Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 1063–1067, February 1983
Genetics

DNA sequence changes in an upstream DNase I-hypersensitive region are correlated with reduced gene expression

(glue protein genes/*Drosophila*/chromatin structure)

WILLIAM MCGINNIS*, ANTONY W. SHERMOEN, JILL HEEMSKERK, AND STEVEN K. BECKENDORF†

Department of Molecular Biology, University of California, Berkeley, California 94720

Communicated by Robley C. Williams, November 22, 1982

ABSTRACT **Previous experiments have identified a region that is required for the expression of the *Drosophila* glue protein gene *Sgs-4* and is located 300–500 base pairs upstream from the structural gene. The chromatin in this region changes conformation and becomes hypersensitive to DNase I digestion when the gene becomes active, a change that apparently induces additional conformational changes near the site of transcription initiation. To learn more about the DNA sequence requirements for the function of this region, we analyzed three naturally occurring *Sgs-4* underproducers. In two of these strains, a single base pair change within the hypersensitive region is correlated with a 50% reduction in the amount of *Sgs-4* RNA produced. Another strain, which has multiple 5′ lesions, is severely reduced in *Sgs-4* expression and in the DNase hypersensitivity of the upstream region. Several of the sequence changes in this extreme underproducer lie near hypersensitive sites, suggesting that they inhibit the appearance of the normal DNase hypersensitive conformation.**

During eukaryotic development, the activity of many structural genes is limited to specific tissues and specific developmental times. The sequence elements involved in this control have yet to be defined. However, the sequences associated with sites in chromatin that are hypersensitive to digestion with endonucleases like DNase I are good candidates. Many DNase I-hypersensitive sites are associated with the 5′ ends of genes (1, 2) and, for tissue-specific genes, are only found in cells in which the gene is active (3–5).

To explore this possibility further, we have chosen to study naturally occurring variants of the *Sgs-4* locus of *Drosophila*, several of which alter the pattern of DNase hypersensitivity upstream from the gene. *Sgs-4* is expressed only in larval salivary glands and codes for one of the *Drosophila* glue proteins used to attach the pupal case to a dry surface during metamorphosis (6–8). Because there are seven other major glue proteins, *Sgs-4* is dispensable for the fly. Therefore, many *Drosophila* strains that produce little or no *Sgs-4* protein have been identified (6, 7). These naturally occurring mutants have the great advantage that the disrupted function of the gene can be studied in a natural setting, in its normal chromosomal location, and in the tissue in which it is normally expressed during development.

Previous work has shown that *Sgs-4* and all of its *cis*-acting control sequences are limited to a 16- to 19-kilobase (kb) region of the X chromosome (9), and a recent report reveals that naturally occurring deletions of DNA in the region 300–500 base pairs (bp) upstream from *Sgs-4* are correlated with a reduction or loss of gene expression (10). This region normally contains a complex of DNase I-hypersensitive sites that are found only in the chromatin of cells that are actively transcribing *Sgs-4* (5). In this report we find that two different *Sgs-4* underproducers have only one (and indeed the very same) base pair change in this upstream region. Another variant strain has many small changes 5′ to the gene, with the apparent result being a loss of DNase hypersensitivity over the altered sequences and a concomitant severe reduction in expression. The results are consistent with the hypothesis that the *Sgs-4* gene has a remote 5′ control region, 300–500 bp from transcription initiation, which must be in a special chromatin conformation in order to exert its effect on the downstream transcription unit.

MATERIALS AND METHODS

***Drosophila* Strains.** The Oregon R-B strain of *D. melanogaster* (obtained from W. Petri, Boston College) was used as the standard for DNA sequence arrangement near *Sgs-4*. It has been separated from the Oregon R-S strain (obtained from D. Hogness, Stanford Univ.) for at least 30 yr. The Canton S strain (obtained from Caltech) is derived from flies collected in Canton, Ohio, before 1940. Strain D323 (obtained from W. Marks, University of California, Davis) was isolated at the Gundlach–Bundschu Winery in Sonoma, California, in 1976. The Daekwanryeong strain (Daek, obtained from D. Grace, University of Oregon) was isolated in Korea. The X chromosomes of these stocks were made isogenic before the experiments reported here were conducted.

