Volume 9 Number 19 1981 Nucleic Acids Research

Nucleotide sequences at the boundaries between gene and insertion regions in the rDNA of Drosophila melanogaster

Igor B.Dawid and Martha L.Rebbert

Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205, USA

Received 26 June 1981

ABSTRACT

Ribosomal RNA genes interrupted by type 1 insertions of 1 kb and 0.5 kb have been sequenced through the insertion region and compared with an uninterrupted gene. The 0.5 kb insertion is flanked by a duplication of a 14 bp segment that is present once in the uninterrupted gene; the 1 kb insertion is flanked by a duplication of 11 of these 14 bp. Short insertions are identical in their entire length to downstream regions of long insertions. No internal repeats occur in the insertion. The presence of target site duplications suggests that type 1 insertions arose by the introduction of transposable elements into rDNA. Short sequence homologies between the upstream ends of the insertions and the 28S' boundaries of the rRNA coding region suggest that short type 1 insertions may have arisen by recombination from longer insertions.

We have sequenced both boundaries of two molecules containing type 2 insertions and the upstream boundary of a third; the points of interruption at the upstream boundary (28S' site) differ from each other in steps of 2 bp. Between the boundary in the 0.5 kb type 1 insertion and the type 2 boundaries there are distances of 74, 76, and 78 bp. At the downstream boundary (28S" site) the two sequenced type 2 insertions are identical. The rRNA coding region of one molecule extends across the insertion without deletion or duplication, but a 2 bp deletion in the RNA coding region is present in the second molecule. Stretches of 13 or 22 adenine residues occur at the downstream (28S") end of the two type 2 insertions.

INTRODUCTION

In Drosophila melanogaster more than half of the rDNA repeating units are interrupted by noncoding regions, named ribosomal insertions. Two nonhomologous sequence types have been found, each in several size classes (reviewed in ref. 1). Genes interrupted by both types of insertions are transcribed at very low rates implying that these genes are not effective in rRNA production (2,3). The major type of ribosomal insertion in D. virilis is of type 1 and has a target site duplication of 14 bp at the point where the rRNA coding region is interrupted (4). We have sequenced the boundaries of type 1 and type 2 insertions in D. melanogaster rDNA to determine the

relations between these structures.

MATERIALS AND METHODS

The cloned rDNA molecules used in this study have been described previously (3,5). The insertion boundary regions had been placed approximately with respect to restriction sites by heteroduplex and RNA/DNA hybrid mapping in the electron microscope. Detailed restriction maps of the relevant regions of the molecules were derived by double and triple digestion, and by the end labeling methods of Botchan et al. (6).

DNA sequence determination was carried out by the method of Maxam and Gilbert (7), using in most cases an alternate A+G reaction suggested by J. Shine (see 8). Endlabeling of restriction fragments was achieved mostly by polynucleotide kinase at 5' ends, but 3'-terminal labeling with terminal transferase, DNA polymerase or reverse transcriptase was also carried out (7,9). Secondary digestion of restriction fragments followed by electrophoretic separation, and strand separation were carried out according to the protocols of Maxam and Gilbert (7).

RESULTS

Type 1 Insertions

Among the several size classes of type 1 insertions those of 5 kb are the most abundant while those of 0.5 and 1 kb are moderately frequent (5,10). We have chosen to study one example of each of the two short insertions. Figure 1 shows maps of these cloned rDNA molecules and indicates the regions that were sequenced. Figure 2 presents the sequence of about 450 bp of the uninterrupted rRNA gene Dmr Y22 and the entire sequences of the two insertions. The following points emerge from these data. (i) The rRNA coding region of the cloned molecule Y22 is identical to the sequence of the uninterrupted, independently cloned rDNA molecule reported by Roiha et al. (11), with a single difference. The rRNA coding regions of interrupted genes is highly homologous with the uninterrupted genes, with only four substitutions in a total of 700 bp that were compared. (ii) The downstream 550 bp of the 1 kb insertion are highly homologous to the entire 0.5 kb insertion, with only four substitutions. This finding extends earlier data based on hybridization and heteroduplex experiments (5,12). (iii) The 0.5 kb insertion is flanked by a target site duplication of the same 14 bp that have been found previously at the ends of the D. virilis 10 kb insertion (4). The 1 kb insertion has a duplication of only 11 bp. These relations are summarized in

