EVOLUTION AND EXPRESSION AT THE 68C GLUE GENE CLUSTER OF *DROSOPHILA*

Thesis by

Christopher Hayes Martin

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1988

(Submitted December 7, 1987)

ACKNOWLEDGEMENTS

First, I would like to thank my thesis advisor, Elliot Meyerowitz, for his support and for providing such an excellent lab in which to learn and work.

Secondly, I thank the members of the Meyerowitz lab for their advice, assistance, and friendship. In particular, I would like to thank Mike Palazzolo and K. Vijay Raghavan for their wealth of very entertaining stories; Lynn Crosby for teaching me the joys of fly genetics; Robert Pruitt and Peter Mathers for being great friends to share a lab bay with; and Carol Mayeda for her excellent assistance.

Thirdly, I would like to thank my mother, without whom it would not have been possible.

ABSTRACT

This thesis describes investigations on the evolution of a region containing a cluster of three glue genes located at chromosomal site 68C in the *Drosophila melanogaster* genome. These studies have used a set of five closely related *Drosophila* species, all members of the *melanogaster* species subgroup. The first chapter serves as an introduction and summarizes this work. The second chapter describes the initial characterization of the glue gene clusters and the surrounding regions in the five *Drosophila* species. The third chapter describes the characterization at the sequence level of the boundary that was found between adjacent blocks of rapidly and slowly evolving sequences located at the 68C glue gene cluster. The fourth chapter describes the evolution of the largest of the three glue genes in the 68C glue gene cluster: *Sgs-3*. Together, these studies reveal that this region of the genome is evolving as a mosaic, with adjacent regions evolving at different rates and in very different ways.

TABLE OF CONTENTS

Chapter 1

Summary:
Mosaic Evolution in the *Drosophila* Genome

Christopher H. Martin and Elliot M. Meyerowitz
Division of Biology
California Institute of Technology
Pasadena, CA 91125

(submitted to *Bioessays*)

The tools of molecular biology have allowed the study of evolution at the molecular level. An important goal in the field of molecular evolution is to determine how the processes of mutation and selection produce evolutionary change. Our studies have focused on a region containing a small gene cluster in five closely related *Drosophila* species. We find that this region is evolving as a mosaic: adjacent subregions display strikingly different patterns of evolution. We argue that some differences between the rates of evolution seen in the individual subregions are due to local variations in the rate of mutation, while others are primarily due to local variations in selection pressure. There is also evidence that mutation can occur by a number of mechanisms and that the contribution of different mechanisms can vary in different subregions. Our studies indicate that one must be cautious of inferences as to the functional significance of a given subregion of DNA based solely upon the level of sequence conservation. For example, in the small portion of the *Drosophila* genome that we have examined, we find examples of very rapidly evolving subregions that are likely to be essential protein coding domains while also finding a large, slowly evolving subregion that has no apparent function. The results from studies in a number of laboratories indicates that the mosaic pattern of evolution seen in this one genomic region may be a general feature in the *Drosophila* genome.

Our investigations have dealt with the evolution of the glue gene cluster located at cytological site 68C in the polytene chromosomes of *Drosophila melanogaster*. The three

genes in this cluster, *Sgs-3*, *Sgs-7*, and *Sgs-8*, code for components of a proteinaceous glue that serves to affix the animal in place for the duration of metamorphosis (Meyerowitz & Hogness, 1982; Crowley *et al.*, 1983). These and other glue proteins are produced in the salivary gland of the animal during much of the third (and final) larval phase (Beckendorf & Kafatos, 1976). The three glue genes, along with over 30 kilobase pairs (kb) of flanking regions, have been cloned from *D. melanogaster* (Meyerowitz & Hogness, 1982) and also from the closely related *D. simulans*, *D. erecta*, *D. yakuba*, and *D. teissieri* (Meyerowitz & Martin, 1984). These species are all members of the *melanogaster* species subgroup, one of 11 species subgroups defined for the *melanogaster* species group (Lemeunier *et al.*, 1986). The cloned regions from the five *Drosophila* species are shown in Figure 1. Each of the clusters contains either three or four regions that are homologous to one of the three glue genes located at 68C in *D. melanogaster*. Each of these regions also hybridizes to an abundant transcript in the late larval salivary gland from the same species (Meyerowitz & Martin, 1984).

From a comparison of the restriction maps of the cloned regions and the determination of the RNA sizes and directions of transcription, it was found that the glue gene region is evolving rapidly by a number of mechanisms. These include frequent point mutation, insertion and/or deletion, gene inversion, gene duplication, and the gain or loss of repetitive elements. In contrast, a region of at least 13 kb that lies to the left (see Figure 1) of the glue gene

cluster appears to be evolving much more slowly. This became apparent from comparison of the restriction maps of the homologous glue gene cluster regions: sites are much more conserved among these species in this region as compared to the adjacent region that contains the glue genes (Meyerowitz & Martin, 1984). This was confirmed by experiments that determined the melting temperature depression of interspecies hybrids between restriction fragments from either the rapidly or slowly evolving region.

The molecular nature of this boundary between rapidly and slowly evolving sequences was investigated by DNA sequence analysis. The region thought to contain the boundary was sequenced in three species: *D. melanogaster*, *D. erecta*, and *D. yakuba* (Martin & Meyerowitz, 1986). Alignment of the sequences reveals that the boundary between slowly and rapidly evolving sequences in the three species is abrupt: a 5- to 10-fold change in the frequency of nucleotide substitutions occurs over a distance of less than 50 nucleotides. In contrast to this dramatic change in nucleotide substitution rate, the frequency of insertion/deletion events remains nearly constant across the boundary. Normally, highly conserved sequences are associated with an important functional domain, since mutations that occur in a conserved domain are presumed to be eliminated by the process of selection. In this instance, this would seem to require that the function of the well conserved region is sensitive to point mutations but not to small insertions and deletions. This would be an unusual functional entity.

There is as yet no evidence for a functional role of the conserved region. No conserved open reading frames have been found in the sequenced portion of the conserved region. The breakpoint of a chromosomal inversion, *In(3L)HR15*, lies within the conserved region; animals homozygous for this inversion are viable and without a visible phenotype (Ashburner, 1972; Crosby & Meyerowitz, 1986a). Further, a mutagenesis experiment designed to locate lethals and semi-lethals in the region surrounding the 68C glue gene cluster revealed no such mutations in the conserved region (Crosby & Meyerowitz, 1986b). This region also does not appear to be involved in the regulation of the nearby glue gene cluster: P-factor-mediated transformation experiments show that normal patterns of glue gene expression are seen with constructs that contain none of the sequences of the conserved region (Richards *et al.*, 1983; Bourouis & Richards, 1985; Crosby & Meyerowitz, 1986a; Vijay Raghavan *et al.*, 1986).

An alternative explanation is that the difference between the conserved and non-conserved regions lies not with the effects of selection but with a difference in the rate of mutation (and/or the efficiency of repair) between the two regions. Under this model, whatever process is responsible for the majority of point mutations is strongly affected by some property that abruptly changes at the boundary. In contrast, the primary mechanism responsible for short insertion/deletion events is not affected by the boundary. Models have been proposed that attribute small insertions and deletions to slippage of short direct repeats

during DNA replication (Efstratiadis *et al.*, 1980). Several of the insertion/deletions seen in the sequenced regions occur in such short direct repeats. It is possible that this large, relatively well conserved region may not be a functional entity that is being preserved by the forces of selection. If this is the case, the boundary would reflect a sharp change in a subset of the mutational forces that are acting on the DNA sequences located at different parts of the 68C locus.

The boundary between rapidly and slowly evolving sequences that is found at the 68C locus may be representative of a general feature of the evolution of the *Drosophila* genome. Evidence for this has been found from the results of studies on the reassociation kinetics of single copy sequences from related *Drosophila* species (Hunt *et al.*, 1981; Zwiebel *et al.*, 1982; Schultze & Lee, 1986). These experiments demonstrate the presence of two classes of sequences. The first class consists of sequences that will cross-hybridize, with an average melting temperature depression that is characteristic of the level of sequence change. The second class consists of sequences that do not cross-hybridize under the conditions used in these experiments: these sequences represent relatively rapidly evolving sequences of the *Drosophila* genome. Further, the fraction of non-cross-hybridizing sequences is greater between species pairs that have higher average melting temperature depressions in the fraction of sequences that do cross-hybridize (Schultze & Lee, 1986). It has also been found that these rapidly and slowly evolving sequences are

present as interspersed blocks of sequence that are, on average, greater than 500 base pairs in length (Zwiebel *et al.*, 1982).

A second set of investigations deals with the evolution of the largest of the three glue genes: *Sgs-3* (Martin *et al.*, 1987) The structure of this gene is shown in Figure 2. The three glue genes reside within a 5 kb region. The two smaller genes, *Sgs-7* and *Sgs-8*, are divergently transcribed and produce mRNAs that are 320 nt and 360 nt nucleotides in length, respectively. The larger *Sgs-3* gene produces a 1120 nt transcript (Meyerowitz & Hogness, 1982). A 6.7 kb region that contains these three genes has been sequenced in *D. melanogaster* (Garfinkel *et al.*, 1983). This sequence reveals that the three genes possess related amino acid sequences in both the amino- and carboxy-terminal domains. However, the larger *Sgs-3* gene differs from the other two by the presence of a third protein coding domain: a central, threonine-rich region that contains 37 tandem repeats of a five amino acid sequence with the consensus sequence of pro-thr-thr-thr-lys. The amino terminal domain of all three proteins contains a 23 amino acid hydrophobic leader that is removed during processing of the protein. The carboxy-terminal domain is 50 amino acids in length and contains seven cysteine residues that are conserved among all three of the glue genes at 68C. These genes are also related by the presence of a short (73 nt in *Sgs-3*) intron after the first base of the codon for the tenth amino acid. It is therefore likely that this gene cluster arose from the duplication and subsequent divergence of a single gene.

The investigation of evolution in this gene had two goals: first, to locate potential regulatory sequences by searching for conserved regions upstream of the *Sgs-3* gene and, second, to look at the evolution of the *Sgs-3* gene, particularly that of the central, repeat-containing region. Towards this goal, the *Sgs-3*-homologues have been sequenced in three other species: *D. simulans*, *D. erecta*, and *D. yakuba*. A comparison of these sequences shows that the *Sgs-3* gene is also evolving as a mosaic: different regions in and around the gene evolve in different ways and at different rates.

The most conserved regions comprise two of the three protein coding domains: the amino-terminal domain that contains the hydrophobic leader, and the carboxy-terminal domain that contains the conserved cysteine residues. No insertion/deletion events are apparent in these regions, thus conserving the length and reading frame of these two domains. However, several nucleotide and amino acid substitutions are apparent. Although change is occurring in these regions, certain features of the regions are strictly maintained. The hydrophobicity of the amino-terminal domain and the number and position of the cysteine residues in the carboxy-terminal domain are conserved among the four species. These cysteine residues are thus conserved in this way in at least six glue genes: the *Sgs-7*, *Sgs-8*, and *Sgs-3* genes of *D. melanogaster* and in the three *Sgs-3* homologues of *D. simulans*, *D. erecta*, and *D. yakuba*. These features of the two domains are presumably important for their function; within these constraints however, the region can change

appreciably in both nucleotide and amino acid sequence. This pattern of evolutionary change, with limited insertion/deletion events and moderate nucleotide substitution, is typical of the protein coding portions of many genes (*e.g.*, in *Drosophila*, Bodmer & Ashburner, 1984; Blackman & Meselson, 1986; Schaeffer & Aquadro, 1987).

Introns are often examples of relatively rapidly evolving sequences that are located within more highly conserved protein coding regions of a gene. In the *Sgs-3* intron, however, the level of nucleotide substitutions is not much greater than that seen in the two most conserved protein coding regions of the gene (22.2±6.5% in the intron vs. 18.6±3.0% in the hydrophobic region). This is unusual, but could be explained by the following: (1) the amino- and carboxy-domains of the *Sgs-3* protein are evolving moderately rapidly because several sequences are compatible with proper function and, (2), the intron, being relatively short, may have a significant fraction of its nucleotides that are directly involved in splicing. As the *Sgs-3* gene is heavily transcribed in the salivary gland during third instar, efficient interaction of intron sequences with the splicing machinery may be required for the processing of this message. The intron region does show a number of insertion/deletions, a type of change not seen in either the amino- or carboxy-domains of the protein. We again find a sharp boundary between adjacent regions where the pattern of evolutionary change alters abruptly. In this case, nucleotide substitution frequencies are similar on either side of the boundary while the rate of insertion/deletion

changes abruptly. In contrast to the previous example, this one is of a familiar sort, and very likely the result of changes in selection forces.

The regions flanking the coding region are evolving moderately rapidly and are subject to numerous nucleotide substitution events and both large and small insertion/deletion events. Thus, both types of mutation are occurring and neither is being strongly selected against. However, several small islands of relatively well-conserved sequences can be identified, primarily in the 5' flanking region of the *Sgs-3* gene that has been implicated in the regulation of this gene. This conservation is particularly evident in the 130 bp just 5' of the mRNA start site in *D. melanogaster*. This region is capable of directing tissue and time-specific gene expression in the absence of any additional upstream sequences, although at levels reduced from normal (Vijay Raghavan, 1986).

The most rapidly changing region that has been found in the 68C glue gene cluster is the central threonine-rich repeat-containing region of the glue gene. The region varies greatly in size: from 139 amino acids in *D. simulans* to 250 amino acids in *D. erecta* (Martin *et al.*, 1988). The length of this region is also known to vary among strains of *D. melanogaster* (Mettling *et al.*, 1985; Crosby & Meyerowitz, 1986a). In addition to the change in the length of the region, both the nucleotide and amino acid sequence of the region are evolving very rapidly. The repeat-containing region has changed to such an extent that meaningful sequence alignments of these sequences cannot be generated,

thus making it difficult to quantitate the levels of nucleotide substitution and insertion/deletion that are occurring. However, it is apparent that the region is evolving by both point mutation and insertion/deletion events. There do appear to be certain constraints imposed by selection on this region. First, the central regions of each species are rich in threonine, proline, and lysine and contain few acidic residues. Second, each of these regions begin with a threonine-rich part that is followed by a tandem array of a five amino acid sequence.

The individual five amino acid repeat motifs are very similar within a species, but vary substantially between species. A likely mechanism for this is that of unequal crossover (Smith, 1976). This process would be capable of both maintaining homogeneity of the repeats within a species, while allowing for the rapid divergence of this region between species. Frequent unequal crossover events would also explain the dramatic length variation seen in this region. This region represents a case in which the rapid evolution appears to be driven by the sequence of the region itself. Such tandemly repeated sequences lend themselves to mutation by a mechanism that is relatively rare in the absence of this type of sequence. For example, the non-coding regions that surround the *Sgs-3* gene are not evolving as rapidly as the central tandem-repeat region. Although the bulk of these sequences are unlikely to be under strong selection pressures, the evolutionary mechanism that results in the rapid evolution seen in the central tandem-repeat region cannot act on these non-coding

sequences because they lack significant amounts of tandemly repeated sequences. Similar examples of the rapid evolution of tandemly repeated sequences have been found in both non-coding regions, for example in the non-transcribed spacer between the ribosomal genes of *Xenopus* (Brown *et al.*, 1972) and in coding regions, for example in the Balbiani ring genes of *Chironomus* (Lendahl *et al.*, 1987).

The 68C glue gene cluster region is thus evolving as a series of sharply delimited regions, each with its own pattern of evolution. This is summarized in Figure 3, where the amounts of nucleotide substitution and insertion/deletion are shown for each of the regions discussed. Some regions show both types of change, while others show notably less of one or the other of these types of mutational change. Certain patterns of evolution that are found at the 68C glue gene cluster are somewhat surprising. For example, the most rapidly evolving region consists of a protein-coding region of the *Sgs-3* gene. This region is likely to code for a functionally important domain of the protein, as it is present in each of the species studied. Also, several properties of this region are highly conserved. It seems that the proper function of this protein domain can be performed by many possible amino acid sequences. In contrast, the least rapidly evolving region has no known function and is apparently not protein coding. Furthermore, this region is evolving slowly only in terms of single nucleotide substitutions; it is evolving relatively rapidly in terms of insertion/deletion events. This pattern of evolution is the opposite of that seen in the moderately

well conserved protein coding domains of *Sgs-3*, where insertion/deletions are absent and nucleotide substitutions are fairly common.

The results of these studies demonstrate the importance of recognizing the potential respective contributions of mutational and selectional processes to the overall pattern of evolution that is seen in a particular region of the genome. Various combinations of these two processes appear to lead to very different patterns of evolution. The relative influences of these processes on the resulting evolution pattern can change abruptly over a very short region of sequence, resulting in the mosaic pattern of evolution that is seen in this glue gene cluster region. It is likely that the rate of evolution found in a local region of the genome is itself under evolutionary control. Such alterations in the rate of evolution could be responsible for some of the discrepancies found when applying the molecular clock hypothesis to certain genomic regions (Zuckerkandl & Pauling, 1962; Margoliash, 1963; Britten, 1986). Future investigations on the patterns of evolution in other portions of the genome and in other phyla should extend our understanding of the processes that underlie the piecemeal evolution of genomes.

Acknowledgements

We thank the members of the Meyerowitz lab for their helpful suggestions on the manuscript. This work was supported by grants GM28075 and 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.

References

Ashburner, M. (1972). New mutants: report of M. Ashburner. *Dros. Inf. Serv.* **49**, 34.

Beckendorf, S.K. & Kafatos F.C. (1976). Differentiation in the salivary glands of *Drosophila melanogaster*: Characterization of the glue proteins and their developmental appearance. *Cell* **9**, 365-373.

Blackman, R.K. & Meselson, M. (1986). Interspecific nucleotide sequence comparisons used to identify regulatory and structural features of the *Drosophila hsp82* gene. *J. Mol. Biol.* **188**, 499-515.

Bodmer, M. & Ashburner, M. (1984). Conservation and change in the DNA sequences coding for alcohol dehydrogenase in sibling species of *Drosophila*. *Nature* **309**, 425-430.

Bourouis, M. & Richards, G. (1985). Remote regulatory sequences of the *Drosophila* glue gene *sgs3* as revealed by P-element transformation. *Cell* **40**, 349-357.

Britten, R.J. (1986). Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**, 1393-1398.

Brown, D.D., Wensink, P.C. & Jordan, E. (1972). A comparison of the ribosomal DNAs of *Xenopus laevis* and *Xenopus*

mulleri: The evolution of tandem genes. *J. Mol. Biol.* **63**, 57-73.

Crosby, M.A. & Meyerowitz, E.M. (1986a). *Drosophila* glue gene *Sgs-3*: Sequences required for puffing and transcriptional regulation. *Dev. Biol.* **118**, 593-607.

Crosby, M.A. & Meyerowitz, E.M. (1986b). Lethal mutations flanking the 68C glue gene cluster on chromosome 3 of *Drosophila melanogaster*. *Genetics* **112**, 785-802.

Crowley, T.E., Bond, M.W. & Meyerowitz, E.M. (1983). The structural genes for three *Drosophila* glue proteins reside at a single polytene chromosome puff locus. *Mol. Cell. Biol.* **3**, 623-634.

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., Belchl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.S. & Proudfoot, N.J. (1980). The structure and evolution of the human b-globin gene family. *Cell* **21**, 653-668.

Garfinkel, M.D., Pruitt, R.E. & Meyerowitz, E.M. (1983). DNA sequences, gene regulation and modular protein evolution in the *Drosophila* 68C glue gene cluster. *J. Mol. Biol.* 168, 765-789.

Hunt, J.A., Hall, T.J. & Britten, R.J. (1981). Evolutionary distances in Hawaiian *Drosophila* measured by DNA reassociation. *J. Mol. Evol.* **17**, 361-367.

Lemeunier, F., David, J.R., Tsacas, L. & Ashburner, M. (1986). The *melanogaster* species group in *The Genetics and Biology of Drosophila*, eds. Ashburner, M., Carson, H.L. & Thompson, J.N., Jr., (Academic, London), Vol. 3E, pp. 147-256.

Lendahl, U., Saiga, H., Hoog, C., Edstrom, J.-E. & Wieslander, L. (1987). Rapid and concerted evolution of repeat units in a Balbiani ring gene. *Genetics* **117**, 43-49.

Margoliash, E. (1963). Primary structure and evolution of cytochrome c. *Proc. Natl. Acad. Sci. USA* **50**, 672-679.

Martin, C.H. & Meyerowitz, E.M. (1986). Characterization of the boundaries between adjacent rapidly and slowly evolving genomic regions in *Drosophila*. *Proc. Natl. Acad. Sci., USA* **83**, 8654-8658.

Martin, C.H., Mayeda, C.A. & Meyerowitz, E.M. (1988). Evolution and expression of the *Sgs-3* glue gene of *Drosophila*. *J. Mol. Biol.*, in press.

Mettling, C., Bourouis, M. & Richards, G. (1985). Allelic variation at the nucleotide level in *Drosophila* glue genes. *Mol. Gen. Genet.* **201**, 265-268.

Meyerowitz, E.M. & Hogness, D.S. (1982). Molecular organization of a *Drosophila* puff site that responds to ecdysone. *Cell* **28**, 165-176.

Meyerowitz, E.M. & Martin, C.H. (1984). Adjacent chromosomal regions can evolve at very different rates: Evolution of the *Drosophila* 68C glue gene cluster. *J. Mol. Evol.* **20**, 251-264.

Richards, G., Cassab, A., Bourouis, M., Jarry, B. & Dissous, C. (1983). The normal developmental regulation of a cloned *sgs3* 'glue' gene chromosomally integrated in *Drosophila melanogaster* by P element transformation. *EMBO J.* **2**, 2137-2142.

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.

Schultze, D.H. & Lee, C.S. (1986). DNA sequence comparison among closely related *Drosophila* species in the *Mulleri* complex. *Genetics* **113**, 287-303.

Smith, G.P. (1976). Evolution of repeated DNA sequences by unequal crossover. *Science* **191**, 528-535.

Vijay Raghavan, K., Crosby, M.A., Mathers, P.H. & Meyerowitz, E.M. (1986). Sequences sufficient for correct regulation of *Sgs-3* lie close to or within the gene. *EMBO J.* **5**, 3321-3326.

Zuckerkandl, E. & Pauling, L. (1962). Molecular disease, evolution and genic heterogeneity, in: Horizons in Biochemistry (M. Kasha and B. Pullman, eds.), Academic Press, New York, pp. 189-225.

Zwiebel, L.J., Cohn, V.H., Wright, D.W. & Moore, G.P. (1982). Evolution of single-copy DNA and the ADH gene in seven Drosophilids. *J. Mol. Evol.* **19**, 62-71.

Figure 1. The cloned sequences of the 68C-homologous glue gene cluster regions. Restriction enzyme abbreviations are: *Bam*HI (B), *Bgl*II (Bg), *Eco*RI (R), *Hin*dIII (H), *Sal*I (S), *Sac*I (Sc), *Xba*I (Xb), and *Xho*I (Xh). Sites in parentheses are present in some strains (*D. melanogaster*) or clones (*D. teissieri*) but not in others (Meyerowitz & Martin, 1984). The arrows below each map indicate the extent and direction of transcription of the glue gene transcription units. In cases where the exact position of the gene is not known, the region that is hybridized by cDNA made from salivary gland mRNA is shown by a solid black bar. Hatched bars show the extent of the regions in these glue gene clusters that have been sequenced. The maps are aligned by the positions of the conserved restriction sites found left of the RNA coding regions. The vertical dashed line shows the boundary between the conserved region at the left and the nonconserved region at the right.

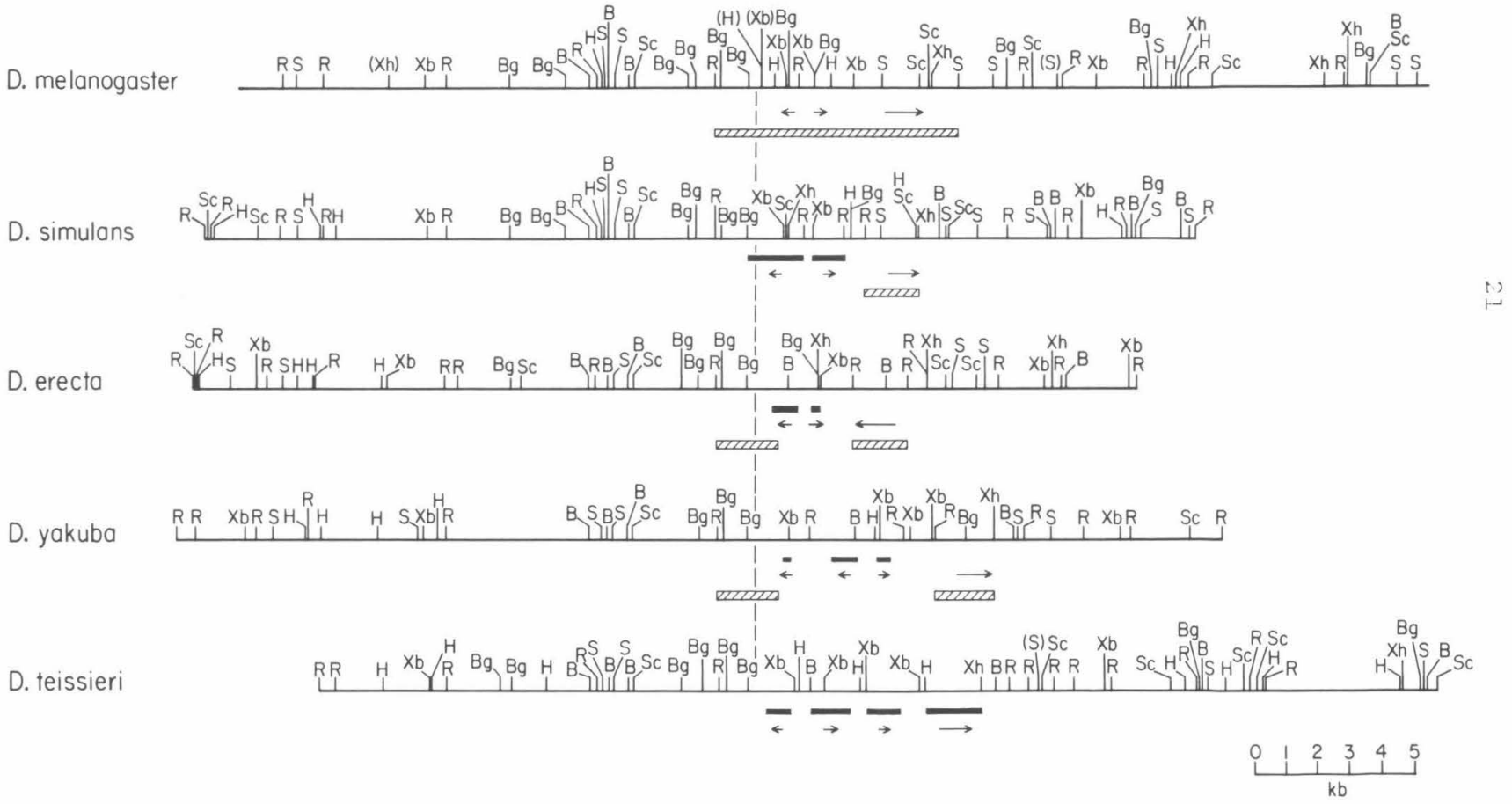

D. melanogaster
D. simulans
D. erecta
D. yakuba
D. teissieri
0 1 2 3 4 5
kb

Figure 2. The structure of the *Sgs-3* glue gene of *Drosophila melanogaster*. The numbering scheme used is relative to the transcription start site of the gene (Garfinkel *et al.*, 1983). The arrow indicates the extent of the mRNA transcript of the *Sgs-3* gene and the location of the 73 nucleotide intron, near the 5' end. Above the map, a block diagram shows the major subdivisions of this region that were revealed from comparisons of the *Sgs-3* genes from four related *Drosophila* species.

