Molecular population genetics of sequence length diversity in the *Adh* region of *Drosophila pseudoobscura*

STEPHEN W. SCHAEFFER*
Department of Biology, and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, PA 16802, USA

(*Received 12 April 2002 and in revised form 21 August 2002*)

Summary

Positive and negative selection on indel variation may explain the correlation between intron length and recombination levels in natural populations of *Drosophila*. A nucleotide sequence analysis of the 3·5 kilobase sequence of the alcohol dehydrogenase (*Adh*) region from 139 *Drosophila pseudoobscura* strains and one *D. miranda* strain was used to determine whether positive or negative selection acts on indel variation in a gene that experiences high levels of recombination. A total of 30 deletion and 36 insertion polymorphisms were segregating within *D. pseudoobscura* populations and no indels were fixed between *D. pseudoobscura* and its two sibling species *D. miranda* and *D. persimilis*. The ratio of Tajima's *D* to its theoretical minimum value (D_{min}) was proposed as a metric to assess the heterogeneity in *D* among *D. pseudoobscura* loci when the number of segregating sites differs among loci. The magnitude of the D/D_{min} ratio was found to increase as the rate of population expansion increases, allowing one to assess which loci have an excess of rare variants due to population expansion versus purifying selection. *D. pseudoobscura* populations appear to have had modest increases in size accounting for some of the observed excess of rare variants. The D/D_{min} ratio rejected a neutral model for deletion polymorphisms. Linkage disequilibrium among pairs of indels was greater than between pairs of segregating nucleotides. These results suggest that purifying selection removes deletion variation from intron sequences, but not insertion polymorphisms. Genome rearrangement and size-dependent intron evolution are proposed as mechanisms that limit runaway intron expansion.

1. Introduction

How selection acts on insertion and deletion variation to expand or contract the sizes of genomes has been examined with studies of processed pseudogenes, retroposons and introns. Comparisons of paralogous functional and pseudogenes indicated that deletions tend to outweigh insertion events (Graur *et al.*, 1989; Saitou & Ueda, 1994; Ogata *et al.*, 1996; Ophir & Graur, 1997). Comparisons of *Drosophila* dead-on-arrival retroposons with functional elements also showed a strong bias towards deletions and against insertions (Petrov *et al.*, 1996; Petrov & Hartl, 1998). This indel bias suggested that either deletions are favoured by selection or insertions are selected against when genetic information is replicated to new locations in the genome by either gene duplication or transposition.

A negative correlation between intron size and local recombination rate in *Drosophila* has suggested that selection acts on intron size variation differently in the genome (Carvalho & Clark, 1999; Comeron & Kreitman, 2000). The 41 000 introns within the *D. melanogaster* genome occupy 20 Mb of DNA (Adams *et al.*, 2000) and vary in size from 40 bp to 70 kb. The dominant size class, however, varies between 59 and 63 bp (Mount *et al.*, 1992; Deutsch & Long, 1999; Adams *et al.*, 2000). Carvalho & Clark (1999) have suggested that short (<60 bp) and long (>80 bp) intron sizes are selected against in regions of high recombination, but larger and smaller introns are located in regions of low recombination because selection is less effective in removing non-optimally sized introns due to the Hill–Robertson effect (Hill & Robertson, 1966). Alternatively, Comeron & Kreitman

* Department of Biology, The Pennsylvania State University, 208 Mueller Labs, University Park, PA 16802-5301, USA. Tel: +1 (814) 8653269. Fax: +1 (814) 8659131. e-mail: sws4@psu.edu

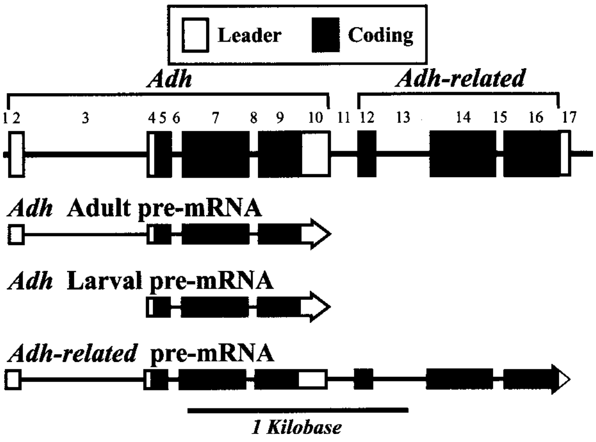

Fig. 1. The *Adh* region of *D. pseudoobscura*. Two genes, *Adh* and *Adhr*, are encoded in this region and the transcripts from the two genes are shown. *Adh* is expressed from a distal promoter in adults and a proximal promoter in larvae. *Adhr* is expressed from the *Adh* promoters and is produced as a dicistronic message (Brogna & Ashburner, 1997). The numbers above the gene region label 17 functional domains in the *Adh* region: (1) 5′ flanking region; (2) *Adh* adult leader; (3) *Adh* adult intron (770 bp); (4) *Adh* larval leader; (5) *Adh* exon 1; (6) *Adh* intron 1 (59 bp); (7) *Adh* exon 2; (8) *Adh* intron 2 (64 bp); (9) *Adh* exon 3; (10) 3′ UTR; (11) intergenic; (12) *Adhr* exon 1; (13) *Adhr* intron 1 (304 bp); (14) *Adh* exon 2; (15) *Adh* intron 2 (57 bp); (16) *Adh* exon 3; and (17) 3′ UTR and 3′ flanking.

(2000) thought that selection favours long introns in regions of low recombination because larger introns would act as positive modifiers of recombination in the face of a strong deletion bias (Graur *et al.*, 1989; Saitou & Ueda, 1994; Ogata *et al.*, 1996; Petrov *et al.*, 1996; Petrov & Hartl, 1998). Recently, Comeron & Kreitman (2002) have shown that introns can reduce the effects of the Hill–Robertson effect by increasing the distance among weakly selected sites.

The alcohol dehydrogenase (*Adh*) region of *D. pseudoobscura* provides an excellent model system to test how selection operates on intron size. Two intron-containing genes are located in this region: *Adh* and *Adh-related* (*Adhr*) (Schaeffer & Aquadro, 1987; Brogna & Ashburner, 1997). These two genes were derived from a common ancestral gene that pre-dates the formation of the genus *Drosophila* (Russo *et al.*, 1995). The *Adh* region is on the fourth chromosome of *D. pseudoobscura* (Schaeffer & Aquadro, 1987) and is in a region that experiences extensive recombination (Schaeffer & Miller, 1993). The two genes have five introns that fall into two classes of intron size: small (<80 bp) or large (>80 bp) (Mount *et al.*, 1992). The *Adh* adult intron (770 bp) and the *Adhr* intron 1 (304 bp) make up the large intron class, while *Adh* intron 1 (59 bp), *Adh* intron 2 (64 bp) and *Adhr* intron 2 (57 bp) make up the small intron class (Fig. 1). This study presents a molecular population genetic analysis of indel variation in the introns of *Adh* and *Adhr* to determine how selection acts on intron length variation for a gene in a region with extensive recombination.

2. Materials and methods

(i) *Nucleotide sequences and GenBank accession numbers*

Fig. 1 shows the fragment of DNA sequenced in this study and the fine structure of the sequenced *Adh* region. A total of 139 sequences of the *D. pseudoobscura Adh* region were analysed and the details of the sequencing methods are found in (Schaeffer & Miller, 1991, 1992*a*, *b*, 1993). The population location, strain names and the EMBL/GenBank Data Library (Benson *et al.*, 2002) accession numbers for these 139 nucleotide sequences have been published previously (Schaeffer & Miller, 1993; Schaeffer *et al.*, 2001). The sequences of the *Adh* region in *D. persimilis* (GenBank accession numbers AF006564–AF006568; Wang *et al.*, 1997) and *D. miranda* (GenBank accession number M60998; Schaeffer & Miller, 1991) were used as outgroups to polarize indel mutations within *D. pseudoobscura*. The strain PSU 271 (GenBank accession number U64535) from the Kaibab National Forest in Arizona was updated to include the 1105 bp transposable element insertion that was excluded from previous analyses.