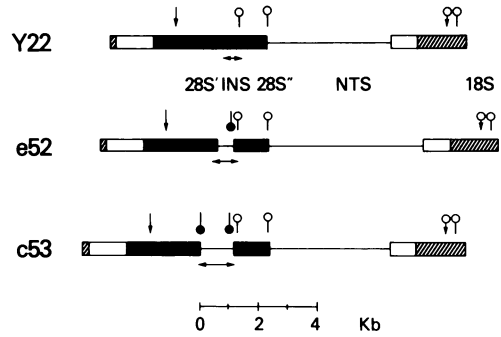


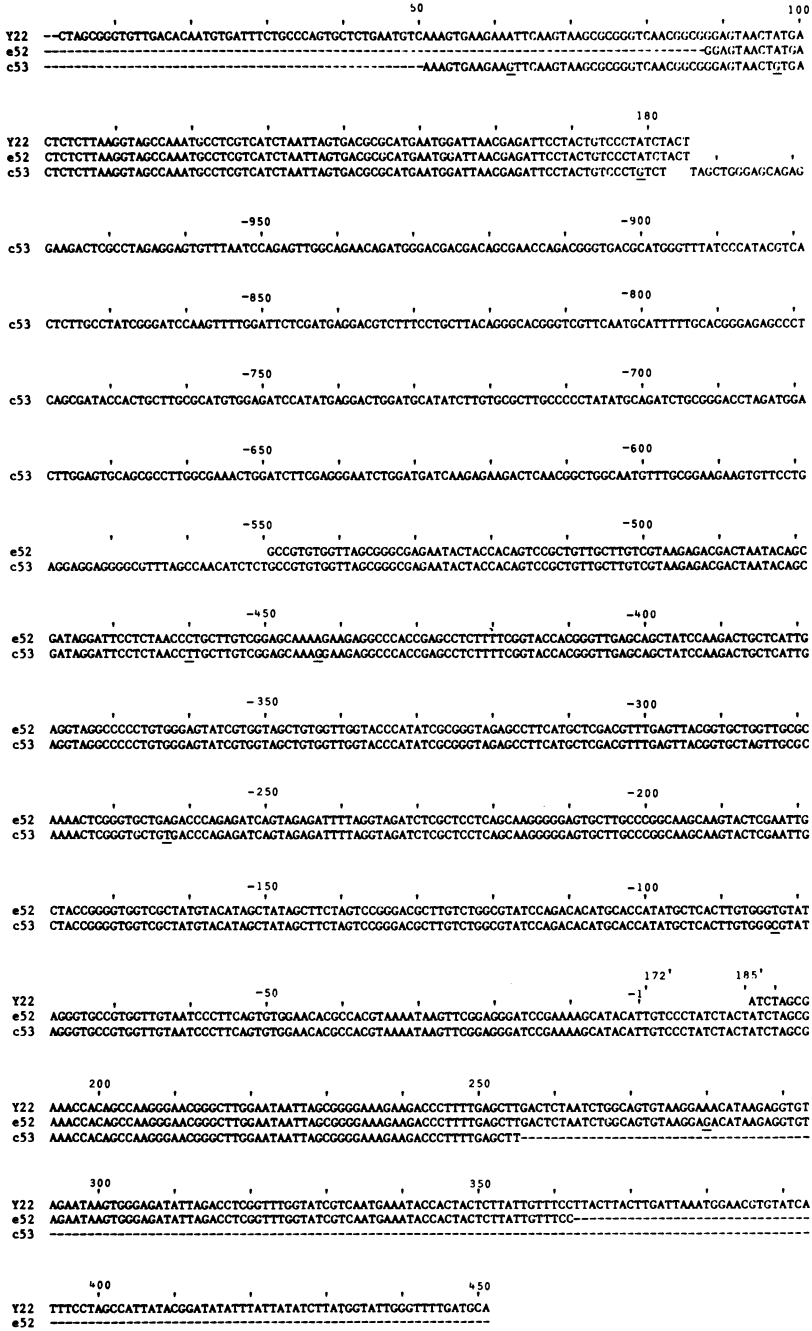
Figure 1. Maps of rDNA fragments used for sequence analysis. The cloned molecule Dmr Y22 contains an uninterrupted rRNA gene from the Y chromosome. Dmr e52 contains a 0.5 kb and Dmr c53 contains a 1 kb type 1 insertion. All fragments are terminated by EcoRI sites, as described previously (5). The following restriction sites are shown: ( ) HindIII; ( ) BamHI; ( ) SmaI. The horizontal double arrows below each map define the regions that have been sequenced.