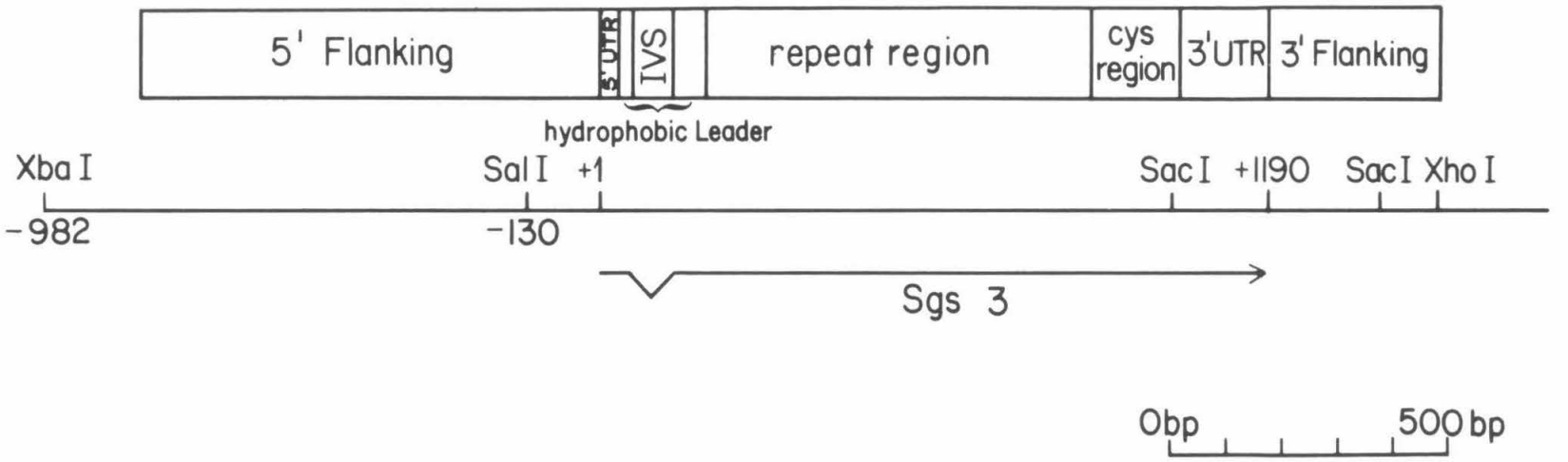

5' Flanking
5'UTR
IVS
repeat region
cys region
3'UTR
3' Flanking
hydrophobic Leader
Xba I
Sal I
+1
Sac I
+1190
Sac I
Xho I
-982
-130
Sgs 3
0bp
500 bp

Figure 3. Change at the 68C glue gene cluster region. The numbers represent the average percent of either nucleotide substitution or insertion/deletion that are occurring between three pairs of species: *D. melanogaster vs. D. erecta*, *D. melanogaster vs. D. yakuba* and *D. yakuba vs. D. erecta*. The levels of change in the repeat region were not quantitated because meaningful sequence alignments of these rapidly evolving regions could not be generated.

% substitution
% deletion
%substitution or %deletion
0
10
20
30
40
consv
nonconsv
5' utr
amino
intron
thr-rich
carboxy
3' utr
not measured
appears very high
border region
Sgs-3 region

Chapter 2

Adjacent Chromosomal Regions Can Evolve at Very Different Rates: Evolution of the *Drosophila* 68C Glue Gene Cluster

Elliot M. Meyerowitz and Christopher H. Martin
Division of Biology
California Institute of Technology
Pasadena, CA 91125

(published in the *Journal of Molecular Evolution*)

J Mol Evol (1984) 20:251–264

Journal of
Molecular Evolution

Adjacent Chromosomal Regions Can Evolve at Very Different Rates: Evolution of the *Drosophila* 68C Glue Gene Cluster

Elliot M. Meyerowitz and Christopher H. Martin

Division of Biology 156-29, California Institute of Technology, Pasadena, California 91125, USA

Summary. The 68C puff is a highly transcribed region of the *Drosophila melanogaster* salivary gland polytene chromosomes. Three different classes of messenger RNA originate in a 5000-bp region in the puff; each class is translated to one of the salivary gland glue proteins sgs-3, sgs-7, or sgs-8. These messenger RNA classes are coordinately controlled, with each RNA appearing in the third larval instar and disappearing at the time of puparium formation. Their disappearance is initiated by the action of the steroid hormone ecdysterone. In the work reported here, we studied evolution of this hormone-regulated gene cluster in the *melanogaster* species subgroup of *Drosophila*. Genome blot hybridization experiments showed that five other species of this subgroup have DNA sequences that hybridize to *D. melanogaster* 68C sequences, and that these sequences are divided into a highly conserved region, which does not contain the glue genes, and an extraordinarily diverged region, which does. Molecular cloning of this DNA from *D. simulans, D. erecta, D. yakuba,* and *D. teissieri* confirmed the division of the region into a slowly and a rapidly evolving portion, and also showed that the rapidly evolving region of each species codes for third instar larval salivary gland RNAs homologous to the *D. melanogaster* glue mRNAs. The highly conserved region is at least 13,000 bp long, and is not known to code for any RNAs.

Key words: *Drosophila* — Genome evolution — 68C Glue gene cluster — *Drosophila melanogaster* species subgroup

Offprint requests to: E.M. Meyerowitz

Introduction

Puffs are regions of polytene chromosomes that are actively undergoing transcription (Pelling 1964). One of the largest puffs found in the *Drosophila melanogaster* salivary gland polytene chromosomes is the 68C puff, on the left arm of the third chromosome. This puff is present through much of the third larval instar, regressing several hours before the time of puparium formation (Ashburner 1967). The regression of the 68C puff is a direct result of an increase in the level of the steroid hormone ecdysterone in the larval hemolymph (Ashburner 1973, 1974; Ashburner and Richards 1976). Several lines of evidence indicate that 68C puff regression results from the binding of ecdysterone, itself bound to a protein receptor, to the DNA of the 68C region (Gronemeyer and Pongs 1980; Dworniczak et al. 1983).

Molecular cloning of 68C puff DNA followed by DNA, RNA, and protein sequence analysis has shown that the puff contains DNA sequences that are transcribed to produce three different polyadenylated messenger RNAs whose accumulation is under ecdysterone control (Meyerowitz and Hogness 1982; Crowley et al. 1983; Garfinkel et al. 1983; Crowley and Meyerowitz 1984). Each of these RNAs is translated in the salivary gland, and each of the resulting polypeptides is one of the salivary gland secretion proteins. This is a group of at least seven polypeptides that are produced in salivary gland cells during the third larval instar; they are secreted into the lumen of the salivary gland near the end of this developmental stage, and are expelled from the luman through the salivary gland duct to the larval substrate at the end of the third instar. The secretion then hardens to form a strong glue that binds the newly formed puparial case to a solid surface (Fraen-

kel and Brookes 1953). The three salivary gland secretion proteins coded for in the 68C puff are sgs-3, sgs-7, and sgs-8 (Crowley et al. 1983). These three proteins are related in their amino acid sequences and thus form a small, clustered gene family. The differences among the proteins have arisen both from single-nucleotide substitutions and from the appearance in one of the proteins (sgs-3) of a module of 234 amino acids not present in the other two (Garfinkel et al. 1983).

There are two major reasons for examining the DNA sequence topography of the 68C puff in species of *Drosophila* other than *D. melanogaster*. One is to understand the evolution of the three members of this diverged gene family, in particular to investigate the mechanism of modular evolution that gave rise to sgs-3. The other is to find those elements of the puff DNA that are conserved in evolution, in the expectation that the regions of sequence that are relatively conserved in related species of *Drosophila* will include those that interact with regulatory proteins coded elsewhere in the genome. Identification of conserved sequences would thus be a step toward understanding the relation of the 68C DNA sequences to the regulated expression of the 68C RNAs.

In the experiments reported here we have begun the analysis of the 68C glue gene cluster in the group of closely related *Drosophila* species making up the *melanogaster* species subgroup. By interspecies DNA hybridization and molecular cloning, we show that five of these species contain regions homologous to the 68C region of *D. melanogaster* that code for abundant salivary gland RNAs related to the *D. melanogaster* 68C glue messengers. We also discover that the 68C-homologous region is divided into adjacent blocks of sequence that evolve at very different rates, with the gene cluster found in a region that is evolving with extraordinary rapidity. Our results show that the rate of sequence evolution is a local property of chromosomal regions, and also serve as a first step toward understanding the relation of DNA sequence structure and its evolution to the regulated expression of the 68C glue gene cluster.

Materials and Methods

Materials. Restriction endonucleases were purchased from New England Biolabs. The large proteolytic fragment of *Escherichia coli* DNA polymerase I was from either New England Biolabs or New England Nuclear, and T4 DNA polymerase was purchased from New England Nuclear. T4 DNA ligase and *E. coli* DNA polymerase I were gifts of Dr. S. Scherer. ^{32}P-Labeled nucleoside triphosphates were from either Amersham or ICN. Avian Myeloblastosis Virus reverse transcriptase was a gift of G. Duyk. Nitrocellulose was purchased from Schleicher & Schuell.

The *Drosophila melanogaster* strains used were the homozygous third chromosome strain OR16f (Meyerowitz and Hogness 1982) and the Canton-S wild-type from the California Institute of Technology stock collection. The other fly species—*D. mauritiana, D. simulans* (*jv st pe*), *D. erecta, D. yakuba,* and *D. teissieri*—were from the California Institute of Technology *Drosophila* stock collection. Flies were cultured on standard food (Lewis 1960) at 18° or 22°C.

Nucleic Acid Preparations. Drosophila DNA was extracted from adult flies by freezing the flies in liquid N_2 and then powdering them in a mortar. The powder from 0.1–2 g of flies was added to 2.5 ml 0.2 M Tris–HCl, pH 8.0, 0.2 M ethylenediaminetetraacetate (EDTA), 1% sodium N-lauroyl sarcosine, 100 μg/ml proteinase K (Merck). This was incubated with gentle shaking at 48°C for 1 h, and then centrifuged at 10,000 rpm for 5 min in a Sorvall SS-34 rotor. The supernatant was brought to 4.0 ml with 10 mM Tris-HCl, pH 8.0, 1 mM EDTA; then 3.7 g CsCl and 0.4 ml 10 mg/ml ethidium bromide were added. This mixture was centrifuged at 53,000 rpm for 20 h in a Beckman VTi65 rotor, and then the ultraviolet-fluorescent band was removed with a syringe and gently butanol extracted four times to separate the DNA from the ethidium bromide. Following this, the DNA was precipitated by addition of 2 volumes of ethanol followed by gentle hand centrifugation. The DNA pellet was washed with 70% ethanol, air dried, and resuspended overnight at 4°C without agitation. The resulting DNA was pure and over 150,000 bp in fragment length.

Plasmid and bacteriophage DNA preparations were performed as described by Davis et al. (1980), with occasional modifications that did not affect the results. *Drosophila* salivary gland RNA was prepared from hand-dissected salivary glands by two phenol–chloroform extractions followed by two chloroform extractions and ethanol precipitation. Polyadenylated RNA was separated on an oligo(dT) cellulose (Collaborative Research) column as described by Maniatis et al. (1982).

Nucleic Acid Labeling. Nick translations followed the method of Rigby et al. (1977). 3′-End labeling of restriction fragments using T4 DNA polymerase was done as described by Maniatis et al. (1982). ^{32}P-Labeled cDNA was made from poly(A)$^+$ RNA hybridized to an oligo(dT) (Collaborative Research) primer in the presence of 100 μg/ml Actinomycin-D, in a modification of the reaction described by Lis et al. (1978).

Nuclease Digestions. Restriction endonuclease digestions of DNA were done as described by Davis et al. (1980).

Recombinant DNA. All *D. melanogaster* genomic libraries used are described by Meyerowitz and Hogness (1982). The genomic libraries from *D. simulans, D. erecta, D. yakuba,* and *D. teissieri* were produced by performing partial EcoRI digestion on high-molecular-weight adult DNA (see above) to give DNA with a mean size of 15,000–20,000 bp (15–20 kb). This DNA was subjected to sedimentation in a 10% to 40% sucrose gradient in 0.2 M sodium acetate, 10 mM Tris, 10 mM EDTA, pH 7.6, at 35,000 rpm in a SW41 swinging bucket rotor at 4°C for 15–20 h, and DNA in the size range 15–20 kb was isolated. This DNA was ligated to purified EcoRI arms of the vector λSep6 (Meyerowitz and Hogness 1982) using T4 DNA ligase (see Davis et al. 1980). The ligated DNA was treated with a λ in vitro packaging extract prepared using the *E. coli* strains NS428 and NS433 (Sternberg et al. 1977) and a modification of the procedure of Collins and Hohn (1978). The phage particles were plated on L agar in L soft agar with K802 cells (see Davis et al. 1980), without any amplification step, and screened as described below. Positive plaques were single-plaque purified twice more before proceeding. Restriction fragments of these λ clones were subcloned in plasmid or λ vectors by standard methods (Davis et al. 1980; Maniatis et al. 1982).

The system of recombinant clone nomenclature, which originated in the laboratory of D.S. Hogness, is as follows. All λ clones are prefixed with the letter λ, followed by a letter indicating the λ vector used: a for λ647 (Murray et al. 1977), b for λSep6, c for Charon 4 (Blattner et al. 1977), or e for λgt 10 (R. Davis, personal communication). Following this are two letters: Dm for *Drosophila melanogaster*, Ds for *D. simulans*, De for *D. erecta*, Dy for *D. yakuba*, or Dt for *D. teissieri*. Last is a number identifying the specific clone. Plasmid subclones begin with a single letter: a for pBR322 (Bolivar et al. 1977), f for pBR325 (Bolivar 1978), or q for DOA-1, a kanamycin-resistant, high-copy-number plasmid with multiple cloning sites (R.E. Pruitt and E.M. Meyerowitz, unpublished). The rest of the clone designation is as for the λ clones.

Gel Electrophoresis. Double-stranded DNA was subjected to electrophoresis in agarose gels cast and run in Tris–borate–EDTA buffer (Peacock and Dingman 1968). DNA was strand separated on gels as described by McDonell et al. (1977) and Meyerowitz and Hogness (1982). RNA was subjected to electrophoresis in horizontal agarose gels buffered with 40 mM sodium 3-(N-morpholino)propanesulfonate, pH 7.0, 5 mM sodium acetate, 1 mM EDTA, and containing 6% formaldehyde. Running buffer was the same as the gel buffer, but without formaldehyde. Prior to electrophoresis RNA was treated at 55°C for 15 min in 50% formamide, 6% formaldehyde, and RNA gel running buffer.

Size standards on Tris–borate–EDTA gels were restriction fragments of bacteriophage λ DNA. RNA gel size standards were single-stranded DNA prepared from HinfI-digested pBR322 and formaldehyde treated as was RNA, except that the temperature was 70°C.

Filter Binding and Hybridization of Nucleic Acids. DNA gels were denatured and neutralized as described by Southern (1975), and were transferred to nitrocellulose by the Southern procedure, using 20× SSPE (180 mM NaCl, 10 mM Na H_2PO_4, 8 mM NaOH, 1 mM Na_2 EDTA pH 7.0, Davis et al. 1980). *Drosophila* genome blots had 1.5 μg restricted DNA per lane. RNA gels were either rinsed in water; or rinsed and then treated for 45 min with 50 mM NaOH, 100 mM NaCl then neutralized in 100 mM Tris–HCl, pH 7.5, before blotting. The blot procedure was as for DNA gels. RNA extracted from three salivary gland lobes was used in each lane. Plaque and colony filters were prepared as described by Davis et al. (1980).

All hybridizations were in 50% formamide, 5× SSPE, 100 μg/ml sonicated and boiled salmon testis DNA (Sigma), 1× Denhardt's solution (0.02% Ficoll, 0.02% polyvinylpyrrolidone, 0.02% bovine serum albumin; Denhardt 1966), 0.1% sodium dodecyl sulfate (SDS), at 43° C. After hybridization, filters were washed in 1× SSPE, 0.1% SDS, at room temperature, unless otherwise indicated.

DNA Sequencing. DNA sequencing was performed by the modifications of the Maxam and Gilbert chemical method (1980) described by Garfinkel et al. (1983).

Results

Gel Blot Experiments

The first step in our analysis of evolution of the 68C puff was the demonstration that species of *Drosophila* other than *D. melanogaster* contain DNA sequences homologous to the 68C genes. The species chosen were those of the *melanogaster* species subgroup, which contains *D. melanogaster* and at least six other species: *D. simulans, D. mauritiana, D. yakuba, D. teissieri, D. erecta,* and *D. orena.* These are the *Drosophila* species most closely related to *D. melanogaster* in morphology, mitotic chromosome karyotype, and polytene chromosome banding pattern (Bock and Wheeler 1972; Lemeunier and Ashburner 1976; Lemeunier et al. 1978). DNAs from adults of all of these species except *D. orena* were purified and digested to completion with the restriction endonuclease EcoRI, and the resulting restriction fragments were separated by electrophoresis through an agarose gel. The gel pattern was transferred to a nitrocellulose filter and this genome blot filter was hybridized with a series of ^{32}P-labeled cloned 68C probes from *D. melanogaster.* Autoradiography revealed the presence and size of EcoRI fragments from each of the species that were homologous to the *D. melanogaster* probes. The intensity of hybridization and persistence of the signal through successively more stringent filter washes gave an estimate of the sequence divergence between the *D. melanogaster* probes and the homologous sequences from the other species.

The cloned *D. melanogaster* sequences used are depicted in Fig. 1. The first probe was λaDm1501-10, a genomic clone containing all or a substantial part of *D. melanogaster* EcoRI fragments 3.8, 4.7, 3.7, 2.6, and 7.0 kb long. The control *D. melanogaster* lane on the filter showed strong autoradiographic signals resulting from hybridization to bands of these sizes. The lanes containing DNAs from the other species showed hybridization to a smaller number of bands. For *D. simulans,* bands of 4.7, 3.8, and 3.7 kb showed strong signals, and weak hybridization was seen to EcoRI fragments of 2.6 and 1.4 kb. *D. mauritiana* also showed strong signals on bands of 4.7, 3.8, and 3.7 kb. *D. erecta* showed hybridization to bands of 4.3, 4.0, and 3.7 kb; *D. yakuba* to 8.4- and 4.2-kb fragments; and *D. teissieri* to bands at 4.7, 3.8, and 3.7 kb. Since the *D. melanogaster* DNA in λaDm1501-10 extends for almost 18 kb, it can be seen that only a fraction of the λaDm1501-10 insert hybridizes strongly to the genomic DNAs of the other species. This indicates that only part of the *D. melanogaster* probe has a high degree of similarity to sequences in the other species, with another part of the probe hybridizing weakly.

To determine which regions of the λaDm1501-10 insert were responsible for the strong cross-hybridization, several additional *D. melanogaster* clones were used as probes. λbDm2031 contains *D. melanogaster* DNA repesenting the leftward part of the DNA cloned into λaDm1501-10 (see Fig. 1). The *D. melanogaster* EcoRI fragments represented in λbDm2031 are 3.8, 4.7, and 3.7 kb long, the same size as the EcoRI fragments of *D. simulans, D.*

mauritiana, and *D. teissieri* that were strongly labeled with the λaDm1501-10 probe. A species genome blot filter hybridized with labeled λbDm2031 DNA gave the expected *D. melanogaster* pattern, and in the lanes containing DNA from the other species gave the same pattern of strongly labeled bands as that obtained using λaDm1501-10 as the probe. Thus, the *D. melanogaster* sequences with a high degree of similarity to sequences from the other species are those represented in λbDm2031, and not those that contain the salivary gland secretion protein genes.

This conclusion was tested by using λcDm2021, which includes all of the sequences found in λbDm2031 as well as several kilobases of additional DNA (Fig. 1). This clone hybridized strongly with the same bands as λbDm2031 did, and with smaller, additional bands that presumably were hybridized by those *D. melanogaster* sequences present in λcDm2021 and not in λbDm2031. Both the λbDm2031 hybridized filter and the λcDm2021 filter were washed at successively higher stringencies after the initial autoradiographic exposure. The first washes were in 0.01 × SSPE at 47°C. On both filters, the hybridization patterns and intensities were unchanged by this treatment. After the first wash and autoradiographic exposure was a second wash of each filter in 0.01 × SSPE at 52°C. This caused a uniform reduction in signal in all bands on both filters, with the signal change in the control *D. melanogaster* lanes paralleling that in the lanes representing the other species. Thus, the strongly hybridized regions of the 68C sequences in all of the species are not detectably diverged in DNA sequence, if one uses the melting temperature of filter-bound DNA duplexes as a crude divergence assay.

One further *D. melanogaster* probe from the conserved region of the 68C clones was used, aDm2003. This plasmid clone contains most of the rightmost fragment of the conserved region of *D. melanogaster,* the 3.7-kb EcoRI fragment, and much of a 2.6 kb EcoRI fragment, which is the leftmost fragment of the less-conserved RNA coding region. When hybridized to a species genome blot filter, the aDm2003 probe gave a strong signal at 3.7 kb and a weak signal at 2.6 kb in the *D. simulans* lane, a strong signal at 3.7 kb and a weak one at 4.0 kb in the *D. mauritiana* lane, strong hybridization to a 3.7-kb band and weak hybridization to a 4.2-kb fragment in the *D. erecta* track, and strong hybridization to a 3.7-kb *D. teissieri* band. Thus, not only is the 3.7-kb EcoRI band highly conserved by duplex melting criteria, but all of the species except *D. yakuba,* which was not tested in this experiment, appear to have the same EcoRI sites, spaced equally, surrounding this DNA. This indicates a very high degree of sequence conservation. That the remaining signal, from the 2.6-kb EcoRI fragment that includes one of the glue genes, was weak confirms that the boundary between conserved and less conserved sequences is approximately at the EcoRI site separating the 3.7-kb fragment from the adjacent EcoRI fragment to the right.

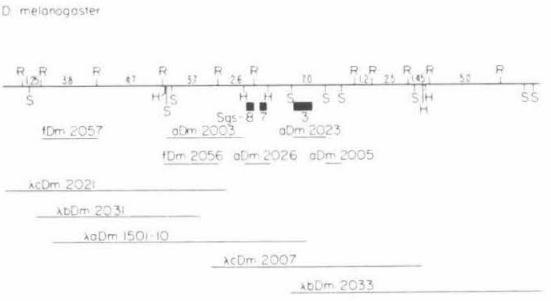


Fig. 1. *Drosophila melanogaster* cloned sequences used. The cloned *D. melanogaster* sequences used in this study are shown beneath a restriction endonuclease map of a portion of the 68C puff DNA of this species. The derivations of the map and clones are described in Meyerowitz and Hogness (1982) and in Garfinkel et al. (1983). The restriction endonuclease cleavage sites shown are those of EcoRI (R), HindIII (H), and SalI (S). The numbers on the restriction map are the sizes of each EcoRI fragment in kilobase pairs. The solid bars under the map show the positions and extents of the DNA coding for the glue proteins sgs-8, -7, and -3.

Several probes from the relatively unconserved glue RNA coding region of the 68C puff were also used in gel blot experiments. aDm2026 contains a 1.65-kb HindIII fragment from *D. melanogaster* that includes the *Sgs-7* and *Sgs-8* genes. When ^{32}P-labeled aDm2026 DNA was hybridized to a species genome blot filter, the hybridization to *D. melanogaster* sequences was much stronger than that to DNA of any of the other species. Each of the other species did show binding of the labeled probe to EcoRI fragments of various sizes. A series of washes of the filter in 0.01 × SSPE at temperatures of 43°, 49°, and 52.5°C, with autoradiography performed after each new wash, showed that the aDm2026-homologous sequences of *D. yakuba, D. teissieri,* and *D. erecta* lost all binding to aDm2026 DNA between 43° and 49°C, whereas the signals in the *D. mauritiana* and *D. simulans* lanes were considerably weakened after the 49°C wash, but not removed. A 52.5°C wash sufficed to remove almost all of the signal in the *D. mauritiana* and *D. simulans* lanes, while having little effect on *D. melanogaster* self-hybridization. Thus, the *D. melanogaster* sequences containing the *Sgs-7* and *Sgs-8* genes are more diverged from their homologous sequences in the other species than are the *D. melanogaster* sequences to the left of the glue genes, sequences that are not known to code for any *Drosophila* RNAs (Meyerowitz and Hogness 1982).

The next *D. melanogaster* DNA used as a labeled

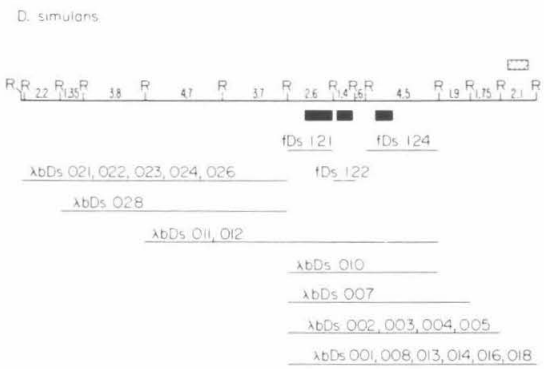


Fig. 2 Cloned sequences of *Drosophila simulans*. The original λ clones obtained from the *D. simulans* genomic library and several plasmid subclones used are shown in relation to a composite EcoRI restriction map of the cloned *D. simulans* DNA. The numbers on the map are distances between adjacent EcoRI sites in kilobase pairs. The hatched bar above the map indicates the maximum extent of the middle repetitive DNA element found in the 68C-homologous region in *D. simulans*. The solid bars below the map indicate the restriction fragments hybridized by cDNA made from abundant polyadenylated RNAs from the salivary glands of third instar larval *D. simulans*. All of the recombinant phage except λbDs007 are consistent with the composite map. λbDs007 has, in addition to the sequences shown, a 7.2-kb EcoRI fragment that is not present in any of the other clones and is not hybridized by *D. melanogaster* 68C sequences. It seems certain that this 68C-unrelated fragment was ligated to the 68C-homologous *D. simulans* DNA during the construction of the recombinant phage library, and that it derives from a random, noncontiguous region of the *D. simulans* genome

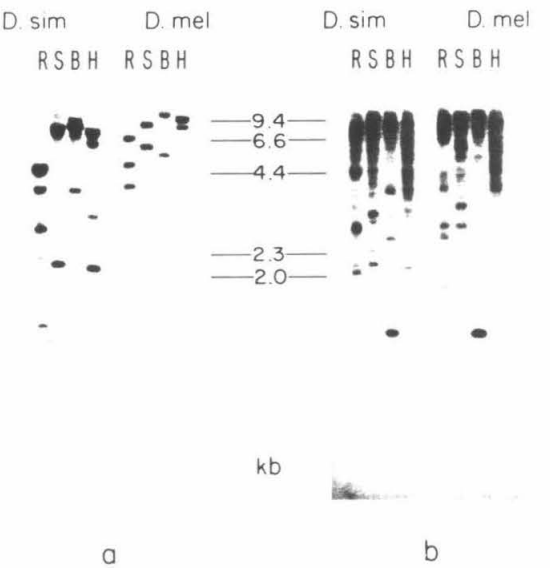


Fig. 3a,b. Hybridization of *Drosophila simulans* clones to *D. simulans* and *D. melanogaster* genomic DNA. *D. simulans* (D. sim) genomic DNA was digested with EcoRI (R), SalI (S), BamHI (B), or HindIII (H) in four separate reactions, each with 1.5 μg DNA. The digested samples were loaded in four adjacent lanes of a 0.9% agarose gel. The nearby group of lanes was loaded similarly with identical digests of *D. melanogaster* (D. mel) DNA. After electrophoresis the DNA in the gel was denatured and blotted to a nitrocellulose filter, and the filter was hybridized with a ^{32}P-labeled *D. simulans* λ clone probe and then autoradiographed. In **a** the probe was λbDs011. In **b** the signal was washed from the filter in a with boiling 0.01× SSPE, and the filter was rehybridized with λbDs001. The size standards are from λc-I857S7 DNA digested with HindIII

probe was aDm2023, a 2.4-kb SalI fragment containing the *Sgs-3* gene. Again, the *D. melanogaster* self-hybridization gave much stronger autoradiographic signals than did the hybridization of *D. melanogaster* sequences to the homologous sequences of other *Drosophila* species. A wash in 0.01× SSPE at 48°C reduced but did not eliminate the hybridization in the *D. yakuba, D. teissieri,* and *D. erecta* lanes. Thus, at least for some species, the RNA-coding region of the 68C puff is again less conserved in evolution than is the adjacent DNA. A final probing of a species genome blot was performed, using ^{32}P-labeled λcDm2007. The bands hybridized by this λ clone, which overlaps aDm2026 and aDm2023, included those hybridized by those two plasmid clones; the initial hybridization and subsequent washes at higher stringency confirmed the results obtained with those clones. Since λcDm2007 includes *D. melanogaster* sequences to the right of the RNA-coding region as well as the sequences containing the glue genes, there was hybridization to fragments not seen in the aDm2026 or aDm2023 experiments. These fragments were as intensely hybridized as *D. melanogaster* self-hybridized fragments of the same sizes, and after filter washes of either 48°C or 52.5°C in 0.01× SSPE these bands were still equal in intensity to the *D. melanogaster* bands. This indicates that the sequences to the right of the RNA-coding region are not highly diverged. No further experiments that analyzed this rightward region were performed.