(ii) *DNA sequence alignment*

The 139-nucleotide sequences were aligned manually with the Eyeball Sequence Editor (ESEE, version 2.00a; Cabot & Beckenbach, 1989). The alignments were determined by minimizing the number of mismatches and gaps assumed in the sequences. The alignment is different from that used in Schaeffer *et al.* (2001) because several gap regions were shifted to minimize the numbers of mismatches. The aligned sequences are available from the POPSET database within GenBank or from the author in other file formats.

(iii) *Analysis of sequence length polymorphisms*

Gaps in the sequence alignment of the 139 strains of *D. pseudoobscura* were assumed to result from insertion or deletion events. Indels that involved a single nucleotide site within the alignment could easily be classified as an insertion or deletion by examination of the *D. miranda* and *D. persimilis* sequences for the presence or absence of sequences (Fig. 2). If *D. persimilis* and *D. miranda* lacked nucleotides at a site or sites, but *D. pseudoobscura* had bases segregating in the population, then the indel was classified as an insertion event. If *D. persimilis* and *D. miranda* had nucleotides at a site or sites, but some strains of *D. pseudoobscura* were missing bases, then the indel was inferred to be a deletion event. If an indel was segregating in both

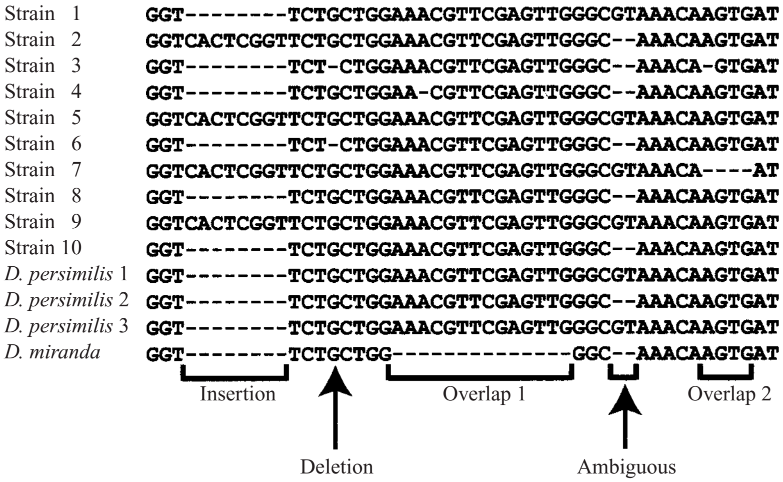
Strain 1 GGT--------TCTGCTGGAAACGTTCGAGTTGGGCGTAAACAAGTGAT
Strain 2 GGTCACTCGGTTCTGCTGGAAACGTTCGAGTTGGGC--AAACAAGTGAT
Strain 3 GGT--------TCT-CTGGAAACGTTCGAGTTGGGC--AAACA-GTGAT
Strain 4 GGT--------TCTGCTGGAA-CGTTCGAGTTGGGC--AAACAAGTGAT
Strain 5 GGTCACTCGGTTCTGCTGGAAACGTTCGAGTTGGGCGTAAACAAGTGAT
Strain 6 GGT--------TCT-CTGGAAACGTTCGAGTTGGGC--AAACAAGTGAT
Strain 7 GGTCACTCGGTTCTGCTGGAAACGTTCGAGTTGGGCGTAAACA----AT
Strain 8 GGT--------TCTGCTGGAAACGTTCGAGTTGGGC--AAACAAGTGAT
Strain 9 GGTCACTCGGTTCTGCTGGAAACGTTCGAGTTGGGCGTAAACAAGTGAT
Strain 10 GGT--------TCTGCTGGAAACGTTCGAGTTGGGC--AAACAAGTGAT
D. persimilis 1 GGT--------TCTGCTGGAAACGTTCGAGTTGGGCGTAAACAAGTGAT
D. persimilis 2 GGT--------TCTGCTGGAAACGTTCGAGTTGGGC--AAACAAGTGAT
D. persimilis 3 GGT--------TCTGCTGGAAACGTTCGAGTTGGGCGTAAACAAGTGAT
D. miranda GGT--------TCTGCTGG--------------GGC--AAACAAGTGAT
Insertion
Overlap 1
Overlap 2
Deletion
Ambiguous

Fig. 2. Scheme for inferring insertions and deletions within *D. pseudoobscura* using the outgroup sequences of *D. persimilis* and *D. miranda*. Dashes indicate sequences missing from a sequence. Two overlapping indel events are shown to illustrate characteristics of the observed data. Overlapping indel 1 is an insertion event that was followed by a deletion event. Overlapping indel 2 shows two independent deletions: a four-base deletion and a one-base deletion. An ambiguous event is shown where the apparent insertion is segregating in both *D. pseudoobscura* and *D. persimilis*.

D. persimilis and *D. pseudoobscura*, then the indel could not be classified and its state in *D. pseudoobscura* was considered ambiguous. Indels that included contiguous bases were assumed to be the result of a single event rather than multiple events if adjacent sites had the same phylogenetic partition (for a definition of phylogenetic partition see Stephens, 1985).

Some indels events within *D. pseudoobscura* overlapped, requiring assumptions to be made about the polarity of successive mutations. Overlap 1 in Fig. 2 shows a case where *D. miranda* is missing 14 bases that the majority of *D. pseudoobscura* strains have. Without definitive sequence in *D. miranda*, it is unclear whether the original derived mutation within *D. pseudoobscura* was 13 or 14 bases. The *D. persimilis* sequences were used to clarify the ancestral state as 14 bases. Overlap 2 in Fig. 2 shows two independent deletion events. In this case, the ancestral state is clearly known and the two derived mutations are easily inferred. An ambiguous event is also shown where the two-base indel is segregating in both *D. persimilis* and *D. pseudoobscura*.

Each indel event was classified into one of five categories based on the repetitive nature of the sequence. An insertion or deletion was classified as a direct repeat if the sequence flanking was identical to the added or dropped bases. A homopolymer repeat was inferred if the flanking sequence had a run of bases identical to the inserted or deleted base. An indel was considered to be a microsatellite if two to seven bases were repeated three or more times in the region of the sequence duplication or loss (Charlesworth *et al.*, 1994). A sequence was classified as non-repetitive if the inserted or deleted bases were unrelated to the flanking sequence. The last category of repeat is those sequences that show similarity to known transposable elements.

(iv) *Tests for clustering of indel variation*

Two approaches were used to test for uniform distribution of indels across the *Adh* region. The *Adh* region can be divided into 17 function domains (Fig. 1). A chi-square goodness of fit test was used to determine whether the frequency of indels was directly related to the number of bases in the respective functional domains. The goodness of fit test asked whether insertions and deletions were distributed uniformly among the non-coding domains. Tang & Lewontin (1999) suggested that a cumulative distribution function, $G(x)$, be used to find significant clustering of genetic diversity in a genomic region. $G(x)$ is defined as:

$$G(x_k) = \frac{k}{S} - \frac{x_k}{N},$$

where k is the ordinal rank of the segregating site, S is the total number of segregating sites, x_k is the nucleotide number of the kth segregating site, and N is the total number of nucleotides in the sequence. For the indel variants, x_k was the 5′ nucleotide of the insertion or deletion in the sequence alignment that had all gaps and non-coding sequence removed or 1656 nucleotides. Regions where segregating sites are clustered show a monotonic increase in $G(x)$, while regions that lack segregating sites show a monotonic decrease in $G(x)$. The statistical significance of the segments that showed the longest increases or decreases in $G(x)$ were evaluated with a random permutation test. For each simulation, the total number of indel segregating sites were randomly assigned to the 1656 aligned nucleotide sites in the *Adh* region and $G(x)$ was estimated from the randomized data. The monotonic increases or decreases were estimated and ranked from each permuted data set. A total of 100 000 simulations were performed and the significance level was determined by comparing the observed monotonic increases or decreases with that of the rank-ordered simulation distributions.

