Volume 10 Number 23 1982 Nucleic Acids Research

Human globin ψB_2 is not a globin-related sequence

Shi-Hsiang Shen and Oliver Smithies

Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA

Received 4 August 1982; Revised and Accepted 29 October 1982

ABSTRACT

We have determined the complete nucleotide sequence of 3.4 kilobase pairs of DNA covering the region of the human β globin gene cluster where a human globin-related sequence $\psi\beta_2$ was thought to lie (Fritsch, Lawn, and Maniatis (1980) Cell 19, 959-972). Analysis of the resulting data reveals no evidence for any globin-related sequences in this region. The region does, however, contain several stretches of poly (dA-dT).

We have confirmed the observations of Fritsch et al. that DNA from the $\psi\beta_2$ region hybridizes to the poly (dA-dT)-tailed human fetal globin cDNA plasmid, pJW151 (Wilson et al., (1978) Nucl. Acids Res. 5, 563-581) under conditions of low stringency, but we find that this hybridization is abolished by the addition of poly(rA). We conclude that $\psi\beta_2$ is not a globin pseudogene, and that the earlier investigators were probably misled by hybridization between the poly (dA-dT) stretches within the $\psi\beta_2$ region and the tails used in constructing the cDNA plasmid.

INTRODUCTION

Cloning and hybridization studies of the human β-like globin gene cluster by Fritsch et al. (1) revealed two sequences which weakly hybridized to human β or γ globin cDNA plasmids. One sequence, designated $\psi\beta_1$, is located in the $^{A}\gamma$ to δ intergenic region, and was detected using conventional hybridization conditions. Another sequence, $\psi\beta_2$, was detected only when low stringency hybridization conditions were used. Sequence analysis has shown that $\psi\beta_1$ is a globin pseudogene (P. Jagdeeswaran, J. Pan and S. Weissman, personal communication). In this paper we report the nucleotide sequence of 3.4 kbp (kilobase pairs) of human DNA which includes the entire region in which $\psi\beta_2$ was thought to lie. In analyzing the sequence data we do not detect any globin-like sequences. The probable reasons why Fritsch et al. detected hybridization in the region are revealed by the sequence data and by experimental tests. We conclude that there is no globin pseudogene in this region.

0305-1048/82/1023-7809 $2.00/0

MATERIALS AND METHODS

The procedures used to construct the recombinant phage molecules and subclone fragments from them into plasmids have been described elsewhere (2). Clone IX.10 containing the $\psi\beta_2$ region was isolated by Pat Powers, University of Wisconsin, from a library constructed by Dr. Jerry Slightom, University of Wisconsin, in the vector Charon 28 (3) from Bam HI partial digests of human DNA. The human DNA was obtained from cultured fibroblasts derived from a human embryo (2). A 3.4 kbp Hind III fragment from clone IX.10 was subcloned into pBR322 as p5-36 (see Figure 1, below).

DNA sequencing was performed by the method of Maxam and Gilbert (4). Reaction times and conditions were varied slightly according to fragment lengths. The 5' end labeling of fragments with T4 polynucleotide kinase (P-L Biochemicals) and the 3' end labeling of fragments with the large fragment of E. coli Klenow polymerase (Boehringer Mannheim) have been described previously (5).

For blotting and hybridization studies, DNA was transfered from agarose gels to nitrocellulose filters (Schleicher and Schuell) by the method of Southern (6). After transfer, the filters were baked at 68° C for at least 2 hours, and were then prewashed for 1 hour with gentle shaking at 68° C in 6X saline/sodium citrate/Denhardt's solution as previously described (7). The hybridizations were conducted for 20 hours at 68° C with gentle shaking in sealed plastic bags in 6X saline/sodium citrate/Denhardt's solution in the presence of 40 µg per ml of denaturated sonicated salmon sperm DNA and 1 X 10^6 cpm ^{32}P labeled probe per ml. The probe was ^{32}P nick-translated $^{G}\gamma$ globin cDNA plasmid, pJW151 (8). In some experiments, 40 µg per ml of poly(rA) (Sigma) was added to the hybridization solution as noted in figure legends. After hybridization, the filters were washed twice for 30 minutes under low stringency conditions at room temperature in 3X saline/sodium citrate containing 0.5% $NaDodSO_4$. The dried filters were exposed to Kodak XR-5 film with Dupont Quanta III intensifier screens.