Molecular Cloning

The general picture of 68C puff evolution gained from the genome blot experiments is of a highly diverged set of sequences containing the three glue genes, surrounded by highly conserved sequences that are not known to have any function in glue gene expression. To learn more about the evolution of this region and to establish a basis for DNA sequencing studies, the DNAs homologous to the *D. melanogaster* 68C sequences in *D. simulans, D. erecta, D. yakuba,* and *D. teissieri* were cloned. The first clones obtained were from *D. simulans,* the species most closely related to *D. melanogaster* (Sturtevant 1919). The strain of *D. simulans* used

256

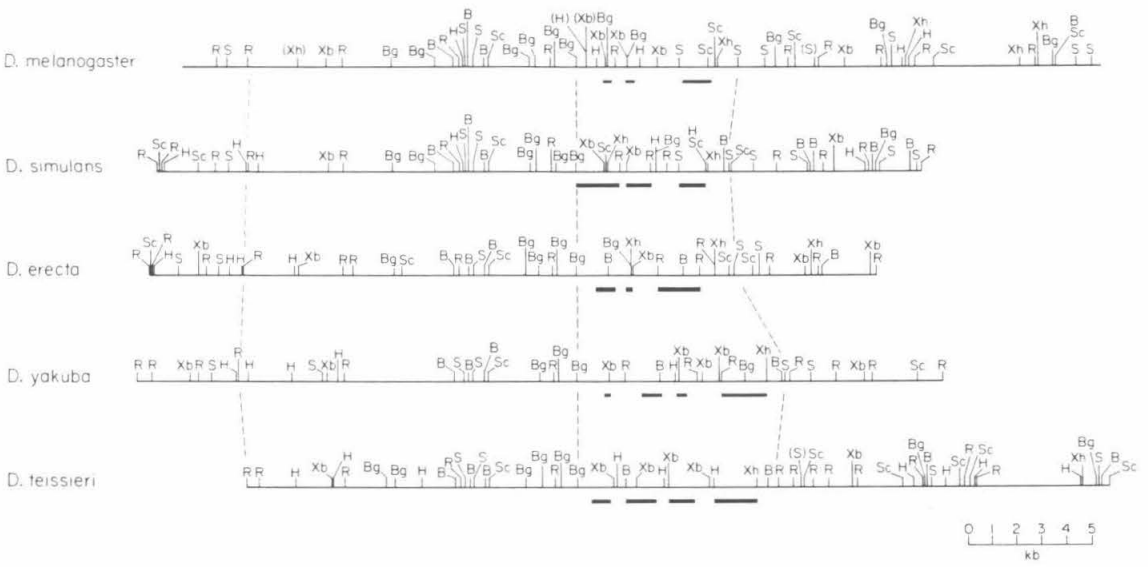


Fig. 4. Restriction maps of the cloned 68C-homologous sequences. All known BamHI (B), BglII (Bg), EcoRI (R), HindIII (H), SalI (S), SacI (Sc), XbaI (Xb), and XhoI (Xh) sites are depicted except for a single EcoRI site in *D. erecta*, which is omitted for the reason detailed in the caption to Fig. 5. Sites in parentheses on the *D. melanogaster* map are found in chromosomes of some wild-type strains, but not in others (Meyerowitz and Hogness 1982; Garfinkel et al. 1983). The parenthetical SalI site on the *D. teissieri* map is present in λbDt9100 but not in the overlapping λbDt9200. Beneath each map are solid bars showing the extents of the restriction fragments hybridized by [^{32}P]cDNA derived from third instar larval salivary gland polyadenylated RNA from each of the species (see Fig. 8 for more details). The maps are aligned by the series of common restriction endonuclease sites found to the left of the RNA-coding region. Three vertical dashed boundaries separate the maps into a leftward conserved region and a rightward RNA-coding region. The leftmost boundary is set at the EcoRI site that marks the left end of fDm2057 hybridization to the clones of each species, and is set at this point only because *D. teissieri* cloned sequences extend no farther to the left. Whether the conserved region continues beyond this point is not known, although comparison of the four species whose restriction maps do include DNA to the left of this boundary indicates that it probably does. The central boundary is the common BglII site just to the left of the RNA-coding region, this site separates conserved from diverged sequences by criteria of hybridization and restriction mapping. The rightward boundary marks the right end of the restriction fragments hybridized by aDm2023, where each species except *D. teissieri* has a SalI site. In *D. teissieri*, the boundary between hybridization of aDm2023 and aDm2005 is within the BamHI–EcoRI fragment that includes the RNA-coding-region boundary

was a homozygous third chromosome strain with the recessive third chromosome markers *jv, st,* and *pe*. DNA from adult flies was partially digested with EcoRI and 15- to 20-kb fragments were selected by sucrose gradient sedimentation and cloned into the EcoRI cloning vector λSep6 (see Materials and Methods). The resulting recombinant DNA library was not amplified, but was directly plated and screened by the plaque lift method, using the ^{32}P-labeled *D. melanogaster* clones aDm2026 and aDm2023 as probes. Fourteen independent clones were isolated that hybridized to both probes. These are λDs001, 002, 003, 004, 005, 007, 008, 010, 011, 012, 013, 014, 016, and 018. Figure 2 is a simple restriction endonuclease map of the DNA represented in these clones, showing the relation of the cloned segments. To obtain clones representing the highly conserved DNA adjacent to the RNA coding region, more recombinant phage were screened, using the *D. melanogaster* clone fDm2057 (see Fig. 1) as a labeled probe. The recombinant clones λbDs021, 022, 023, 024, 026, and 028 are all hybridized by fDm2057, and are also shown in Fig. 2.

That this collection of phage represents the sequence organization actually found in the *D. simulans* genome and does not result from the artificial joining of separate EcoRI fragments during the cloning procedure or from any other cloning artifact is shown by several facts. First, all regions of the composite restriction map, and all EcoRI junctions, were cloned more than once from a library of independent clones. In addition, the EcoRI fragments in the clones that hybridize to the *D. melanogaster* probes aDm2026 and aDm2023 are the same size as the EcoRI fragments of whole-genome *D. simulans* DNA hybridized by the same probes. Finally, two of the *D. simulans* λ clones were ^{32}P-labeled by nick translation and used as probes of genome blot filters containing *D. melanogaster* whole-genomic DNA digested with BamHI, EcoRI, HindIII, and SalI in different lanes and *D. simulans* genomic DNA similarly digested in a separate set of gel tracks. The

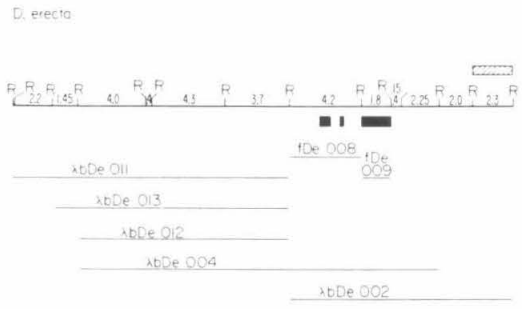


Fig. 5. Cloned sequences of *Drosophila erecta*. A composite restriction map of the cloned *D. erecta* sequences is depicted with the distances in kilobases between adjacent EcoRI (R) sites shown. The interval with the two numbers 0.15 and 0.4 contains two EcoRI fragments of these sizes; their order is unknown. Below the map are indicated the positions and extents of the λ clones and plasmid subclones used. Immediately below the map, solid bars show the location of the restriction fragments of the cloned DNA that hybridize to cDNA derived from abundant poly(A)$^+$ RNAs isolated from third instar salivary glands of *D. erecta*. Above the map, a hatched bar indicates the maximal extent of the middle repetitive DNA element found in the 68C-homologous region of *D. erecta*

first probe used was λbDs011. It hybridized to *D. melanogaster* fragments of the sizes expected from the known restriction map of the 68C region in this species. This indicates that no additional fragments from other genomic regions were incorporated into this phage during its construction, and that the *D. simulans* 68C region does not contain the breakpoints of any large inversions or translocations relative to the *D. melanogaster* sequence. λbDs011 also hybridized to the *D. simulans* restriction fragments expected from the restriction map of the phage, confirming that no large deletions or rearrangements occurred during the cloning of the *D. simulans* DNA (Fig. 3a). The second *D. simulans* λ probe used in hybridization to the genome blot filter was λbDs001. The autoradiogram resulting from this hybridization showed a multiplicity of labeled fragments in all lanes from both species, demonstrating that some part of λbDs001 contains a repetitive element present in numerous copies in both the *D. simulans* and the *D. melanogaster* genome (Fig. 3b). This element was localized to the 1.25-kb SalI–BamHI fragment internal to the 2.1-kb EcoRI fragment of λbDs001 by annealing ^{32}P-labeled, single-stranded *D. simulans* genomic DNA to a gel blot filter with lanes containing λbDs001 DNA digested with EcoRI, BamHI, and SalI. The autoradiogram of this filter showed strong labeling of the 2.1-kb EcoRI fragment, a 1.5-kb BamHI fragment, and a 1.65-kb SalI fragment; much weaker hybridization to the other *D. simulans* insert fragments; and no hybridization to phage vector DNA. Thus, the repetitive element is in the position shown in Fig. 2. Figure 4 shows a detailed restriction map of the 68C region of *D. simulans*, including the BamHI and SalI sites just mentioned. To show that the repetitive λbDs001 does represent contiguous *D. simulans* sequence and that the restriction fragments of this clone correspond to those at 68C in *D. melanogaster*, a blot filter with lanes of EcoRI-, BamHI-, and SalI-digested λbDs001 was hybridized with ^{32}P-labeled DNA of λbDm2033, a *D. melanogaster* clone (see Fig. 1) representing approximately the same region of 68C as λbDs001 does. All λbDs001 bands were hybridized except the leftmost two EcoRI fragments (1.4 and 0.6 kb), which are not expected to be represented in λbDm2033. The λbDs001 fragments containing the repetitive element hybridized less than did the others, as would be expected if a substantial portion of these DNA pieces contained sequences not present in the probe.

The next species whose 68C-homologous sequences were cloned was *D. erecta*. Production and screening of the *D. erecta* libraries was done as for *D. simulans*. The first screening used aDm2026 and aDm2023 as probes of duplicate plaque filters; two different clones hybridized by both probes were obtained. These are λbDe002 and λbDe004. A different set of clones from the *D. erecta* library was then probed with fDm2057, and three more positive plaques were obtained. These contained the phage λbDe011, λbDe012, and λDe013. A simple restriction map of the *D. erecta* sequences in these five clones and the relation of these clones to this composite map are shown in Fig. 5. A detailed restriction map of the *D. erecta* 68C-related region is in Fig. 4. Several results were obtained that show that the restriction map derived from the cloned *D. erecta* segments does correspond to the restriction map of the same sequences in the *D. erecta* genome. The first is that the restriction maps of all the overlapping regions of the λ clones are identical, thus eliminating rare cloning artifacts as a possibility in these regions. Also, hybridization of the ^{32}P-labeled *D. melanogaster* clones aDm2023 and aDm2026 to filter-bound EcoRI fragments of λbDe002 and λbDe004 showed that the sizes of the EcoRI fragments homologous to these probes are the same as the sizes of the *D. erecta* genomic EcoRI fragments hybridized by these *D. melanogaster* probes in the earlier species genome blot experiments. In addition, when λbDe004 DNA was ^{32}P-labeled and hybridized to a genome blot filter containing lanes with *D. erecta* genomic DNA digested with BamHI, EcoRI, HindIII, or SalI, the labeled bands on the filter correspond with the restriction fragment sizes expected from the clone restriction map. ^{32}P-Labeled λbDe004 DNA was also annealed to a genome blot

filter with separate lanes of *D. melanogaster* genomic DNA digested with BamHI, EcoRI, HindIII, or SalI. In each case the labeled restriction fragments were of the sizes expected from the known *D. melanogaster* 68C restriction map. Hybridizations to *D. erecta* and *D. melanogaster* genomic DNA digested with the same set of enzymes were done using ^{32}P-labeled λbDe002 as a probe as well. These showed, in addition to the expected bands, a distinct background smear in all lanes for both *D. erecta* and *D. melanogaster*. This smear, not seen in the λbDe004-hybridized genome blot filter, implies that λbDe002 contains a repetitive DNA element. Hybridization of ^{32}P-labeled *D. erecta* genomic DNA to a filter blotted from a gel containing lanes of λbDe002 DNA digested with EcoRI, XbaI, BglII, SacI, and a combination of EcoRI and each of the other enzymes showed that the repetitive element is entirely within the 2.1-kb EcoRI–XbaI fragment found in the 2.3-kb EcoRI fragment of λbDe002. Figure 5 shows the location of the repetitive element relative to the *D. erecta* restriction map.

The next DNA cloned was from *D. yakuba*. The first labeled *D. melanogaster* clone DNA used as a probe of the *D. yakuba* λ library was fDm2056; using this probe six positive plaques were obtained. These contained the *D. yakuba* genomic clones λbDy101, 102, 103, 104, 105, and 107. An additional clone, λbDy110, was subsequently obtained using aDm2023 as a probe of a separate portion of the *D. yakuba* library. Figure 6 shows the restriction map of the *D. yakuba* 68C-homologous region derived from the maps of these clones and the overlap of each λ clone insert with this map. Figure 4 shows a more detailed map of the *D. yakuba* sequences. As with the other species, the correspondence of the restriction map derived from the clones with that in the genomic DNA was shown in a variety of ways. As before, all regions of the phage clones that overlapped the same region of the composite restriction map showed identical restriction sites, eliminating the possibility of cloning artifacts in these areas. The sizes of EcoRI restriction fragments of *D. yakuba* genomic DNA hybridized by the *D. melanogaster* clone λcDm2021 in the earlier species genome blot experiments are all found in the *D. yakuba* clones that cover the left end of the composite restriction maps, and the sizes of the *D. yakuba* genomic EcoRI fragments hybridized by λcDm2007 in the species genome blots correspond to the sizes of the EcoRI fragments found in λbDy110. λbDy103, when ^{32}P-labeled and annealed to *D. yakuba* DNA that had been digested with BamHI, EcoRI, HindIII, or SalI, subjected to electrophoresis, and then transferred to nitrocellulose, showed hybridization to fragments of the same sizes as are present in the clones in all cases. When λbDy103 was used as a labeled probe

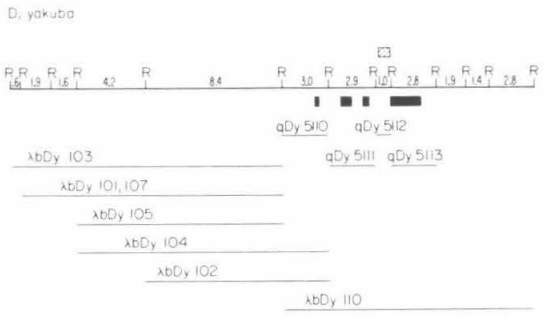


Fig. 6. Cloned sequences of *Drosophila yakuba*. A composite EcoRI (R) map of the cloned 68C-homologous DNA of *D. yakuba* is shown. Above the map, a hatched bar indicates the site of repetitive DNA; below the map, solid bars show the restriction fragments in the cloned DNA that hybridize to cDNA made from poly(A)$^+$ salivary gland RNA from *D. yakuba*. Below this, lines depict the extent of *D. yakuba* DNA in the plasmid and λ clones indicated

of *D. melanogaster* DNA digested with BamHI, EcoRI, HindIII, or SalI, the fragments hybridized were those that would be expected if the 68C regions of the two species are colinear. The results were different when λbDy110 was used as a labeled probe of genome blot filters containing restriction-endonuclease-digested *D. yakuba* and *D. melanogaster* genomic DNA samples. In this case, all *D. yakuba* lanes showed dark smears with a number of distinct bands superimposed. Thus, λbDy110 contains some DNA sequences related to sequences repeated many times in the *D. yakuba* genome. The *D. melanogaster* lanes do not show the dark smear that indicates hybridization of repetitive DNA; therefore the *D. yakuba* repetitive element is not highly repeated in the *D. melanogaster* genome. To localize the repetitive DNA in λbDy110, DNA of this λ clone was digested with EcoRI, XbaI, and a combination of both enzymes, and then subjected to electrophoresis in an agarose gel. The gel was blotted to nitrocellulose, and the resulting blot filter was hybridized with ^{32}P-labeled *D. yakuba* DNA. The λbDy110 fragments strongly hybridized were the 1.0-kb EcoRI piece and the 0.75-kb XbaI fragment wholly contained in this EcoRI fragment. The repetitive DNA was therefore shown to reside within this small XbaI fragment, in the position shown in Fig. 6.

The final λ library made used DNA from adult *D. teissieri*. A portion of these clones was screened on duplicate plaque lift filters, with both aDm2003 and aDm2023 as ^{32}P-labeled probes. One phage clone, λbDt9000, hybridized to aDm2003 and not aDm2023; one other clone, λbDt9008, hybridized to both probes. More clones were screened using aDm2005 as a labeled probe; λbDt9100 and λbDt9200 were thus obtained. A last set of λ clones

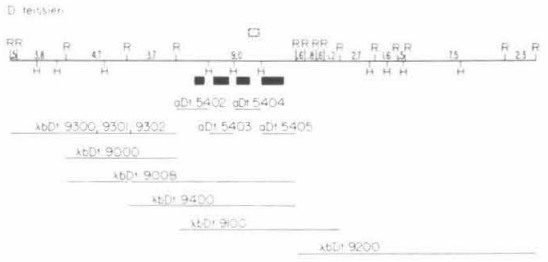


Fig. 7. Cloned sequences of *Drosophila teissieri*. A restriction endonuclease map of the cloned *D. teissieri* DNA is shown, with EcoRI (R) and HindIII (H) sites marked. The size of each EcoRI fragment in kilobase pairs is indicated. The hatched bar above the map shows the location of a repetitive DNA element; below the map are solid bars showing the location of the restriction fragments of the cloned DNA that hybridize to cDNA made from *D. teissieri* salivary gland poly(A)$^+$ RNA. Below this are lines showing the *D. teissieri* DNA represented in each of the clones used

was probed with fDm2056 and fDm2057 on duplicate plaque lift filters. λbDt9300, 9301, and 9302 were selected by both probes. The restriction maps of all of the clones were consistent with the map shown in Fig. 7, with three exceptions. One exception is the appearance of two EcoRI fragments (3.4 kb and 3.5 kb) at the right end of the insert of λbDt9000. These fragments are not hybridized by *D. melanogaster* 68C clones, and do not correlate with the fragments found in the analogous locations in λbDt9008, λbDt9400, and λbDt9100. We therefore conclude that these fragments were ligated to the 68C-homologous DNA in λbDt9000 during the λ clone construction, and do not represent the genomic DNA of the *D. teissieri* 68C-equivalent region. The second exception is the existence of an additional 0.7 kb of DNA, including a BglII site, in λbDt9100 and centered 1–2 kb to the left of the EcoRI site marking the right end of the 9.0-kb EcoRI fragment. This additional DNA is not present in λbDt9008 or λbDt9400, and may represent a polymorphism found in the population of *D. teissieri* flies from which the DNA was obtained. Finally, there is a single SalI site present in λbDt9100 but absent in the overlapping region of λbDt9200 (see Fig. 4). The usual battery of tests to determine if the restriction map derived from the clones was the same as that in the *D. teissieri* genome was applied: λbDt9200, λbDt9300, and λbDt9400 were each ^{32}P-labeled and used as hybridization probes of both *D. teissieri* and *D. melanogaster* genomic DNA that had been digested with the restriction endonucleases BamHI, EcoRI, HindIII, or SalI and then subjected to electrophoresis in an agarose gel and blotted to a nitrocellulose filter. λbDt9300 and λbDt9200 hybridized to fragments of the expected sizes in both *D. teissieri* and *D. melanogaster*, although with some enzymes both probes showed faint extra bands in the *D. teissieri* lanes. This is probably due to the presence of restriction-fragment-length polymorphism in the *D. teissieri* fly population from which the DNA was derived. λbDt9400 gave a highly repetitive signal (a dark smear with numerous discrete bands superimposed upon it) when hybridized to each of the four *D. teissieri* restriction digest lanes, indicating the presence of a repetitive DNA element in this clone. The *D. melanogaster* lanes did not show a repetitive pattern; rather, they showed fragments of the sizes found in the *D. melanogaster* 68C glue puff. The λbDt9400 repetitive element was localized to the position shown in Fig. 7 by hybridization of ^{32}P-labeled genomic DNA from *D. teissieri* to a blot filter with lanes of λbDt9400 DNA digested with both PvuII and XbaI. The only strongly labeled band was at the position of a 0.7-kb PvuII–XbaI fragment (see Fig. 8).

RNA-Homologous Regions of the Cloned Sequences

To find out if the 68C-equivalent regions in the species other than *D. melanogaster* contain DNA sequences that code for abundant polyadenylated third instar larval salivary gland RNAs, ^{32}P-labeled cDNA corresponding to the salivary gland poly(A)$^+$ RNA of each species was produced using oligo(dT) primers and reverse transcriptase. This labeled cDNA was than annealed to gel blot filters containing various restriction digests of cloned DNA from the same species. The DNA sequences hybridized by the cDNA were detected by autoradiography, and the sizes of the fragments hybridized indicated the DNA sequences that might code for RNA in each species. The restriction endonucleases used and the results are shown in Fig. 8. Since the reverse transcriptase reaction was performed in limiting amounts of one nucleotide (the labeled one, dCTP), it is unlikely that each RNA transcript was fully copied into cDNA. Rather, the 3′ ends are likely to be relatively overrepresented in the labeled cDNA; thus the sites indicated in the figure may not show the coding position of the 5′ end of each RNA. The results clearly show that the DNA homologous to the glue genes of *D. melanogaster* hybridizes to salivary gland cDNA in each species. It was also found that while *D. melanogaster, D. simulans,* and *D. erecta* each have three noncontiguous DNA regions hybridized by the cDNA, *D. yakuba* and *D. teissieri* each have four. It thus appears that the *D. yakuba* and *D. teissieri* 68C-homologous regions may contain an extra gene as compared with the other species, and therefore that the 68C gene family may have changed in size since the divergence of the species under study. This evidence alone does not exclude the possibility that the extra RNA-coding region in *D. yak-*

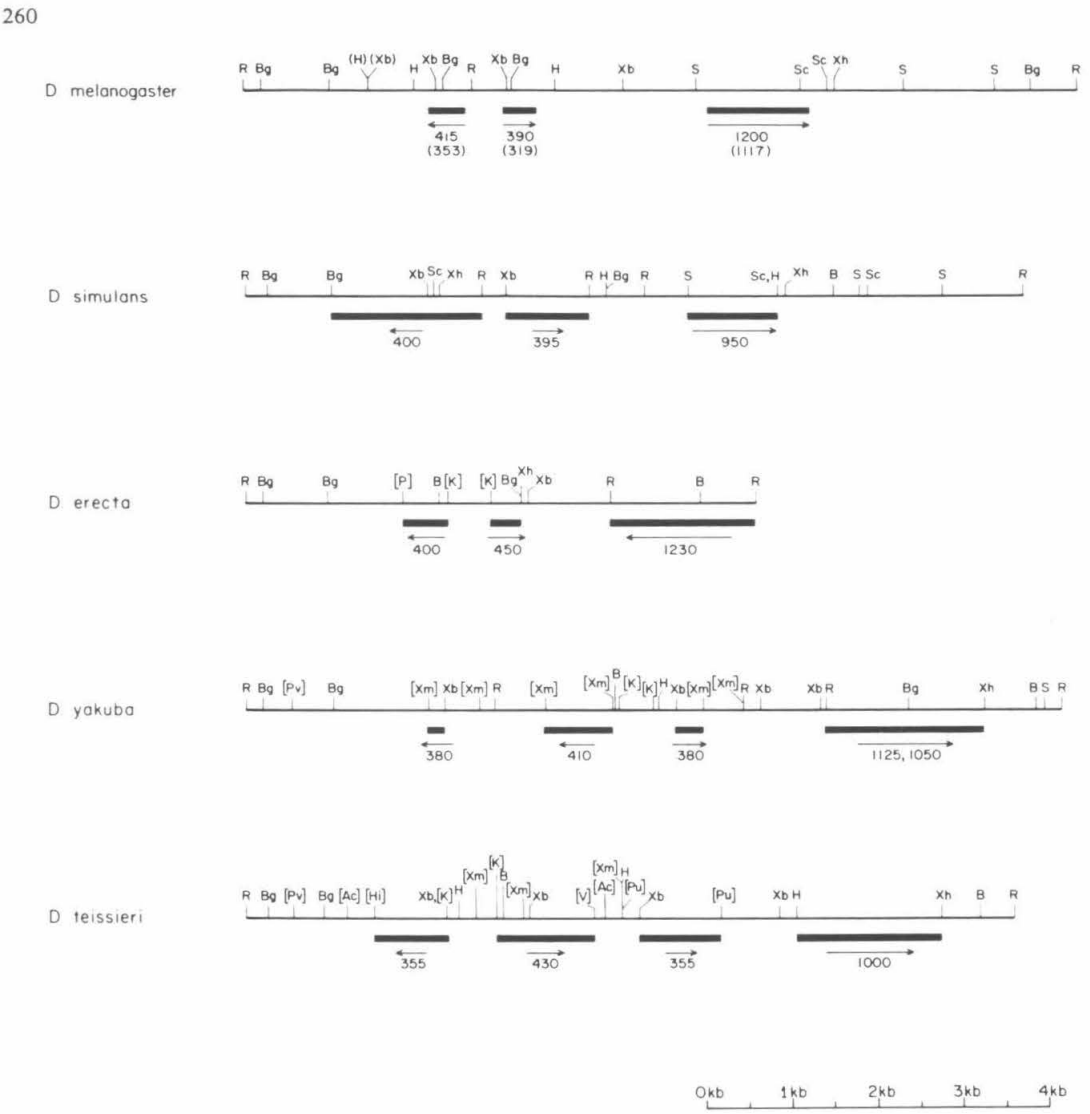


Fig. 8. Detailed restriction maps of 68C-homologous sequences coding for abundant third instar salivary gland RNAs. Restriction enzyme abbreviations are those used in Fig. 4, with the following additions: AccI (Ac), HincII (Hi), KpnI (K), PstI (P), PvuII (Pu), PvuI (Pv), EcoRV (V), and XmnI (Xm). Sites in parentheses are used as in Fig. 4. Sites in brackets indicate that only a subset of the sites recognized by the indicated enzyme are shown. The maps are aligned by the EcoRI site at the left edge. Filled bars below the maps indicate those restriction fragments that hybridize to ^{32}P-cDNA, as in Fig. 4. Arrows indicate the sizes of the RNAs hybridized by each of these regions and the direction of transcription of each of these RNAs. Below each arrow the size of each poly(A)$^+$ RNA is expressed in nucleotides. For *D. melanogaster* the sizes of the RNAs were determined using single-stranded DNA size standards. The extent of each *D. melanogaster* RNA was derived from DNA sequencing results (Garfinkel et al. 1983) and the predicted size of each mRNA [minus any poly(A) tail] is shown in parentheses. For the other species, the size of each transcript was determined using both single-stranded DNA and the *D. melanogaster* 68C RNAs as size standards. Two bands of approximately equal intensity were observed for the largest RNA of *D. yakuba*. This may be due to allelic variation in the *D. yakuba* stock used

uba and *D. teissieri* is due to entry of a new intervening sequence into a preexisting glue gene. This possibility would be excluded by finding that RNAs of different sizes are coded for by each of the four RNA regions in *D. yakuba* and *D. teissieri*, or by determining that all adjacent RNA-hybridized regions are transcribed in opposite directions.