(v) *Statistical tests for departures from neutrality*

The frequency spectrum of the indel variation was tested for departures from an equilibrium neutral model of molecular evolution with Tajima's D (1989) statistic. The Tajima test was performed on deletion and insertion variation separately. Coalescent simulations were used to determine the probabilities of observing more extreme values of the D statistics given the number segregating sites and sample size in the two classes of mutation. A total of 10 000 coalescent simulations were used to estimate the probability of the observed Tajima's D. A significant negative value of Tajima's D can result from strong purifying selection or population expansion, while a significant positive value of D can result from balancing selection.

Population demography is expected to act on loci in the genome homogeneously, while purifying selection is expected to act in a heterogeneous manner (Hahn *et al.*, 2002). Twenty-one of 22 loci within *D. pseudoobscura* have negative Tajima's D values (Wang & Hey, 1996; Wang *et al.*, 1997; Hamblin & Aquadro, 1999; Kovacevic & Schaeffer, 2000; Schaeffer *et al.*, 2001; Machado *et al.*, 2002), but it is difficult to assess heterogeneity among D estimates at different loci when sample sizes and numbers of segregating sites vary. Hahn *et al.* (2002) used Fu & Li's (1996) D to test for significant excesses in rare variants in synonymous and non-synonymous sites and used a difference statistic (ΔD) between these two types of sites to assess heterogeneity in coalescent simulations.

I propose using the ratio of Tajima's (1989) D to its theoretical minimum value $D_{\min}$ to assess heterogeneity among loci. The expectation of $D/D_{\min}$ will be shown to be similar for loci with different numbers of segregating sites. The magnitude of $D/D_{\min}$ will be shown to decrease as the rate of population expansion increases. Tajima's D will have its minimum value ($D_{\min}$) when all segregating sites have a frequency of 1 in the sample, giving a minimum number of pairwise differences, $k_{\min}$:

$$k_{\min} = S\left(1 - \sum_{i=1}^{2} p_i^2\right) = S\left(\frac{2(n-1)}{n^2}\right),$$

where S is the number of segregating sites, p_i is the frequency of the ith allele, n is the sample size, and all sites have two nucleotides segregating. From equation 38 in Tajima (1989), $D_{\min}$ is defined as:

$$D_{\min} = \left|\frac{d}{Var(d)}\right| = \left|\frac{k_{\min} - \frac{S}{a}}{Var(d)}\right|.$$

The ratio of D to $D_{\min}$ is:

$$\frac{D}{D_{\min}} = \frac{\frac{d}{Var(d)}}{\left|\frac{d_{\min}}{Var(d)}\right|} = \frac{k - \frac{S}{a}}{\left|k_{\min} - \frac{S}{a}\right|}.$$

The absolute value of $D_{\min}$ is used to preserve the sign of D in the ratio. We used coalescent simulations to estimate the mean and the 95% confidence interval on the $D/D_{\min}$ ratio assuming various exponential growth parameters (Nr) (Slatkin & Hudson, 1991), where N is the current effective population size and r is exponential growth rate. Lower values of Nr approach the expectations of a constant population size, while higher values of Nr indicate populations that are expanding at faster rates. When populations are expanding, genealogies are expected to become more star-like so that most loci will have an excess of rare variants and the $D/D_{\min}$ ratio will approach -1.

(vi) *Pairwise linkage disequilibrium*

All pairs of indels in the *Adh* region were tested for significant non-random association with Fisher's exact test (Sokal & Rohlf, 1981). Only comparisons capable of generating a significant result with Fisher's exact test were performed (Lewontin, 1995). A sequential Bonferroni correction was used to overcome the multiple comparison problem (Rice, 1989). The strength of linkage disequilibrium between indel sites was assessed with the r^2 coefficient (Hill & Robertson, 1968).

3. Results

(i) *Indel diversity in the* Adh *region*

A total of 84 indel mutations have occurred in the *Adh* region since *D. pseudoobscura* and *D. miranda*/*D. persimilis* diverged (Table 1). A total of 30 deletion and 36 insertion polymorphisms were segregating within *D. pseudoobscura* populations, while 18 indels could not be unambiguously classified as either event. The deletions vary in size from 1 to 43 bp with an average loss of 6·7 bp, while the insertions vary in length from 1 to 1105 bp with an average gain of 33·4 bp. The 1105 bp insertion is a rare mobile element that significantly inflates the estimate of mean insertion size. If the single insertion of 1105 bp is removed from the mean estimate, then the average insertion size is 2·8 bp.

Average indel heterozygosity can be estimated from the number of segregating indels (Θ) (Watterson, 1975) or from the average number of indel differences between two randomly chosen genes (π) (Nei, 1987). The average heterozygosity based on segregating indels is $\Theta = 15{\cdot}2$ per locus. The average heterozygosity based on pairwise differences is lower than the segregating site estimates, where $\pi = 7{\cdot}9$ per locus for all indels. Indel heterozygosity is one-fifth of nucleotide heterozygosity, $\Theta = 75{\cdot}9 \pm 308{\cdot}0$ and $\pi = 37{\cdot}1 \pm 262{\cdot}9$ (Schaeffer *et al*., 2001).

We wanted to determine whether the inserted or deleted nucleotides were similar to flanking sequences. If the indel mutations result from mechanisms that involve repetitive DNA, then one might expect that adjacent nucleotides will be similar in sequence. The observed estimate of flanking sequence similarity was obtained by comparing the indel nucleotides to the sequences immediately 5′ or 3′ to the indel site. For instance, the two-base insertion of GT at positions 132 and 133 in the overall alignment was compared with the two bases 5′ (GT) and 3′ (AA) of the site to determine the maximum similarity estimate. The 5′ sequence was identical to the inserted bases while the 3′ sequence shared no nucleotides in common with the indel (Table 1), which gives a maximum fraction of similar bases of 1·0. A mean fraction of similar nucleotides was estimated for insertions and deletions separately.

We asked what the expected similarity of flanking sequences is given the sizes of events within the two classes of sequence length variants. One might expect small insertions or deletions to have higher similarity values because of the reduced number of states possible in the flanking sequences. A bootstrap analysis was used to derive the random expectation for flanking sequence similarity in the insertion and deletion class of mutations. This analysis did not take into account that the indels could be related to a more distant flanking sequence, only the adjacent sequence. For each bootstrap replicate, the maximum similarity was derived over all events within both mutation classes. For each event of size *n* bases, a random nucleotide site was drawn from the *Adh* region and *n* nucleotides were compared with the 5′ and 3′ adjacent sequences to derive a maximum similarity. A mean similarity was estimated for each bootstrap replicate and a 95% confidence interval (CI) was determined from 10 000 bootstrap replicates. The observed similarity of flanking sequences in 30 deletions was 0·634, which was significantly greater than the expected similarity for a randomly chosen sequence (mean = 0·400, 95% CI = 0·278 to 0·530). The observed similarity of flanking sequences in 35 insertions, excluding the transposon, was 0·782, which was significantly greater than the random expectation (mean = 0·432, 95% CI = 0·299 to 0·572).

The observed similarity of flanking sequences is slightly higher for insertions than deletions. A bootstrap analysis was used to derive 95% CIs for the mean flanking sequence similarities in deletions and insertions. The observed data were resampled with replacement up to the total number events within each mutation class and the mean sequence similarity was estimated within a bootstrap replicate. The 95% CI was estimated from 10 000 bootstrap replicates. Although the flanking sequence similarity was higher in insertions compared with deletions, this difference was not significantly different (deletion, 95% CI = 0·491 to 0·773; insertions, 95% CI = 0·687 to 0·922).