RESULTS

Sequencing strategy and data

Figure 1 presents a linkage map (1) of the human β globin gene cluster, with the sequenced $\psi\beta_2$ region enlarged in the lower portion of the figure. Sites for several six base pair restriction enzymes are indicated. The figure also shows the location and extent of clone IX.10 from which a subcloned 3.4 kbp Hind III fragment was prepared for

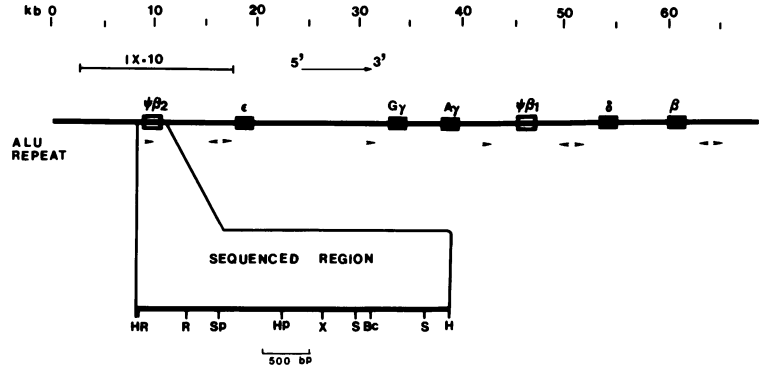

Figure 1. Map of the human β globin gene cluster (1). The productive genes are indicated by solid boxes. Two non-productive sequences, $\psi\beta_1$ and $\psi\beta_2$, described by Fritsch et al. (1) are indicated by open boxes. The 3.4 kbp Hind III fragment of DNA, here sequenced from a subclone p5-36, is shown below the map on an enlarged scale together with verified sites for the restriction enzymes H (HindIII), R (EcoRI), Sp (SphI), Hp (HpaI), X (XbaI), S (SacI), Bc (BclI). The arrow heads indicate the positions and orientations of members of the Alu family of repeated DNA (25) and present work. The extent of the phage clone IX.10 referred to in the text is shown.

sequencing the $\psi\beta_2$ region. Over 90% of the 3.4 kbp fragment was sequenced on both DNA strands, except for 120 bp at the beginning of our sequence, and about 150 bp in the middle, between the Hpa I and Xba I restriction sites; all parts were sequenced at least twice from different sites to overlap and confirm the sequence.

The nucleotide sequence of the 3.4 kbp fragment, which covers the entire $\psi\beta_2$ region, is presented in Figure 2 with various interesting features emphasized. These features include: three stretches of poly (dT), a member of the Alu family of repeated DNA (9), direct repeats flanking the Alu sequence, and two stretches of the repeated dinucleotide $(CA)_n$. We comment on these sequences below.

Lack of Homology of the $\psi\beta_2$ region with globin gene sequences

To compare the nucleotide sequence of the human $\psi\beta_2$ region with that of the human globin genes, we used a dot matrix program (10) in which significant regions of homology appear as dots which fall on diagonal lines. Nucleotide differences appear as gaps in these lines. Insertions or deletions cause a relative displacement of different portions of the diagonal lines. This method of analysis allows visualization of relationships between compared sequences which might otherwise be