Figure 5. In view of the presence of 14 bp duplications in the 0.5 kb and the D. virilis insertions one may think of the situation in the 1 kb insertion as a 14 bp duplication followed by a 3 bp deletion.

We have searched both strands of the available insertion sequences for protein reading frames by the computer program of Queen and Korn (13). The longest coding stretch between a methionine and a termination codon corresponds to 104 amino acids, and additional reading frames between 80 and 100 amino acids exist. We have no data to suggest that any of these reading frames are actually translated into proteins.

Type 2 Insertions

About 15% of all rDNA repeating units are interrupted by type 2 insertions (14). Among these sequences, size classes of 1.4 and 3.5 kb are the most common (3). In addition, there exist rare rDNA molecules which have the sequence arrangement 28S' - type 2 insertion - type 1 insertion (3,15). Figure 3 shows maps of one cloned molecule with each of these structures, and Figure 4 presents the results of sequence analysis of these molecules. As in the case of type 1 molecules the short (Dmr 249) and long (Dmr 303) type 2 insertions are precisely homologous at their downstream (28S") side, while at their upstream (28S') side there is no detectable homology. No duplication occurs at the ends of either of the two sequenced type 2 insertions. The upstream points of interruption of the rRNA coding sequence differ for all



three genes that were studied. Dmr 249, which carries a short insertion of 1.4 kb, is interrupted at a position 74 bp upstream from the type 1 interruption in Dmr e52. The RNA coding region in 249 continues at the downstream site without any deletion. In contrast, Dmr 303, which carries a 3.5 kb insertion, is interrupted 2 bp upstream from the 249 breakpoint. Since the downstream boundaries are identical between 249 and 303, 2 bp from the RNA coding region are missing in Dmr 303 (see also Figure 5). Dmr 205 carries a type 2 insertion which is followed by a type 1 insertion; whether a gene region follows further downstream is not known because the cloned molecule ends. Therefore we could analyze the upstream breakpoint only in Dmr 205, and find that it occurs 4 bp upstream from the 249 breakpoint. Two cloned rDNA molecules interrupted by type 2 insertions that were analyzed by Roiha et al. (11) are interrupted at the same position as our clone 249 (see also Discussion).

A characteristic feature of type 2 insertions is the presence of a stretch of A residues at the downstream boundary. Dmr 303, and Dm 264 of Roiha et al. (11) carry 22 As, while Dmr 249 has only 13 As. It is notable that there is perfect homology between all the type 2 insertions upstream of the stretch of As and again in the gene region downstream from the insertion site in spite of the different number of As present. This finding suggests that the A stretch may have originated separately from the insertion itself.

Figure 2. Sequences of uninterrupted rDNA (Y22), and type 1 insertions of 0.5 kb (e52) and 1 kb (c53). The sequences were obtained from fragments of the cloned molecules shown in Figure 1. The boundary regions were sequenced across both strands, but the sequences of some segments away from the boundaries are derived from one strand only. The sequences were lined up for presentation as follows. Actually determined sequences are shown explicitely. Corresponding regions in other cloned molecules that have not been sequenced are indicated by dashed lines; e.g., Dmr e52 does extend from position 1 to 50 but we have not determined the sequence in this region. Empty spaces indicate that the particular clone does not extend through that region, the space being introduced to allow line-up of homologous sequences. Thus, the sequence of Dmr Y22 continues from the T at position 185 to the A at position 186; the space was introduced to allow showing the insertions in their appropriate positions. The gene sequence is numbered with positive values from a HaeIII site in the 28S' RNA region. The nucleotides duplicated at the downstream end of the insertion have been assigned "prime" numbers. The insertions are given negative numbers, starting with the nucleotide immediately upstream from the duplicated segment. Nucleotides differing from Y22 in the gene region, and from e52 in the insertion, are underlined.