Table 2 shows the frequencies of the five classes of insertion and deletion observed in the *Adh* region. The observed data show that direct repeats, homopolymers and microsatellites are more frequently involved in insertion events than in deletion events, but a chi-square test of homogeneity for the four classes of indel event, excluding transposable elements, fails to show a significant difference in the frequency of the four classes of indel event in the *Adh* region ($\chi^2 = 7{\cdot}50$; df = 3; $P > 0{\cdot}05$). If direct repeats, homopolymers and microsatellites are combined into a single repetitive DNA classification, then a chi-square homogeneity test shows that insertions tend to occur more frequently in repetitive DNA sequences than in non-repetitive sequences. Alternatively, deletions occur more often in non-repetitive than in repetitive DNAs (two-tailed Fisher's exact test, $P = 0{\cdot}011$).

(ii) *Distribution of indel variation*

The indel variation was tested for non-random distribution among the 11 non-coding domains in the *Adh* region with a chi-square goodness of fit test. Neither insertions nor deletions showed a significant non-random distribution among the non-coding regions; however, the combined set of indels were non-randomly distributed among the 10 non-coding domains (deletions: $\chi^2 = 11{\cdot}7$, df = 10, $P = 0{\cdot}300$; insertions: $\chi^2 = 15{\cdot}4$, df = 10, $P = 0{\cdot}118$; indels: $\chi^2 = 23{\cdot}4$, df = 10, $P = 0{\cdot}009$). The cumulation distribution function, $G(x)$, of Tang & Lewontin (1999) was used to determine whether indels were clustered within any of the non-coding domains (Fig. 3). The *Adh* region had significant clusters of indels within non-coding sequences that were indicated by significant monotonic increases in $G(x)$ and had regions devoid of segregating indels that were indicated by significant monotonic decreases in $G(x)$ (Fig. 3). Two non-coding segments, *Adh* adult intron and *Adh* UTR + Intergenic, show significant clustering of indel variation.

(iii) *Tests for departures from selective neutrality*

The frequency spectra of insertions and deletions are shown in Fig. 4*A*. The relative abundances of indels of different sizes are shown in Fig. 4*B*. Small indels were more frequent than large indels, with one-base insertions or deletions having the greatest abundance. In general, large indels had lower frequencies than small indels. The Tajima (1989) test shows that the frequency spectrum of insertions and deletions both reject a neutral model of molecular evolution because of an excess of rare variants (Table 3). I partitioned the

Table 1. *Insertion and deletion events in the* Adh *region of* D. pseudoobscura

Align site	*Dmir* site	Type	Bases	I/D sequence	Frequency	Class	Match
Adh adult intron							
130–131	130–131	D	2	GT	1	nr	0·500
132–133	131/132	I	2	GT	4	dr	1·000
143–144	140/141	I	2	CG	1	nr	0·000
150	145/146	I	1	A	1	nr	0·000
168	162/163	I	1	G	1	hp	1·000
170–177	163/164	A	8	CACTCGGT	28	dr	1·000
237–243	223–229	D	7	AAAAACA	1	nr	0·143
241	227	D	1	A	1	hp	1·000
266–279	251/252	D	16	AAACGTTCGAGAGTTG	2	nr	0·312
268	251/252	D	1	A	2	hp	1·000
301–319	273–289	D	17	AAAATATCAAATAAGAG	2	nr	0·588
301–307	273–279	D	7	AAAATAT	1	dr	1·000
314	286	D	1	A	4	hp	1·000
315	286/287	I	1	A	5	hp	1·000
316	286/287	I	1	A	3	hp	1·000
325–329	289/290	D	5	ATCTT	1	nr	0·200
336	292/293	I	1	A	10	hp	1·000
354	309/310	I	1	T	1	hp	1·000
374–375	329–330	D	2	AA	1	hp	1·000
468	422/423	A	1	A	10	hp	1·000
469–1573	422/423	I	1105	ACGAGG…GGTATA	1	te	ND
1608	457	D	1	G	1	nr	0·000
1646–1668	495–517	D	23	AGTAATGCCCTCGCTCTCTGTTA	1	nr	0·522
1678	526/527	I	1	T	1	nr	0·000
1782	629/630	I	1	A	1	hp	1·000
1830	677/678	D	1	A	22	mcs	1·000
1831	677/678	A	1	C	28	mcs	1·000
1832	677/678	I	1	A	9	mcs	1·000
1833	677/678	I	1	C	7	mcs	1·000
1834	677/678	A	1	G	8	nr	0·000
1892–1896	734/735	I	5	ACCGA	1	dr	0·800
Adh intron 1							
2164–2167	1002–1005	A	4	ACGA	12	nr	0·500
2168–2186	1006–1022	A	17	GAGAGAGAGCAATCCCT	19	nr	0·294
2174–2175	1012–1013	A	2	GA	12	mcs	1·000
2176–2177	1013/1014	I	2	GA	1	mcs	1·000
2209	1044/1045	I	1	A	10	hp	1·000
2219	1054	D	1	T	3	hp	1·000
2220	1054/1055	I	1	C	7	nr	0·000
2233	1067	D	1	T	2	hp	1·000
2237	1071/1072	I	1	T	1	hp	1·000
Adh intron 2							
2660	1493	D	1	G	1	nr	0·000
2671–2672	1503/1504	A	2	AA	31	hp	0·612
3′ UTR							
2977–2978	1806/1807	I	2	AT	9	nr	0·500
3030	1858	A	1	T	4	hp	1·000
3050–3061	1877/1878	I	12	ACATACATAAGA	1	nr	0·417
3095–3112	1897–1914	A	18	TTCTCTTTTATGGAATGAATGA	12	nr	0·500
Intergenic							
3180	1982	D	1	A	3	nr	0·000
3191–3194	1992/1993	I	4	TTTT	6	hp	1·000
3195	1992/1993	I	1	T	4	hp	1·000
3196	1992/1993	I	1	T	2	hp	1·000
3208	2004	D	1	A	5	hp	1·000
3209	2004/2005	I	1	A	4	hp	1·000
3231–3237	2023/2024	I	7	CAGTGGT	4	dr	0·857
3238–3240	2023/2024	A	3	CGT	24	nr	0·722
3241–3250	2024–2033	D	10	CAGTGGTGGT	5	dr	0·900
3251–3253	2033/2034	A	3	GGG	11	nr	0·758
3261	2040/2041	I	1	G	8	nr	0·000

Table 1. (*cont.*)

Align site	*Dmir* site	Type	Bases	I/D sequence	Frequency	Class	Match
Adhr intron 1							
3446–3456	2225–2232	A	8	TAGAGTGG	5	nr	0·750
3451–3453	2229/2230	A	3	AGG	69	nr	0·725
3457–3461	2233–2237	A	5	TGTAG	3	dr	0·600
3468–3474	2244–2250	D	7	GAGAGTG	5	dr	1·000
3475–3484	2251–2260	A	10	TTCCGAAGTG	49	nr	0·500
3475–3478	2251–2254	A	4	TTCC	1	nr	0·000
3516–3522	2292–2298	D	7	AGTCTCT	14	mcs	1·000
3523–3529	2298/2299	I	7	AGTCTCT	74	mcs	1·000
3530–3536	2298/2299	I	7	AGTCTCT	24	mcs	1·000
3537–3543	2298/2299	I	7	AGTCTCT	5	mcs	1·000
3544–3550	2298/2299	I	7	AGTCTCT	4	mcs	1·000
3551–3557	2298/2299	I	7	AGTCTCT	1	mcs	1·000
3564	2305	D	1	C	1	nr	1·000
3624–3636	2364–2376	D	13	CTGACTTTTGCTG	6	nr	0·385
3690	2430	D	2	G	1	hp	1·000
3694–3743	2434–2477	D	43	TACTTTCG…TAGATACCAAG	2	nr	0·295
3711–3715	2450/2451	I	5	TCGAA	7	dr	0·800
3732	2466/2467	I	1	C	1	nr	0·000
3757–3761	2491/2492	A	5	TCATA	32	dr	1·000
3784	2513	D	1	G	3	nr	0·000
3785	2513/2514	I	1	G	2	nr	1·000
3786	2514	D	1	C	4	nr	1·000
Adhr intron 2							
4230	2957/2958	I	1	C	1	hp	1·000
4239–4243	2966–2970	D	5	TTCCA	1	nr	0·600
3′ UTR and 3′ flanking							
4593–4606	3320–3333	D	14	CATGTCTTGATCCA	5	nr	0·071
4652	3378/3379	I	1	T	6	hp	1·000
4655–4662	3381–3388	D	8	GGGTCTGG	2	nr	0·500

