

# Characterization of the boundaries between adjacent rapidly and slowly evolving genomic regions in *Drosophila*

(glue protein genes/genome structure/evolution)

CHRISTOPHER H. MARTIN AND ELLIOT M. MEYEROWITZ

Division of Biology, California Institute of Technology, Pasadena, CA 91125

Communicated by Roy J. Britten, August 4, 1986

**ABSTRACT** The site of a dramatic change in the rate of DNA sequence evolution exists near the 68C glue gene clusters of several *Drosophila* species. We have previously determined the approximate location of this transition site by comparison of restriction maps of the regions flanking the 68C-like glue gene cluster of five members of the *melanogaster* species subgroup. In the present work we report the sequence of the transition region in three of these *Drosophila* species: *D. melanogaster*, *D. yakuba*, and *D. erecta*. Using a best-fit alignment of these sequences, we find that the site of transition from slowly to rapidly evolving sequences occurs abruptly within a region <50 nucleotides in length. Although frequency of nucleotide substitutions changes as much as 10-fold across this boundary, frequency of small insertion/deletion events stays nearly constant.

The 68C puff of *Drosophila melanogaster* contains three genes that code for components of a protein glue that affixes the puparial case to a solid substrate (1–3). These genes are expressed abundantly in the salivary glands of third instar larvae and are controlled by the steroid hormone ecdysterone (4–7). The homologous gene clusters from five closely related species of *Drosophila*—*D. melanogaster*, *D. simulans*, *D. erecta*, *D. yakuba*, and *D. teissieri*—have been cloned. These species are all members of the *melanogaster* species subgroup, which is one of eleven species subgroups defined for the *melanogaster* species group (24). Comparison of the restriction maps of these cloned sequences revealed two adjoining regions with dramatically different levels of homology (8). That this genomic segment contains adjacent regions that have evolved at different rates is confirmed by experiments that demonstrate very different melting temperature depressions ( $\Delta t_m$ s) of inter-species hybrids of restriction fragments from each of the two adjoining regions (8). The relatively nonconserved region, which is  $\approx 6$  kb (kilobase pairs) in length, contains the glue gene cluster and appears to be evolving by a number of mechanisms: point mutations, insertions and deletions, inversions, duplications, and the gain or loss of repetitive sequences (8). In contrast, the conserved region, which consists of  $\geq 10$  kb of single-copy sequence, is not known to contain any genes and evolves through relatively infrequent point mutations and small insertions and deletions. To learn more about the boundary between the two regions and about any possible functions of the conserved DNA, we determined the DNA sequences of the regions from three members of the *D. melanogaster* species subgroup.

## MATERIALS AND METHODS

**Materials.** Restriction endonucleases were obtained from Boehringer Mannheim and New England Biolabs. The large

proteolytic fragment of *Escherichia coli* DNA polymerase I was obtained from Boehringer Mannheim. T4 DNA polymerase was obtained from New England Nuclear. T4 DNA ligase was a gift from S. Scherer.  $^{32}\text{P}$ -labeled nucleoside triphosphates were obtained from Amersham. Deoxynucleotides and dideoxynucleotides were obtained from Pharmacia.

**Clones for DNA Sequencing.** Clones for sequencing were prepared by inserting fragments from previously cloned *Drosophila* sequences into vectors M13mp18 and M13mp19 (9); M13 vectors were a gift of G. Siu. The *D. melanogaster* clones were constructed by inserting the 1.95-kb *EcoRI*–*HindIII* restriction fragment from clone aDm2003 (8) into vectors M13mp18 and M13mp19. For *D. erecta*, the 2.25-kb *EcoRI*–*BamHI* restriction fragment from clone fDe009 (8) was inserted into both M13 vectors. For *D. yakuba*, the 2.9-kb *EcoRI* restriction fragment from clone qDy5110 (8) was cloned in both orientations into M13mp18. Cloning was done by standard procedures described by Davis *et al.* (10) and Maniatis *et al.* (11).