The RNA sizes were determined by subjecting total RNA from third instar salivary glands of each species to electrophoresis in agarose–formaldehyde gels and then transferring the RNA to nitrocellulose filters by blotting. The resulting filters were hybridized with a ^{32}P-labeled restriction fragment from cloned DNA derived from the appropriate species,

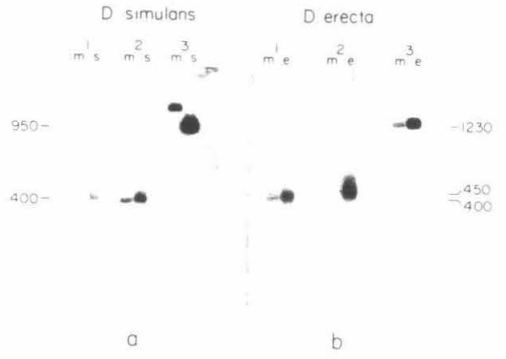


Fig. 9a,b. Hybridization of cloned probes to species RNAs. **a** *D. simulans* nick-translated probes hybridized to *D. melanogaster* (m) and *D. simulans* (s) total third instar salivary gland RNA. Probes used were (1) fDs121, (2) fDs122, and (3) fDs124. Both the fDs122 and fDs124 probes cross-hybridized to the expected *D. melanogaster* RNAs (sgs-7 and sgs-3, respectively). The lack of cross-hybridization by the fDs121 probe to any *D. melanogaster* RNA is probably due to the low level of the sgs-8 transcript produced by the OR16f *D. melanogaster* strain used (Crowley and Meyerowitz 1984). Numbers on the side indicate the lengths, in nucleotides, of the *D. simulans* RNAs. **b** *D. erecta* end-labeled probes hybridized to *D. melanogaster* (m) and *D. erecta* (e) total third instar salivary gland RNA. Probes used were (1) EcoRI insert of λeDe5020, (2) EcoRI insert of λeDe5021, and (3) EcoRI insert of fDe009. The probe containing the large RNA-coding region (fDe009) shows noticeable cross-hybridization to the sgs-3 transcript of *D. melanogaster*. The extent of cross-hybridization appears to be less than that in the similar *D. simulans* vs *D. melanogaster* experiment in **a**. The cross-hybridization observed using the λeDe5020 probe is to the sgs-7 transcript of *D. melanogaster* as determined by the size of the RNA, rather than to the expected sgs-8 RNA. It is not known if this indicates inversion of the region coding for the small RNAs of *D. erecta* in comparison with that of *D. melanogaster*, in addition to the *D. erecta* inversion that includes the gene coding for the large sgs-3 homologous RNA. Numbers on the side indicate the lengths, in nucleotides, of the *D. erecta* RNAs

Table 1. λgt10 subclones used for transcription direction mappings

Clone	Source
D. erecta	
λeDe5020	2.4-kb *Eco*RI–*Kpn*I fragment of fDe008
λeDe5021	1.4-kb *Kpn*I–*Eco*RI fragment of fDe008
D. yakuba	
λeDy5120	2.9-kb *Eco*RI insert of qDy5110
λeDy5121	1.5-kb *Eco*RI–*Kpn*I fragment of qDy5111
λeDy5122	1.1-kb *Kpn*I–*Eco*RI fragment of qDy5111
λeDy5123	2.7-kb *Eco*RI insert of qDy5113
D. teissieri	
λeDt5420	2.5-kb *Eco*RI–*Hind*III insert of aDt5402
λeDt5421	1.9-kb *Hind*III insert of aDt5403
λeDt5422	2.0-kb *Hind*III insert of aDt5404
λeDt5423	2.6-kb *Hind*III–*Eco*RI insert of aDt5405

Where necessary, *Eco*RI linkers were ligated onto the blunt-ended fragment. The fragments were all cloned into the *Eco*RI site of λgt10

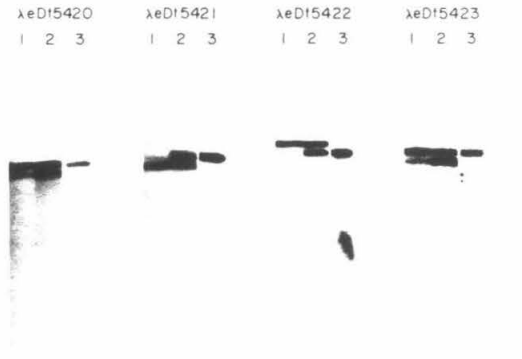


Fig. 10. Transcription direction mapping of *D. teissieri* RNAs. Three hundred nanograms of each of the indicated λgt10 subclones (see Table 1) was denatured with 0.1 N NaOH at 37°C for 10 min, loaded onto a wide lane of a 0.4% agarose gel [2 mM EDTA, 40 mM Tris-acetate (pH 8.3)], and separated by electrophoresis at 0.75 V cm^{-1} for approximately 24 h. Since each of the λgt10 subclones used contains only a single RNA-coding block, the direction of transcription of each of the genes can be determined by observing which of the two separated strands hybridizes to a single-stranded cDNA probe synthesized from poly(A)$^+$ third instar salivary gland RNA. It is also necessary to determine the 5′-to-3′ orientation of the insert DNA of each of the separated strands relative to the map shown in Fig. 8. This was done utilizing a restriction fragment of DNA homologous to the λgt10 insert DNA labeled at only one of its 3′ ends. This probe hybridizes to the strand that has a 5′-to-3′ direction opposite to that of the end-labeled strand of the probe. After electrophoresis, the gel was denatured in 1.5 M NaCl, 0.5 M NaOH for 1 h and then neutralized in 1 M Tris–HCl (pH 8.0), 1.5 M NaCl for 1 h. The gel was then blotted to a nitrocellulose filter. Each lane of the baked filter was then cut into three strips. The leftmost strip (strip 1) was hybridized with a probe labeled at a single 3′ end. The middle strip (strip 2) was hybridized with a probe labeled at both of its 3′ ends and was used to register the location of both of the separated strands. The rightmost strip (strip 3) was hybridized with the cDNA probe. The probes used and the transcription directions obtained are shown in Table 2 and in Fig. 8

from each of the RNA-coding blocks. Adjacent to each RNA lane on each blot filter was *D. melanogaster* salivary gland RNA included to detect cross-hybridization between the DNA of each species and the specific glue RNAs of *D. melanogaster*. Single-stranded DNA size standards were also included in each gel. Autoradiograms from the *D. simulans* and *D. erecta* experiments are shown in Fig. 9. The sizes of the RNAs are shown in Fig. 8. The cloned DNA probes used in these RNA blot experiments often hybridized weakly to the *D. melanogaster* 68C glue RNAs that were in the lanes adjacent to the strongly labeled RNAs from the same species as the probe (Fig. 9). Thus, the *D. melanogaster* glue RNAs appear to be homologous to, though quite diverged from, the similar RNAs of each of the other species.

The transcription direction of each of the RNAs

Table 2. Summary of transcription direction mapping experiments

Species	Single-end probe[a]	Strand hybridized by end-labeled probe	Strand hybridized by cDNA	Direction[b]
D. simulans	2.2-kb *Eco*RI*–*Xho*I of fDs121	Top	Bottom	R to L
	1.0-kb *Xba*I–*Eco*RI* of fDs122	Top	Bottom	L to R
	2.2-kb *Eco*RI*–*Bam*HI of fDs124	Top	Top	L to R
D. erecta	1.4-kb *Bgl*II–*Eco*RI* of λeDe5020[c]	Top	Top	R to L
	0.4-kb *Eco*RI*–*Xba*I of λeDe5021[c]	Top	Top	L to R
	1.0-kb *Eco*RI*–*Bam*HI of fDe009	Bottom	Top	R to L
D. yakuba	1.9-kb *Bgl*II–*Eco*RI* of qDy5110	Top	Top	R to L
	1.4-kb *Eco*RI*–*Bam*HI of qDy5111	Bottom	Top	R to L
	0.9-kb *Hin*dIII–*Eco*RI* of qDy5111	Top	Bottom	L to R
	1.8-kb *Eco*RI*–*Xho*I of qDy5113	Top	Top	L to R
D. teissieri	1.6-kb *Bgl*II–*Hin*dIII* of aDt5402	Top	Top	R to L
	1.5-kb *Bam*HI–*Hin*dIII* of aDt5403	Bottom	Top	L to R
	0.9-kb *Pvu*II–*Hin*dIII* of aDt5404	Top	Bottom	L to R
	1.7-kb *Hin*dIII*–*Xho*I of aDt5405	Top	Top	L to R

[a] Asterisk indicates 3′ end-labeled site

[b] "R to L" indicates right to left; "L to R" indicates left to right. See Fig. 8

[c] *Eco*RI ends of λgt10 subclones were generated during cloning.

was also determined. The strategy used is shown in Fig. 10. The DNA fragments and probes used, and the results obtained, are listed in Tables 1 and 2 and depicted in Fig. 8. The transcription direction of the rightmost *D. erecta* RNA was also determined by DNA sequencing to confirm that it is indeed inverted relative to the orientation of transcription of the similar RNA from all of the other species. The EcoRI fragment coding for this RNA was labeled at the 3′ ends of both strands, using the large proteolytic fragment of *E. coli* DNA polymerase I to add ^{32}P-labeled residues. After digestion with BamHI, the larger of the two resulting fragments (see Figs. 4 and 8), now labeled at only one end, was sequenced using the chemical sequencing method of Maxam and Gilbert (1980). The sequence obtained showed clearly that the DNA adjacent to the EcoRI site is homologous to the 3′ end of the *D. melanogaster Sgs-3* gene sequence and that it is indeed inverted relative to the *D. melanogaster* orientation. This sequence will be presented at a later date (C. Martin and E. Meyerowitz, work in progress). The results from all of the hybridizations show that adjacent coding regions in all of the species code for RNAs of different sizes or of opposite transcription directions. This eliminates the possibility that the enlargement of the RNA coding region of *D. yakuba* and *D. teissieri* is due solely to addition of one or more new intervening sequences.

Discussion

Several conclusions are possible from the results presented. The first is that the five closely related *Drosophila* species studied all do contain DNA sequences hybridized by the 68C glue gene cluster of *D. melanogaster*, and that in all of the species these sequences contain DNA that is transcribed to give several abundant polyadenylated RNA species in third instar larval salivary glands. This is consistent with previous work showing that *D. simulans*, *D. yakuba*, and *D. teissieri* all have puffs similar to that found at 68C in *D. melanogaster*, at the analogous chromosomal position (Ashburner and Lemeunier 1972; Ashburner and Berendes 1978). Although all of the species have a 68C-homologous gene cluster, it is clear that the 68C gene family has evolved since the divergence of the species studied. That the DNA sequence of the genes has changed is evidenced by the difference in restriction endonuclease sites within the individual genes, by the different sizes of the RNAs in the different species, and by the weak cross-hybridization between the genes of the various other species to the *D. melanogaster* glue RNAs. Divergence is also shown by the difference in the number of genes (some of which may in fact be pseudogenes) in the species, and by the inversions of certain of the genes relative to the others. This divergence includes more than just the RNA-coding DNA; the entire region of chromosomal DNA that includes the gene family and all of the flanking sequences is remarkably different in the *Drosophila* species studied. This is shown both by the thermal elution of labeled *D. melanogaster* DNA probes from the DNA representing this region in genome blot experiments and by the virtual absence of any conserved restriction endonuclease sites in the entire region, as shown in Figs. 4 and 8. In striking contrast is the adjacent DNA to the left (toward the telomere in *D. mela-*

nogaster). This set of sequences shows an extraordinary degree of conservation from one species to the other for a distance of at least 13 kb, as demonstrated both by thermal elution of cross-species hybrids on genome blots and by a high degree of restriction site conservation.

A quantitative estimate of nucleotide divergence can be obtained from comparison of restriction maps (Nei and Li 1979). Using Equation (16) of these authors and comparing *D. melanogaster* and *D. simulans* in both the conserved region (where *D. melanogaster* has 18 restriction sites and *D. simulans* 19, with 18 shared) and the RNA-coding region (where *D. melanogaster* has 13 sites and *D. simulans*, 15, with 6 apparently shared) it can be estimated that the sequences in the conserved regions of these species are diverged by less than 0.5%, whereas in the RNA-coding region the mean frequency of nucleotide substitution per position is about 18%. Thus, at least at the 68C glue locus, evolutionary divergence occurs at very different rates in adjacent blocks of chromosomal DNA sequence.

It also appears that the processes leading to the observed divergence are different in the adjacent regions. In all five species the spacing between shared restriction sites is the same in most of the conserved region, with different unshared sites appearing or disappearing against an otherwise constant background. This implies that the primary process in divergence is single-nucleotide substitution, although of course tiny deletions or substitutions of small blocks of nucleotides would not have been detected in our experiments. In contrast, the rapidly evolving RNA-coding region is subject to insertions, deletions, inversions, and an extraordinary number of apparent single-site changes. In fact, these processes are so evident that it is meaningless to try to estimate the levels of nucleotide divergence in this region between any of the species except for the closely related siblings *D. melanogaster* and *D. simulans*, since the available methods for making such calculations (Nei and Li 1979; Engels 1981) assume that only single-nucleotide substitutions have occurred.

That *Drosophila* genomes contain large interspersed blocks of rapidly evolving and more slowly evolving DNA has been predicted from the results of thermal elution studies on interspecies hybrids of single-copy genomic DNA (Zwiebel et al. 1982). The work reported here confirms this in a specific instance, and also shows that the rapidly evolving DNA can code for messenger RNAs. The possibility that mammalian genomes may also contain interspersed blocks of DNA with different evolutionary rates has been shown by analysis of the mouse major histocompatibility complex (Hood et al. 1983). The evolutionary mechanisms that result in disparate rates of sequence divergence in adjacent domains of chromosomal DNA are unknown. Whether the blocks of rapidly and slowly evolving DNA bear any relation to the bands and interbands of the polytene chromosomes, or to other chromatin features, is also unknown.

What is clear is the use to which this divergence pattern can be put in studying the relation of DNA sequence to gene regulation in the 68C glue gene cluster. The RNA-coding region seems to serve the same function and to be transcribed in response to the same tissue- and time-specific signals in all of the studied species. Proof that the DNA of each species does indeed respond to identical intracellular signals will be sought in interspecies P-factor-mediated transformation experiments. It is already known that the cloned *Sgs-3* gene of *D. melanogaster* (including none of the conserved region sequences) is expressed normally after P-factor-mediated reintroduction to the *D. melanogaster* genome (Crosby 1983; M. Crosby and E. Meyerowitz, work in progress). If the genes of the other species are expressed normally when integrated into the *D. melanogaster* genome, DNA sequencing studies should reveal which, if any, regions of the genic and intergenic sequences have been conserved in evolution and thus may be functionally significant, since conserved islands of sequence should be evident against the remarkably diverged background of the surrounding DNA. These studies will also show if all of the genes are capable of coding for proteins, and if the proteins coded for are similar to the 68C glue polypeptides of *D. melanogaster*. Sequencing studies should also point to the nature of the events that result in extremely rapid divergence in defined chromosomal segments, and may show any special features of the DNA sequences found at the sharp boundary between the rapidly and slowly evolving regions.

Acknowledgments. We would like to acknowledge Mark Garfinkel's construction and screening of a portion of the *D. teissieri* library, and to thank Lynn Crosby, Tom Crowley, Mark Garfinkel, Pete Mathers, and Bob Pruitt for their comments on the manuscript. This work was supported by NIH grant GM28075 to E.M.M. C.H.M. was supported by a National Science Foundation predoctoral fellowship.

References

Ashburner M (1967) Patterns of puffing activity in the salivary gland chromosomes of *Drosophila*. I. Autosomal puffing patterns in laboratory stock of *Drosophila melanogaster*. Chromosoma 21:398–428

Ashburner M (1973) Sequential gene activation by ecdysone in polytene chromosomes of *Drosophila melanogaster*. I. Dependence upon ecdysone concentration. Dev Biol 35:47–61

Ashburner M (1974) Sequential gene activation by ecdysone in polytene chromosomes of *Drosophila melanogaster*. II. The

effects of inhibitors of protein synthesis. Dev Biol 39:141–157

Ashburner M, Berendes HD (1978) Puffing of polytene chromosomes. In: Ashburner M, Wright TRF (eds) The genetics and biology of *Drosophila*. vol 2b. Academic Press, London pp 315–395

Ashburner M, Lemeunier F (1972) Patterns of puffing activity in the salivary gland chromosomes of *Drosophila*: VII. Homology of puffing patterns on chromosome arm 3L in *D. melanogaster* and *D. yakuba*, with notes on puffing in *D. teissieri*. Chromosoma 38:283–295

Ashburner M, Richards G (1976) Sequential gene activation by ecdysone in polytene chromosomes of *Drosophila melanogaster*. III. Consequences of ecdysone withdrawal. Dev Biol 54:241–255

Blattner FR, Williams BG, Blechl AE, Thompson KD, Faber HE, Furlong LA, Grunwald DJ, Kiefer DO, Moore DD, Schumm JW, Sheldon EL, Smithies O (1977) Charon phages: safer derivatives of bacteriophage lambda for DNA cloning. Science 196:161–169

Bock IR, Wheeler MR (1972) The *Drosophila melanogaster* species group. University of Texas Publication 7213. University of Texas, Austin

Bolivar F (1978) Construction and characterization of new cloning vehicles. III. Derivatives of plasmid pBR322 carrying unique EcoRI sites for selection of EcoRI generated recombinant DNA molecules. Gene 4:121–136

Bolivar F, Rodriguez RL, Greene PJ, Betlach MC, Leyneker HL, Boyer HW, Crossa JH, Falkow S (1977) Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gene 2:95–113

Collins J, Hohn B (1978) Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage λ heads. Proc Natl Acad Sci USA 75:4242–4246

Crosby MA (1983) Determination of sequences necessary for regulated expression of the Sgs-3 gene. In: Caltech Biology Annual Report. California Institute of Technology, Pasadena, pp 58–59

Crowley TE, Meyerowitz EM (1984) Steroid regulation of RNAs transcribed from the *Drosophila* 68C polytene chromosome puff. Dev Biol 102:110–121

Crowley TE, Bond MW, Meyerowitz EM (1983) The structural genes for three *Drosophila* glue proteins reside at a single polytene chromosome puff locus. Mol Cell Biol 3:623–634

Davis RW, Botstein D, Roth JR (1980) Advanced bacterial genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

Denhardt DT (1966) A membrane-filter technique for the detection of complementary DNA. Biochem Biophys Res Commun 23:641–646

Dworniczak B, Kobus S, Schaltmann-Eiteljorge K, Pongs O (1983) Ecdysterone, ecdysterone receptor, and chromosome puffs. In: Roy AK, Clark JH (eds) Gene regulation by steroid hormones II. Springer-Verlag, New York, pp 79–91

Engels WR (1981) Estimating genetic divergence and genetic variability with restriction endonucleases. Proc Natl Acad Sci USA 78:6329–6333

Fraenkel G, Brookes VJ (1953) The process by which the puparia of many species of flies become fixed to a substrate. Biol Bull 105:442–449

Garfinkel MD, Pruitt RE, Meyerowitz EM (1983) DNA sequences, gene regulation and modular protein evolution in the *Drosophila* 68C glue gene cluster. J Mol Biol 168:765–789

Gronemeyer H, Pongs O (1980) Localization of ecdysterone on polytene chromosomes of *Drosophila melanogaster*. Proc Natl Acad Sci USA 77:2108–2112

Hood L, Steinmetz M, Malissen B (1983) Genes of the major histocompatibility complex of the mouse. Annu Rev Immunol 1:529–568

Lemeunier F, Ashburner M (1976) Relationships within the *melanogaster* species subgroup of the genus *Drosophila* (*sophophora*). II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences. Proc R Soc Lond [Biol] 193:275–294

Lemeunier F, Dutrillaux B, Ashburner M (1978) Relationships within the *melanogaster* subgroup species of the genus *Drosophila* (*sophophora*). Chromosoma 69:349–361

Lewis EB (1960) A new standard food medium. Dros. Inf. Ser. 34:117–118

Lis JT, Prestidge L, Hogness DS (1978) A novel arrangement of tandemly repeated genes of a major heat shock site in *D. melanogaster*. Cell 4:901–919

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

Maxam AM, Gilbert W (1980) Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol 65: 499–560

McDonell MW, Simon MN, Studier FW (1977) Analysis of restriction fragments of T7 DNA and determination of molecular weights by electrophoresis in neutral and alkaline gels. J Mol Biol 110:119–146

Meyerowitz EM, Hogness DS (1982) Molecular organization of a *Drosophila* puff site that responds to ecdysone. Cell 28: 165–176

Murray NE, Brammar WJ, Murray K (1977) Lambdoid phages that simplify the recovery of in vitro recombinants. Mol Gen Genet 150:53–61

Nei M, Li W-H (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273

Peacock C, Dingman CW (1968) Molecular weight estimation and separation of ribonucleic acid by electrophoresis in agarose–acrylamide composite gels. Biochemistry 7:668–674

Pelling C (1964) Ribonukleinsäure-Synthese der Riesenchromosomen. Autoradiographische Untersuchungen an *Chironomus tentans*. Chromosoma 15:71–122

Rigby PWJ, Dieckmann M, Rhodes C, Berg P (1977) Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. J Mol Biol 113:237–251

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503–517

Sternberg N, Tiemeier D, Enquist L (1977) In vitro packaging of a λ Dam vector containing EcoRI fragments of *Escherichia coli* and phage P1. Gene 1:255

Sturtevant AH (1919) A new species closely resembling *Drosophila melanogaster*. Psyche (Stuttg) 26:153–155

Zwiebel LJ, Cohn VH, Wright DR, Moore GP (1982) Evolution of single-copy DNA and the ADH gene in seven drosophilids. J Mol Evol 19:62–71

Chapter 3

Characterization of the Boundaries between Adjacent Rapidly and Slowly Evolving Genomic Regions in *Drosophila*

Christopher H. Martin and Elliot M. Meyerowitz
Division of Biology
California Institute of Technology
Pasadena, CA 91125

(published in the *Proceedings of the National Academy of Sciences, USA*)

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.

1. Introduction

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 (Meyerowitz & Hogness, 1982; Crowley *et al.*, 1983; Garfinkel *et al.*, 1983). These genes are expressed abundantly in the salivary glands of third instar larvae and are controlled by the steroid hormone ecdysterone (Ashburner, 1973, 1974; Ashburner & Richards, 1976; Crowley & Meyerowitz, 1984). 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 (Lemeunier *et al*, 1986). Comparison of the restriction maps of these cloned sequences revealed two adjoining regions with dramatically different levels of homology (Meyerowitz & Martin, 1984). That this genomic segment contains adjacent regions that have evolved at different rates is confirmed by experiments that demonstrate very different melting temperature depressions (Δt_ms) of inter-species hybrids of restriction fragments from each of the two adjoining regions (Meyerowitz & Martin, 1984). The relatively nonconserved region, which is ≈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 (Meyerowitz & Martin, 1984). In contrast, the conserved

region ,which consists of ≥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 *melanogaster* species.

2. MATERIALS AND METHODS

(a) 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}P-labelled nucleoside triphosphates were obtained from Amersham. Deoxynucleotides and dideoxynucleotides were obtained from Pharmacia.

(b) Clones for DNA sequencing.

Clones for sequencing were prepared by inserting fragments from previously cloned *Drosophila* sequences into M13mp18 and M13mp19 (Norrander *et al.*, 1983); M13 vectors were a gift of G. Siu. The *D. melanogaster* clones were constructed by inserting the 1.95 kb *Eco*RI-*Hind*III restriction fragment from aDm2003 (Meyerowitz & Martin, 1984) into vectors M13mp18 and M13mp19. For *D. erecta*, the 2.25 kb *Eco*RI-*Bam*HI restriction fragment from clone fDe009 (Meyerowitz & Martin, 1984) was inserted into both M13 vectors. For *D. yakuba*, the 2.9 kb *Eco*RI restriction fragment from qDy5110

(Meyerowitz & Martin, 1984) was cloned in both orientations into M13mp18. Cloning was done by standard procedures described by Davis *et al.* (1980) and Maniatis *et al.* (1982).

(c) Sequencing.

DNA sequencing was performed by the dideoxy chain-termination method of Sanger *et al.*(1977). 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.* (1986). All sequences were determined on both strands.

(d) Computer Analysis.

DNA sequences were analyzed using programs written by one of the authors (CHM) for an IBM PC-XT computer. DNA sequences were aligned using the algorithm of Gotoh (1982) as implemented by R. Pruitt on an Apple Macintosh computer.

3. RESULTS

The border between regions of high and low conservation was located by inspection of the restriction maps of the regions containing and adjacent to the cloned glue gene clusters. The broken vertical line in Figure 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 (1984) 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 used is shown in Figure 2. All sequences start at the *Eco*RI site (R) that lies at least 1000 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 Figure 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 constant. This is apparent from the similar frequencies of deletions observed on both sides of the boundary.

Furthermore, near the boundary the ≈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 ± 0.5% A+T (all values are ± SD) vs. an average of 67.5 ± 0.2% A+T in the remaining 1153 bases of the conserved region, and an average of 61.0 ± 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* (Laird & McCarthy, 1968); this A+T-rich region tends to contain stretches of adenines and thymines as opposed to interspersed adenines and thymines. In the three species, ApA and TpT dinucleotides make up 50.4 ± 0.8% of this region, whereas ApT and TpA dinucleotides comprise only 19.4 ± 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 this region is only 0.7 ± 0.6% vs. 2.9 ± 0.8% in the remaining conserved 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 ± 0.03. This is comparable to the ratio of 0.75 seen in the 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 (Bodmer & Ashburner, 1984). 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.