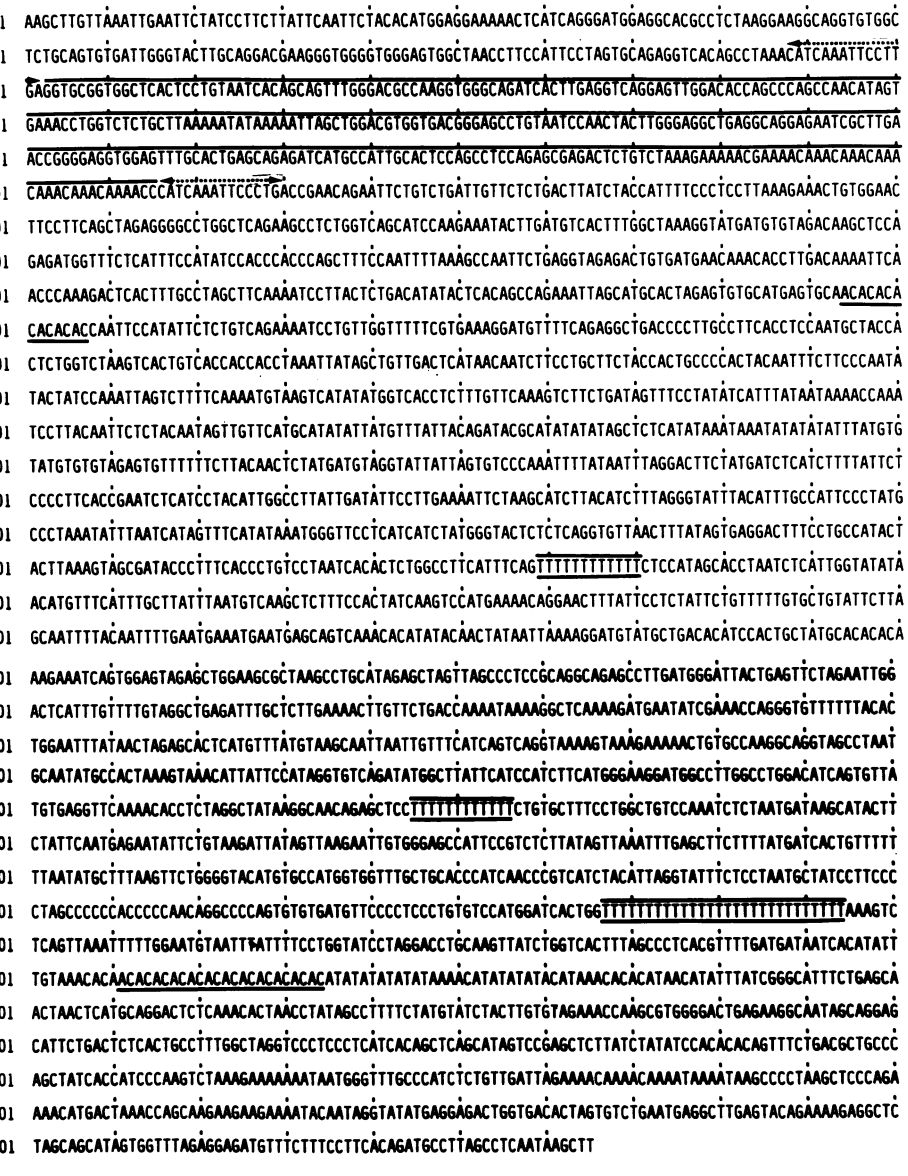

Figure 2. The nucleotide sequence of the human $\psi\beta_2$ region. Three stretches of poly(T) referred to in the text are indicated by over- and underlines. Two short repeats (dotted arrows) flanking a member of the Alu-family of repeated DNA (overlined) are indicated. Stretches of poly(AC)$_n$ are underlined.

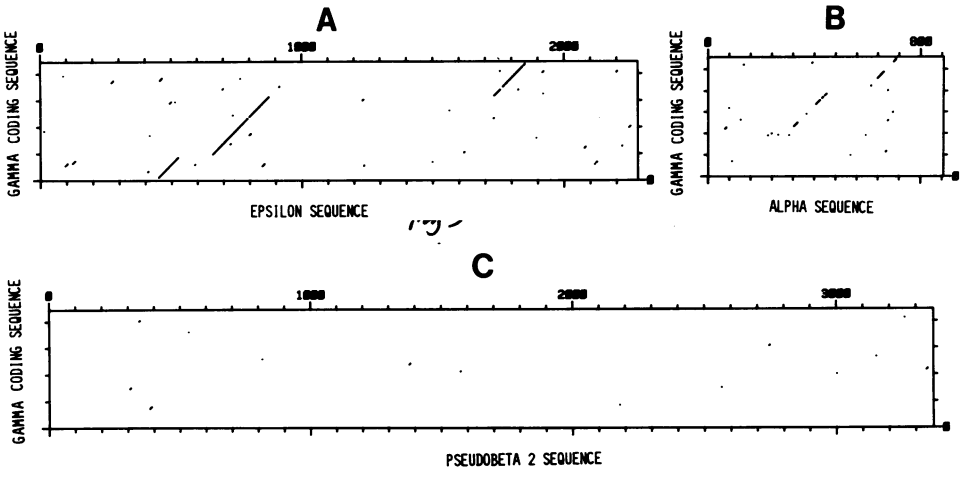

Figure 3. Dot matrix comparisons of several sequences. The sequences compared by a dot matrix program (10) are indicated on the axes of each panel of the figure. Each dot represents 14 or more identities per 20 nucleotides of compared sequences.