RNA Isolation and Blotting. RNA was isolated from late third-instar *Drosophila* larvae essentially as described by Muskavitch and Hogness (8). Blots of RNA separated on agarose gels were prepared as described by Thomas (11). DNA probes were labeled by nick translation as described by Weinstock *et al.* (12).

Quantitative Dot Blots. Dot blots of total late third-instar larval RNA were done as described (11). Dilutions of Oregon R-B RNA were dotted onto nitrocellulose and probed with the nick-translated *Hin*dIII/*Xho* I 1.4-kb fragment that contains the 3′ end of the *Sgs-4* coding sequences, a region whose length is invariant among the strains tested. The resulting radioactive dots were cut out and assayed by scintillation counting.

A plot of radioactivity (cpm) versus the amount of RNA applied to the dots was linear between 0.25 and 5 μg of RNA. The relative abundance of *Sgs-4* RNA in Canton S, D323, and Oregon R-S was determined by comparison with the Oregon R plot.

Cloning and Sequence Determination. Whole genomic libraries from each of the mutant and standard strains were constructed with the λ590 vector of Murray *et al.* (13). Clones containing the 5′ end of the *Sgs-4* coding region and the 5′ flanking region were isolated by the method of Benton and Davis (14) with the pR1.5 clone as a probe (described in ref. 5). The *Hin*dIII fragments from these isolated clones were subcloned into pBR322,

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

Abbreviations: kb, kilobase(s); bp, base pair(s).

* Present address: Biozentrum, University of Basel, Basel, Switzerland.
† To whom reprint requests should be addressed.

and the sequence of the 5′ region was determined by the method of Maxam and Gilbert (15), with the *Eco*RI, *Xho* I, and *Bam*HI restriction sites just upstream from the *Sgs-4* structural genes (see Fig. 1) used for end labeling.

Assay for DNase I-Hypersensitive Sites. Salivary gland nuclei were analyzed for preferential DNase I digestion near *Sgs-4* as described by Shermoen and Beckendorf (5). In brief, isolated salivary gland nuclei were treated for 3 min with an appropriate amount of DNase I, and the digestion was stopped by addition of EDTA (final concentration, 20 mM). DNA was then purified, digested with *Sal* I, and analyzed by Southern blotting.

RESULTS

***Sgs-4* Phenotypes of the Variant Strains.** Variation in the amount of *Sgs-4* expression is a common phenotype among wild strains of *D. melanogaster* (7, 16). For the current experiments, we chose three strains, Canton S, D323, and Daek and compared them with our standard, the Berkeley strain of Oregon R (Oregon R-B), and with a strain of Oregon R from Stanford (Oregon R-S) whose *Sgs-4* locus has been analyzed by Muskavitch and Hogness (10). As shown by RNA blot analysis and by semiquantitative assays, Canton S and D323 produced about half as much *Sgs-4* RNA as did Oregon R-B, whereas Daek produced very small amounts (Fig. 1). By these methods, the amounts of *Sgs-4* RNA in Oregon R-S were indistinguishable from those in Oregon R-B. This blot also showed the difference in size of the Canton S and D323 transcripts which, as pointed out by Muskavitch and Hogness (10), is caused by a difference in the number of copies of an intragenic tandem repeat sequence.