4. 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 (Figure 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) (Crosby & Meyerowitz, 1986a). (*iii*) An experiment designed to saturate the region surrounding the 68C glue gene cluster for lethal and semi-lethal mutations did not reveal any such mutations in the conserved region (Crosby & Meyerowitz, 1986b). 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 (Crosby & Meyerowitz, 1986a). 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 (Efstratiadis *et al.*, 1980). 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 interspersion of 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 (Hunt *et al.*, 1981; Zwiebel *et al.*, 1982; Schulze & Lee, 1986). The experiments of Zweibel et al. (1982) 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 (Δt_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 (1986) 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 sequence 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.

ACKNOWLEDGEMENTS

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.

REFERENCES

Ashburner, M. & Richards, G. (1976) *Dev. Biol.* **54**, 241-255.

Ashburner, M. (1973) *Dev. Biol.* **35**, 47-61.

Ashburner, M. (1974) *Dev. Biol.* **39**, 141-157.

Bodmer, M. & Ashburner, M. (1984) *Nature* **309**, 425-430.

Crosby, M.A. & Meyerowitz, E.M. (1986a) *Dev. Biol.*,**118**, 593-607

Crosby, M.A. & Meyerowitz, E.M. (1986b) *Genetics* **112**, 785-802.

Crowley, T.E. & Meyerowitz, E.M. (1984) *Dev. Biol.* **102**, 110-121.

Crowley, T.E., Bond M.W. & Meyerowitz, E.M. (1983) *Mol. Cell. Biol.* **3**, 623-634.

Davis, R.W., Botstein, D. & Roth, J.R. (1980) *Advanced Bacterial Genetics* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).

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.

Garfinkel, M.D., Pruitt, R.E. & Meyerowitz, E.M. (1983) *J. Mol. Biol.* **168**, 765-789.

Gotoh, O. (1982) *J. Mol. Biol.* **162**, 705-708.

Hunt, J.A., Hall, T.J. & Britten, R.J. (1981) *J. Mol. Evol.* **17**, 361-367.

Laird, C.D. & McCarthy, B.J. (1968) *Genetics* **60**, 303-322.

Lemeunier, F. & Ashburner, M. (1984) *Chromosoma* **89**, 343-351.

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.

Maniatis, T., Fritsch, E.F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).

Meyerowitz, E.M. & Hogness, D.S. (1982) *Cell* **28**, 165-176.

Meyerowitz, E.M. & Martin, C.H. (1984) *J. Mol. Evol.* **20**, 251-264.

Norrander, J., Kempe, T. & Messing, J. (1983) *Gene* **26**, 101-106.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467.

Schulze, D.H. & Lee, C.S. (1986) *Genetics* **113**, 287-303.

Strauss, E., Kobori, J.A., Siu, G. & Hood, L.E. (1986) *Anal. Biochem.* **154**, 353-360.

Zwiebel, L.J., Cohn, V.H., Wright, D.R. & Moore, G.P. (1982) *J. Mol. Evol.* **19**, 62-71.

Table 1. Change in the conserved versus the nonconserved regions. 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.

sequences compared	% mismatch		% dels + mismatches		% deletions	
	consv	non-consv	consv	non-consv	consv	non-consv
D. mel vs. D. yak	3.2	18.2	4.8	19.0	1.7	1.6
D. yak vs. D. ere	1.9	19.6	2.7	19.7	0.8	1.5
D. ere vs. D. mel	2.6	23.9	4.2	23.0	1.8	1.9

Figure 1. Restriction maps of the cloned 68C-homologous sequences. All known *Bam*HI(B), *Bgl*II (Bg), *Eco*RI (R), *Hin*dIII (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 Meyerowitz & Martin, 1984). Sites in parentheses are present in some strains (*D. melanogaster*) or clones (*D. teissieri*) and not in others (Meyerowitz & Martin, 1984). 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)$^+$ RNA (Meyerowitz & Martin, 1984). 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.

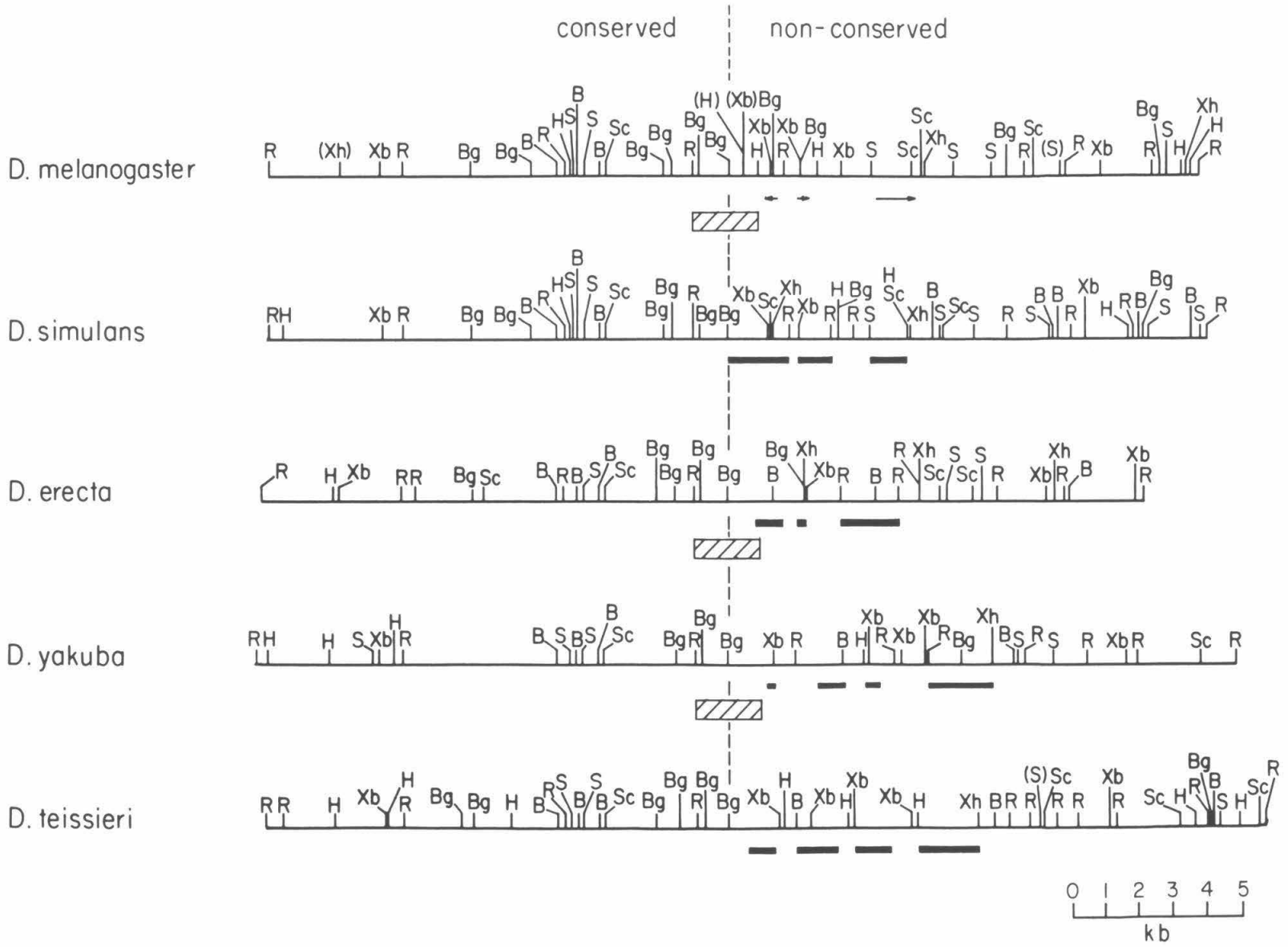

conserved
non-conserved
D. melanogaster
D. simulans
D. erecta
D. yakuba
D. teissieri
0 1 2 3 4 5
kb

Figure 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 start 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 Figure 1. *D. melanogaster*, Dm; *D. erecta*, De; *D. yakuba*, Dy.

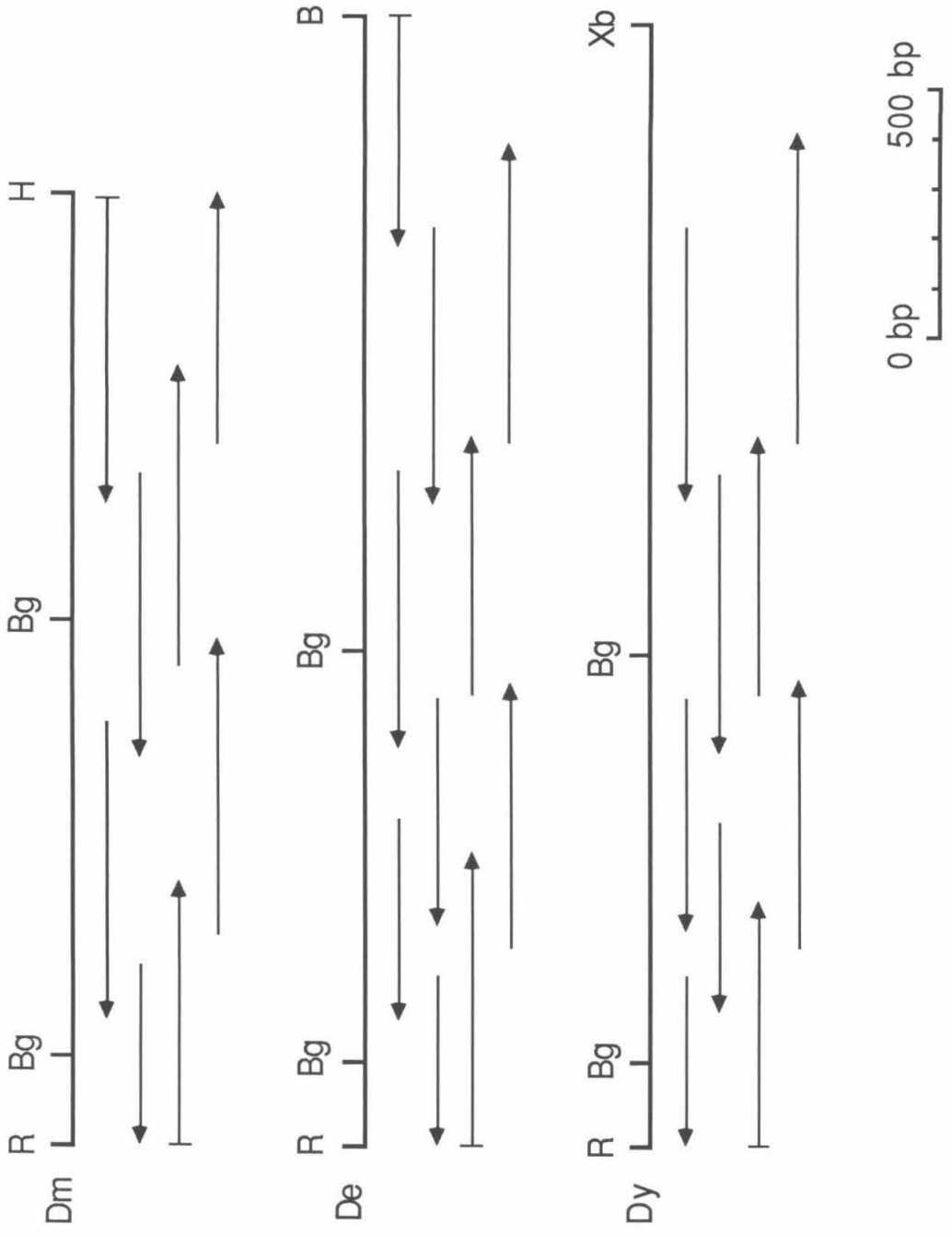
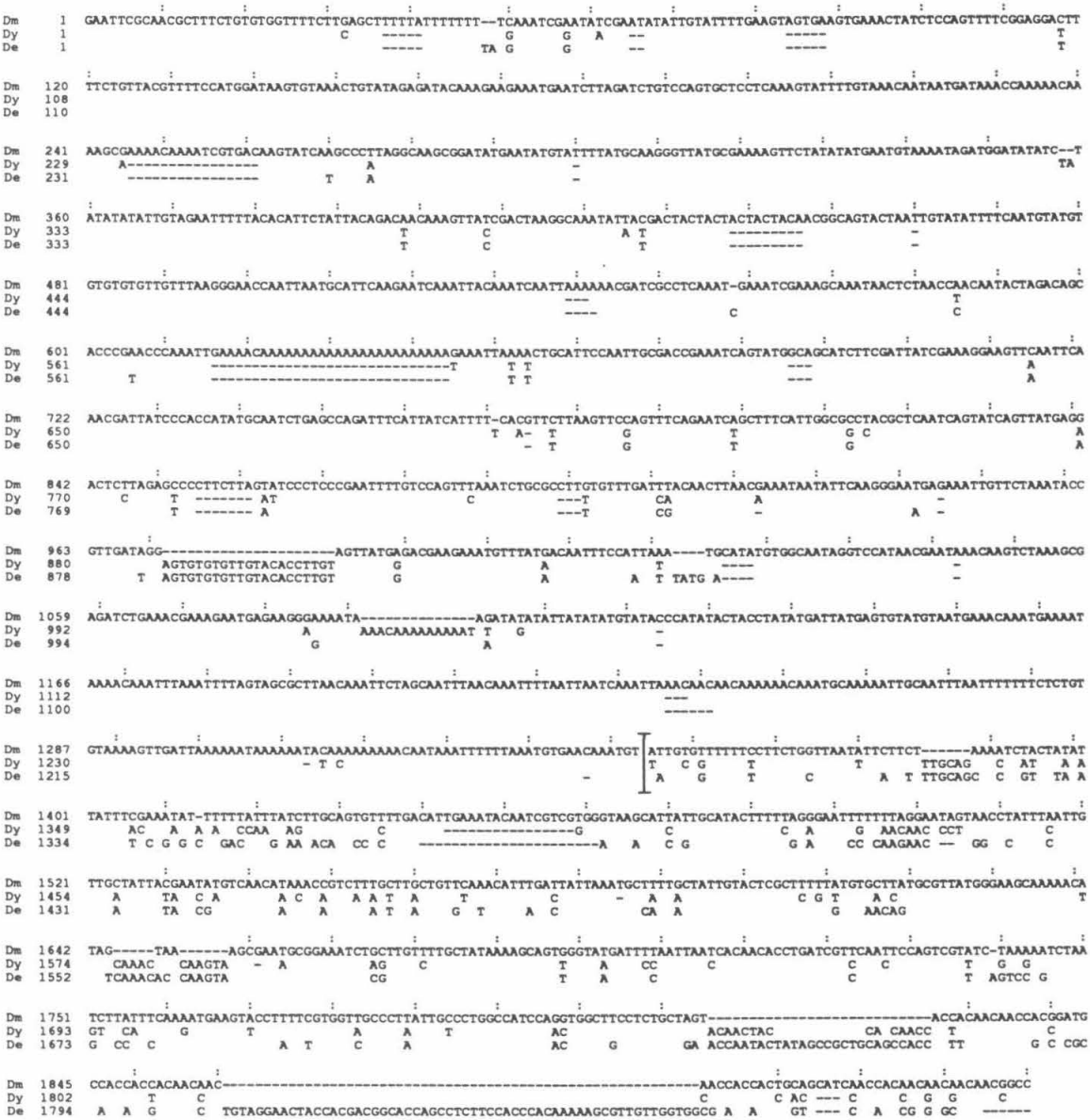

H
Bg
R
Bg
Dm
B
Bg
R
Bg
De
Xb
Bg
R
Bg
Dy
500 bp
0 bp

Figure 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 (1982). The mismatch penalty was 10, the start deletion penalty was 40, and the deleted base penalty was 5. The three-way alignment shown was generated by hand from the pair-wise alignments. Colons mark every tenth base in the *D. melanogaster* sequence. Spaces indicates 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.



Chapter 4

Evolution and Expression of the *Sgs-3* Glue Gene of *Drosophila*

Christopher H. Martin, Carol A. Mayeda and
Elliot M. Meyerowitz
Division of Biology
California Institute of Technology
Pasadena, CA 91125

(in press in the *Journal of Molocular Biology*)

1. Abstract

A cluster of three glue genes is present at chromosomal site 68C in the *Drosophila melanogaster* genome. In this study, we have used a comparative approach to investigate both the regulation and the evolution of the largest of these three genes, *Sgs-3*. The homologous genes from two related *Drosophila* species (*D. erecta* and *D. yakuba*) have been introduced into the *D. melanogaster* genome by P-factor mediated transformation. When the resulting transformant lines were assayed for expression of the introduced genes, near normal patterns of expression were seen. This demonstrates that the *cis*-acting regulatory sequences of the introduced *Sgs-3*-homologous glue genes are capable of effectively interacting with the transcriptional machinery of *D. melanogaster*. We have also determined the sequences of the *Sgs-3*-homologous glue genes from *D. simulans*, *D. erecta* and *D. yakuba*. These sequences were compared and used in two ways. The first was to locate conserved sequence elements in regions known to be involved in regulation of the gene. Several such elements were found; they represent potential sites of *cis*-acting regulatory sequences. Second, we looked at the evolution of the glue gene protein-coding regions. A very rapidly evolving central region of the protein coding sequences was found; this region contains a striking series of tandem repeats of a five amino acid sequence in all four

species. Also, a number of conserved aspects of the *Sgs-3*-homologous proteins were found; these features may be essential to their function as a glue.

2. Introduction

The salivary gland of *D. melanogaster* is the exclusive site for the production of a proteinaceous glue (Korge, 1977a). This glue, which consists of at least eight different polypeptides, serves to affix the animal to a solid surface for the duration of metamorphosis (Fraenkel & Brookes, 1953; Crowley *et al.*, 1983). Glue protein synthesis occurs throughout much of the third larval instar period of the animal's life (Beckendorf & Kafatos, 1976). The chromosomal locations of several of the glue genes are known to coincide with the sites of some of the prominent intermolt puffs which can be seen on the polytene salivary gland chromosomes of mid- to late-third instar animals (Korge, 1975; Akam *et al.*, 1978; Muskavitch & Hogness, 1980; Velissariou & Ashburner, 1980, 1981; Crowley *et al.*, 1983; Guild & Shore, 1984). One of these puffs, located on the left arm of the third chromosome at 68C, contains the genes coding for three of the glue proteins: sgs-8, sgs-7 and sgs-3 (Meyerowitz & Hogness, 1982; Crowley *et al.*, 1983). These genes are regulated by the steroid hormone ecdysterone, which appears to be necessary for both induction and shut-off of glue gene transcription (Hansson & Lambertsson, 1983; Crowley & Meyerowitz, 1984). These three glue genes are subject to strict tissue, time and hormonal regulation. These features make this glue gene cluster an appealing subject for the investigation of regulated gene expression.

Several studies have been aimed at defining the molecular limits of the sequences necessary for the proper regulation of the 68C glue genes. Many of these studies have

focused on the largest glue gene in the cluster, *Sgs-3* (Richards *et al.*, 1983; Bourouis & Richards, 1985; Meyerowitz *et al.*, 1985; Crosby & Meyerowitz, 1986; Vijay Raghavan *et al.*, 1986). Classical genetic experiments, utilizing chromosomal deletions and inversions, have limited the required sequences to a 20 kb region which includes the 5 kb glue gene cluster itself (Crosby & Meyerowitz, 1986). The use of P factor-mediated transformation has further delimited the necessary sequences: transformants carrying 2.27 kb of sequences upstream from the *Sgs-3* mRNA cap site express the gene normally. In contrast, transformants containing only 130 bp of upstream sequence express the introduced gene at 10 to 40 fold lower levels when compared to endogenous expression; however, the regulation of tissue and time of expression appear to be normal (Crosby & Meyerowitz, 1986; Vijay Raghavan *et al.*, 1986). Further recombinant DNA constructs, in which the *Sgs-3* upstream sequences are fused to a β-galactosidase gene, are also properly expressed with respect to tissue and time (Vijay Raghavan *et al.*, 1986). The *Sgs-3* glue gene sequences in these constructs extend to 948 bp beyond the mRNA start site. Thus, the *cis*-acting sequences sufficient for proper time and tissue of expression are located between -130 bp and 948 bp relative to the mRNA start site, while an element (or elements) necessary for high levels of expression lies within the region -2270 bp to -130 bp.

Aside from its regulatory features, there are several notable evolutionary aspects of the 68C glue gene cluster. The sequence of the glue gene cluster at 68C has been

determined (Garfinkel *et al.*, 1983). The two small gene products (sgs-7 and sgs-8, which are 74 aa and 75 aa in length, respectively) are related in amino acid sequence (47% amino acid identity). The large *Sgs-3* gene (307 aa in length) is related to the two small genes in amino acid sequence at both the amino and carboxy terminal regions. However, the large gene contains an additional central region which codes for a very threonine-rich region of the protein. This threonine-rich region contains 37 tandem repeats of a five amino acid sequence. The structure of the 68C-homologous glue gene cluster in four closely related *Drosophila* species is similar to that in *D. melanogaster*: there are two or three small genes and a single large gene present in each species (Meyerowitz & Martin, 1984). These glue gene clusters reside within a relatively rapidly evolving block of sequence that is adjacent to a region that evolves at a much slower rate; the boundary between the two regions is relatively abrupt and lies within 0.85 kb of the left end (as shown in Figure 1) of the glue gene cluster in *D. melanogaster* (Martin & Meyerowitz, 1986).

As an approach to learning more about both the regulation and the evolution of the glue gene cluster, we have undertaken a comparative study of the *Sgs-3*-homologous genes in four species of *Drosophila*. The four species used in this study, *D. melanogaster*, *D. simulans*, *D. erecta* and *D. yakuba*, are all members of the *melanogaster* species subgroup, which is one of eleven species subgroups defined for the *melanogaster* species group (Lemeunier, et al., 1986). The *Sgs-3*-like glue genes of *D. erecta* and *D. yakuba*

have been introduced into the *D. melanogaster* genome by P factor-mediated transformation. The ability of these constructs to be expressed in the foreign *D. melanogaster* background has been characterized. Additionally, the sequences of the *Sgs-3*-homologous glue genes from *D. simulans*, *D. erecta* and *D. yakuba*, along with upstream and downstream regions, have been determined. Together, these studies yield insights into the processes of evolution acting upon these glue genes and the regulatory elements required for their proper expression.

3. Materials and Methods

(a) *P-factor Transformations*

P-factor transformation experiments were performed using standard methods (Spradling & Rubin, 1982; Rubin & Spradling, 1982) and are described further in Crosby & Meyerowitz (1986). The host strain used in these transformations was *ry*^{506}. Transformants were selected by screening the progeny of injected individuals for *ry*$^{+}$ animals. Several lines were established from injections of each of the three constructs. The inserts were mapped to a specific chromosome using standard techniques. The lines were made homozygous by use of appropriate balancer chromosomes.

(b) *RNA gel blots*

RNA extraction, electrophoresis, blotting and probing were all done under standard conditions as described in Crosby & Meyerowitz (1986). Probes were removed from filters

for rehybridization by 3x five minute washes in boiling 0.01x SSPE, 0.1% SDS.

(c) *DNA Sequencing*

The majority of the DNA sequences presented in this paper were determined by the dideoxy chain termination method of Sanger *et al.* (1977). Custom oligodeoxynucleotides, used to prime sequencing reactions from sites in the interior of the cloned inserts, were obtained from the California Institute of Technology Division of Biology Microchemical Facility. These primers, ranging in length from 15 to 22 nucleotides, were purified and used as described in Strauss et al. (1986).

The clones used for dideoxy sequencing were constructed by inserting previously cloned DNA fragments (described in Meyerowitz & Martin, 1984), into the vectors M13mp18 and M13mp19 (Norrander et al., 1983).

Two of the *D. simulans* clones were derived from the 1.3 kb *Sal*I/*Xho*I fragment of fDs024, which was inserted into M13mp18 in both orientations. The upstream regions (contained in the 0.53 kb *Eco*RI/*Sal*I fragment of fDs024) were sequenced directly using the chemical method of Maxam & Gilbert (1977, 1980).

The *D. erecta* sequences were from the 1.8 kb *Eco*RI fragment of fDe009 inserted into the vector M13mp18. Only one orientation of this fragment was recovered. For this clone, a series of deletions were generated by the exonuclease III digestion method of Henikoff (1984). Five overlapping clones were selected from those generated and

were used to determine the entire 1.8 kb of *D. erecta* sequence. In order to obtain clones suitable for determination of the sequence of the other strand, the two *Eco*RI/*Bam*HI fragments (1.15 kb and 0.65 kb) of fDe009 were directionally cloned into M13mp18 and M13mp19, respectively.

The *D. yakuba* gene sequence was derived from the 1.95 kb *Eco*RI/*Xho*I insert fragment of qDy5113 inserted directionally into M13mp18 and M13mp19.

(d) *DNA Sequence Analysis*

DNA sequences were analyzed using programs written by one of the authors (CHM) for an IBM PC-XT computer. DNA sequences were aligned using the algorithm of Gotoh (1982) as implemented by Robert E. Pruitt for the Apple Macintosh computer.

4. Results and Discussion

(a) *Inter-species transformations of the* Sgs-3*-like gene*

One approach towards determining if the *cis*-acting regulatory sequences required for proper regulation of a gene are well conserved over the course of evolution is to characterize the ability of a gene derived from one species to function when introduced into another species. In the experiments described here, *Sgs-3*-homologous genes from *D. erecta* and *D. yakuba* were introduced into the *D. melanogaster* genome by the technique of P-factor mediated transformation (Spradling & Rubin, 1982; Rubin & Spradling,

1982). The resulting transformants were assayed for their ability to express the foreign glue genes.

The three different glue gene-containing plasmids constructed for P-factor-mediated transformation of *D. melanogaster* are shown in Figure 1. Previous experiments have shown that the *Sgs-3* glue gene of *D. melanogaster* is properly regulated and expressed when 2.27 kb of upstream sequence is present (Crosby & Meyerowitz, 1986). It was considered to be possible that the pGXDy4.9a construct, which contains the glue gene from *D. yakuba* with 3.0 kb of upstream sequence and 1.0 kb of downstream sequence, would contain the sequences required for proper expression. A second *D. yakuba* construct, pGXDy1.8, contains only 529 bp of upstream and ~356 bp of downstream sequence (the amount of downstream sequence is approximate due to the uncertainty in the exact position of the 3' end of the mRNA). The third construct, pGXDe1.7b, contains 532 bp upstream and ~37 bp downstream of the *D. erecta Sgs-3*-like glue gene.

For each of the three recombinant DNA constructs, five independent *D. melanogaster* transformant strains were established. The activity of the introduced glue genes in the salivary gland was assayed by RNA gel blot hybridization. These blots were probed with nick-translated plasmids that contain the introduced foreign glue gene. The extensive divergence between the *Sgs-3*-like glue genes of the three species (*D. melanogaster*, *D. erecta* and *D. yakuba*) make these probes species-specific under moderately stringent hybridization conditions (Meyerowitz & Martin, 1984). The results of these hybridizations are shown in

Figure 2. It is apparent that the introduced genes are expressed at a lower level in D. melanogaster as compared to the level seen in the species from which the gene was originally obtained. However, this difference does not appear to be greater than about 2-fold.