Align site, location of insertion or deletion event in the alignment of 139 *D. pseudoobscura* sequences and the *D. miranda* sequence. *Dmir* site, location of insertion or deletion in the *D. miranda* sequence. The '/' indicates that an insertion occurred between the two base pairs in the *D. miranda* sequence. Type, type of indel mutation event (A, ambiguous; D, deletion; I, insertion). Bases, number of bases in the indel. I/D sequence, the consensus nucleotide sequence of the indel. Frequency, the frequency of the indel with *D. pseudoobscura*. Class, the classification of the indel mutation (dr, direct repeat; hp, homopolymer repeat; mcs, microsatellite repeat; nr, non-repetitive; te, transposable element). Match, maximum percentage similarity of the indel sequence to the 5′ and 3′ flanking sequence.

indels into large introns and small introns to determine whether the frequency spectrum departed from neutral expectations in the two classes of intervening sequence. Deletion and insertion variation in large introns showed a significant excess of rare variants while indel diversity in small introns did not (Table 3).

Fig. 5 shows the expected $D/D_{\min}$ ratio with 95% confidence intervals for the segregating site numbers observed in this study (Table 3) and for three values of *Nr*. Several observations are worth noting. First, the expected $D/D_{\min}$ ratio is homogeneous for different numbers of segregating sites within a given growth model, but the variances are largest for the smallest numbers of segregating sites. Second, as the exponential growth rate increases, the $D/D_{\min}$ ratio asymptotes to a value of -1. The observed $D/D_{\min}$ ratios from Table 3 are also shown in Fig. 5 for the three growth parameters. At $Nr = 1$, some observed values are outside the 95% confidence interval and reject a neutral model because of an excess of rare variants. At $Nr = 100\,000$, most observed values reject neutrality because of too many higher-frequency variants. The $D/D_{\min}$ ratio decreases slightly with increasing sample size if numbers of segregating sites are held constant, suggesting that comparisons of the $D/D_{\min}$ ratio among loci with different sample sizes may not be appropriate (data not shown). The statistical properties of the $D/D_{\min}$ ratio and its sensitivity to changes in sample sizes will be considered in future publications.

I estimated the value of *Nr* that best fits the $D/D_{\min}$ ratios for two regions that are expected to be neutral in the *Adh* region: non-coding nucleotides and the synonymous sites in *Adhr* (Schaeffer *et al.*, 2001) (Table 3). Synonymous sites in *Adh* were not used because Akashi & Schaeffer (1997) had shown that weak selection acts on synonymous codons of the *Adh* gene. When *Nr* is 7, the average value of $D/D_{\min}$ ratio is

Table 2. *Frequency of five classes of indels in the* Adh *region of* D. pseudoobscura

Indel type	Deletion	Insertion	Ambiguous
Direct repeat	3	4	3
Homopolymer	8	14	3
Microsatellite	2	8	2
Non-repetitive sequence	17	9	10
Transposable element	0	1	0
Repetitive versus non-repetitive			
Repetitive	13	27	8
Non-repetitive	17	9	10

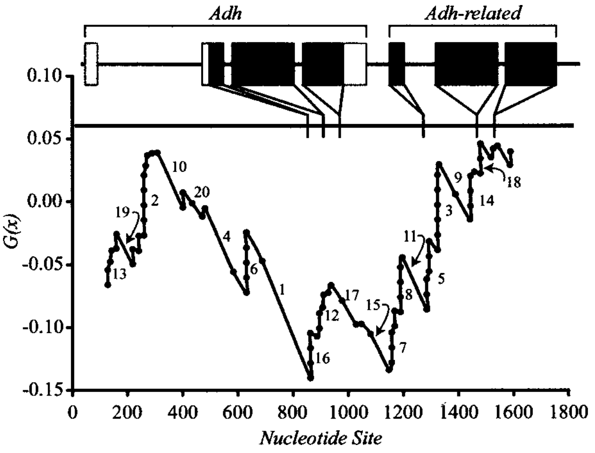

Fig. 3. Cumulative distribution function $G(x)$ of Tang & Lewontin (1999) for indel variation in the *Adh* region. The *Adh* region is shown above the $G(x)$ curves. The largest monotonic increases or decreases in $G(x)$ that are significant ($P<0{\cdot}05$) are indicated on the curves in rank order. Significant increases are indicated to the right of the curves while significant decreases are shown to the left of the curves.

$-0{\cdot}565$, which is the average $D/D_{\min}$ ratio for non-coding nucleotides and *Adhr* synonymous sites (Table 3). In addition, the average $D/D_{\min}$ ratio for X-linked genes is also consistent with a growth parameter of $Nr=7$ (data not shown; Kovacevic & Schaeffer, 2000). Given this exponential growth rate ($Nr=7$), I determined whether the $D/D_{\min}$ ratios of any of the observed types of variation were significantly different from the mean with a two-tailed test ($P<0{\cdot}05$). Deletions had a significant excess of rare variants when population expansion was taken into account, while the slight excess of rare variants in all other types of variation in the *Adh* region can be explained by population expansion. The negative Tajima's D in *Adh* synonymous and non-synonymous sites appears to be consistent with population expansion and not purifying selection as was previously reported (Schaeffer *et al.*, 2001).

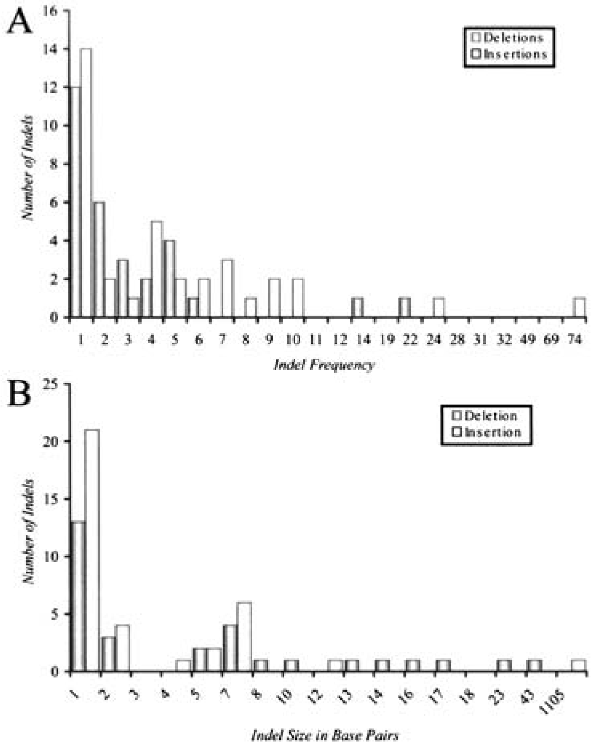

Fig. 4. (*A*) The frequency spectrum for insertions and deletions in the *Adh* region. (*B*) The frequencies of indels of different sizes.

When *Adh* synonymous codons are partitioned into preferred and unpreferred mutations, then unpreferred mutations have a significant excess of rare variants consistent with purifying selection while preferred mutations do not reject neutral evolution with population growth (data not shown).