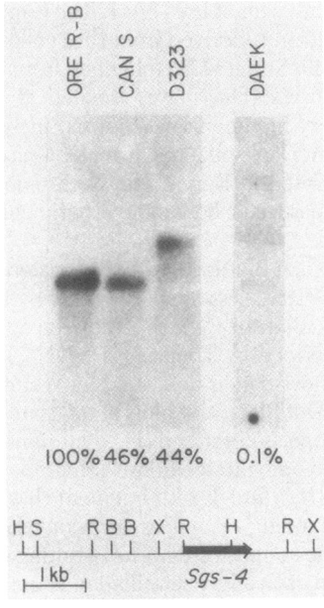

FIG. 1. (*Upper*) *Sgs-4* RNA from the variant strains. (*Upper Left*) Autoradiographic signals from a RNA blot of Oregon R-B (Ore R-B), Canton S (Can S), and D323 RNAs (5 μg each) from late third-instar larvae. The blot, prepared from RNA separated on a 1.2% agarose gel, was hybridized with nick-translated DNA from the plasmid pRH0.75. This clone contains the 5′ three-quarters of the *Sgs-4* coding sequence, which is bounded by an *Eco*RI and a *Hin*dIII site. (*Upper Right*) Autoradiographic signals from a RNA blot of 20 μg of Daek late third-instar salivary gland RNA, also probed with pRH0.75. Percentages beneath the lanes indicate the relative abundance of *Sgs-4* RNA in the four strains. Values for Can S and D323 were determined from dot blots. The value for Daek was obtained by using the gel blot procedure and comparing the Daek signal with that from dilutions of Oregon R-B RNA from the same gel. (*Lower*) Restriction map of the *Sgs-4* gene and immediate surrounding region in Oregon R-B DNA. The thick arrow indicates the location of the transcribed region and the direction of transcription. H, *Hin*dIII; S, *Sal* I; R, *Eco*RI; B, *Bam*HI; X, *Xho* I.

Sequence Changes 5′ to *Sgs-4* in the Variant Strains. Previous studies had shown that *Sgs-4* expression is disrupted by deletions of 50–100 bp in the region 300–500 bp upstream from the structural gene (5, 10). It then became of interest to examine the effects of smaller changes in the DNA sequence of this region. Canton S, D323, and Daek were chosen for this study because Southern blot analysis of genomic DNA from these strains showed no change in the size of restriction fragments within 1.45 kb 5′ to *Sgs-4*. For further analysis, the 3.4-kb *Hin*dIII fragment containing most of the *Sgs-4* structural gene and its 5′ flanking sequences was cloned into λ590 from Oregon R-B and from the three variant strains. The DNA sequence of each strain was then determined from at least 660 bp 5′ to the gene to at least 230 bp within the structural gene. Fig. 2 shows the sequence from position −600 to −1 for these four strains and for the previously analyzed Oregon R-S (10, 17).

The two Oregon R strains have accumulated nine apparently neutral sequence differences in these 600 bp. This result suggests that, in several parts of this region, a specific sequence is not required for normal *Sgs-4* expression. In the light of this result, it is perhaps surprising that two of the variants, Canton S and D323, showed the same sequence for this entire region though their derivation from wild populations occurred at different times and places. That our Canton S and D323 strains are not simply the result of cross-contamination in the laboratory or reisolation of the same strain is shown by the fact that the flanking regions are not identical. Canton S and D323 sequences differ in the size of restriction fragments 1.5 kb upstream from the gene, in DNA sequence at position −638 (data not shown), and in the size of the *Sgs-4* structural gene (ref. 10; see also Fig. 1). The other striking fact about the Canton S/D323 sequence is that it shows only two differences from the Oregon R-B sequence, one at position −344 and one at −520. The C present at −520 is found also in the Oregon R-S strain, which suggests that it is neutral. This leaves the T at position −344 as the only sequence change between −638 and +159 that is unique to these two underproducers.

The sequence of Daek, which had very low *Sgs-4* expression, is more drastically changed from the Oregon R sequences. It has nine alterations—seven of them single-base pair substitutions, one a single-base pair insertion and one a two-base pair deletion—which were not found in either *Sgs-4*$^+$ strain. At the nine sites of disagreement between the two Oregon R sequences, the Daek sequence has six sites that agree with those of Oregon R-S and three that agree with those of Oregon R-B.

Chromatin Conformation in the Underproducers. As described in detail (5), the chromatin upstream from *Sgs-4* undergoes a change in conformation when the gene is active. This change is detected as a tissue-specific complex of five DNase I-hypersensitive sites, three of which are clustered upstream from the gene at positions −480, −405, and −330 and two of which are near the site of transcript initiation at positions −70 and +30. Analysis of deletion mutants suggests that these regions interact, the more distal region being required for the formation of the sites nearer the gene (5).