The hybridized probe was removed (see experimental procedures) from the filters and the filters were then hybridized with a probe specific for the *D. melanogaster Sgs-3* glue gene. The level of expression of the endogenous gene in the transformed lines was similar to that seen with the wild-type strain OR16f (Meyerowitz and Hogness, 1982); an example for one set of transformants is shown in Figure 3. It was also necessary to determine if equivalent amounts of RNA were present in each lane. This was done by stripping the blots of any hybridized material and then probing with nick-translated λbDm103, a phage lambda clone containing the 18S and 28S ribosomal genes of *D. melanogaster*. It was found that similar amounts of RNA were present in all lanes by visual inspection of the autoradiograms (data not shown).

In order to determine the tissue specificity of expression of the introduced glue genes during third instar, the carcasses remaining after removal of the salivary glands were also used to prepare RNA blots. No reproducible expression of either the endogenous or introduced glue genes was seen (data not shown). Occasionally, very low level expression of both the introduced and endogenous genes was seen in a particular lane; however this is not thought to result from aberrant expression of the genes in other tissues. It is instead believed to result from occasional

incomplete removal of the salivary gland from the animal, because the expression is not reproducible between duplicate experiments. In summary, the introduced glue genes are expressed at levels which approach normal and this expression is limited to the proper tissue during third instar, namely the salivary gland.

The regulation of developmental stage of expression was also examined. RNA was prepared from second instar, early third instar, late third instar, white prepupae, tan prepupae and adult animals; this RNA was used to prepare RNA blots. Prominent expression was seen only in late third instar animals (Figure 4). Faint signal was sometimes seen in white prepupae and, in lines transformed with the *D. erecta Sgs-3* gene, in tan prepupae as well. This could result from continued expression of the genes past what is seen for the *Sgs-3* glue gene of *D. melanogaster*.

Although the levels of expression were not quantitated, it appears that the introduced genes are expressed at less than full levels. Three possible explanations are: (1) there are further regulatory elements which lie outside the regions used in the constructions which are required for full levels of expression, (2) that the regulatory sequences, although in general well conserved, have mutated such that they interact less efficiently with the trans-acting factors present in *D. melanogaster*, or (3) that the foreign mRNA is less stable in the *D. melanogaster* background.

These experiments demonstrate that the *D. yakuba* and *D. erecta* glue gene constructs that were introduced into

D. melanogaster contain sequence elements sufficient to yield near-normal patterns of gene expression. In addition, they show that the transcriptional regulatory sequences of these genes are capable of interacting with the regulatory factors of *D. melanogaster*.

(b) *Introduction to sequence comparisons*

The comparison of homologous sequences which have been subjected to the processes of mutation and selection can be useful in discerning those features of a gene which are essential for proper expression, processing and protein function. Such an approach is similar to site-directed mutagenesis in that one goal is to identify a set of sequence changes that are compatible with gene expression. Additionally, one may gain further understanding of the processes of evolution which are acting upon the gene cluster.

We have determined the nucleotide sequences of the *Sgs-3* homologues in three species of *Drosophila*: *D. simulans*, *D. erecta* and *D. yakuba*. The previously determined *D. melanogaster* sequence extends 4456 bp upstream of the *Sgs-3* mRNA start site (this regions also contains the *Sgs-7* and *Sgs-8* genes) and ~1104 bp downstream of the transcript (Garfinkel *et al.*, 1983). In *D. simulans*, the sequence of a region from 645 bp of upstream to ~273 bp downstream was determined; in *D. erecta* from 532 bp upstream to ~36 bp downstream; and in *D. yakuba* from 529 bp upstream to ~356 bp downstream. The sequenced regions and their relationships to the glue gene clusters are shown in Figure 1.

The *Sgs-3* gene of *D. melanogaster* can be divided into a series of regions, as shown in Figure 5. First, there are the 2270 bp of sequence 5' of the gene that are implicated in regulation (Crosby & Meyerowitz, 1986; Vijay Raghavan *et al.*, 1986). Next comes the 5' untranslated portion of the mRNA; this region is short (29 bp). The protein begins with a hydrophobic leader peptide which is removed during processing to yield the mature polypeptide chain (Crowley & Meyerowitz, 1983). The region coding for this leader is interrupted after the first base of the tenth codon by a 73 bp intron. Following the sequences coding for the leader peptide is a region coding for a threonine rich region of the mature protein that is 49 amino acids in length. This is followed by the tandem repeats, which are also threonine rich. The sgs-3 protein is known to be heavily glycosylated; these threonine residues are presumed to be the attachment sites for the carbohydrate moieties. The size of the threonine-rich region varies among different strains of *D. melanogaster* (Crosby & Meyerowitz, 1986; Mettling *et al.*, 1985). The 50 amino acid carboxy-terminal domain of the sgs-3 protein is cysteine rich; the position of the cysteine residues is conserved among the three related *D. melanogaster* glue genes at 68C. The termination codon is followed by a 3' untranslated portion of the mRNA and sequences 3' to the poly A addition site. What follows is a region by region comparison of the sequences of this gene and its homologues in three other species.

(c) *5' Flanking and 5' Untranslated Regions*

An alignment of the 5' flanking and 5' untranslated regions is shown in Figure 6. There are several relatively conserved 'islands' of sequence interspersed among areas of lesser conservation. One notable feature is the large deletion present in each of the three species relative to the *D. melanogaster* sequence. Bases Dm -491 to Dm -157 (this notation gives the species name abbreviation followed by the base number relative to the RNA start site in that species) are absent in *D. erecta* and *D. yakuba*. Given the ability of the *D. erecta* and the *D. yakuba* genes to express in a *D. melanogaster* background, it is clear that these deleted sequences are not necessary for proper expression of the glue gene homologues. These sequences which are deleted in the other species are thus not likely to be essential to the proper expression of the *D. melanogaster Sgs-3* glue gene. The sequences beyond these deletions (*i.e.*, towards the gene), between Dm -156 and Dm +29, are relatively well conserved: the amount of change varies from 10.6% (De *vs.* Dy) to 18.8% (Ds *vs.* Dy). This correlates with the likely involvement of these sequences in the regulation of the gene. In contrast, the sequences upstream of the deletion (Dm -815 to Dm -493) are somewhat less conserved: the amount of change varies from 11.5% (Dm *vs.* Ds) to 31.2% (De *vs.* Dy). However, there are islands that are well conserved; some examples are the sequences lying at Dm -518 to Dm -493, Dm -580 to Dm -548 and Dm -756 to Dm -732. These data should be useful in the design of experiments to further delimit

the sequences required for normal levels of expression of the *Sgs-3* glue gene.

Several regions upstream of the *Sgs-3* glue gene have been proposed as regulatory sites. Hoffman & Corces (1986) have observed that sites upstream of the *Sgs-3*, *Sgs-4*, *Sgs-7* and *Sgs-8* glue genes are similar to a sequence in the *Hsp27* gene that has been implicated in ecdysterone-mediated regulation. This sequence is very AT rich; the sequence in *Sgs-3* is 95% AT (19 out of 20 bases). The sequence resides at Dm -328 to Dm -309, a region deleted in each of the other three species. Another sequence, identified by Shermoen & Beckendorf (1982) by homology to sequences at a DNase I hypersensitive site upstream of *Sgs-4*, is located at Dm -433 to Dm -420; this sequence is also deleted in the other species. The loss of these sites suggests that they are not necessary for the expression of the *Sgs*-3 gene.

The locations of DNase I hypersensitive sites in salivary gland chromatin surrounding the *Sgs-3* gene have been determined by Ramain *et al.* (1986). Their study found four such sites upstream of the gene, centered around the locations Dm -750, Dm -600, Dm -470 and Dm -75 (these locations are approximate and Ramain *et al.* suggest an accuracy of ±30bp). DNase I hypersensitivity is considered to be an indicator of alterations in the chromatin structure at or near the area of the site, e.g., the displacement of histones and associated proteins by the binding of a regulatory factor. The site at Dm -75 is both centered in a region of relatively high conservation and within a region strongly implicated in the regulation of the *Sgs-3* gene

(Vijay Raghavan *et al.* 1986). The site at Dm -750 also lies in an island of relatively well conserved sequence. In contrast to these well conserved sites, the sequences at Dm -470, while present in *D. simulans*, are deleted in both *D. erecta* and *D. yakuba*. The sequences at Dm -600 were considered to be particularly significant by Ramain *et al.* due to the correlation of the times at which the site is DNAse I hypersensitive with the expression of the *Sgs-3* gene. The evolution of this region is relatively complex. In *D. melanogaster*, there is a short (ten nucleotide) direct repeat. Most or all of one of the repeats is deleted in *D. simulans* and *D. erecta;* in *D. yakuba* the direct repeat structure is less clear due to the presence of several base substitutions. However, because the precise location of the DNAse I hypersensitive site is not known, it is possible that the sequences responsible for the Dm -600 DNAse I hypersensitive site are actually located near but not at Dm -600. The sequences up to 30 bp 5' of Dm -600 are poorly conserved. In contrast, the nearby sequences centered around Dm -565 are relatively well conserved. These regions should be attractive targets for site directed mutagenesis experiments aimed at determining the actual sequences which are required for the proper regulation of the *Sgs-3* glue gene.

In summary, the results of the evolutionary comparisons of the upstream regions argues against the importance of several possible regulatory sites that have been proposed solely on the basis of sequence similarity. In contrast, two out of four of the sites located by the mapping of DNase I

hypersensitive sites do correlate with the location of evolutionarily conserved regions; an additional site is near such a region. Also, the presence of the large deletions found in the other species with respect to the *D. melanogaster* sequence, in combination with the ability of these foreign genes to express properly in *D. melanogaster*, suggests that a large block of *D. melanogaster* sequences are not necessary for proper regulation of the *Sgs-3* gene.

(d) *The Hydrophobic Leader and Adjacent Protein Coding Regions*

The first part of the protein coding region to be discussed starts at the proposed initiator methionine codon at Dm +30 and extends for 28 amino acids up to the beginning of the AC-rich region (at Dm +187). The alignment of this region is shown in Figure 7. In *D. melanogaster*, the first 23 amino acids are known to be cleaved from the protein during processing. In all four species, there is an intron present after the first base of the tenth codon; the evolution of the intron sequences is discussed in the next section.

Some of the features of the region are well conserved among the four species. First, there are no apparent insertions or deletions; this preserves both the reading frame and the overall length. This may be necessary to preserve a functional leader sequence. Second, the relative abumdance of hydrophobic, polar, basic and acidic residues is relatively constant; this is summarized Table 1. In general, signal sequences are primarily hydrophobic and

contain one or two basic residues near the beginning of the leader sequence (von Heijne, 1985). These properties are preserved, while several nucleotide and amino acid substitutions have occurred (see Table 2).

(e) *The Intron*

An alignment of the introns is shown in Figure 8. The introns range in length from 73 bp to 77 bp. In contrast to the surrounding sequences, which code for the hydrophobic leader peptide, there is evidence of insertions and deletions within the intron. There are three portions of an intron that are currently known to be necessary for proper RNA splicing to occur. The 5' splice site is the first of these. The four nucleotides before and the seven nucleotides after the 5' splice site are identical in all four of the species. The sequence matches the consensus sequence of (C/A)AGGT(G/A)AGT determined by Mount (1982) in seven out of nine positions. The 3' splice site shows greater change in the surrounding sequence than does the 5' splice site. However, all species posess the highly characteristic AG sequence adjacent to the 3' splice site. The species match the consensus sequence $((C/T)_nN(C/T)AGG$, where n = 11, Mount (1982)) at 11/16 positions (*D. erecta*) and 15/16 positions (*D. simulans* and *D. yakuba*). The third sequence is the lariat junction site, typically located a few tens of nucleotides upstream from the 3' splice site (Ruskin et al., 1984). A consensus sequence for the 3' splice signal, (C/T)T(A/G)A(T/C), has been derived by Keller & Noon (1985) based upon a comparison of the sequences of 39 *Drosophila*

introns. Matches to the consensus were usually found between 18 and 35 nucleotides 5' of the 3' splice site. In their survey, the third and fifth bases of the sequence showed instances of any base, while no deviations from the consensus occurred in the other three positions. Using these criteria, matches to the sequence (C/T)TNAN were searched for in the appropriate region upstream of the 3' splice site; these potential lariat junction sites are underlined in Figure 8. Although matches are found in all four species within this region, no one site is sufficiently conserved in all species to match the consensus sequence. The determination of the actual lariat junction sites used during processing of the transcripts awaits further experiments.

The level of nucleotide substitution in the introns is comparable to those seen in some of the surrounding protein-coding regions (see Table 2). One possible explanation is that the intron sequences are under moderate selection pressures. While the assumption that intron sequences are free to evolve relatively rapidly may be valid when dealing with larger introns, there may be a significant fraction of the intron sequence present in small introns such as these which is required for proper processing. The *Sgs-3* glue gene is transcribed at a very high rate during mid-third instar; it seems reasonable to expect that the primary transcript must be capable of efficient interaction with the components of the splicing machinery. This requirement could serve to place constraints on the changes which would be allowable in these intron sequences.

(f) *The AC-rich Region*

In *D. melanogaster*, the region coding for a hydrophobic portion of the protein is soon followed by a large region which is very AC-rich on the RNA-like strand. This region can be subdivided into two parts: the first is very threonine rich and contains little, if any, repeating structure; the second is also threonine rich and consists of a variable number (37 in the OR16f strain of *D. melanogaster*) of repeats of a five amino acid sequence with the consensus pro-thr-thr-thr-lys (Garfinkel, *et al.*, 1983). It is this central AC-rich region which distinguishes the sgs-3 protein from the otherwise similar sgs-7 and sgs-8 proteins. Several questions about this region can be asked, including how this region arose, what function the repeats serve, how the repeats evolve, and what features of the region are conserved during evolution.

A comparison among the corresponding regions of the four genes reveals several aspects which are conserved. First, all of the regions are noticeably A+C-rich in the RNA-like strand, varying from 72% A+C for *D. erecta* to 85% A+C for *D. simulans*. The amino acid composition of the region is similar in the four species; this is summarized in Table 3. Most prominent is threonine, which makes up from 36% (*D. erecta*) to 56% (*D. melanogaster*) of the region. Proline is also abundant, along with the basic residues lysine, arginine and histidine. Acidic residues are either absent (in *D. melanogaster* and *D. simulans*) or relatively rare (in *D. erecta* and *D. yakuba*). Glycine, isoleucine,

phenylalanine, methionine, tryptophan and tyrosine are not present in the AC-rich regions of any of the species. Several aspects of these amino acid biases correlate with the nucleotide bias present in the regions. For example, the codons for threonine are ACN, which, given the prevalence of threonine, should require a large proportion of A's and C's. Similarly, the proline codons are CCN and the lysine codons AA(A/G). Some of the non-represented amino acids are coded for by GT-rich codons, including glycine, phenylalanine, methionine and tryptophan (however, the absence of isoleucine and tyrosine do not fit this correlation).

Another conserved aspect is that of the general structure of the conserved region: a threonine rich region followed by a repeat region which consists of tandem repeats of a five amino acid consensus sequence.

In contrast to these conserved aspects, there are several noticeable differences among the species in this region. The region, unlike the surrounding protein coding regions, is variable in length (from 139 amino acids in *D. simulans* to 250 amino acids in *D. erecta*). The sequences of the AC-rich regions are difficult to align: computer generated alignments for this region are sensitive to small changes in the mismatch or gap penalties used by the algorithm. This is due to a high rate of nucleotide substitution and insertion/deletion events. This region is evolving much more rapidly than the regions either 5' or 3' of the central AC rich sequences; however, the level of change is difficult to quantitate. The AC-rich regions are shown individually in Figure 9. The prominent repeat-

containing regions contained within each of the AC-rich regions are shown in Figure 10.

The consensus core repeats are shown in Figure 11. It is apparent that within the constraints imposed by the AC-rich character of the region and the propensity toward threonine, proline, lysine and arginine residues, this core sequence is very different among the species.

One feature is apparent in the AC-rich region of the species *D. simulans*, *D. erecta* and *D. yakuba* which is not present in *D. melanogaster*: this region contains repeats which are not five amino acids in length. These motifs are most apparent in *D. simulans* (see Figure 10b) and *D. erecta* (see Figure 10c).

The evolution of this region appears complex, with relatively rapid changes in overall size and in both DNA and protein sequence One likely mechanism is unequal crossover; the many tandem repeats present could be highly susceptible to these events. This process would be capable of producing the rapid change in the length of the region while maintaining homology within a given species, and at the same time allowing for the rapid divergence seen between the species (Cooke, 1975; Smith, 1976; Wayne & Willard, 1986). However, certain features (some of which may be essential for proper protein function) are maintained in the midst of this rapid change.

Many genes code for proteins which contain tandem repeats. One example is the *Sgs-4* gene, located at chromosomal site 3C in *D. melanogaster*; this gene also produces a component of the salivary gland glue. The protein

contains 19 (in the Oregon-R strain) to 31 (in the Hikone-R strain) repeats of a 7 amino acid sequence with the consensus: Thr-Cys-Lys-Thr-Glu-Pro-Pro (Muskavitch & Hogness, 1982). This sequence is rich in threonine and proline, as is the *Sgs-3* repeat sequence. The DNA consensus sequences is also very AC-rich on the RNA-like strand: 76%.

A glue protein which has been the subject of much research is that produced by several marine mussels, including *Mytilus californianus* and *Mytilus edulis* (Waite *et al.*, 1985; Waite, 1986). The protein contains a high proportion of the modified amino acids hydroxyproline (Hyp) and 3,4-dihydroxyphenylalanine (Dopa). Digests of the isolated protein from *M. edulis* yield a 10 amino acid sequence: Ala-Lys-Pro-Ser-Tyr-Hyp-Hyp-Thr-Dopa-Lys (Waite, 1983). There are apparently about 75 repeats of this sequence in the protein; however, it is not yet known if the repeats are arranged in a tandem array.

Several structural proteins are known to contain tandem repeats. A few of the many examples include silk fibroin from the moth *Bombyx mori* (Sprague *et al.*, 1979; Gage & Manning, 1980; Manning & Gage, 1980), zein, the seed storage protein of maize (Geraghty *et al.*, 1981; Pedersen *et al.*, 1982) and the human involucrin gene (Eckert & Green, 1986).

An example of a striking tandem repeat motif in a non-structural protein has been found in the *RPO21* gene of *S. cerevisiae* (Allison *et al.*, 1985). This gene codes for the largest subunit of RNA Polymerase II; it contains 26 tandem repeats of a heptapeptide with the consensus sequence: Pro-Thr-Ser-Pro-Ser-Tyr-Ser. It would be

interesting to compare the evolution of this repeat with that of the *Sgs-3* gene; if this RNA polymerase II subunit is under strong selection pressures, then the *RPO21* repeat region should evolve much more slowly than the repeat region of *Sgs-3* (Allison *et al.*, 1985).

(g) *Carboxy-terminal Protein Region*

Following the AC-rich region in each of the *Sgs*-3-homologous glue genes are the sequences which code for the 50 amino acids which make up the carboxy terminal portion of the protein. An alignment of the sequences of this region is shown in Figure 12. This region is easily distinguished from the preceding region: it is not noticeably AC-rich, it is easily aligned among the species and it contains no apparent insertion/deletion events. The region also does not have a repeating structure. The most notable feature of this region is the complete conservation of all cysteine residue positions. Not only are the number and position of these residues conserved in the *Sgs-3*-like glue genes of each of the four species, but cysteines are present at the same positions in the corresponding carboxy terminal portion of the *Sgs-7* and *Sgs-8* glue genes of *D. melanogaster* as well (Garfinkel *et al.*, 1983). These cysteines may play a role in the function of these proteins as a glue: oxidation of the -SH groups could lead to cross-linking of the proteins and create an insoluble plug of glue proteins. The evolution rate of this region is similar to, although greater than, that of the hydrophobic leader sequences: see Table 2.

(h) *3' Untranslated and 3' Flanking Regions*

The final sequences to be compared encompass the 3' untranslated and 3' flanking regions. An alignment of the sequences of this region is shown in Figure 13. The poly(A) consensus sequence, AAUAAA, is found in three of the four species between 142 bp (*D. melanogaster*) and 166 bp (*D. yakuba*) downstream of the termination codon. However, the homologous sequence in *D. simulans* is GAUAAA. This particular variant was not listed among several known deviations of the canonical poly(A) sequence, including those known to function and those known to impair or prevent function, compiled in the review by Birnstiel *et al.* (1985).

The lengths of the poly(A)+ mRNA transcripts derived from these genes has been determined (Meyerowitz and Martin, 1984); the lengths of the poly(A)- message can be deduced from the sequence of the genes and homology to the 5' and 3' ends of the *D. melanogaster Sgs-3* gene. The resulting prediction for the poly(A) tail lengths yield similar sizes for the *D. melanogaster* (~80 nt) and *D. simulans* (~90 nt) genes. This argues for the utilization of the GAUAAA sequence as a poly(A) addition site in *D. simulans*. A reasonable size for the *D. yakuba* tail (~100 nt) is obtained if one assumes that the larger of the two alleles found in the *D. yakuba* strain used was sequenced (see Meyerowitz & Martin, 1984). The predicted tail length for the *D. erecta* gene (~40 nt) is noticeably shorter than the others; this could represent the true tail length, however, other possibilities include inaccuracies in the determination of the length of the *D. erecta Sgs-3*-like glue gene transcript

or, alternatively, the utilization of a different transcription start or different poly(A) addition site. The region surrounding the poly(A) addition site is relatively well conserved when compared to sequences further downstream, which display both more frequent insertion/deletion and base substitution events.

(i) *Comparison of amounts of change in the* Sgs-3 *gene*

The amount of change seen in each region of the *Sgs-3* gene in summarized in Table 2. The table does not include any data for the central, AC-rich region. This region is evolving so rapidly that no quantitation was possible. However, it is apparent from the sequence data that this region is evolving more rapidly than any of the other sequenced regions in and around the *Sgs-3* genes. The next most rapidly changing regions include those sequences 5' and 3' to the gene, along with the intron sequences. The most conserved regions are the 5' and 3' protein coding portions of the genes.

The presence of such a rapidly evolving component within the coding sequences of a gene is unusual. Other *Drosophila* genes have been used in similar evolutionary studies but have not been found to evolve at the high rates seen in *Sgs-3* (Bodmer & Ashburner, 1984-*Adh;* Blackman & Meselson, 1986-*Hsp82*). A summary of some of the data from two of these other comparisons are shown in Table 4. It can be seen that in both *Adh* and *Hsp82*, there is much less nucleotide substitution in the coding regions as compared to the introns. This not the case with *Sgs-3*, where the rates

of change are either similar or actually higher in some portions of the coding sequences. It appears that the *Sgs-3* coding sequences are under much less stringent selection pressures than are the *Adh* and *Hsp82* genes; the rate of amino acid substitution in *Sgs-3* is dramatically higher than that seen in the other two genes. One can infer from this that the sgs-3 protein can endure much change and still perform its function. However, there are several aspects of the *Sgs-3* gene which do appear to be conserved through evolution and are likely to be essential for proper function.

5. Conclusions

We have taken two experimental approaches towards learning more about both the expression and the evolution of the *Sgs-3* gene. First, we used P-factor mediated transformation to demonstrate that the *Sgs-3* glue genes of *D. erecta* and *D. yakuba* can function properly when introduced into the *D. melanogaster* genome. Secondly, we determined the nucleotide sequences of the *Sgs-3* homologues from *D. simulans*, *D. erecta* and *D. yakuba*. These sequences were compared to the previously determined *D. melanogaster* sequence and used to discover those features of the gene that are conserved in evolution.

These results lead to three main conclusions about the evolution and expression of the *Sgs-3* gene. First, the *cis*-acting sequences necessary for correct tissue and time of expression of the *Sgs-3* gene in *D. erecta* and *D. yakuba* are sufficiently conserved that they interact appropriately with

the *trans*-acting regulators of this locus that are found in *D. melanogaster*. A corollary of this is that the *trans*-acting regulators of the *Sgs-3*-like genes found in *D. erecta* and *D. yakuba* are functionally conserved in evolution. The combination of sequence comparison with functional assays utilizing P-factor mediated transformation is very useful for studies on the evolution of the elements that control expression of a gene. One could presumably extend the studies described here and determine at what evolutionary distance the amount of accumulated change in control regions is sufficient to abolish proper regulation of the gene. Such investigations would utilize species that are more distantly related to *D. melanogaster* than those used here.

Another conclusion is that sequence homology alone is not a sufficient criterion for the reliable prediction of regulatory sequences. Those proposed as *Sgs-3* regulatory elements have been shown to be deleted in other species; these foreign genes, with the putative regulatory sequences deleted, are expressed normally when introduced into *D. melanogaster*. It is apparent that functional assays are necessary to determine the involvement of a particular sequence in the regulation of a gene.

Finally, we have found a dramatic exception to the commonly seen pattern of relatively well-conserved protein coding sequences which are surrounded by and also contain (in the form of introns) less well conserved sequences. The central AC-rich repeat-containing region of the *Sgs-3* gene has been found to be the most rapidly evolving portion of this gene. While this region is evolving rapidly in length,

nucleotide sequence and amino acid sequence, there are some properties of the region that are conserved. This demonstrates the highly variable behavior of genes as they are subjected to the processes of evolution. The commonly used measures of change, such as nucleotide or amino acid substitution rate, may not fully reflect the importance of a region to the function of a gene.

Acknowledgements

We thank Robert Pruitt, Mark Garfinkel, Charles Rice and Joan Kobori 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 GM28075 from the National Institutes of Health. C.H.M. was supported by a National Science Foundation Fellowship and by a Graduate Fellowship from the General Electric Foundation.

REFERENCES

Akam, M.E., Roberts, D.B., Richards, G.P. & Ashburner, M. (1978). *Cell*, **13**, 215-225.

Allison, L.A., Moyle, M., Shales, M. & Ingles, C.J. (1985). *Cell*, **42**, 599-610.

Beckendorf, S.K. & Kafatos, F.C. (1976). *Cell*, **9**, 365-373.

Birnstiel, M.L., Busslinger, M. & Strub, K. (1985). *Cell*, **41**, 349-359.

Blackman, R.K. & Meselson, M. (1986). *J. Mol. Biol.*, **188**, 499-515.

Bodmer, M. & Ashburner, M. (1984). *Nature*, **309**, 425-430.

Bourouis, M. & Richards, G. (1985). *Cell*, **40**, 349-357.

Cooke, H.J. (1975). *J. Mol. Biol.*, **94**, 87-99.

Crosby, M.A. & Meyerowitz, E.M. (1986). *Dev. Biol*, **118**, 593-607.

Crowley, T.E., Bond, M.W. & Meyerowitz, E.M. (1983). *Mol. Cell. Biol.*, **3**, 623-634.

Crowley, T.E. & Meyerowitz, E.M. (1984). *Dev. Biol.*, **102**, 110-121.

Eckert, R.L. & Green, H. (1986). *Cell*, **46**, 583-589.

Fraenkel, G. & Brookes, V.J. (1953). *Biol. Bull.*, **105**, 442-449.

Gage, L.P. & Manning, R.F. (1980). *J. Biol. Chem.*, **255**, 9444-9450.

Garfinkel, M.D., Pruitt, R.E. & Meyerowitz, E.M. (1983). *J. Mol. Biol.*, **168**, 765-789.

Geraghty, D., Peifer, M.A., Rubenstein, I. & Messing, J. (1981). *Nucl. Acids Res.*, **9**, 5163-5174.

Gotoh, O. (1982). *J. Mol. Biol.*, **162**, 705-708.

Guild, G.M. & Shore, E.M. (1984) *J. Mol. Biol.*, **179**, 289-314.

Hansson, L. & Lambertsson, A. (1983). *Mol. Gen. Genet.*, **192**, 395-401.

Henikoff, S. (1984). *Gene*, **28**, 351-359.