(iv) *Linkage disequilibrium among indel polymorphisms*

A total of 3486 pairwise comparisons are possible for the 84 polymorphic indels in the *Adh* region. Of those, 632 comparisons were capable of rejecting the null hypothesis of no association among indels with Fisher's exact test (Lewontin, 1995). Twenty-nine indels showed non-random association with at least one other sequence length polymorphism. The indel sites were a near-equal mix of insertions (15) and deletions (14). Fig. 6 shows the 16 pairwise comparisons of indels that were in significant pairwise linkage disequilibrium in the *Adh* region. The strength of association among the pairs of indels was assessed with r^2, which varied from $0{\cdot}032$ to $0{\cdot}199$. Several observations are worth noting. First, only $2{\cdot}5$% of the tests showed significant non-random associations among indels. This is a higher level of association than observed for nucleotide substitutions (Schaeffer & Miller, 1993). Second, short distances separated pairs of indels

Table 3. *Estimates of Tajima's* D *for insertions, deletions and nucleotide variation in the* Adh *region of* D. pseudoobscura

Type of variation	S	S/a	k	D	$D_{\min}$	$D/D_{\min}$
Deletions						
All	30	5·4	1·4	−2·18*	2·70	−0·81*
Large introns	21	3·8	1·0	−2·06*	2·58	−0·80
Small introns	3	0·5	0·08	−1·43	1·56	−0·92
Adh introns	16	2·9	0·61	−2·12*	2·47	−0·86*
Adhr introns	9	1·6	0·51	−1·64	2·19	−0·75
Insertions						
All	36	6·5	2·6	−1·81*	2·76	−0·65
Large introns	22	4·0	1·7	−1·62*	2·60	−0·62
Small introns	5	0·9	0·3	−1·41	1·86	−0·76
Adh introns	18	3·3	0·9	−1·99*	2·52	−0·79
Adhr introns	9	1·6	1·1	−0·37	2·19	−0·37
Nucleotides						
Adh synonymous	55	10·0	3·2	−2·11*	2·85	−0·74
Adh non-synonymous	3	0·5	0·04	−1·56*	1·56	−1·00
Adhr synonymous	79	14·3	7·3	−1·54	2·91	−0·53
Adhr non-synonymous	13	2·4	1·5	−0·94	2·38	−0·39
Non-coding nucleotides	239	43·4	19·2	−1·82*	3·01	−0·60

S, number of segregating sites; S/a, nucleotide heterozygosity based on the number of segregating sites (Watterson, 1975); k, average number of pairwise differences; D, Tajima's (1989) test statistic; $D_{\min}$, the absolute value of the minimum possible estimate of D. Significance for $D/D_{\min}$ was determined for a growing population with an exponential growth rate of $Nr = 7$ (Slatkin & Hudson, 1991).
* $P < 0·05$.

that were non-randomly associated. There were only two cases where non-randomly associated indels were separated by longer distances (990 and 1366 bp).

(v) *Large insertion variation*

A large insertion of 1105 bp was observed in a single strain of *D. pseudoobscura* (PS271) that was collected from the Kaibab National Forest in Arizona. The insertion occurs in the adult intron of *Adh*. A BLAST (Altschul *et al.*, 1997) search of GenBank (Benson *et al.*, 2002) was used to determine what the identity of this large insertion might be. This insertion sequence was similar to an abundant class of mobile elements called *mini-me* elements (Wilder & Hollocher, 2001). The *D. pseudoobscura* element has the canonical features of the *mini-me* elements including a 5′ (TA) repeat, 33 bp core sequence and a 3′ (GTCY) repeat (Fig. 7). The *D. pseudoobscura mini-me* element had 861 bp of additional sequence that contained two sets of direct repeats. One repeat was present in three copies (1A, 1B and 1C), while the second repeat was found in two copies (2A and 2B) (Fig. 6). Phylogenetic analysis of the tandem repeats suggests that the two repeats 1A and 2A gave rise to repeats 1B and 2B through a duplication event, then a second duplication of repeat 1B gave rise to repeat 1C (Fig. 7).

The previously described *mini-me* elements did not encode for reverse transcriptase or transposase (Wilder & Hollocher, 2001); however, the size of extra DNA associated with the canonical elements varies substantially, suggesting that some elements might be incomplete. The *mini-me* from the *D. pseudoobscura Adh* region is one of the larger elements in GenBank, suggesting that proteins for transposition might be detected in this element. An analysis of open reading frames finds little evidence for an encoded protein in the *D. pseudoobscura mini-me* element. The average open reading frame size is 23·2 amino acids for the six reading frames of this new *mini-me* element. The largest open reading frame is 87 amino acids. This peptide was compared with the GenBank protein database, where four putative matches were examined; however, none of the matches was similar to a protein associated with transposable elements or with a consistent protein function. Thus, no protein function can yet be attributed to the *mini-me* element.

4. Discussion

(i) *Indel variation and selection*

The analyses presented here used intra- and interspecific nucleotide data in the *Adh* locus to determine whether selection operates on indel variation in introns and non-coding sequences. The number of indels is one-fifth that of segregating sites (84 indels versus 418 segregating sites) suggesting either that the indel

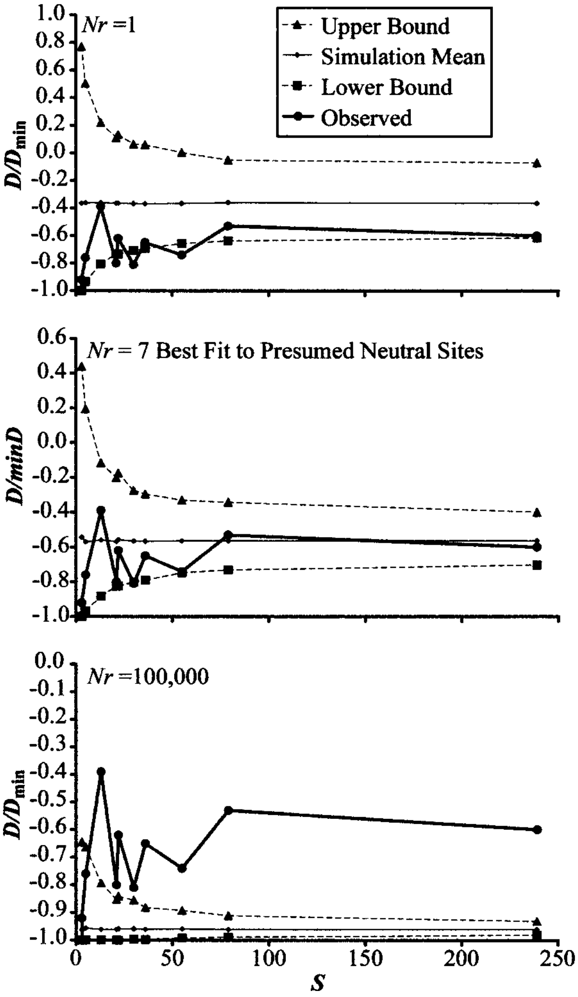

Fig. 5. Ratio of Tajima's (1989) D to its theoretical minimum $D_{\min}$ versus the number of segregating sites plotted for three exponential growth rates (Nr), where N is the initial population size and r is the growth rate (Slatkin & Hudson, 1991). The Best Fit curve is for the Nr value that fits presumed neutral sites, non-coding and synonymous sites in *Adhr*.

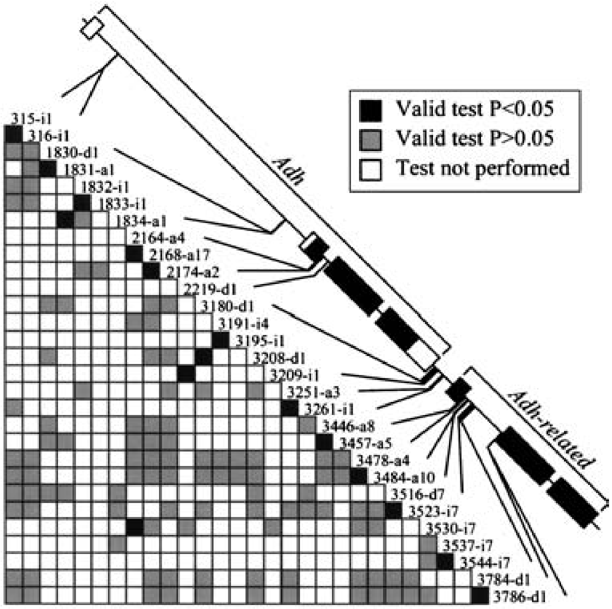

Fig. 6. Plot of significant linkage disequilibrium among pairs of indels. Each box in the diagram indicates a comparison of two indel sites. The indel numbers shown on the diagonal indicate the nucleotide position of the indel, the type of indel (a, ambiguous indel; i, insertion; d, deletion) and the number of bases in the indel sequence. The locations of the indels are shown within the *Adh* region. Only indels that showed significant linkage disequilibrium with at least one other indel are shown on the diagram.

mutation rate is less than the nucleotide mutation rate or that indel events have higher selective constraints than nucleotide changes. When a constant population size model is assumed, Tajima's (1989) test shows that deletions and insertions each had a significant excess of rare variants. Insertions and deletions in large introns also had a significant excess of rare variants, while the two types of indel failed to reject a neutral model for variation in small introns. Neither type of indel had any fixed differences between *D. pseudoobscura* and its two sister species.