Fig. 3 compares the pattern of hypersensitive sites found in salivary gland nuclei from Oregon R-B and the three underproducers. The locations of these sites in Canton S and D323 are the same as those in Oregon R-B, but the assay is not sufficiently quantitative to detect small changes in the sensitivity of the sites. In contrast, sensitivity of the *Sgs-4* 5′ region in Daek is strikingly reduced. Only the normally most prominent site at position −405 is easily detectable. Longer exposures of the autoradiogram suggest that the other four sites are still prefer-

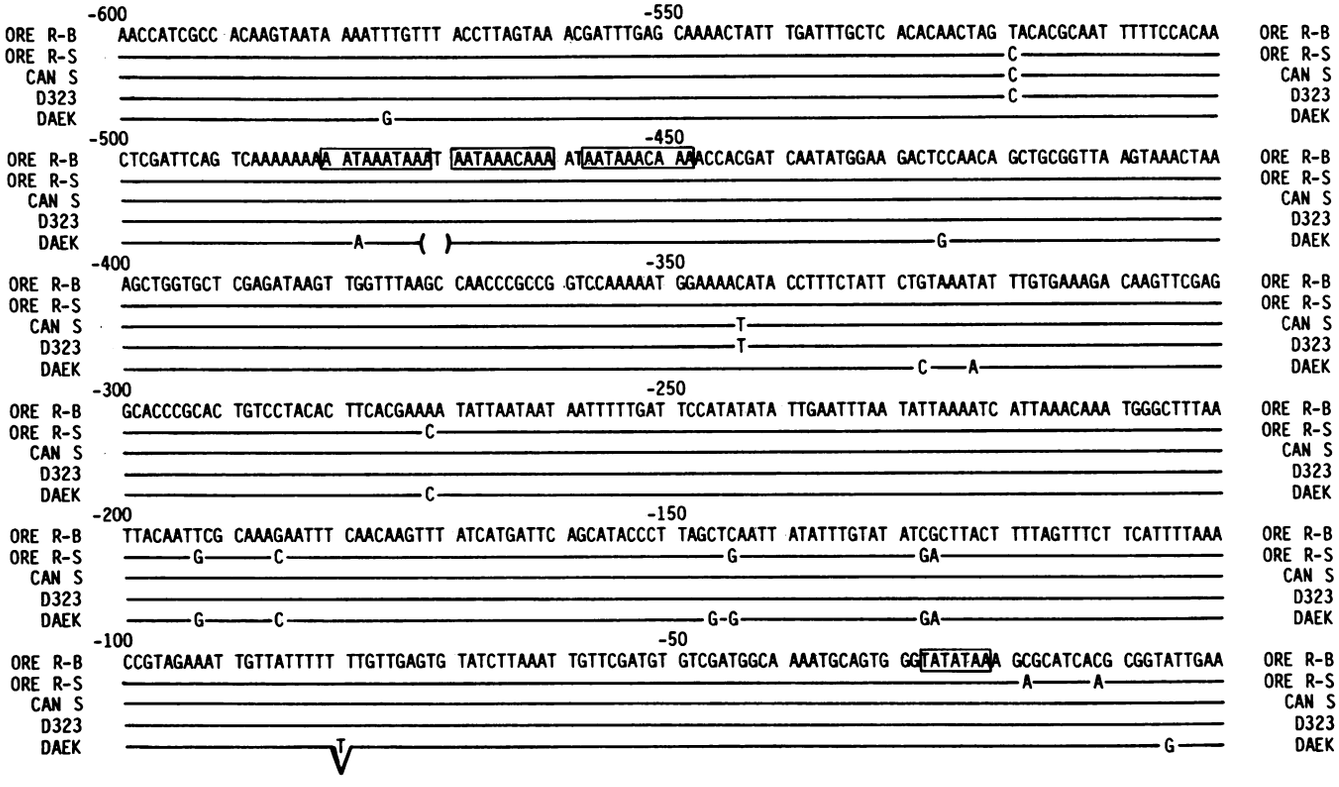

FIG. 2. DNA sequence upstream of *Sgs-4* in Oregon R-B, Oregon R-S, Canton S, D323, and Daek. The numbering system is that used by Muskavitch and Hogness (10), with the −1 position assigned to the first base upstream from transcription initiation. All the sequences shown in this figure were determined in this study except for the Oregon R-S sequence, which is found in Muskavitch and Hogness (10) and Muskavitch (17). Solid lines for the variant strains indicate identity with the Ore R-B DNA sequence. Base substitutions interrupt the line where the variants differ from the Ore R-B sequence. Open parentheses indicate a deletion of sequence encompassed by the parentheses, and the 1-bp insertion in Daek DNA is shown by a base atop a V. The "TATA" sequence from −22 to −29 is boxed, as is a triple tandem repeat of 10 bp located between −449 and −481.

entially sensitive to DNase I, but not by much.

