetic drift. Therefore, the high degree of conservation of the *Adh* protein may be the result of natural selection being able to operate on mutants with small deleterious fitness effects.

Given the high degree of conservation within the species, it is of interest that *Adh* in *D. melanogaster* differs by at least three amino acid replacement substitutions[15] from its sibling species, *D. simulans*. Although this difference can be explained by the accumulation of selectively neutral substitutions, it is also consistent with the view that some of these differences have been selectively favoured. This argument rests on the theoretical result that when the effective population size is large, the probability of fixation of an advantageous mutation, being approximately twice its selective advantage[24], can be much greater than the probability of fixation of a neutral mutation, which remains approximately $1/(2N_e)$. Therefore, even if selectively favoured mutations arise at the *Adh* locus at a lower frequency than 'effectively' neutral mutations, their relative rate of evolution could be higher.

Phylogenetic comparison of *Adh* alleles

Several additional features of the evolution of *Adh* can be inferred from a different type of analysis by comparing the distribution of shared nucleotides among pairs or groups of alleles: this is a phylogenetic comparison of DNA sequences.

The 11 *Adh* sequences are arranged in Table 1 so that those with the largest number of shared nucleotides at the polymorphic sites are placed next to one another. For example *F1-1S* and *Wa-S* share nine nucleotides not present in any of the other nine alleles, but differ at only three polymorphic sites. This preponderance of shared substitutions implies that these two alleles are much more closely related to each other than to the remaining alleles in the sample. In contrast, the two most distantly related alleles, *Wa-s* and *Ja-f*, share no unique polymorphisms and differ at 33 sites. The alignment of genes in Table 1 indicates three distinct ancestrally related groups, with *F1-1S* and *Wa-S* forming one group, *Af-S* and *Fr-S* a second group, and the five *Adh-f* alleles a third group. The remaining two genes, *F1-2s* and *Is-S*, are phylogenetically intermediate. The fact that closely related alleles are broadly distributed geographically provides additional evidence that the effective population size is large.

The *Adh-f* alleles can be distinguished from *Adh-s* alleles by only three of the 49 possible polymorphisms, but as a group are considerably more homogeneous than the *Adh-s* alleles. For example, the six *Adh-s* alleles differ from one another, on average, at 15.5 sites, while *Adh-f* alleles differ by 8.1 sites ($F = 1.4173$, $P < 0.025$). There are two possible explanations for this result. One is that the slow allele is ancestral to the fast and has had more time to accumulate nucleotide variation. The alternative assumes that the two alleles are at equilibrium between mutation and drift but that the frequency of the fast allele has been lower, on average, than that of the slow allele. However, the fact that two sibling species, *D. simulans* and *D. mauritiana*, carry the slow allele substitution at codon 192[13] suggests that *Adh-f* is derived from an *Adh-s* allele.

Although the polymorphic differences distinguishing the 11 strains are clustered along the DNA, as would be expected with tight linkage, two independent intragenic recombination events can be inferred from the distribution of shared polymorphism within both fast and slow alleles. *Fr-s* and *Ja-s* contain the same nucleotide at polymorphic positions 107, 113, 175, 293 and 304 in the adult intron while *Fl-2s* and *Fr-f* contain the other nucleotide at each of these sites. This defines two divergent lineages. Yet each lineage is polymorphic for four sites in exon 3—1,452, 1,518, 1,527 and 1,557. Unless these polymorphisms have arisen independently in both lineages, their presence implies a recombination between the 3′ and 5′ ends of the gene. Similarly, two complementary lineages *Fl-1s–Wa-s* and *Af-s–Wa-f*, which can be defined by polymorphisms at positions 1,068, 1,229, 1,283, 1,354, 1,362, 1,400, 1,405 and 1,431, are each polymorphic for two positions in the 5′ end of *Adh*, −1 and 175.

The results of this study raise several questions that can only be addressed by additional sequence comparison studies. Having demonstrated a high level of silent nucleotide polymorphism for the *Adh* gene within a species, it is now important to determine how much of this variation occurs in single populations. It is also important to investigate other loci, especially to determine whether the estimate of nucleotide polymorphism and the level of purifying selection observed for *Adh* is representative of other genes. The distribution of polymorphism at loci which are already known to contain many electrophoretic alleles, such as xanthine dehydrogenase and esterase-5 in *Drosophila*, will be particularly revealing: it is possible that intragenic recombination will be important in generating a diversity of alleles from a much smaller number of amino acid replacements[25]. Finally, the fact that non-coding regions can be very highly conserved relative to introns challenges the assumption that these regions have no functional importance. However, if this conservation is characteristic of most of the genome, its maintenance would be difficult to explain by purifying selection, which would generate very high levels of mutational genetic load in the species. This suggests that mutation rates may vary within the genome.

I thank W. Gilbert, W. Herr, G. Church, D. Strauss, R. Cate, R. Tizard, C. Benyajati, R. Blackman, S. Orzack, L. Brooks, J. Coyne, R. C. Lewontin, T. Mukai, J. David, T. Prout, L. Harley, D. Goldberg and J. Karn for their help. This work was supported by grants GM21179 and GM29301 awarded by the National Institute of General Medical Sciences to R. C. Lewontin.