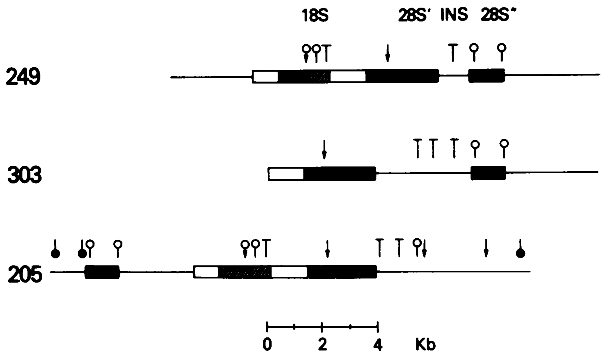


Figure 3. Maps of rDNA fragments containing type 2 insertions. These molecules have been described previously (3). The same restriction sites are shown as in Figure 1 in addition to EcoRI sites ( ).

```
                                                  50                                                100
Y22  --CTAGCGGGTGTTGACACAATGTGATTTCTGCCCAGTGCTCTGAATGTCAAAGTGAAGAAATTCAAGTAAGCGCGGGTCAACGGCGGGAGTAACTATGA
249  ------------------------------------------------------------------------CGCGGGTCAACGGCGGGAGTAACTATGA
303  ---------------------------------------------------------------------AAGCGCGGGTGAACGGCGGGAGTAACTATGA
205  -----------TTGACACAATGTGATTTCTGCCCAGTGCTCTGAATGTCAAAGTGAAGAAATTCAAGTAAGCGCGGGTGAACGGCGGGAGTAACTATGA

Y22  CTCTCTTAAGG
249  CTCTCTTAAGGGCCCGAGATAGATTATTAGCTGAACAAAATGAACTGTTATCGCGTCCGGCAATAGAAAAATATTGGGCGAACAAGTTGTACCTCTCAGT
303  CTCTCTTAA  TTGGGGATCATGGGGTATTTGAGAGCAGAGGGGGAGTATTCTTCTGTAATTCGTAAGTCATATCATATGATGTGCGGAAGGGGAATTTT
205  CTCTCTT    CGCCCTTCGGACTGTCCTTATCCCACAGCTCTACCACAAGTTAGCCCTTGGGAGTGTGGCGATAGGCGTCCTAAGAAAAACTGATAAAC

249  TGATGGTAGCGAACTCCGTGAAGCGGACCAT---------------------------------------------------------------------
303  ACTCTGTAACTCACAAGTCTCTCCTTTACTCAAG------------------------------------------------------------------
205  TTATAAGATATTATGTGCGAAGATGGCTAAATCTTCCGCTGGATGTGCCGATAGCATTTGTTCATGCACCCCCAAAAAGTGGAGGTCTCGGAA-------

249  ~ 1.2kb ----------------AATGAATCGTGCATGCTAGGAAAACTGACCACACACAGTGTTGGCAGACCTAGTATCTTTCGAAGATTTCCATACC
303  ~ 2.3kb -TCGGTCATATAGAGCAATGAATCGTGCATGCTAGGAAAACTGACCACACACAGTGTTGGCAGACCTAGTATCTTTCGAAGATTTCCATACC

                                140
Y22                TAGCCAAATGCCTCGTCATCTAATTAGTGA--------
249  TCCGCGATC(A)13TAGCCAAATGCCTCGTCATCTAATTAGTGA--------
303  TCCGCGATC(A)22TAGCCAAATGCCTCGTCATCTAATTAGTGA--------
```

Figure 4. Sequences of rDNA repeats containing type 2 insertions. The sequences are presented in the same way as in Figure 2, except that the insertions have not been sequenced throughout. At the downstream end the insertion sequences in Dmr 249 and 303 are identical, but at the upstream end these insertions, and that in Dmr 205, share no detectable homology. The multiple A residues at the downstream boundaries are indicated by subscripts. Only the gene regions have been numbered. In Dmr 205 only the upstream boundary was sequenced.

DISCUSSION

Target Site Duplication in Type 1 Insertions

A duplication of the same 14 bp that flank the 10 kb insertion of D. virilis rDNA (4) also occurs at the boundaries of the 0.5 kb insertion of D. melanogaster (Figure 5). These observations suggest that type 1 insertions arose by the introduction of a transposable DNA element into rDNA, since target site duplications are a feature of prokaryotic and eukaryotic transposons (16,17). An interpretation of the origin of type 1 insertion is complicated by the fact that the 1 kb insertion has only an 11 bp duplication, suggesting either a different insertion event, or the loss of 3 bp from the original duplication by deletion. A further complication arises from the observation that the most common insertion in D. melanogaster, with a length of 5 kb, shows no target site duplication but rather a deletion of 9 bp from the RNA coding sequence (11,18). Nevertheless is seems likely that the 5 kb insertion is the basic unit of type 1 sequences, because of its abundance, and because it is also present in duplicated form in D. virilis (4,18). Thus one may have to suggest that the original 5 kb insertion did contain the 14 bp duplication but lost it subsequently.