This figure also shows that an endogenous nuclease that was occasionally present in preparations of salivary gland nuclei recognized some aspects of the 5′ hypersensitive regions. Although its cuts at positions −70 and +30 appeared to be similar to those made by DNase I, it cut within the distal region almost exclusively at or near the −330 site. There is a minor cut near position −380 but no evidence for cleavage at the −405 or −480 DNase I sites.

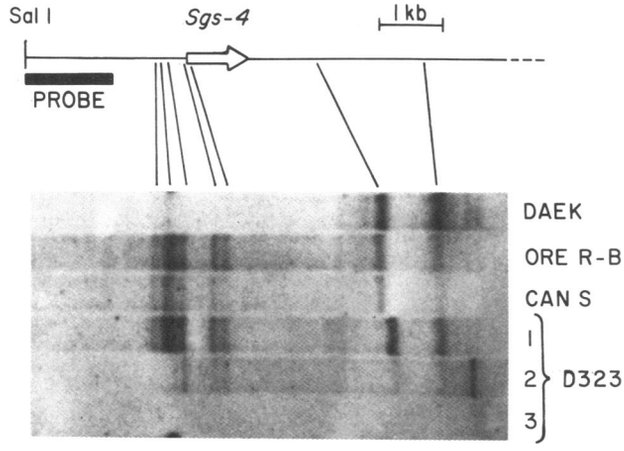

FIG. 3. DNase I-hypersensitive sites in *Sgs-4* underproducers. Salivary gland nuclei were isolated from mid-third-instar larvae and treated with DNase I. DNase I concentrations were 5.8 μg/ml for Daek, 7.8 μg/ml for Oregon R-B and Canton S, and 3.9, 0, and 0 μg/ml for the three D323 samples. Incubations were for 3 min at 25°C for all except the D323-3 sample, which remained at 0°C. DNA was then purified and digested with *Sal* I, and each sample was analyzed by the Southern blotting procedure. The probe used was a 1.35-kb *Sal* I–*Bam*HI fragment isolated from the pSB1.35 clone. The position that corresponds to the *Sgs-4* transcribed region is indicated on the diagram by an open arrow. Guidelines connect the 5′ and 3′ hypersensitive sites to the fragments produced by the DNase I treatment. The fragments produced by the 3′ hypersensitive sites are longer for D323 because its gene is larger (10).

DISCUSSION

Evidence from both genetic and molecular studies indicate that a region just upstream of the *Sgs-4* structural gene is important for its proper expression. In the Samarkand underproducer strain, Korge (16) found one recombinant that at least partially separated its defect in *Sgs-4* expression from the gene. His results locate the Samarkand defect 5′ to the gene, and the rare occurrence of the recombination event places the defect near the 5′ end of the gene. Important parts of the 5′ flanking region were then indicated by a comparison between salivary gland-specific DNAse I-hypersensitive sites and the DNA sequences missing in other *Sgs-4* nonproducers. The top line of Fig. 4 illustrates the correspondence between the three upstream hypersensitive sites and two kinds of deletions that strongly reduce or eliminate gene expression (5, 10). Together these define an ≈200-bp region and suggest that it is required for gene expression.

Comparison of the DNAs of the two Oregon R strains shows that a significant amount of sequence variation can be tolerated 5′ to *Sgs-4* without impairing gene expression. However, although the nine differences between the strains are distributed throughout most of the region shown in Fig. 4, none alter the region containing the three distal hypersensitive sites.