**Sequencing.** DNA sequencing was performed by the dideoxy chain-termination method of Sanger *et al.* (12). Custom oligonucleotides, used to prime sequencing reactions from sites in the interior of a cloned insert, were provided by S. Horvath of the California Institute of Technology Microchemical Facility. These primers were purified and used as described in Strauss *et al.* (13). All sequences were determined on both strands.

**Computer Analysis.** DNA sequences were analyzed using programs written by one of the authors (C.H.M.) for an IBM PC-XT computer. DNA sequences were aligned using the algorithm of Gotoh (14) as implemented by R. Pruitt on an Apple Macintosh computer.

## RESULTS

The border between regions of high and low conservation was located by inspection of the restriction maps of regions containing and adjacent to the cloned glue gene clusters. The broken vertical line in Fig. 1 demarcates the transition from conserved to nonconserved restriction site pattern. This apparent change in relative levels of sequence conservation occurs over a distance of <1 kb. To characterize this transition, we obtained the DNA sequences of this region and compared them for three species: *D. melanogaster*, *D. yakuba*, and *D. erecta*. The phylogenetic relationship between these species has been determined by Lemeunier and Ashburner (15) by comparing differences in chromosomal banding patterns. *D. yakuba* and *D. erecta* seem to be more closely related to each other than either is to *D. melanogaster*.

The sequencing strategy is diagrammed in Fig. 2. All sequences start at the *EcoRI* site (R) that lies at least 1000

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: kb, kilobase pair(s).