Received 5 April; accepted 6 May 1983.

1. Hubby, J. L. & Lewontin, R. C. *Genetics* **54**, 577–594 (1966).
2. Lewontin, R. C. & Hubby, J. L. *Genetics* **54**, 595–609 (1966).
3. Coyne, J. A. *Isozymes: Current Topics in Biological & Medical Research* Vol. 6, 1–32 (Liss, New York, 1982).
4. Maxam, A. M. & Gilbert, W. *Proc. natn. Acad. Sci. U.S.A.* **74**, 560–564 (1977).
5. Sanger, F., Nicklen, S. & Coulson, A. R. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463–5467 (1977).
6. Johnson, F. M. & Schaffer, H. E. *Biochem. Genet.* **10**, 149–163 (1973).

7. Oakshott, J. G. *et al. Evolution* **36,** 86–96 (1982).
8. Benyajati, C., Place, A. R., Powers, D. A. & Sofer, W. *Proc. natn. Acad. Sci. U.S.A.* **78,** 2717–2721 (1981).
9. Retzios, A. D. & Thatcher, D. R. *Biochimie* **61,** 701–704 (1980).
10. Chambers, G. K. *Genetic Studies of Drosophila Populations* (eds Gibson, J. B. & Oakshott, J. G.) 77–94 (Australian National University, Canberra, 1981).
11. Sampsell, B. *Biochem. Genet.* **15,** 971–988 (1977).
12. Wilkes, A. V., Gibson, J. B., Oakshott, J. G. & Chambers, G. K. *Aust. J. biol. Sci.* **33,** 575–585 (1980).
13. Bodmer, M. & Ashburner, M. (personal communication).
14. Borer, P. N., Dengler, B. & Tinoco, I. *J. molec. Biol.* **86,** 843–853 (1974).
15. Holmquist, R., Cantor, C. R. & Jukes, T. H. *J. molec. Biol.* **64,** 145–162 (1972).
16. Moschonas, N., deBoer, E. & Flavell, R. A. *Nucleic Acids Res.* **10,** 2109–2120 (1982).
17. Langley, C. H., Montgomery, E. & Quattlebaum, W. F. *Proc. natn. Acad. Sci. U.S.A.* **79,** 5631–5635 (1982).
18. Kimura, M. *Nature* **217,** 624–626 (1968).
19. Kimura, M. *Genetics* **61,** 893-903 (1969).
20. Ewens, W. J. *Mathematical Population Genetics* (Springer, New York, 1979).
21. Mukai, T. & Cockerham, C. C. *Proc. natn. Acad. Sci. U.S.A.* **74,** 2514–2517 (1977).
22. Voelker, R. A., Shaffer, H. E. & Mukai, T. *Genetics* **94,** 961–968 (1980).
23. Mukai, T. & Yamagushi, O. *Genetics* **76,** 339–366 (1974).
24. Crow, J. F. & Kimura, M. *An Introduction to Population Genetic Theory* (Burgess, Minneapolis, 1970).
25. Strobeck, C. & Morgan, K. *Genetics* **88,** 829–844 (1978).
26. Benyajati, C., Spoerel, N., Haymerle, H. & Ashburner, M. *Cell* **33,** 125–133 (1983).
27. Benton, W. D. & Davis, R. W. *Science* **196,** 180–182 (1977).
28. Maniatis, T. *et al. Cell* **15,** 687-701 (1978).
29. Enquist, L. & Sternberg, N. *Meth. Enzym.* **68,** 281–298 (1979).
30. Karn, J., Brenner, S., Barnett, L. & Cesareni, G. *Proc. natn. Acad. Sci. U.S.A.* **77,** 5172–5176 (1980).
31. Goldberg, D. A. *Proc. natn. Acad. Sci. U.S.A.* **77,** 5794–5798 (1980).
32. Grunstein, M. & Hogness, D. *Proc. natn. Acad. Sci. U.S.A.* **72,** 3961 (1975).
33. Holmes, D. S. & Quigley, M. *Analyt. Biochem.* **114,** 193–197 (1981).
34. Maxam, A. M. & Gilbert, W. *Meth. Enzym.* **65,** 499–560 (1980).

LETTERS TO NATURE

Discovery of a 6.1-ms binary pulsar PSR1953+29

V. Boriakoff*, R. Buccheri† & F. Fauci‡

* National Astronomy and Ionosphere Center, Cornell University, Ithaca, New York 14850, USA
† Istituto di Fisica Cosmica e Informatica, Consiglio Nazionale delle Richerche, Palermo, Italy
‡ Istituto di Fisica, GIFCO–CNR, Universit di Palermo, Palermo, Italy

A systematic search for fast radio pulsars in some of the error boxes of the γ-ray satellite COS B[1] unidentified point sources[2–4] on the 305-m Arecibo Observatory radiotelescope was started 3 years ago. The motivation for the search and the method and limitations are given elsewhere[5]. One of the search areas covered was the γ-source 2CG065+00. We report here the discovery of a 6.13317-ms pulsar in a binary system with a 120-day orbital period inside the error box of this source. The initial search observation (23 July 1980) with a sampling rate of 4.167 ms produced an aliased (76.93 Hz) detection at 430 MHz. Confirmatory observations made in March 1983, showed a variable pulse frequency indicating a Doppler shift due to a periodic orbit around another object[6]. A compilation of measured system parameters and the underlying assumptions are given in Table 1.