This conserved region is altered in the Canton S and D323 sequence by a single-base pair substitution—the C-to-T change

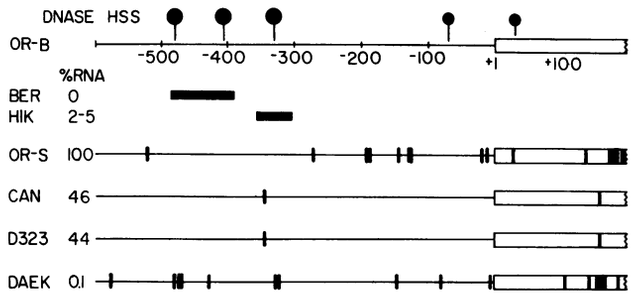

FIG. 4. Summary of chromatin structure and DNA sequence variation 5′ to *Sgs-4*. The top line represents the Oregon R-B (OR-B) DNA sequence for 600 bp 5′ to *Sgs-4* and for about 150 bp of the structural gene. ▼, Position of the five tissue-specific DNase I-hypersensitive sites (DNase HSS); ■, positions of two kinds of deletions, Ber-1 (BER) and Hikone (HIK), characterized by Muskavitch and Hogness (10). Vertical bars on the Oregon R-S (OR-S) line indicate positions at which its sequence differs from that of Oregon R-B. For clarity of presentation, the vertical bars on the Canton S (CAN), D323, and Daek lines indicate only those positions at which the DNA sequence of these variants differs from that of both Oregon R strains. The changes in the Daek sequence at positions −471 and −80 correspond, respectively, to a two-base pair deletion and a one-base pair insertion (see Fig. 2). All of the other bars indicate single-base pair substitutions. With the exception of the initial change within the Oregon R-S gene, all changes indicated in the coding sequence lie within the array of 21-bp tandem repeats (10).

at position −344—which is the only point between positions −638 and +159 at which their sequence differs from those of both Oregon R strains. [A partial sequence of the *Sgs-4* 5′ region of Canton S DNA (−21 to −386) has been published (10). However, because it was compared with the Oregon R-S sequence, which is different at seven places in this region, it was difficult to pick out any particular sequence change as important for gene expression. The serendipitous resemblance of the Oregon R-B DNA sequence to that of Canton S focuses attention on the substitution at position −344.] The extended region of DNA sequence identity between Canton S and D323, their indistinguishable RNA phenotypes, and the fact that their deviation from wild-type DNA sequence occurs within the previously identified hypersensitive region suggest the possibility that the substitution at position −344 represents a true point mutation that disrupts control of *Sgs-4* expression. Nevertheless, it is still possible that sequence changes outside the region we report here could be responsible for the changes in expression. These alternative possibilities can now be directly tested by making *in vitro* recombinants of variant *Sgs-4* loci and reintroducing them into flies through the DNA-mediated gene-transfer method recently developed by Rubin and Spradling (18).

The region 5′ to *Sgs-4* in Daek DNA has many sequence changes not found in the DNAs of the wild-type strains, so that it is difficult to judge which are the most important. However, we note that there are five changes in the conserved, hypersensitive region. Two of these, a G-to-C change position at −328 and a T-to-A change at position −323 lie very near the −330 hypersensitive site and not far from the lesion in Canton S and D323 DNAs. Another two of the Daek DNA changes, a T-to-A substitution at position −479 and a deletion of base pairs at positions −470 and −471, lie near the −480 hypersensitive site. Muskavitch and Hogness (10) noticed a 3-fold direct repeat in this region (boxed in Fig. 2), the distal element of which is eliminated by the two Daek DNA changes. The correspondence of these four sequence changes with hypersensitive sites and the severe reduction in hypersensitivity of this region in Daek DNA suggests to us that these changes inhibit the formation of the active chromatin conformation, which we detect as DNase hypersensitivity; this alteration in chromatin structure not being accomplished, the gene is not efficiently expressed. The correlation between loss of the hypersensitive sites and loss of gene expression was first observed with the deletion mutant BER 1. It lacks sequences corresponding to the two distal hypersensitive sites, forms no hypersensitive sites even though sequences for the three proximal sites are intact, and produces no detectable *Sgs-4* RNA (5, 10). Whether there is a causal relationship between the formation of the hypersensitive sites and activity of the gene can now be tested by DNA-mediated gene transfer.