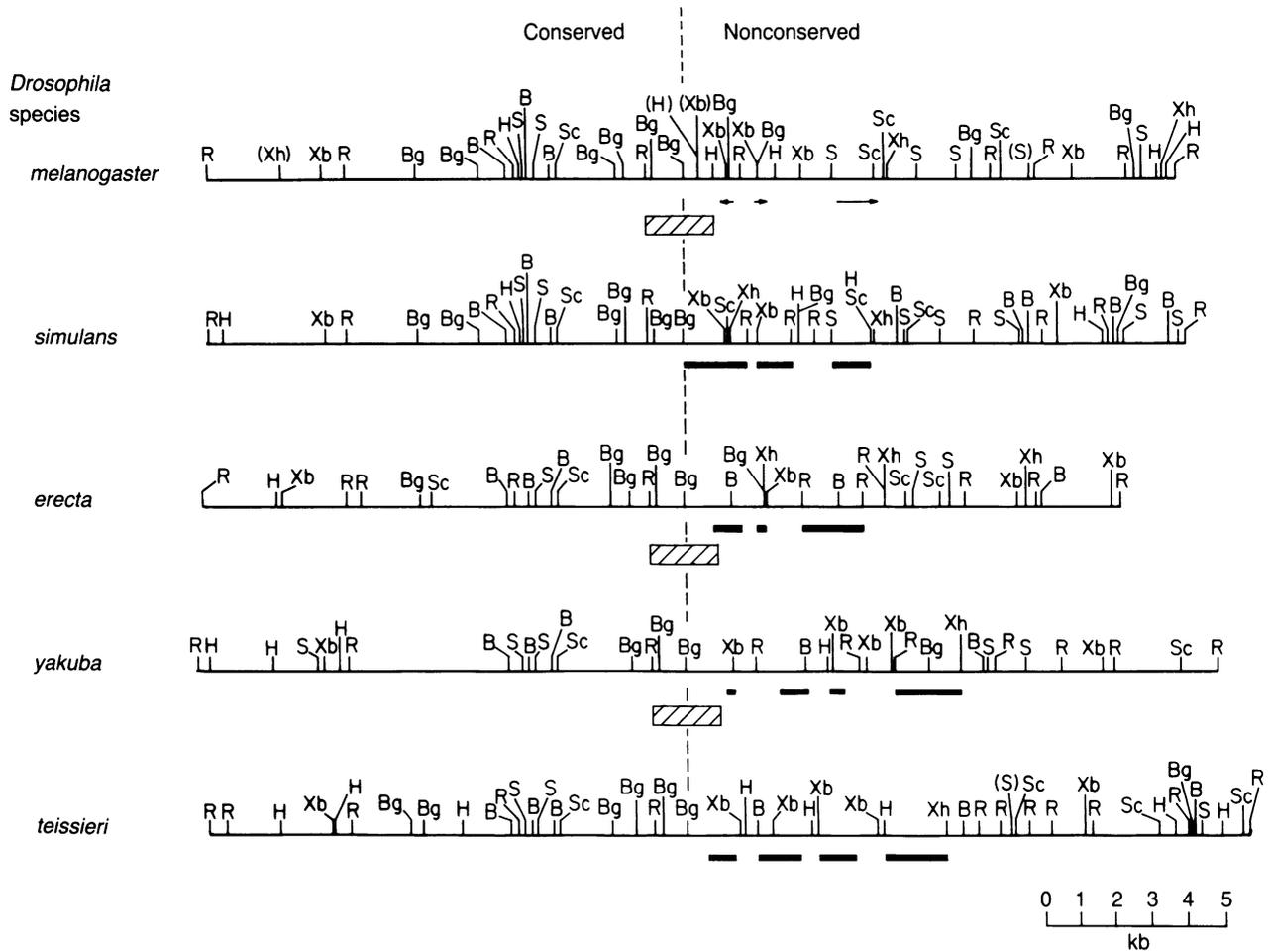


FIG. 1. Restriction maps of the cloned 68C-homologous sequences. All known *Bam*HI (B), *Bgl* II (Bg), *Eco*RI (R), *Hind*III (H), *Sal* I (S), *Sac* I (Sc), *Xba* I (Xb), and *Xho* I (Xh) sites are depicted (except for a single *Eco*RI site in *D. erecta*; see ref. 8). Sites in parentheses are present in some strains (*D. melanogaster*) or clones (*D. teissieri*) and not in others (8). The arrows below the *D. melanogaster* map show the location and direction of transcription of the glue gene transcription units. For the other species, solid bars show the extent of restriction fragments that are hybridized by cDNA made from salivary gland poly(A)<sup>+</sup> RNA (8). The maps are aligned by the positions of the conserved restriction sites found left of the RNA-coding regions. The vertical line shows the boundary between the conserved region at the left and the nonconserved region at the right. Hatched boxes show those regions sequenced.

bases inside the conserved region, and each sequence continues at least 1800 bases toward and into the nonconserved region.

The aligned nucleotide sequences are shown in Fig. 3. Inspection of the alignment reveals a dramatic change in the frequency of nucleotide substitutions that occurs near base 1354 of the *D. melanogaster* sequence. Substitution rates appear to change abruptly: there is no evidence for a region of intermediate divergence between the conserved and nonconserved regions. This site of rapid change can be used to divide the sequenced regions into conserved and nonconserved sections, a useful device in comparing the types and amounts of change that are occurring on each side of the site.

A summary of changes occurring in the two sections is shown in Table 1. Two methods have been used to calculate divergence values (see legend for Table 1). The first method counts only those events in which bases are substituted and ignores any base that is deleted from the other member of the species pair. A dramatic change in the frequency of point mutation occurs across the boundary in all pair-wise comparisons of the three species. Another method of calculation used in Table 1 additionally counts each group of contiguous deleted bases as a single mismatch. While the number of point mutations varies sharply on either side of the boundary, the frequency of small insertion/deletion events is relatively