Origin of Short Type 1 Insertions

Sequence homologies exist between a short segment in the 1 kb insertion

Figure 5. Summary of boundary structures. Y22 is an uninterrupted rRNA gene; rDNA molecules with type 1 and type 2 insertions are indicated to the left and to the right of their sequences. Gene regions are presented in small type, the target site duplication is in large type, and the insertions are shown in italics. Dashes indicate variable lengths of sequence that are not shown.

just upstream of the 28S' end of the 0.5 kb insertion and the segment of rRNA coding region included in the target site duplication. These homologies are shown in Figure 6 of the paper by Rae (18); they involve nucleotides 179 to 185 of Y22, ATCT(A)CT, and nucleotides -545 to -550 of c53, ATCTCT (Figure 2). One may hypothesize that unequal crossing over between rDNA repeats carrying long insertions and uninterrupted repeats may have lead to the generation of short insertions. This hypothesis assumes that a six bp homology is sufficient to allow recombination. Further, in contrast to the expectations of such a scheme, the reciprocal products of recombination, i.e., insertions corresponding to the upstream 4 kb of the 5 kb sequence, or the upstream half of the 1 kb insertion, have never been observed and are either absent or rare.

Heterogeneous Breakpoints in Type 2 Insertions

The upstream breakpoints in three rDNA repeats interrupted by type 2 insertions differ from each other by two and two bp (Figure 5). The type 2 boundary region is 74 to 78 bp upstream from the type 1 breakpoint in the 0.5 kb insertion. Two repeats with type 2 insertions analyzed by Roiha et al. (11) have breakpoints like Dmr 249, thus defining what appears to be a predominant position. Although we cannot exclude the possibility that the two molecules we found that differ might have been rearranged during cloning in *E. coli* this appears unlikely because of the precise conservation of other regions in these clones. Thus it appears that type 2 insertions are associated with deletions in gene regions. This has been observed on a larger scale in earlier experiments (19, unpublished observations).

Homologies in Type 2 Insertions

As in the case of type 1 insertions those of type 2 including those analyzed here and Dm 264 of Roiha et al. (11) show close homology at their downstream side. The short (1.4 kb) insertion is homologous to the downstream part of the longer ones, just as is the case in type 1 insertions.

In contrast, the upstream ends of the three rDNA molecules analyzed here differ from each other and from the sequences determined by Roiha et al. (11). This is so in spite of the fact that our clone 303 and Dm 264 of Roiha belong to the 3.5 kb class of type 2 insertions. The hexamer repeat seen at the upstream end of Dm 264 (11) is not present in the cloned DNAs we analyzed.

The combined information on type 2 insertions shows that no terminal duplications occur in this structure. A stretch of A residues has been found at the downstream boundary of all type 2 insertions analyzed to date. Similar stretches have been observed in a transposable sequence in D. melanogaster (23).

Ribosomal Insertions as Transposable Sequences

The target site duplications observed in D. virilis 10 kb and D. melanogaster 1 and 0.5 kb insertions suggest that these structures arose by the introduction of transposable sequences into rDNA (16,17). No such duplications are associated with type 2 insertions and their origin is therefore poorly understood. If type 1 insertions are considered as transposable sequences it is striking that they differ substantially from some well characterized examples of such elements. Transposable sequences in prokaryotes and eukaryotes frequently contain long terminal direct repeats of several hundred bp (16,20,21). This is clearly not the case for the 0.5 kb or 1 kb type 1 insertions studied here, the 5 kb insertion studied by Roiha et al. (11) and by Rae (18), or any of the type 2 insertions. We wondered whether perhaps a terminal repeat had existed in type 1 insertions at some time but had now diverged; therefore we searched for sequence homologies between the upstream and downstream halves of the 0.5 and the 1 kb insertions using the computer programs of Queen and Korn (13) and J. Maizel and R. Lenk (unpublished; see ref. 22). Only short homologies of marginal statistical significance were found. Therefore we conclude that long terminal repeats may never have been a feature of type 1 insertions. We have recently characterized a class of transposable sequences in Drosophila which do not carry long terminal repeats (23), and it appears likely that ribosomal insertions of type 1 belong to such a class.

Expression of Interrupted rRNA Genes

Type 1 and type 2 insertions are transcribed at very low levels in both wild type and bobbed mutant D. melanogaster (2,3,24). Because of the very high transcription rates required for effective production of rRNA we concluded that interrupted rDNA repeats do not produce rRNA and that the small amount of transcription that is detected results from "leakiness" in the system. In spite of their different positions and boundary structures both types of insertions appear to inactivate rRNA genes in Drosophila.