Figure 1 shows the pulsar pulse frequency–time relationship. Data taken at 430 MHz in the search mode contain simultaneous scanning of the output of 30 contiguous-in-frequency filters. Because the initial confirmation runs were done in the search mode with 0.25-MHz bandwidth per filter (dispersion time in one filter bandwidth is 44% of the pulse period), error in the pulse frequency was relatively large, but this method of data collection was continued until other parameters were determined. Later on dedispersed observations were used with a filter bank of 20 kHz bandwidth per filter with better timing accuracy[7]. At the same time single-pulse arrival timing was started. Because only about half of an orbit has been observed, the orbital parameters determined up to now may have errors. However, initial fittings of a sinusoid to the pulse frequency show a very good approximation, although at present a non-zero orbital eccentricity cannot be excluded. Assuming $e = 0$ a best-fit sinusoid has a period of 120 days and a peak velocity value of 6.02 km s^{-1}. Care was taken to rule out the possibility of an aliased orbital period. The worst case is a 23.74 or a 24.14-h period, all other orbital periods shorter than these will show drastic pulse frequency changes over the maximum possible 2 h of observation. On days when $\mathrm{d}P/\mathrm{d}t$ due to the Doppler shift was maximum, pulse frequencies close to both ends of the 2 h of observation were compared and found to be consistent only with $P_{\mathrm{orb}} = 120$ days.

The July 1980 observation is close to the lowest pulse frequency of the 1983 half-orbital period observations. Assuming that indeed it is the lowest frequency a maximum value of $\dot{P}$ can be computed by taking the difference between the 1980 and 1983 minima. As the 1980 point is not necessarily the minimum, a smaller value of $\dot{P}$ is likely; this includes negative $\dot{P}$ values. We have assumed the shape and size of the orbit to be invariant over the 3 yr interval. A better value of $\dot{P}$ will be obtained when the orbital parameters are sufficiently well

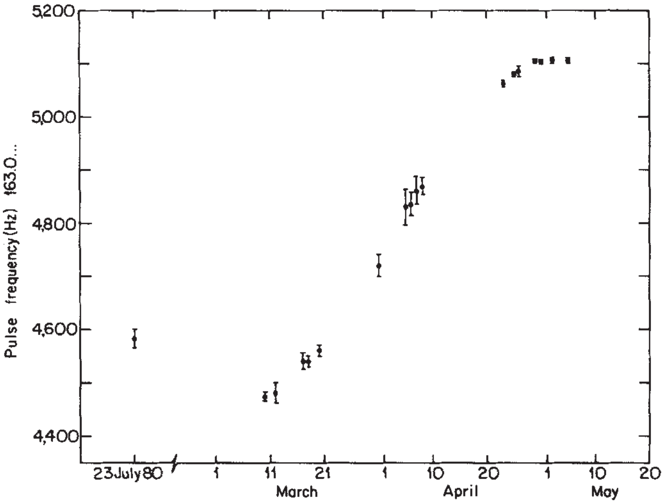

Fig. 1 Frequency of pulses at the Solar System barycentre of PSR1953+29 function of time. Accuracy is dependent on the data taking system used.

Table 1 Preliminary parameters of the millisecond binary pulsar PSR1953+29

VLA position (1950.0):
RA = 19h 53min 26.7s ± 0.24s
Dec. = 29° 00′ 42.0″ ± 3.2″
$l = 65.84°$
$b = +0.443°$
$P = 6.13317$ ms ± 0.00002 on JD = 2445427.96
$\dot{P} \leq 5.8 \times 10^{-16}$ s s^{-1}
Characteristic age = $P/2\dot{P} \geq 1.675 \times 10^{5}$ yr
$B \leq 6.0 \times 10^{10}$ G (assuming $I = 10^{45}$ g cm^{2}, $R = 10^{6}$ cm, $\alpha = \pi/2$)
Velocity of light cylinder radius = 292.6 km
Projected orbital velocity peak value = $K_{\mathrm{p}} = 6.02$ km s^{-1} ± 0.02
Dispersion measure = 104.5 pc cm^{-3} ± 0.03
Distance = 3.5 kpc (assume $\langle \eta_{\mathrm{e}} \rangle = 0.03$ cm^{-3})
Flux at 1,385 MHz as a continuum source = 1.0 mJy ± 0.5
If the orbit is circular ($e = 0$) then
$P_{\mathrm{orb}} = 120$ days ± 4
$a_{\mathrm{p}} \sin i = 9.9 \times 10^{6}$ km ± 0.1

$$\text{Mass function} = \frac{m_{\mathrm{c}}^{3} \sin(i)}{(m_{\mathrm{p}} + m_{\mathrm{c}})^{2}} = 0.00272 M_{\odot}$$