Hoffman, E. & Corces, V. (1986). *Mol. Cell. Biol.*, **6**, 663-673.

Keller, E.B. & Noon, W.A. (1985). *Nucleic Acids Res.*, **13**, 4971-4981.

Korge, G. (1975). *Proc. Nat. Acad. Sci. USA*, **72**, 4550-4554.

Korge, G. (1977a). *Dev. Biol.*, **58**, 339-355.

Korge, G. (1977b). *Chromosoma*, **62**, 155-174.

Lemeunier, F., David, J. R., Tsacas, L. & Ashburner, M. (1986). The *melanogaster* Species Group. In: Ashburner, M., Carson, H. L. & Thompson Jr., J. N. (eds) The Genetics and Biology of *Drosophila*. vol 3e. Academic Press, London pp 147-256.

Lewin, B.M. (1985). Genes, 2nd. ed. John Wiley & Sons, New York.

Manning, R.F. & Gage, L.P. (1980). *J. Biol. Chem.*, **255**, 9451-9457.

Martin, C.H. & Meyerowitz, E.M. (1986). *Proc. Nat. Acad. Sci. USA*, **83**, 8654-8658.

Maxam, A.M. & Gilbert, W. (1977). *Proc. Nat. Acad. Sci., USA*, **74**, 560-564.

Maxam, A.M. & Gilbert, W. (1980). *Methods Enzymol.*, **65**, 499-560.

Mettling, C., Bourouis, M. & Richards, G. (1985). *Mol. Gen. Genet.*, **201**, 265-268.

Meyerowitz, E.M., Crosby, M.A., Garfinkel, M.D., Martin, C.H., Mathers, P.M. & Vijay Raghavan, K. (1985). *Cold Spring Harbor Symp. Quant. Biol.*, **50**, 347-353.

Meyerowitz, E.M. & Hogness, D.S. (1982). *Cell*, **28**, 165-176.

Meyerowitz, E.M. & Martin, C.H (1984). *J. Mol. Evol.*, **20**, 251-264.

Mount, S.M. (1982). *Nucleic Acids Res.*, **10**, 459-472.

Muskavitch, M.A.T. & Hogness, D.S. (1980). *Proc. Nat. Acad. Sci. USA*, **77**, 7362-7366.

Muskavitch, M.A.T. & Hogness, D.S. (1982). *Cell*, **29**, 1041-1051.

Norrander, J., Kempe, T. & Messing, J. (1983). *Gene*, **26**, 101-106.

Pedersen, K., Devereux, J., Wilson, D.R., Sheldon, E. & Larkins, B.A. (1982). *Cell*, **29**, 1015-1026.

Ramain, P., Bourouis, M., Dretzen, G., Richards, G, Sobkowiak, A. & Bellard, M. (1986). *Cell*, **45**, 545-553.

Richards, G., Cassab, A., Bourouis, M., Jarry, B. & Dissous, C. (1983). *EMBO J.*, **2**, 2137-2142.

Rubin, G.M. & Spradling, A.C. (1982). *Science*, **218**, 348-353.

Ruskin, B., Krainer, A.R., Maniatis, T. & Green, M.R. (1984). *Cell* **38**, 317-331.

Sanger, F., Nicklen, S. & Coulson, A.R. (1977). *Proc. Nat. Acad. Sci USA*, **74**, 5463-5467.

Shermoen, A.W. & Beckendorf, S.K. (1982). *Cell*, **29**, 601-607.

Smith, G.P. (1975) *Science*, **191**, 528-535.

Spradling, A.C. & Rubin, G.M. (1982). *Science*, **218**, 341-347.

Sprague, K.U., Roth, M.B., Manning, R.F. & Gage, L.P. (1979). *Cell*, **17**, 407-413.

Strauss, E., Kobori, J.A., Siu, G. & Hood, L.E. (1986). *Anal. Biochem.*, **154**, 353-360.

Velissariou, V. & Ashburner, M (1980). *Chromosoma*, **77**, 13-27.

Velissariou, V. & Ashburner, M. (1981). *Chromosoma*, **84**, 173-185.

Vijay Raghavan, K., Crosby, M.A., Mathers, P.H. & Meyerowitz, E.M (1986). *EMBO J.*, **5**, 3321-3326.

von Heijne, G. (1985). *J. Mol. Biol.*, **184**, 99-105.

Waite, J.H. (1983). *J. Biol. Chem.*, **258**, 2911-2915.

Waite, J.H., Housley, T.J. & Tanzer, M.L. (1985). *Biochemistry*, **24**, 5010-5014.

Waite, J.H. (1986). *J. Comp. Physiol.* B, **156**, 491-496.

Wayne, J.S. & Willard, H.F. (1986). *Mol. Cell. Biol.*, **6**, 3156-3165.

Table 1

Amino acid contents of the hydrophobic leader regions

	hydrophobic	polar	basic	acidic
D. melanogaster	15	7	1	0
D. simulans	16	6	1	0
D. erecta	17	4	2	0
D. yakuba	16	5	2	0

Classes are those defined in Lewin (1985).

Table 2

Comparison of change in regions of the Sgs-3-homologous glue genes

(mismatch + deletions) / (matches + mismatches + deletions) x 100

	5' utr	protein 5'	intron	protein	3'utr
Dm vs. Ds	15.2%	7.1%	10.8%	10.5%	10.5%
Dm vs. De	24.4%	15.5%	30.7%	15.0%	31.5%
Dm vs. Dy	23.7%	19.0%	18.4%	27.4%	25.1%
Ds vs. De	24.4%	16.7%	30.7%	17.6%	29.9%
Ds vs. Dy	23.5%	22.6%	19.5%	26.8%	25.6%
De vs. Dy	24.2%	21.4%	29.5%	19.6%	28.0%
averages	22.6%	17.0%	23.2%	19.5%	25.1%

Species abbreviations are the same as defined in Figure 5. Change is defined as shown in the formula above the column headings. A mismatch is defined as a base that has changed between the two species; this does not include bases that have been deleted. A deletion is any number of contiguous deleted bases. A match is any base which is identical in both species. The 5' utr includes all bases upstream of the first codon of the protein, as shown in Figure 5. The protein 5' region corresponds to the region shown in Figure 6. The intron includes only those bases in the intron itself. No data are shown for the AC-rich regions; because no consistent alignments could be generated, the level of change could not be quantitated. The protein 3' region is as shown in Figure 11. The 3' utr region includes all bases following the termination codon for the *Sgs-3*-homologous genes, as shown in Figure 12.

Table 3

Amino acid compositions of the AC-rich regions

			basic	acidic	
	thr	pro	lys-arg-his	asp-glu	others
Dm	55.9%	17.9%	14.4%	0.0%	11.8%
Ds	52.5%	15.1%	19.4%	0.0%	13.0%
De	35.6%	11.2%	34.4%	0.4%	18.4%
Dy	54.0%	13.0%	18.4%	2.2%	12.4%

Table 4

Comparison of Change in the Sgs-3-homologous genes with change in Adh and Hsp82

	nt change IVS			nt change coding			amino acid change		
	Sgs-3	*Adh*	*Hsp82*	*Sgs-3*	*Adh*	*Hsp82*	*Sgs-3*	*Adh*	*Hsp82*
Dm vs Ds	10.8%	7.6%	8.5%	8.8%	1.2%	1.2%	14.0%	0.8%	0.0%
Dm vs De/Do	30.7%	39.0%	N.D.	15.2%	4.3%	N.D.	15.4%	3.9%	N.D.

Change is computed as defined in the legend to Table 2. The data for *Adh* are derived from the sequences presented in Bodmer & Ashburner (1983). The abbreviation Do represents the species *D. orena*, which is a close relative of *D. erecta;* they are assumed to be equally distant from *D. melanogaster* for the purposes of this comparison. The data for *Hsp82* are derived from the sequences presented in Blackman & Meselson (1986). N.D. indicates that data are not available for the comparison.

Figure 1. The 68C-homologous glue gene clusters. Restriction enzyme abbreviations are: *Bam*HI (B), *Bgl*II (Bg), *Eco*RI (R), *Hind*III (H), *Kpn*I (K), *Pst*I (P), *Pvu*I (Pv), *Sal*I (S), *Sac*I (Sc), *Xba*I (Xb), *Xho*I (Xh) and *Xmn*I (Xm). Sites in parentheses are present in some strains of *D. melanogaster* but not in others (see Meyerowitz & Martin, 1984). Sites in brackets indicate that only a subset of the sites recognized by the particular enzyme are shown. The maps are aligned by the *Eco*RI site present at the left edge. Some genes of the clusters have only been localized to restriction fragments; these restriction fragments are shown by filled bars. Below the bars, arrows showing the length of the poly(A)+ transcript hybridized by the restriction fragment, along with the direction of transcription, are shown. For genes whose location is known from sequencing of the regions, only the arrows are shown. Below each arrow the size of the poly(A)+ RNA is expressed in nucleotides. For further details, see Meyerowitz and Martin, 1984. The hatched bars below the maps indicate the extent of the regions that have been sequenced. The *D. melanogaster* sequence was determined previously (Garfinkel, et al., 1983). The other sequences are presented in this work. The regions of the glue gene clusters that were used to transform *D. melanogaster* animals are shown by lines with vertical segments at each end below the *D. erecta* and the *D. yakuba* maps. The name of the constructs, in which the indicated fragments were inserted into the vector Carnegie-20 (see experimental procedures) is shown at the left. The nomenclature for these constructs is: G for glue; X for xanthine dehydrogenase (rosy) used as a

scorable marker for identification of transformant animals; Dy or De to indicate that the glue gene containing fragment is from *D. yakuba* or *D. erecta*, respectively; followed by a number which indicates the length of the insert fragment, in kilobase pairs; and, optionally, an a or b, indicating orientation of the insert fragment with respect to the vector.

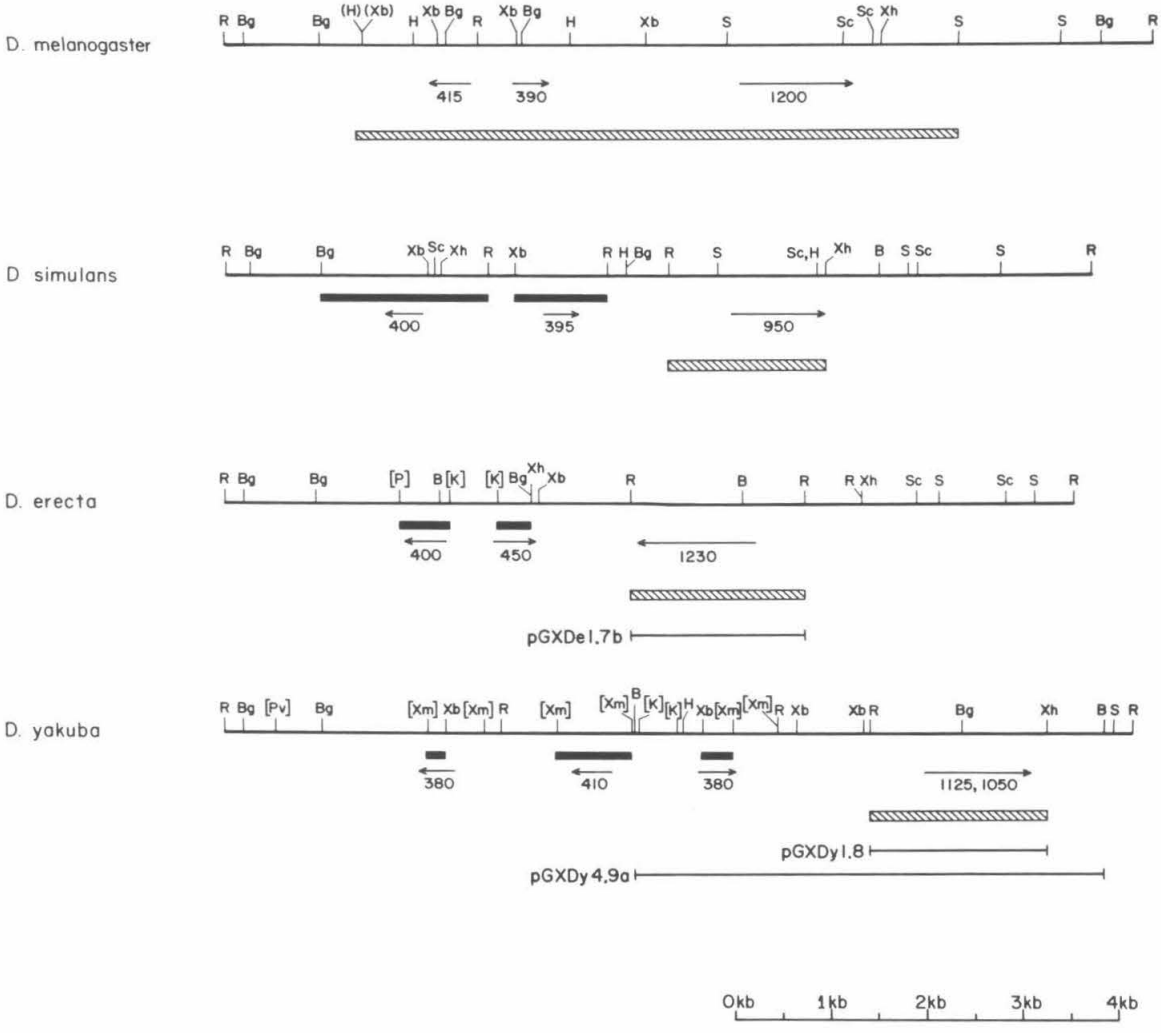
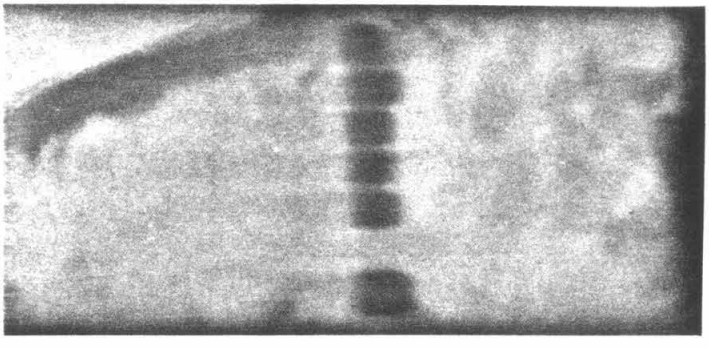

D. melanogaster
R Bg Bg (H) (Xb) H Xb Bg R Xb Bg H Xb S Sc Sc Xh S S Bg R
415
390
1200
D simulans
R Bg Bg Xb Sc Xh R Xb R H Bg R S Sc,H Xh B S Sc S R
400
395
950
D. erecta
R Bg Bg [P] B [K] [K] Bg Xh Xb R B R R Xh Sc S Sc S R
400
450
1230
pGXDe1.7b
D. yakuba
R Bg [Pv] Bg [Xm] Xb [Xm] R [Xm] [Xm] B [K] [K] H Xb [Xm] [Xm] R Xb Xb R Bg Xh B S R
380
410
380
1125, 1050
pGXDy1.8
pGXDy4.9a
0kb 1kb 2kb 3kb 4kb

Figure 2. Assay for expression of the foreign *Sgs-3*-homologous glue genes in the transformant lines. Salivary glands were dissected from late third instar animals. RNA was isolated and fractionated on 1.5% agarose-formaldehyde gels. The gels were blotted to nitrocellulose and hybridized to ^{32}P-labelled DNA probes specific for the foreign glue gene. The first five lanes in each panel contain RNA isolated from each of the five independent transformant lines that were used. The next two lanes are controls which contain RNA derived from non-transformant stocks of the species indicated. A. lines transformed with pGXDy4.9a, probed with nick-translated qDy5113 (see Meyerowitz & Martin, 1984); B. pGXDy1.8 transformant lines, same probe as in panel A; C. pGXDe1.7b transformant lines, probe used was nick-translated fDe009 (see Meyerowitz & Martin, 1984). Note the lower levels of accumulation of the foreign glue genes in the *D. melanogaster* background as compared to levels seen in the control lane (*D. yakuba* or *D. erecta* lanes).

A

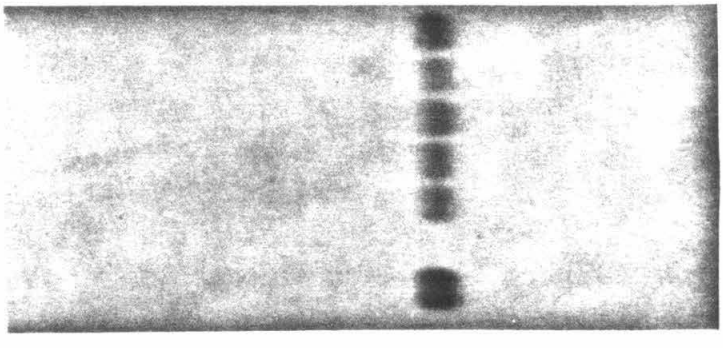
CM1.1.1
CM1.14.1
CM1.15.1
CM1.28.1
CM1.82.1
D. yakuba
D. melanogaster

B



CM2.26.1
CM2.65.1
CM2.66.1
CM2.98.1
CM2.107.1
D. melanogaster
D. yakuba

C



CM3.20.1
CM3.41.1
CM3.47.1
CM3.55.1
CM3.100.1
D. melanogaster
D. erecta

Figure 3. Expression of the endogenous *Sgs-3* glue gene in transformant lines. Hybridizing material was washed off of the nitrocellulose filter shown in Figure 2A and the filter was hybridized to nick-translated aDm2023 (Garfinkel *et al.*, 1983). A similar level of expression is seen in all five transformant lines and in the non-transformant *D. melanogaster* stock; as expected, no expression is seen in the *D. yakuba* stock.

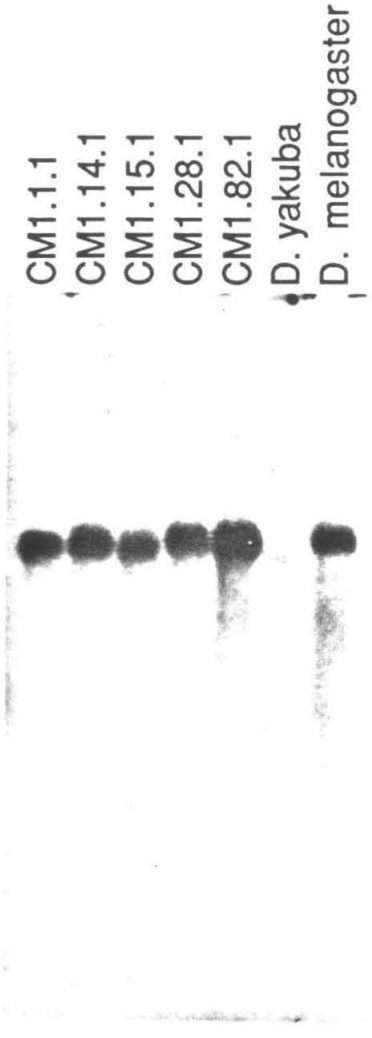

CM1.1.1
CM1.14.1
CM1.15.1
CM1.28.1
CM1.82.1
D. yakuba
D. melanogaster

Figure 4. Developmental expression of the pGXDy4.9a construct in two transformant lines. RNA was prepared from whole animals at the stages indicated above each lane. For second instar, 10 animals were used, for adults 3 animals were used and for all other stages 2 animals were used. The two control lanes, at the right of the blot, contain RNA isolated from 2 late third instar animals from the species indicated. The RNA was fractionated and the blot prepared as in Figure 3. The blot was probed with nick-translated qDy5113, which is specific for the *Sgs-3*-homologous gene of *D. yakuba* (see Meyerowitz & Martin, 1984). The pattern of expression seen is identical to that of the endogenous *Sgs-3* gene of *D. melanogaster* (data not shown).

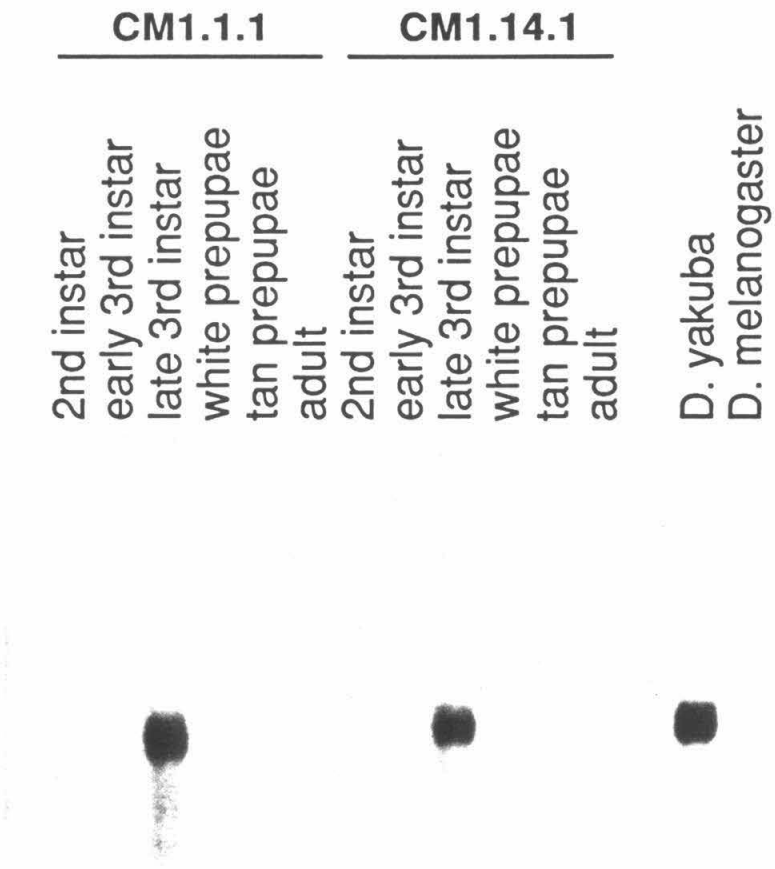

CM1.1.1
CM1.14.1
2nd instar
early 3rd instar
late 3rd instar
white prepupae
tan prepupae
adult
2nd instar
early 3rd instar
late 3rd instar
white prepupae
tan prepupae
adult
D. yakuba
D. melanogaster

Figure 5. The structure of the *Sgs-3* glue gene of *Drosophila melanogaster*. Numbering is relative to the transcription start site of the *Sgs-3* glue gene. A partial restriction map of this region of the *D. melanogaster* genome is shown. Below this map, the arrow indicates the extent of the *Sgs-3* transcript, along with the presence of an intron near the 5' end of the transcript. Above the map, a block diagram shows the major subdivisions of this region that were used in comparisons of the *Sgs-3* genes of the four species used in this study.

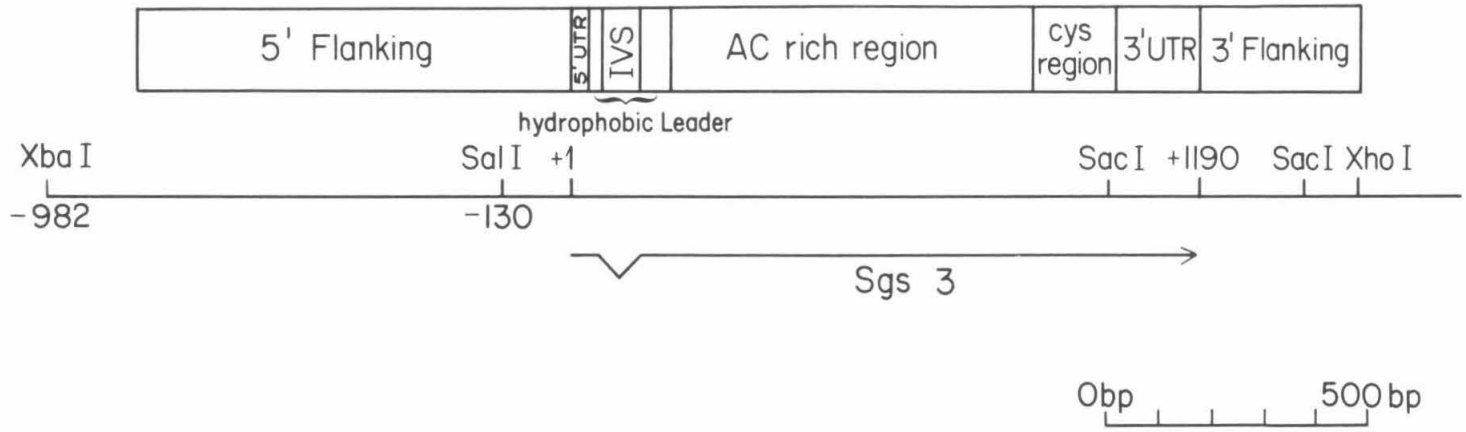

5' Flanking
5' UTR
IVS
AC rich region
cys region
3'UTR
3' Flanking
hydrophobic Leader
Xba I
Sal I
+1
Sac I
+1190
Sac I
Xho I
-982
-130
Sgs 3
0bp
500 bp

Figure 6. The aligned sequences of the 5' flanking and 5' non-coding regions of the *Sgs-3*-homologous glue genes. Species abbreviations are *D. melanogaster* (Dm), *D. simulans* (Ds), *D. erecta* (De) and *D. yakuba* (Dy). All pair-wise alignments were generated by the algorithm of Gotoh (1982). The mismatch penalty was 10, the start gap penalty was 40 and the deleted base penalty was 5. The four-way alignment shown in the figure was generated by hand from the pair-wise alignments. Spaces indicate that the sequence is identical to that of *D. melanogaster* at that point. A dash represents a deleted base. A dot indicates that the sequence was not determined for that base. The numbering is relative to the transcription start site in *D. melanogaster* or to the homologous site, which is presumed to be the transcription start site, in each of the other species. The single space gap in the sequence (between -1 and +1 in all species) shows the location of the mRNA start site.