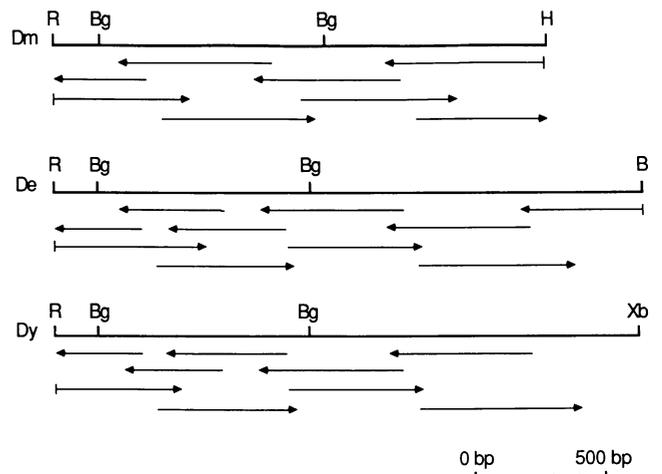


FIG. 2. Sequencing strategy. The arrows show the extent of individual sequencing reactions. All reactions were initiated from synthetic oligonucleotide primers. A short vertical bar at the end of a line indicates that the primer used is complementary to sequences in M13; all other primers are complementary to sequences within the cloned insert. Restriction enzyme symbols are the same as for Fig. 1. *D. melanogaster*, Dm; *D. erecta*, De; *D. yakuba*, Dy.