When population expansion was taken into account, only deletion variation had significant excesses of rare variants, while insertions did not. These data suggest that purifying selection removes deletion variation before it is fixed in populations, but that insertions are capable of fixing through genetic drift. The net effect of this process would be to increase the size of introns over time. These results are consistent with those of Comeron & Kreitman (2000) who showed that insertions had elevated frequencies in regions of high recombination while deletions did not.

The frequency spectra tests of neutrality for indels in the small introns should be viewed with caution because the number of segregating polymorphisms is small for both classes of mutation. The Tajima test has less power to detect selective sweeps, population bottlenecks, population subdivision and purifying selection as the number of segregating sites declines (Simonsen *et al.*, 1995; Akashi, 1999). This can also be seen with the $D/D_{\min}$ ratio, where the variance of $D/D_{\min}$ increases as the number of segregating sites decreases (Fig. 5). This study was not able to adequately test whether selection acts on indel variation in small introns. It is possible that selection acts on variation in small introns, but the potential number of indel polymorphisms that may be segregating within a small intron at any time may be small. As a result, it is unlikely that a frequency spectrum test will detect purifying selection in small introns; however, using methods that combine probabilities over multiple small introns may overcome this problem (Sokal & Rohlf, 1981).

The introns in the *Adh* region of *D. pseudoobscura* are found in a region of the genome where

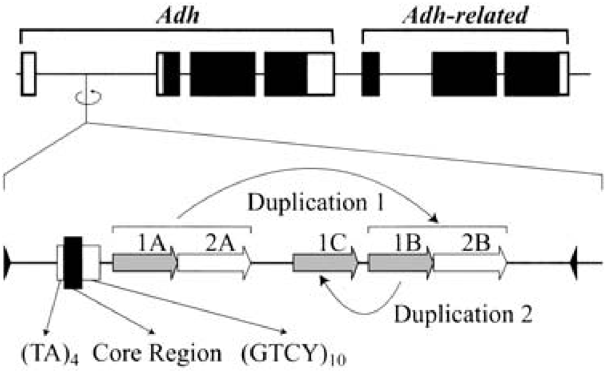

Fig. 7. Structure of the *mini-me* insertion in the *Adh* gene. The top part of the diagram shows the *Adh* region with the location of the *mini-me* insertion in the adult intron. The structure of the *mini-me* element is shown in the lower part of the figure. The main features of the element are: indirect repeats (black triangles) at either end of the element (5′ sequence: TTATACCCGATACT; 3′ sequence: AGTATCGGGTATAA), a core region of 33 bp (black box) that is conserved among dipterans, and two microsatellite progenitor sequences (open boxes). Two sequences of direct repeats indicated by the grey and white arrows were also observed in this element. This pattern of tandem repeats is similar to the direct repeat structure of the *D. miranda* TRIM element (Steinemann & Steinemann, 1993) as noted by Hagemann *et al.* (1998).

recombination is extensive (Schaeffer & Miller, 1993). Carvalho & Clark (1999) predicted that strong selection is maintaining the size of small introns in regions of high recombination by selecting against indels. The data presented here suggest that deletion polymorphisms are selectively removed from large introns leading to a net gain in intron size, but provide no insights into small intron evolution. The implication is that large introns will decrease in size in regions of low recombination due to relaxed selection due to the Hill–Robertson effect (Hill & Robertson, 1966). Population genetics of indels in regions with low genetic exchange should be examined in light of the present study to determine whether selection is relaxed on deletion variation.

A second factor that may influence the evolution of intron length is gene expression. Castillo-Davis *et al.* (2002) found that highly expressed genes tend to have smaller introns than genes with low expression. This may result from selection to enhance transcription speed. *Adh* is expressed at a higher level than *Adhr* (Brogna & Ashburner, 1997), yet we do not have any evidence that purifying selection acts against insertion polymorphism. The data presented here find the opposite result, i.e. deletions, which could decrease the size of introns, are selected against. This suggests that there is a lower limit to intron size even in genes that are highly expressed.

One question that emerges is why the genome does not continue to expand without an obvious mechanism to limit the size of introns. The answer to this question will depend on what forces act on intron sequences in regions of reduced recombination. If selection is less effective in regions of reduced recombination, then selection against deletions would be reduced in regions that experience low levels of genetic exchange (Hill & Robertson, 1966). This would lead to smaller introns in regions of low recombination, a contradiction with the observed data (Carvalho & Clark, 1999; Comeron & Kreitman, 2000). Alternatively, Comeron & Kreitman (2000) have suggested that insertions are positively selected in regions of low recombination, which would enhance the increase in intron size over what genes experience in regions of high recombination. Studies of the linkage maps of closely related species of *Drosophila* have shown that the linear order of genes in the genome is not fixed (Ranz *et al.*, 1997, 2001; Segarra & Aguade, 1992), so that the local recombinational environment that is experienced by a gene is not constant through evolutionary time. Thus, introns within genes located in a region of low recombination may appear to be subject to runaway selection for ever-increasing intron size, but movement of that same gene to a region of increased genetic exchange may relax the selection for intron expansion.

Selection on indels may be size-dependent. As introns increase in size through the fixation of insertions, selection against deletions may relax leading to curtailed growth of the intron. On the other hand, insertions may be favoured when introns become too small.

Pseudogenes and retroposons have been shown to have a strong bias for deletion events (Graur *et al.*, 1989; Saitou & Ueda, 1994; Ogata *et al.*, 1996; Petrov *et al.*, 1996; Petrov & Hartl, 1998). This study found that insertions were more frequent than deletions, suggesting that the mechanisms that act on indels differ for homologous and paralogous events (Charlesworth, 1996; Comeron & Kreitman, 2000). Paralogous events may favour the loss of DNA because loss of redundant information has a lower fitness cost than the gain of DNA. Homologous events, on the other hand, may favour the gain of information because the loss of unique information has more serious fitness costs than the slight gain of sequence.

(ii) *The nature of indel events*

One of the trends that emerged from this study is that insertion sequences tend to be more similar to flanking sequences than are deletions. This suggests that the mechanism for generating insertions may be slightly different from processes that generate deletions. These data are consistent with slip strand repair during DNA replication as a mechanism for generating insertions, because flanking sequences can act as templates for the expansion of nucleotide sequence (Moore, 1983). On the other hand, the lower similarity of deleted sequences to adjacent DNA suggests that deletions may occur independent of DNA replication, although our analysis does not rule out the possibility that sequences away from the immediate vicinity of the deletion play a role in the mutational event.

(iii) *Distribution of indel variation*

The significant clustering of indels in the *Adh* region within non-coding domains may reflect the process that leads to unalignable sequences among species. Chiaromonte *et al*. (2001) have shown that regions that are difficult to align in interspecies comparisons tend to have higher numbers of repeat sequences. The suggestion is that regions with large numbers of repeats are difficult to align because repetitive sequences facilitate the accumulation of point and indel mutations. The 5′ end of the adult intron has accumulated a large number of nucleotide and indel mutations (Fig. 3) (Schaeffer & Miller, 1992*b*), although there is no evidence for a high density of repetitive elements in this region. The data presented here suggest that regions of low functional constraint may accumulate both nucleotide and indel variation causing regions to diverge rapidly among species so that homologous sequences are unalignable.