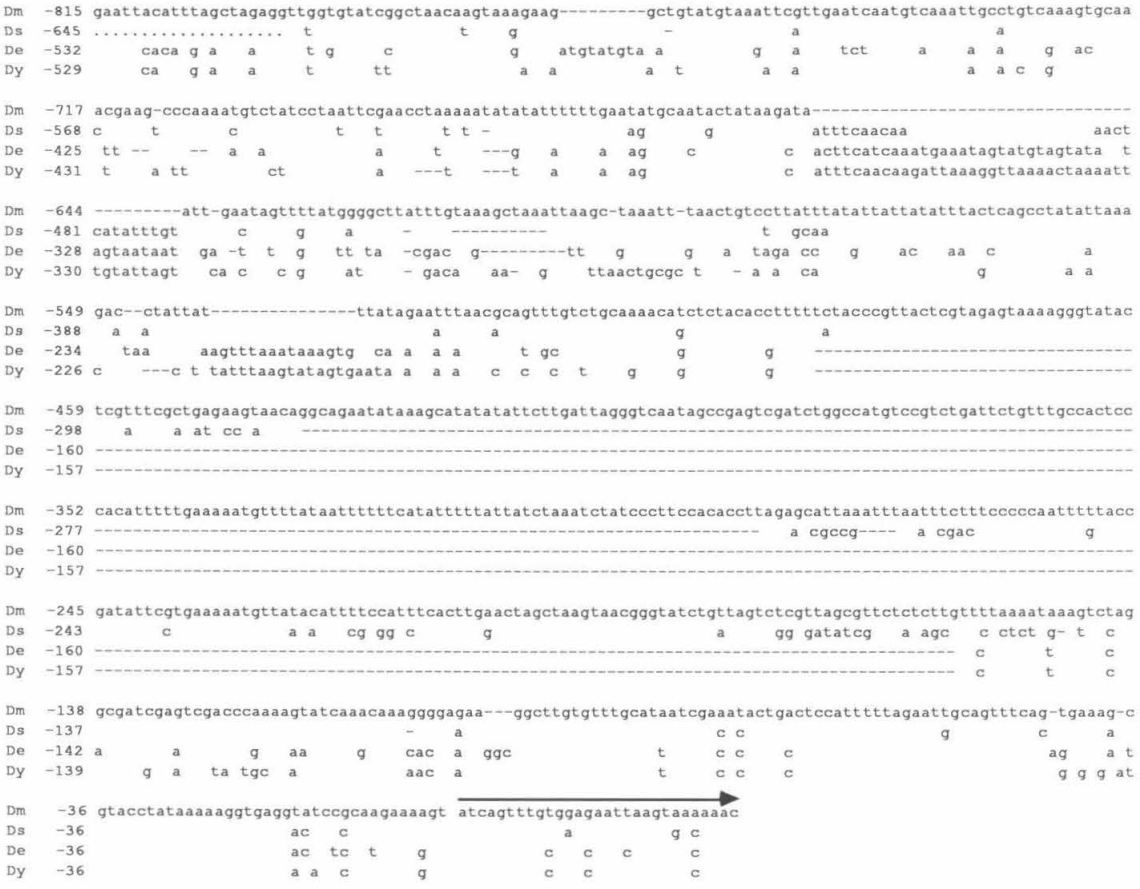

Dm -815 gaattacatttagctagaggttggtgtatcggctaacaagtaaagaag---------gctgtatgtaaattcgttgaatcaatgtcaaattgcctgtcaaagtgcaa
Ds -645 ..... t t g - a a
De -532 caca g a a t g c g atgtatgta a g a tct a a a g ac
Dy -529 ca g a a t tt a a a t a a a a c g
Dm -717 acgaag-cccaaaatgtctatcctaattcgaacctaaaaatatatatttttgaatatgcaatactataagata----------
Ds -568 c t c t t t t - ag g atttcaacaa aact
De -425 tt -- -- a a a t ---g a a ag c c acttcatcaaatgaaatagtatgtagtata t
Dy -431 t a tt ct a ---t ---t a a ag c atttcaacaagattaaaggttaaaactaaaatt
Dm -644 ---------att-gaatagttttatggggcttatttgtaaagctaaattaagc-taaatt-taactgtccttatttatattattatatttactcagcctatattaaa
Ds -481 catatttgt c g a - ---------- t gcaa
De -328 agtaataat ga -t t g tt ta -cgac g---------tt g g a taga cc g ac aa c a
Dy -330 tgtattagt ca c c g at - gaca aa- g ttaactgcgc t - a a ca g a a
Dm -549 gac--ctattat---------------ttatagaatttaacgcagtttgtctgcaaaacatctctacacctttttctacccgttactcgtagagtaaaagggtatac
Ds -388 a a a a g a
De -234 taa aagtttaaataaagtg ca a a a t gc g g ----------
Dy -226 c ---c t tatttaagtatagtgaata a a a c c c t g g g ----------
Dm -459 tcgtttcgctgagaagtaacaggcagaatataaagcatatatattcttgattagggtcaatagccgagtcgatctggccatgtccgtctgattctgtttgccactcc
Ds -298 a a at cc a ----------
De -160 ----------
Dy -157 ----------
Dm -352 cacatttttgaaaaatgttttataatttttcatatttttattatctaaatctatcccttccacaccttagagcattaaatttaatttctttcccccaatttttacc
Ds -277 ---------- a cgccg---- a cgac g
De -160 ----------
Dy -157 ----------
Dm -245 gatattcgtgaaaaatgttatacattttccatttcacttgaactagctaagtaacgggtatctgttagtctcgttagcgttctctcttgttttaaaataaagtctag
Ds -243 c a a cg gg c g a gg gatatcg a agc c ctct g- t c
De -160 ---------- c t c
Dy -157 ---------- c t c
Dm -138 gcgatcgagtcgacccaaaagtatcaaacaaaggggagaa---ggcttgtgtttgcataatcgaaatactgactccatttttagaattgcagtttcag-tgaaag-c
Ds -137 - a c c g c a
De -142 a a g aa g cac a ggc t c c c ag a t
Dy -139 g a ta tgc a aac a t c c c g g g at
Dm -36 gtacctataaaaaggtgaggtatccgcaagaaaagt atcagtttgtggagaattaagtaaaaaac
Ds -36 ac c a g c
De -36 ac tc t g c c c c
Dy -36 a a c g c c c

Figure 7. The aligned sequences of the hydrophobic leader region. Alignments were generated as described in the legend to Figure 6. The predicted amino acid sequence is shown below the DNA sequence. This region begins with the initiator ATG triplet and ends with the last codon previous to the AC-rich region (see text and Figure 9). An intron is present in all four species after the first base of the tenth codon (see text and Figure 8). The cleavage site of the hydrophobic leader in *D. melanogaster*, after the 23rd amino acid, is indicated in the figure.

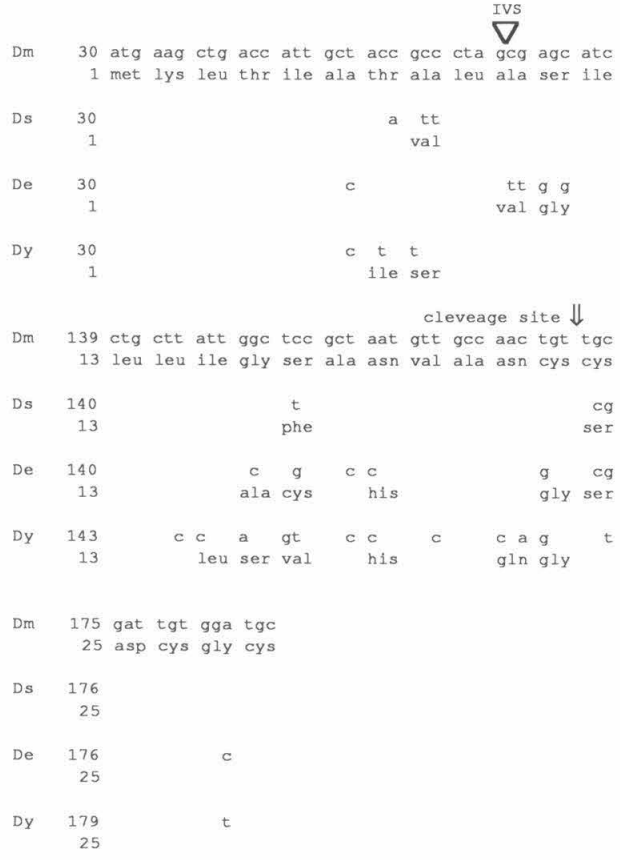

IVS
Dm 30 atg aag ctg acc att gct acc gcc cta gcg agc atc
1 met lys leu thr ile ala thr ala leu ala ser ile
Ds 30 a tt
1 val
De 30 c tt g g
1 val gly
Dy 30 c t t
1 ile ser
cleveage site
Dm 139 ctg ctt att ggc tcc gct aat gtt gcc aac tgt tgc
13 leu leu ile gly ser ala asn val ala asn cys cys
Ds 140 t cg
13 phe ser
De 140 c g c c g cg
13 ala cys his gly ser
Dy 143 c c a gt c c c c a g t
13 leu ser val his gln gly
Dm 175 gat tgt gga tgc
25 asp cys gly cys
Ds 176
25
De 176 c
25
Dy 179 t
25

Figure 8. The intron and surrounding sequences. The 10 base pairs on each end of the intron, along with the sequence of the intron itself, are shown. The splice sites in *D. melanogaster* are indicated; the other species are aligned to this sequence on the basis of homology to the *D. melanogaster* sequence (see Figure 6). The intron lies in the region coding for the hydrophobic leader region of the protein, after the first base of the tenth codon. Potential lariat junction sites (Keller & Noon, 1985) are underlined.

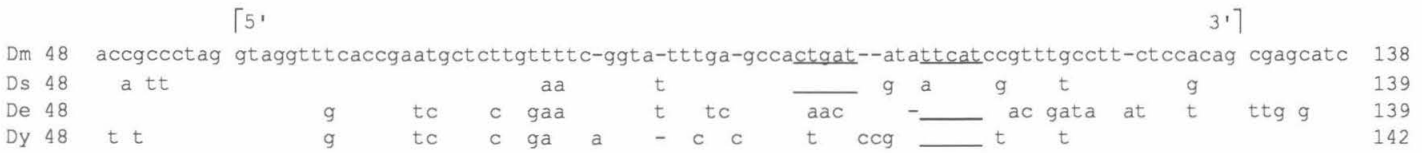

[5'
3']
Dm 48 accgccctag gtaggtttcaccgaatgctcttgttttc-ggta-tttga-gccactgat--atattcatccgtttgcctt-ctccacag cgagcatc 138
Ds 48 a tt aa t _____ g a g t g 139
De 48 g tc c gaa t tc aac -_____ ac gata at t ttg g 139
Dy 48 t t g tc c ga a - c c t ccg _____ t t 142

Figure 9. The AC-rich regions of each of the species. The sequences are shown individually and are not aligned (see text for discussion). The predicted amino acid sequences are shown below the DNA sequences. This region lies between the end of the hydrophobic leader region (see Figure 7) and the carboxy-terminal region (see Figure 12). (a) *D. melanogaster*, (b) *D. simulans*, (c) *D erecta*, and (d) *D. yakuba*.

```
Dm   187 ccc aca act aca act act tgt gcg cca cgt acc acg
      29 pro thr thr thr thr thr cys ala pro arg thr thr

Dm   223 caa cct ccg tgc aca act acg aca aca aca acc aca
      41 gln pro pro cys thr thr thr thr thr thr thr thr

Dm   259 act act tgt gcg cca ccc aca caa caa tct acc acg
      53 thr thr cys ala pro pro thr gln gln ser thr thr

Dm   295 caa cct cca tgc acg aca tct aag ccc acc aca cct
      65 gln pro pro cys thr thr ser lys pro thr thr pro

Dm   331 aag caa act acc acg caa ctt ccg tgc aca aca ccc
      77 lys gln thr thr thr gln leu pro cys thr thr pro

Dm   367 acc acc act aag gcc acc acc acg aag ccc acc acc
      89 thr thr thr lys ala thr thr thr lys pro thr thr

Dm   403 act aaa gcc acc acc act aag gcc acc acc act aag
     101 thr lys ala thr thr thr lys ala thr thr thr lys

Dm   439 ccc acc acc act aag caa act acc acg caa ctt ccg
     113 pro thr thr thr lys gln thr thr thr gln leu pro

Dm   475 tgc aca aca ccc acc acc act aag caa act acc acg
     125 cys thr thr pro thr thr thr lys gln thr thr thr

Dm   511 caa ctt ccg tgc aca aca ccc acc acc act aag ccc
     137 gln leu pro cys thr thr pro thr thr thr lys pro

Dm   547 acc acc acg aag ccc acc acc acg aag ccc acc acc
     149 thr thr thr lys pro thr thr thr lys pro thr thr

Dm   583 act aag ccc acc acc acg aag ccc acc acc acc aag
     161 thr lys pro thr thr thr lys pro thr thr thr lys

Dm   619 ccc acc acc acg aag ccc acc acc act aag ccc acc
     173 pro thr thr thr lys pro thr thr thr lys pro thr

Dm   655 acc acg aag ccc acc acc act aag ccc acc acc acg
     185 thr thr lys pro thr thr thr lys pro thr thr thr

Dm   691 aag ccc acc acc acg aag ccc acc acc act aag ccc
     197 lys pro thr thr thr lys pro thr thr thr lys pro

Dm   727 acc acc acg aag ccc acc acc act aag ccc acc acc
     209 thr thr thr lys pro thr thr thr lys pro thr thr

Dm   763 acg aag ccc acc acc act aag ccc acc acc acg aag
     221 thr lys pro thr thr thr lys pro thr thr thr lys

Dm   799 ccc acc acc act aag ccc acc acc acg aag ccc acc
     233 pro thr thr thr lys pro thr thr thr lys pro thr

Dm   835 acc acg aag ccc acc acc act aag ccc acc aca cct
     245 thr thr lys pro thr thr thr lys pro thr thr pro

Dm   871 aag
     257 lys
```

```
Ds   188 cca acc aag gcc aca act acc tgt gcg cca ccc acg
      29 pro thr lys ala thr thr thr cys ala pro pro thr

Ds   224 aaa cct aca tgc aaa tct act tcc acc aca acc aca
      41 lys pro thr cys lys ser thr ser thr thr thr thr

Ds   260 act aca acc aca acc aca acc aca acc aca act acc
      53 thr thr thr thr thr thr thr thr thr thr thr thr

Ds   296 cgt gcg cca ccc acg aaa cct aca tgc aaa tct act
      65 arg ala pro pro thr lys pro thr cys lys ser thr

Ds   332 tcc acc aca acc aca act acc cgt gcg cca ccc acg
      77 ser thr thr thr thr thr thr arg ala pro pro thr

Ds   368 aaa cct aca tgc aaa tct act tcc acc aca acc aca
      89 lys pro thr cys lys ser thr ser thr thr thr thr

Ds   404 act acc cgt gcg cca ccc aca act act tgc aaa aca
     101 thr thr arg ala pro pro thr thr thr cys lys thr

Ds   440 agt act aca act acc acc aca cac aaa ccc acc aca
     113 ser thr thr thr thr thr thr his lys pro thr thr

Ds   476 cat tcg acc ccc aaa aca aaa ccc acc aaa cat aca
     125 his ser thr pro lys thr lys pro thr lys his thr

Ds   512 acc ccc aaa aca aaa ccc acc aaa cat aca acc ccc
     137 thr pro lys thr lys pro thr lys his thr thr pro

Ds   548 aaa aca aaa ccc acc aaa cat aca acc ccc aca acc
     149 lys thr lys pro thr lys his thr thr pro thr thr

Ds   584 aca acc acc acc aca cct aag
     161 thr thr thr thr thr pro lys
```

```
De   188 ccc aaa aga acc act ccc aag ccc tgc acc aca gca
      29 pro lys arg thr thr pro lys pro cys thr thr ala

De   224 agg cca act tgc gcg cca gta aca acc acc acc tgt
      41 arg pro thr cys ala pro val thr thr thr thr cys

De   260 agg cca ccc aca act act cgc tgc ccg cca ccc aca
      53 arg pro pro thr thr thr arg cys pro pro pro thr

De   296 act act cgc tgc ccg cca ccc aca agg cca gct gaa
      65 thr thr arg cys pro pro pro thr arg pro ala glu

De   332 tgc acc gca aca act aag cgc ccc aca gct agg ccc
      77 cys thr ala thr thr lys arg pro thr ala arg pro

De   368 aca act aga cgc acc aca gtt agg gcc acc act aag
      89 thr thr arg arg thr thr val arg ala thr thr lys

De   404 cgc gcc aca act agg cgc acc act aaa cgc gcc aca
     101 arg ala thr thr arg arg thr thr lys arg ala thr

De   440 act aga cgc acc aca gtt agg gcc aca act aaa cgc
     113 thr arg arg thr thr val arg ala thr thr lys arg

De   476 gcc aca act agg cgc acc aca act aaa cgc gcc cca
     125 ala thr thr arg arg thr thr thr lys arg ala pro

De   512 act agg cgt acc aca act aag cgt gcc aca act agg
     137 thr arg arg thr thr thr lys arg ala thr thr arg

De   548 cgc aac cca acc aga cgc acc aca act agg cgt gcc
     149 arg asn pro thr arg arg thr thr thr arg arg ala

De   584 cca act aag cgt gcc aca act aag cgt gcc aca act
     161 pro thr lys arg ala thr thr lys arg ala thr thr

De   620 agg cgc aac cca act aag cgc aag aca acc aga cgc
     173 arg arg asn pro thr lys arg lys thr thr arg arg

De   656 acc act gtg agg gcc acc aaa aca act aaa cgc gcc
     185 thr thr val arg ala thr lys thr thr lys arg ala

De   692 aca act aag cgt gcc cca act aaa cgc gcc aca act
     197 thr thr lys arg ala pro thr lys arg ala thr thr

De   728 aag cgt gcc cca act aaa cgc gtc aca acc aag cgt
     209 lys arg ala pro thr lys arg val thr thr lys arg

De   764 gcc cca act aag cgt gcc aca act aag cgt gcc cca
     221 ala pro thr lys arg ala thr thr lys arg ala pro

De   800 act aaa cgc gcc aca act aag cgt gcc cca act aag
     233 thr lys arg ala thr thr lys arg ala pro thr lys

De   836 cgt gcc aca act aag cgc gcc cca act aaa cgc gcc
     245 arg ala thr thr lys arg ala pro thr lys arg ala

De   872 aca acc aag cgt gcc cca act aag cgt gcc aca act
     257 thr thr lys arg ala pro thr lys arg ala thr thr

De   908 aag cgt gcc aca gct agg ccc acc agc aag
     269 lys arg ala thr ala arg pro thr ser lys
```

```
Dy   191 ccc acg acc tct ccc aag ccc tgc caa aca acg gta
          29 pro thr thr ser pro lys pro cys gln thr thr val

Dy   227 ccg act tgt gcg cca aca aca acc act aca aca aca
          41 pro thr cys ala pro thr thr thr thr thr thr thr

Dy   263 acc act tgt gcg cca ccc aca agg cca cct cca cct
          53 thr thr cys ala pro pro thr arg pro pro pro pro

Dy   299 cca tgc aca gac gcc cca acg aca act aag agg aca
          65 pro cys thr asp ala pro thr thr thr lys arg thr

Dy   335 acg gaa aaa tcc acc aca agg aga acc acc aca aca
          77 thr glu lys ser thr thr arg arg thr thr thr thr

Dy   371 act aga caa aca aca act aga cct aca aca act aca
          89 thr arg gln thr thr thr arg pro thr thr thr thr

Dy   407 acc aca acc acc aca act aga cgt cca aca act agg
         101 thr thr thr thr thr thr arg arg pro thr thr arg

Dy   443 tct aca aca act aga cat aca aca act aca acc acc
         113 ser thr thr thr arg his thr thr thr thr thr thr

Dy   479 aca act aga cgt cca aca act aca acc acc aca act
         125 thr thr arg arg pro thr thr thr thr thr thr thr

Dy   515 aga cgt cca aca act aca acc acc aca act aga cgt
         137 arg arg pro thr thr thr thr thr thr thr arg arg

Dy   551 cca aca act aca acc acc aca act aga ctt cca aca
         149 pro thr thr thr thr thr thr thr arg leu pro thr

Dy   587 act aga tct aca aca act aga cat aca act aaa tcc
         161 thr arg ser thr thr thr arg his thr thr lys ser

Dy   623 acc aca tct aag cgt cca aca cat gag acc acc acc
         173 thr thr ser lys arg pro thr his glu thr thr thr

Dy   659 aca tct aag cgt cca aca caa gag acc acc aca acc
         185 thr ser lys arg pro thr gln glu thr thr thr thr

Dy   695 act aga cgt gca aca caa gca acc acc aca cct aaa
         197 thr arg arg ala thr gln ala thr thr thr pro lys

Dy   731 ccc acc aac aag cct
         209 pro thr asn lys pro
```

Figure 10. The prominent repeat portions of the AC-rich regions. A portion of each of the AC-rich regions, that consists of easily recognizable amino acid repeat motifs, is shown. The sequences are aligned to emphasize the structure of the repeats. (a) *D. melanogaster*, (b) *D. simulans*, (c) *D. erecta* and (d) *D. yakuba*.

```
Dm   73  pro thr thr pro lys
Dm   78  gln thr thr thr gln
Dm   83  leu pro cys thr thr
Dm   88  pro thr thr thr lys
Dm   93  ala thr thr thr lys
Dm   98  pro thr thr thr lys
Dm 103 (ala thr thr thr lys) x 2
Dm 113  pro thr thr thr lys
Dm 118  gln thr thr thr gln
Dm 123  leu pro cys thr thr
Dm 128  pro thr thr thr lys
Dm 133  gln thr thr thr gln
Dm 138  leu pro cys thr thr
Dm 143 (pro thr thr thr lys) x 22
Dm 253  pro thr thr pro lys
```

```
Ds  37    ala pro pro thr lys pro thr cys lys ser thr ser (thr) x 16 arg
Ds  66    ala pro pro thr lys pro thr cys lys ser thr ser (thr) x 6  arg
Ds  85    ala pro pro thr lys pro thr cys lys ser thr ser (thr) x 6  arg
Ds 104    ala pro pro thr thr     thr cys lys     thr ser (thr) x 6  his
```

```
Ds 121    lys pro thr thr his     ser thr pro lys thr
Ds 131    lys pro thr lys his     thr thr pro lys thr
Ds 141    lys pro thr lys his     thr thr pro lys thr
Ds 151    lys pro thr lys his     thr thr pro thr thr
```

De 89 thr thr arg arg thr thr val arg ala thr thr lys arg ala thr thr arg arg thr thr lys arg ala

De 112 thr thr arg arg thr thr val arg ala thr thr lys arg ala thr thr arg arg thr thr thr lys arg ala

De 136 pro thr arg arg thr thr thr lys arg ala thr thr arg arg asn pro thr arg arg thr thr thr arg arg ala

De 161 pro thr lys arg ala thr thr lys arg ala thr thr arg arg asn pro thr lys arg lys thr thr arg arg thr

De 186 thr val arg ala thr lys

De 192 thr thr lys arg ala

De 197 thr thr lys arg ala pro thr lys arg ala

De 207 thr thr lys arg ala pro thr lys arg val

De 217 thr thr lys arg ala pro thr lys arg ala

De 227 thr thr lys arg ala pro thr lys arg ala

De 237 thr thr lys arg ala pro thr lys arg ala

De 247 thr thr lys arg ala pro thr lys arg ala

De 257 thr thr lys arg ala pro thr lys arg ala

De 267 thr thr lys arg ala

Dy	85	thr	thr	thr	thr	thr	arg	gln	thr	thr	thr	arg pro
Dy	97	thr	thr	thr	thr	thr						
Dy	102	thr	thr	thr	thr	thr	arg	arg	pro	thr	thr	
Dy	112						arg	ser	thr	thr	thr	
Dy	117						arg	his	thr	thr	thr	
Dy	122	thr	thr	thr	thr	thr	arg	arg	pro	thr	thr	
Dy	132	thr	thr	thr	thr	thr	arg	arg	pro	thr	thr	
Dy	142	thr	thr	thr	thr	thr	arg	arg	pro	thr	thr	
Dy	152	thr	thr	thr	thr	thr	arg	leu	pro	thr	thr	
Dy	162						arg	ser	thr	thr	thr	

Figure 11. The consensus repeat sequences for each of the species. These sequences were derived from inspection of the repeats present in each of the species (see Figure 9). All are very AC-rich on the RNA-like strand (from 73.3% in *D. erecta* to 96.7% in *D. simulans*).

D. melanogaster

ccc	acc	acc	act	aag	ccc	acc	acc	acg	aag
pro	thr	thr	thr	lys	pro	thr	thr	thr	lys

D. simulans

aca	acc	ccc	aaa	aca	aaa	ccc	acc	aaa	cat
thr	thr	pro	lys	thr	lys	pro	thr	lys	his

D. erecta

aca	act	aag	cgt	gcc	cca	act	aaa	cgc	gcc
thr	thr	lys	arg	ala	pro	thr	lys	arg	ala

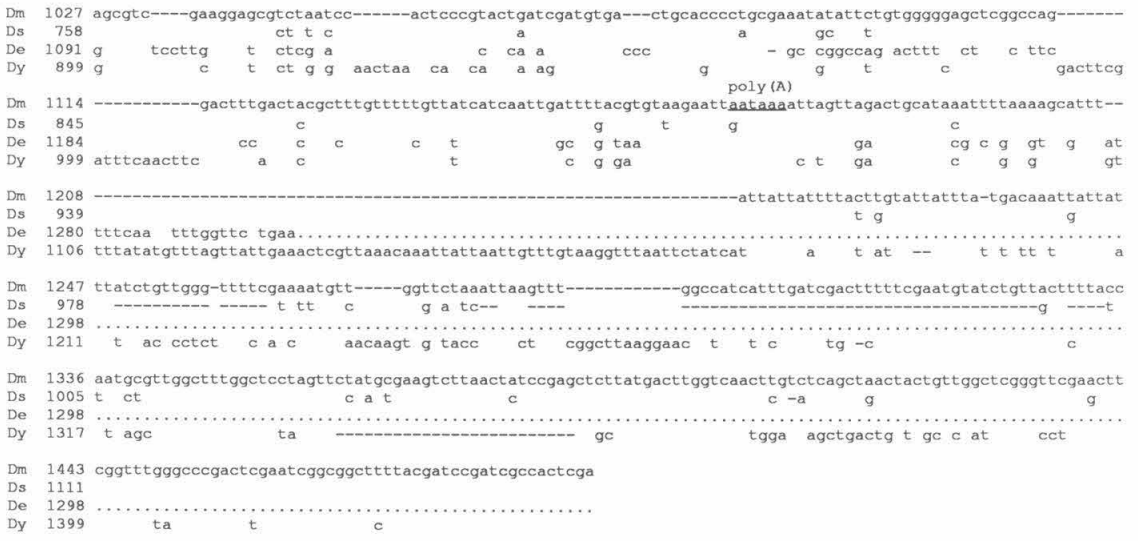
D. yakuba

aca	acc	acc	aca	act	aga	cgt	cca	aca	act
thr	thr	thr	thr	thr	arg	arg	pro	thr	thr

Figure 12. The carboxy-terminal protein coding domain. The sequence alignments are generated as described in Figure 6. The eight cysteine residues of *D. melanogaster*, which are conserved in position (relative to the last amino acid of the protein) in all four genes, are shown in bold type. The '***' codon represents a translation termination signal.

```
Dm   874 ccg tgc ggt tgc aag agc tgc ggt cct gga gga gag
     258 pro cys gly cys lys ser cys gly pro gly gly glu

Ds   605       t                                     c
     168

De   938   a       c          cc                    c
     279                      pro

Dy   746 ggc   t   c          tc                    c ata
     214 gly                                         ile


Dm   910 cca tgc aat gga tgt gct aag agg gat gca ctg tgc
     270 pro cys asn gly cys ala lys arg asp ala leu cys

Ds   641   t       a          g          ag
     180         lys          gly        ser

De   974   c     c a            c
     291         gln

Dy   782   c     gg  a c      c c        c g  g
     226         gly ser      pro        gln gly


Dm   946 cag gat ctt aac ggc gta ctc cgc aat ctg gag cgc
     282 gln asp leu asn gly val leu arg asn leu glu arg

Ds   677           c          t                        g
     192                      leu

De  1010       c   g      aa  a
     303                  asn ile

Dy   818 ac    a          aa  t          c g
     238 thr glu          asn leu        gln


Dm   982 aag atc cgt caa tgc gtc tgc ggt gaa ccg caa tgg
     294 lys ile arg gln cys val cys gly glu pro gln trp

Ds   713 c   g                          c c       g t
     204 gln val                          gln     asp

De  1046     g     c   g                c       c   g
     315     val

Dy   854  g  g     c g g                c c     c gtg
     250 arg val     glu                  gln     val


Dm  1018 ttg ctg tga
     306 leu leu ***

Ds   749
     216

De  1082 c   t    a
     327

Dy   890     t    a
     262
```

Figure 13. The aligned sequences of the 3' untranslated and 3' flanking regions. Alignments and formats are as described in Figure 6. The poly(A) addition signal, beginning at Dm +1180, is underlined in the *D. melanogaster* sequence.


poly(A)