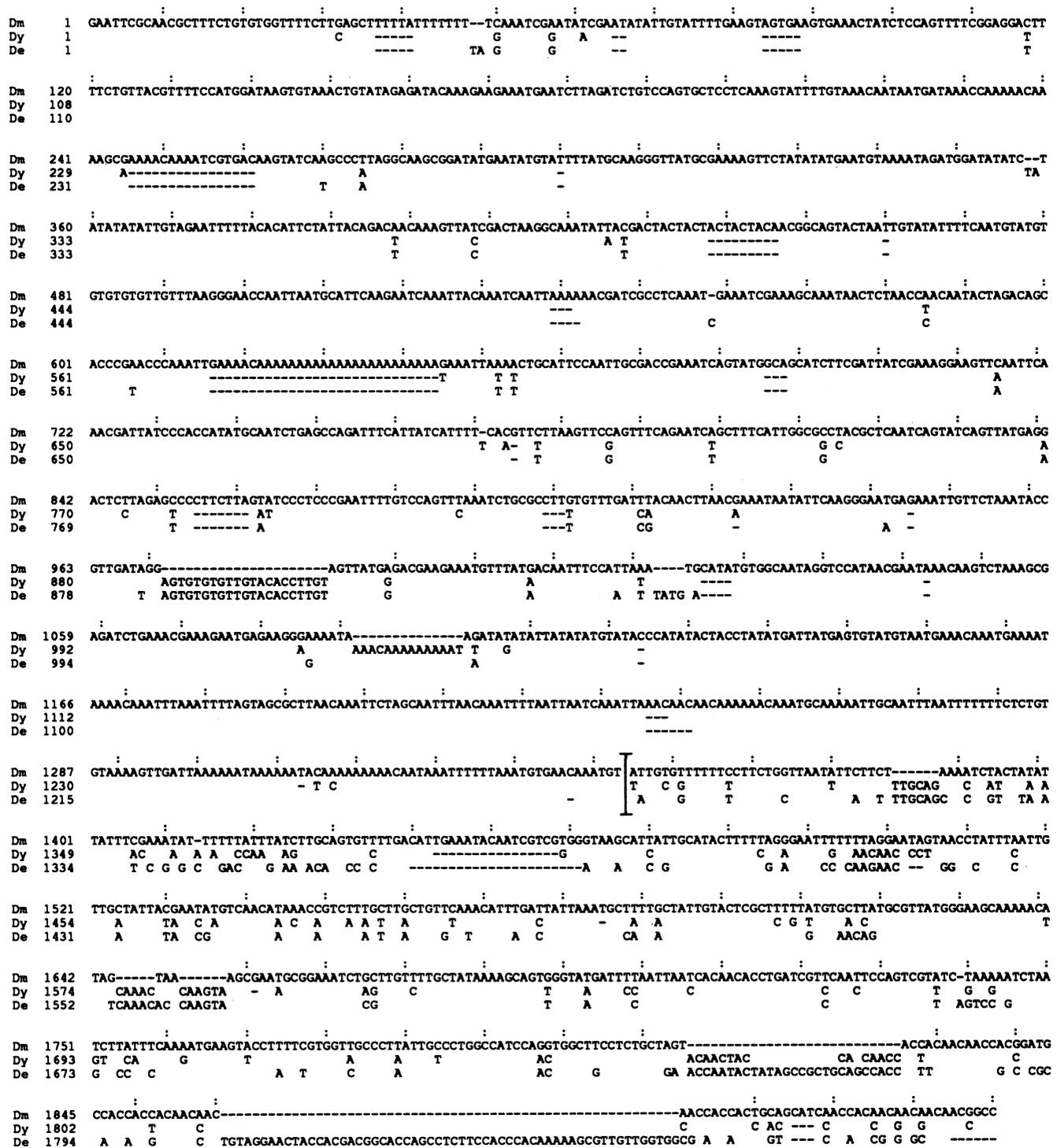


FIG. 3. The aligned DNA sequences of the border region of *D. melanogaster* (Dm), *D. yakuba* (Dy), and *D. erecta* (De). All pair-wise alignments were generated by the algorithm of Gotoh (14). The mismatch penalty was 10, the start deletion penalty was 40, and the deleted base penalty was 5. The three-way alignment was generated by hand from the pair-wise alignments. Colons mark every tenth base in the *D. melanogaster* sequence. Spaces indicate that the sequence is identical to that of *D. melanogaster*. A dash represents a deleted base. The dark vertical bar following base 1353 of *D. melanogaster* locates the boundary between the two regions that evolve at different rates.

constant. This is apparent from the similar frequencies of deletions observed on both sides of the boundary.

Furthermore, near the boundary the  $\approx 200$  bases just preceding the start of the nonconserved region (bases 1154 through 1353 in the *D. melanogaster* sequence) are very rich in A+T. This sequence shows an average of  $83.4 \pm 0.5\%$  A+T (all values are  $\pm$  SD) vs. an average of  $67.5 \pm 0.2\%$  A+T in the remaining 1153 bases of the conserved region, and an average of  $61.0 \pm 3.3\%$  A+T in the nonconserved region. The

value of 83.4% is far above the average of 55% A+T found in total DNA from *D. melanogaster* (16); this A+T-rich region tends to contain stretches of adenines and thymines as opposed to interspersed adenines and thymines. In the three species, A+A and T+T dinucleotides make up  $50.4 \pm 0.8\%$  of this region, whereas A+T and T+A dinucleotides comprise only  $19.4 \pm 0.4\%$  of the region. Also, the A+T-rich region has even fewer point mutations than the rest of the conserved region (the average point mutation frequency in

Table 1. Changes in the conserved versus the nonconserved regions

Sequences compared	% mismatch		% deletions + mismatches		% deletions	
	Conserved	Nonconserved	Conserved	Nonconserved	Conserved	Nonconserved
<i>melanogaster</i> vs. <i>yakuba</i>	3.18	18.25	4.76	18.99	1.74	1.63
<i>yakuba</i> vs. <i>erecta</i>	1.89	19.63	2.72	19.71	0.85	1.53
<i>erecta</i> vs. <i>melanogaster</i>	2.56	23.93	4.25	23.00	1.82	1.89

Three types of calculation were used to describe the differing types of change. The % mismatch calculation, which shows the frequency of nucleotide substitution only, is calculated as the number of mismatched bases divided by the total number of bases that are aligned to another base in the other sequence (matched or mismatched). Any bases that are deleted in either sequence are not counted in this calculation. The % deletion + mismatch calculation is a more general measure of divergence that also takes into account insertions and deletions. This second calculation (% deletions + mismatches) is calculated as the sum of the number of mismatches and the number of contiguous blocks of deleted bases divided by one-half of the sum of the total number of bases in both of the compared sequences. The % deletions calculation shows the relatively constant rate of insertion/deletion events in both regions. The number of deletion events per 100 bases (% deletions) is calculated as the number of contiguous blocks of deleted bases divided by one-half the total number of bases that are in both sequences. The conserved region included bases 1 through 1353 in the *D. melanogaster* sequence.

this region is only  $0.7 \pm 0.6\%$  vs.  $2.9 \pm 0.8\%$  in the remaining region).