I would like to thank the following people who provided excellent technical support and without whom this work would not have been possible: Ellen L. Miller, C. Scott Walthour, Donna M. Toleno and Anna T. Olek. I would also like to thank Andrew Clark, Hiroshi Akashi and two anonymous reviewers for thoughtful comments on the analyses and manuscript. A grant from the National Institutes of Health GM-42472 was used to generate the sequence data for this paper.

References

Adams, M. D. and 195 others (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.

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

Akashi, H. & Schaeffer, S. W. (1997). Natural selection and the frequency distributions of 'silent' polymorphism in *Drosophila*. *Genetics* **146**, 295–307.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. & Wheeler, D. L. (2002). GenBank. *Nucleic Acids Research* **30**, 17–20.

Brogna, S. & Ashburner, M. (1997). The Adh-related gene of *Drosophila melanogaster* is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms. *European Molecular Biology Organization Journal* **16**, 2023–2031.

Cabot, E. L. & Beckenbach, A. T. (1989). Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. *Computer Applications in the Biosciences* **5**, 233–234.

Carvalho, A. B. & Clark, A. G. (1999). Intron size and natural selection. *Nature* **401**, 344.

Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V. & Kondrashov, F. A. (2002). Selection for short introns in highly expressed genes. *Nature Genetics* **31**, 415–418.

Charlesworth, B. (1996). The changing sizes of genes. *Nature* **384**, 315–316.

Charlesworth, B., Sniegowski, P. & Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**, 215–220.

Chiaromonte, F., Yang, S., Elnitski, L., Yap, V. B., Miller, W. & Hardison, R. C. (2001). Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. *Proceedings of the National Academy of Sciences of the USA* **98**, 14503–14508.

Comeron, J. M. & Kreitman, M. (2000). The correlation between intron length and recombination in *Drosophila*. Dynamic equilibrium between mutational and selective forces. *Genetics* **156**, 1175–1190.

Comeron, J. M. & Kreitman, M. (2002). Population, evolutionary and genomic consequences of interference selection. *Genetics* **161**, 389–410.

Deutsch, M. & Long, M. (1999). Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Research* **27**, 3219–3228.

Fu, Y.-X. (1996). New statistical tests of neutrality for DNA samples from a population. *Genetics* **143**, 557–570.

Graur, D., Shuali, Y. & Li, W.-H. (1989). Deletions in processed pseudogenes accumulate faster in rodents than in humans. *Journal of Molecular Evolution* **28**, 279–285.

Hagemann, S., Miller, W. J., Haring, E. & Pinsker, W. (1998). Nested insertions of short mobile sequences in *Drosophila P* elements. *Chromosoma* **107**, 6–16.

Hahn, M. W., Rausher, M. D. & Cunningham, C. W. (2002). Distinguishing between selection and population expansion in an experimental lineage of bacteriophage T7. *Genetics* **161**, 11–20.

Hamblin, M. T. & Aquadro, C. F. (1999). DNA sequence variation and the recombinational landscape in *Drosophila pseudoobscura*: a study of the second chromosome. *Genetics* **153**, 859–869.

Hill, W. G. & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–294.

Hill, W. G. & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226–231.

Kovacevic, M. & Schaeffer, S. W. (2000). Molecular population genetics of X-linked genes in *Drosophila pseudoobscura*. *Genetics* **156**, 155–172.

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

Machado, C. A., Kliman, R. M., Markert, J. A. & Hey, J. (2002). Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Molecular Biology and Evolution* **19**, 472–488.

Moore, G. P. (1983). Slipped-mispairing and the evolution of introns. *Trends in Biochemical Sciences* **8**, 411–414.

Mount, S. M., Burks, C., Hertz, G., Stormo, G. D., White, O. & Fields, C. (1992). Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Research* **20**, 4255–4562.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Ogata, H., Fujibuchi, W. & Kanehisa, M. (1996). The size differences among mammalian introns are due to the accumulation of small deletions. *Federation of European Biochemical Societies Letters* **390**, 99–103.

Ophir, R. & Graur, D. (1997). Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* **205**, 191–202.

Petrov, D. A. & Hartl, D. L. (1998). High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Molecular Biology and Evolution* **15**, 293–302.

Petrov, D. A., Lozovskaya, E. R. & Hartl, D. L. (1996). High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**, 346–349.

Ranz, J. M., Segarra, C. & Ruiz, A. (1997). Chromosomal homology and molecular organization of Muller's elements D and E in the *Drosophila repleta* species group. *Genetics* **145**, 281–295.

Ranz, J. M., Casals, F. & Ruiz, A. (2001). How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Research* **11**, 230–239.

Rice, W. R. (1989). Analyzing tables of statistical tests. *Evolution* **43**, 223–225.

Russo, C. A., Takezaki, N. & Nei, M. (1995). Molecular phylogeny and divergence times of drosophilid species. *Molecular Biology and Evolution* **12**, 391–404.

Saitou, N. & Ueda, S. (1994). Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. *Molecular Biology and Evolution* **11**, 504–512.

Schaeffer, S. W. & Aquadro, C. F. (1987). Nucleotide sequence of the *Adh* gene region of *Drosophila pseudoobscura*: evolutionary change and evidence for an ancient gene duplication. *Genetics* **117**, 61–73.

Schaeffer, S. W. & Miller, E. L. (1991). Nucleotide sequence analysis of *Adh* genes estimates the time of geographic isolation of the Bogota population of *Drosophila pseudoobscura*. *Proceedings of the National Academy of Sciences of the USA* **88**, 6097–6101.

Schaeffer, S. W. & Miller, E. L. (1992*a*). Estimates of gene flow in *Drosophila pseudoobscura* determined from nucleotide sequence analysis of the alcohol dehydrogenase region. *Genetics* **132**, 471–480.

Schaeffer, S. W. & Miller, E. L. (1992*b*). Molecular population genetics of an electrophoretically monomorphic protein in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. *Genetics* **132**, 163–178.

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

Schaeffer, S. W., Walthour, C. S., Toleno, D. M., Olek, A. T. & Miller, E. L. (2001). Protein variation in ADH and ADH-RELATED in *Drosophila pseudoobscura*: linkage disequilibrium between single nucleotide polymorphisms and protein alleles. *Genetics* **159**, 673–687.

Segarra, C. & Aguade, M. (1992). Molecular organization of the X chromosome in different species of the *obscura* group of *Drosophila*. *Genetics* **130**, 513–521.

Simonsen, K. L., Churchill, G. A. & Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429.

Slatkin, M. & Hudson, R. R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562.

Sokal, R. R. & Rohlf, F. J. (1981). *Biometry*. New York: W. H. Freeman.

Steinemann, M. & Steinemann, S. (1993). A duplication including the Y allele of Lcp2 and the TRIM retrotransposon at the *Lcp* locus on the degenerating neo-Y chromosome of *Drosophila miranda*: molecular structure and mechanisms by which it may have arisen. *Genetics* **134**, 497–505.

Stephens, J. C. (1985). Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Molecular Biology and Evolution* **2**, 539–556.

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

Tang, H. & Lewontin, R. C. (1999). Locating regions of differential variability in DNA and protein sequences. *Genetics* **153**, 485–495.

Wang, R. L. & Hey, J. (1996). The speciation history of *Drosophila pseudoobscura* and close relatives: inferences from DNA sequence variation at the period locus. *Genetics* **144**, 1113–1126.

Wang, R. L., Wakeley, J. & Hey, J. (1997). Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. *Genetics* **147**, 1091–1106.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276.

Wilder, J. & Hollocher, H. (2001). Mobile elements and the genesis of microsatellites in Dipterans. *Molecular Biology and Evolution* **18**, 384–392.