Analysis of the *Sgs-4* 5′ sequence of Daek DNA and especially of Canton S and D323 DNAs shows that small changes in the hypersensitive region from −300 to −500 are correlated with reductions in gene expression. If the sequence changes actually cause the reduced expression, then the specific sequence of the control region, not just some more general property like its base composition, must be important for *Sgs-4* expression. In contrast, several parts of the sequence surrounding this region can be changed without affecting expression. This result implies that the region 5′ to *Sgs-4* is a mosaic composed of regions whose precise sequence is unimportant, interspersed with sequences important for expression—the cap sequence, TATAA box, and the hypersensitive region. Another region that may be important is that near position −70, which is associated with a DNase I-hypersensitive site and has a sequence that is related to the G-G-$\frac{C}{T}$-C-A-A-T-C-T sequence found upstream from several other eukaryotic genes (10, 19).

Tissue-specific DNase I-hypersensitive sites have been found near several other developmentally regulated genes, including the chicken α- and β-globin genes (4, 20, 21) and the rat preproinsulin gene (3). Their similar location and limitation to cells in which the adjacent gene is actively expressed suggest that their role may be similar to those associated with *Sgs-4*. The region 5′ to the chicken adult β-globin gene has been analyzed by digestion with several other endonucleases in addition to DNase I. Each showed a unique pattern of hypersensitive sites spread through an ≈200-bp region (21). This result suggested that the entire 200-bp region might be accessible to nucleases and that each enzyme picked out its own preferred sequences within the region. Two aspects of our results suggest that the upstream hypersensitivity at the *Sgs-4* locus may result from a similar exposed region of ≈200-bp and that the discrete nature of the three DNase I-hypersensitive sites we detect may not arise from three separate openings in the chromatin structure. First, the −405 and −330 sites are not always as clearly separated as those shown in Fig. 3. In some preparations a continuous region of hypersensitivity extends proximally from the prominent −405 site to about −300 (see figure 3 of ref. 5). Second, the salivary gland endogenous nuclease has a different pattern of preferential cleavage within the region, ignoring the −480 and −405 sites, cleaving at position −330, and having a minor site at about −380 (Fig. 3).

We thank Marc Muskavitch and David Hogness for continual communication of their results of *Sgs-4* organization and for providing us with several *Sgs-4* cloned sequences. We thank Barbara Kellogg for helping to prepare the figures and for excellent typing. This research was supported in part by Basil O'Connor Starter Research Grant 5-222 from the March of Dimes Birth Defects Foundation, by Grant HD12866 from the National Institute of Child Health and Human Development, and by Grant PCM80-21834 from the National Science Foundation.

1. Wu, C. (1980) *Nature (London)* **286**, 854–860.
2. Keen, M. A., Corces, V., Lowenhaupt, K. & Elgin, S. C. R. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 143–146.
3. Wu, C. & Gilbert, W. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 1577–1580.
4. Stalder, J., Larsen, A., Engel, J. D., Dolan, M., Groudine, M. & Weintraub, H. (1980) *Cell* **20**, 451–460.

5. Shermoen, A. W. & Beckendorf, S. K. (1982) *Cell* **29,** 601–607.
6. Korge, G. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 4550–4554.
7. Beckendorf, S. K. & Kafatos, F. C. (1976) *Cell* **9,** 365–373.
8. Muskavitch, M. A. T. & Hogness, D. S. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 7362–7366.
9. McGinnis, W., Farrell, J., Jr., & Beckendorf, S. K. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 7367–7371.
10. Muskavitch, M. A. T. & Hogness, D. S. (1982) *Cell* **29,** 1041–1051.
11. Thomas, P. S. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 5201–5205.
12. Weinstock, R., Sweet, R., Weiss, M., Cedar, H. & Axel, R. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 1299–1303.
13. Murray, N. E., Brammer, W. J. & Murray, K. (1977) *Mol. Gen. Genet.* **150,** 53–61.
14. Benton, W. D. & Davis, R. W. (1977) *Science* **196,** 180–182.
15. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
16. Korge, G. (1981) *Chromosoma (Berlin)* **84,** 373–390.
17. Muskavitch, M. A. T. (1980) Dissertation (Stanford Univ., Stanford, CA).
18. Rubin, G. M. & Spradling, A. C. (1982) *Science* **218,** 348–353.
19. Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50,** 349–383.
20. Weintraub, H., Larsen, A. & Groudine, M. (1981) *Cell* **24,** 333–344.
21. McGhee, J. D., Wood, W. I., Dolan, M., Engel, J. D. & Felsenfeld, G. (1981) *Cell* **27,** 45–55.