There is no evidence that the conserved region, despite its evolutionary conservation, codes for a protein. The frequency of transitions is consistently less than the frequency of transversions in both regions, with an average ratio of  $0.71 \pm 0.03$ . This is comparable to the ratio of 0.75 seen in noncoding regions of alcohol dehydrogenase genes (ADH) cloned and sequenced in four members of the *melanogaster* species subgroup; a different pattern is seen in the ADH coding regions, where the ratio of transitions to transversions is 1.38 (17). In addition, a search for potential protein coding regions does not reveal any large open reading frame that is present in all three species. The largest open reading frame found would produce a protein 98 amino acids in length starting at base 162 of the *D. yakuba* sequence; however, the homologous open reading frames in *D. melanogaster* and *D. erecta* are 28 and 74 amino acids in length, respectively. Similar wide disparities in potential protein products were seen in the other open reading frames present.

## DISCUSSION

The nucleotide sequences of a region containing a transition from slowly evolving to rapidly evolving sequences have been determined. The existence of this boundary was inferred from the analysis of cloned sequences homologous to the 68C glue gene cluster of *D. melanogaster* from four closely related species. The alignment of the sequences (Fig. 3) reveals a sharp boundary between the two regions: a 5- to 10-fold change in the frequency of nucleotide substitution occurs over a stretch of <50 nucleotides. Additionally, while the frequency of base substitution undergoes a dramatic change across this boundary, the frequency of insertion/deletion events stays approximately the same.

One explanation for the high level of conservation of the conserved region is that it has been subjected to strong selection. However, this region probably does not code for a protein product: (i) No large open reading frames are found in the sequenced portion of the conserved region. (ii) One of the breakpoints of the chromosomal inversion *In(3L)HR15*, which is viable and without a visible phenotype when homozygous, lies within the conserved region (but beyond the sequenced section) (18). (iii) An experiment designed to saturate the region surrounding the 68C glue gene cluster for lethal and semilethal mutations did not reveal any such mutations in the conserved region (19). Thus, there is as yet no evidence that the region is being maintained because of its coding capacity.

Another explanation is that the conserved sequences regulate the glue gene cluster that is located only a few hundred bases away from the end of the conserved region.

However, P-factor-mediated transformation experiments of the glue gene cluster using constructs lacking sequences from the conserved region show normal tissue, time, and level of expression (18). The observations argue against any major role for these sequences in the regulation of the glue gene cluster. Thus, while the slow rate of evolution in the conserved region could be due to selection, a strong pressure to maintain these sequences is not apparent.

A third possibility is that the mutation rate is markedly different in the two regions. Thus, the high level of conservation seen would not be due to strong selection, but instead to the relative lack of mutation. This could be due to more efficient repair locally or to a physical protection of the sequences—e.g., by the complexing of these sequences with proteins. In contrast to the protection from point mutations, the rate of insertions and deletions seems to be constant across the boundary. Models have been proposed that suggest many insertion/deletion mutations arise from slippage of short repeated sequences during DNA replication (20). Many of the deletions seen in the aligned sequences can be explained by this model (e.g., the deletions in *D. yakuba* and *D. erecta* starting at base 1236 in *D. melanogaster*). Thus, while the processes responsible for the generation of point mutations are strongly influenced by some property that undergoes a sharp change at the boundary, little, if any, effect on the process that generates insertions and deletions can be seen.