Concerted Evolution in Drosophila rDNA

Observations of high homology between different members of repeated gene families have lead to postulates for the existence of a "correction mechanism" in the evolution of such families (25-29). Drosophila rDNA poses a special problem in that the different repeat units are not identical but carry different insertions. However, insertions of one sequence class are highly conserved, as are the rRNA coding regions and even the nontranscribed spacers (11,14,18,30,31, and this report). The correction mechanisms proposed pre-

viously may not be sufficient to explain these facts and additional features seem to be required.

ACKNOWLEDGMENT

We thank J.V. Maizel, Jr., and C. Queen for help in computer analysis of DNA sequences; P.M.M. Rae for communicating unpublished results; F. Gallegos for technical assistance; and P.P. DiNocera, B.K. Kay, P.M.M. Rae and M.F. Singer for comments on the manuscript.

REFERENCES

1. Long, E.O. and Dawid, I.B. (1980) Annu. Rev. Biochem. 49, 727-764.
2. Long, E.O. and Dawid, I.B. (1979) Cell 18, 1185-1196.
3. Long, E.O., Rebbert, M.L. and Dawid, I.B. (1980) Cold Spring Harbor Symp. Quant. Biol. 45, in press.
4. Rae, P.M.M., Kohorn, B.D. and Wade, R.P. (1980) Nucleic Acids Res. 8, 3491-3505.
5. Dawid, I.B., Wellauer, P.K. and Long, E.O. (1978) J. Mol. Biol. 126, 749-768.
6. Botchan, P., Reeder, R.H. and Dawid, I.B. (1977) Cell 11, 599-607.
7. Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
8. Hosbach, H.A., Silberklang, M. and McCarthy, B.J. (1980) Cell 21, 169-178.
9. Tu, C.-P.D. and Cohen, S.N. (1980) Gene 10, 177-183.
10. Dawid, I.B. and Wellauer, P.K. (1978) Cold Spring Harbor Symp. Quant. Biol. 42, 1185-1194.
11. Roiha, H., Miller, J.R., Woods, L.C. and Glover, D.M. (1981) Nature 290, 749-753.
12. Wellauer, P.K. and Dawid, I.B. (1978) J. Mol. Biol. 126, 769-782.
13. Queen, C.L. and Korn, L.J. (1980) Methods Enzymol. 65, 595-609.
14. Wellauer, P.K., Dawid, I.B. and Tartof, K.D. (1978) Cell 14, 269-278.
15. Glover, D.M. (1977) Proc. Natl. Acad. Sci. USA 74, 4932-4936.
16. Calos, M.P. and Miller, J.H. (1980) Cell 20, 579-595.
17. Dunsmuir, P., Brorein, W.J., Jr., Simon, M.A. and Rubin, G.M. (1980) Cell 21, 575-579.
18. Rae, P.M.M. (1981) Submitted to Nucleic Acids Res.
19. Roiha, H. and Glover, D.M. (1980) J. Mol. Biol. 140, 341-355.
20. Levis, R., Dunsmuir, P. and Rubin, G.M. (1980) Cell 21, 581-588.
21. Cameron, J.R., Loh, E.Y. and Davis, R.W. (1979) Cell 16, 739-751.
22. Konkel, D.A., Maizel, J.V., Jr. and Leder, P. (1979) Cell 18, 865-873.
23. Dawid, I.B., Long, E.O., DiNocera, P.P. and Pardue, M.L. (1981) Cell, in press.
24. Long, E.O., Collins, M., Kiefer, B.I. and Dawid, I.B. (1981) Molec. Gen. Genet., in press.
25. Brown, D.D., Wensink, P.C. and Jordan, E. (1972) J. Mol. Biol. 63, 57-73.
26. Brown, D.D. and Sugimoto, K. (1973) J. Mol. Biol. 78, 397-415.
27. Wellauer, P.K., Dawid, I.B., Brown, D.D. and Reeder, R.H. (1976) J. Mol. Biol. 105, 461-486.
28. Smith, G.P. (1973) Cold Spring Harbor Symp. Quant. Biol. 38, 507-513.
29. Dover, G. and Coen, E. (1981) Nature 290, 731-732.
30. Long, E.O. and Dawid, I.B. (1979) Nucleic Acids Res. 7, 205-215.
31. Long, E.O., Rebbert, M.L. and Dawid, I.B. (1981) Proc. Natl. Acad. Sci. USA 78, 1513-1517.