Evidence for the interspersed blocks of rapidly and slowly evolving sequence in the *Drosophila* genome has been obtained from experiments on the reassociation kinetics of interspecies hybrids of single-copy sequences (21–23). The experiments of Zwiebel *et al.* (22) reveal two classes of sequences in the *Drosophila* genome. The first consists of sequences that cross-hybridize under stringent solution hybridization conditions; this cross-hybridizing DNA remelts with an average melting temperature depression ( $\Delta t_m$ ) characteristic of the species pair involved. The second class contains sequences that do not cross-hybridize under the conditions used, implying the presence of sequences that have evolved extensively since the divergence of closely related species. In addition, Schulze and Lee (23) have demonstrated that the amount of nonhybridizable sequences present between two species is correlated with the average melting temperature depression found for those sequences that do cross-hybridize.

As a complementary approach to these studies, we have characterized a boundary between adjacent sequences evolving at very different rates. The boundary is abrupt; if this single boundary is characteristic, then the *Drosophila* genome consists of adjacent blocks of sequences that not only evolve at different rates but also are sharply delimited.

It will require further efforts to show any general correlation between the location of genes and that of blocks of differing rates of evolution. That the genome contains the ability to differentially regulate the rate of evolution of DNA sequences in different chromosomal locations is an interesting possibility.

We thank Joan Kobori, Erich Strauss, and Frank Calzone for discussions of sequencing techniques. We also thank the members of the Meyerowitz lab for their helpful suggestions on the manuscript. This work was supported by Grant GM20927 from the National Institutes of Health. C.H.M. was supported by a National Science Foundation Predoctoral Fellowship and by a Graduate Fellowship from the General Electric Foundation.

1. Meyerowitz, E. M. & Hogness, D. S. (1982) *Cell* **28**, 165-176.
2. Crowley, T. E., Bond, M. W. & Meyerowitz, E. M. (1983) *Mol. Cell. Biol.* **3**, 623-634.
3. Garfinkel, M. D., Pruitt, R. E. & Meyerowitz, E. M. (1983) *J. Mol. Biol.* **168**, 765-789.
4. Ashburner, M. (1973) *Dev. Biol.* **35**, 47-61.
5. Ashburner, M. (1974) *Dev. Biol.* **39**, 141-157.
6. Ashburner, M. & Richards, G. (1976) *Dev. Biol.* **54**, 241-255.
7. Crowley, T. E. & Meyerowitz, E. M. (1984) *Dev. Biol.* **102**, 110-121.
8. Meyerowitz, E. M. & Martin, C. H. (1984) *J. Mol. Evol.* **20**, 251-264.
9. Norrander, J., Kempe, T. & Messing, J. (1983) *Gene* **26**, 101-106.
10. Davis, R. W., Botstein, D. & Roth, J. R. (1980) *Advanced Bacterial Genetics* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
11. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
12. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467.
13. Strauss, E., Kobori, J. A., Siu, G. & Hood, L. E. (1986) *Anal. Biochem.* **154**, 353-360.
14. Gotoh, O. (1982) *J. Mol. Biol.* **162**, 705-708.
15. Lemeunier, F. & Ashburner, M. (1984) *Chromosoma* **89**, 343-351.
16. Laird, C. D. & McCarthy, B. J. (1968) *Genetics* **60**, 303-322.
17. Bodmer, M. & Ashburner, M. (1984) *Nature (London)* **309**, 425-430.
18. Crosby, M. A. & Meyerowitz, E. M. (1986) *Dev. Biol.* **118**, in press.
19. Crosby, M. A. & Meyerowitz, E. M. (1986) *Genetics* **112**, 785-802.
20. Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980) *Cell* **21**, 653-668.
21. Hunt, J. A., Hall, T. J. & Britten, R. J. (1981) *J. Mol. Evol.* **17**, 361-367.
22. Zwiebel, L. J., Cohn, V. H., Wright, D. R. & Moore, G. P. (1982) *J. Mol. Evol.* **19**, 62-71.
23. Schulze, D. H. & Lee, C. S. (1986) *Genetics* **113**, 287-303.
24. Lemeunier, F., David, J. R., Tsacas, L. & Ashburner, M. (1986) in *The Genetics and Biology of Drosophila*, eds. Ashburner, M., Carson, H. L. & Thompson, J. N., Jr., (Academic, London), Vol. 3E, pp. 147-256.