Panel A: The coding sequence of the human $^{G}\gamma$ fetal globin coding region (2) is compared to the complete sequence of the human ε gene (11), including its introns.

Panel B: The fetal globin coding region is compared to the complete sequence of the human α globin gene (13).

Panel C: The fetal globin coding region is compared to the complete sequence of the human $\psi\beta_2$ region given in Figure 2. Note the absence of any diagonal lines of dots in this panel.

difficult to detect. We illustrate in Figure 3 the type of result obtained by this method with gene sequences of known relatedness and with the $\psi\beta_2$ sequence.

In panel A of Figure 3 the sequence of the human fetal $^{G}\gamma$ globin gene coding sequence (2) is compared to that of the human embryonic ε gene (11); these two sequences are believed to have had separate histories for about 100 million years (12). A strong diagonal line indicating considerable homology is visible. It is divided into three parts by the two intervening sequences which are present in the ε gene but not in the $^{G}\gamma$ globin coding sequence.

Panel B of Figure 3 shows that homology is detected when even more distantly related sequences are compared - the human $^{G}\gamma$ globin coding sequence and the sequence of the human α globin gene (13). Despite the evolutionary divergence of these two sequences over the course of about 500 million years (12) their relatedness can still be visualized with this

analysis, although only exons two and three give clear diagonals.

Panel C of Figure 3 presents a comparison of the coding sequence for the $^{G}\gamma$ globin gene with the sequence we have obtained for the $\psi\beta_2$ region. No homology can be detected. A similar comparison using the complementary DNA strand of the $\psi\beta_2$ region was likewise negative.

Dots in the matrices of Figure 3 represent a minimum of 14 identities out of 20 nucleotides between the compared sequences. Similar tests using stringencies of 12 identities out of 20 also failed to reveal any homology between the $^{G}\gamma$ and $\psi\beta_2$ sequences. Nor was any found when the $\psi\beta_2$ region was compared to the human α, β, δ and ϵ gene sequences (13, 14, 15, 11).

Blot hybridization

Our failure to find homology between the sequence of the $\psi\beta_2$ region and the sequence of any human globin genes led us to inspect our sequence data and the methods used by Fritsch et al. (1) in order to find the basis for their original observations.

In analysing the sequence data, we found that the $\psi\beta_2$ region has some very A-T rich stretches of sequence, including several poly (dT) tracts. The three poly (dT) tracts which are the longest are indicated by over- and underlines in the sequence presented in Fig. 2. These are: two tracts of $(dT)_{12}$ at nucleotide positions 1659 through 1670 and at 2345 through 2356, and a tract of $(dT)_{28}$ at nucleotide positions 2667 through 2694. We suspected that these tracts could account for the hybridization results obtained by Fritsch et al. (1), since their probe was made from the γ globin cDNA clone, pJW151, which was constructed with poly (dA-dT) tailing (8). It seemed likely that the poly (dA-dT) stretches in the $\psi\beta_2$ region hybridize to the poly (dA-dT) tails of pJW151, but that there is no hybridization between the $\psi\beta_2$ region and the γ globin coding sequence portion of the plasmid. To test this idea we hybridized the cloned $\psi\beta_2$ region to the γ globin cDNA probe in the presence and absence of poly(rA). Poly(rA) would be expected to prevent hybridization between the poly(dA-dT) tracts in $\psi\beta_2$ and the poly(dA-dT) tails of the probe, but should not greatly influence the hybridization of globin coding sequences.

The results of this experiment are shown in Figure 4. Strip 1 is the autoradiograph following electrophoresis of a digest of $\psi\beta_2$ DNA from bacteriophage IX.10 (see figure 1) after blotting and hybridization in the absence of added poly(rA) to ^{32}P nick-translated $^{G}\gamma$ globin cDNA plasmid pJW151 . Strip 2 is identical to strip 1 except that poly(rA) was added to the hybridization mixture. Strip 3 is a control experiment with the

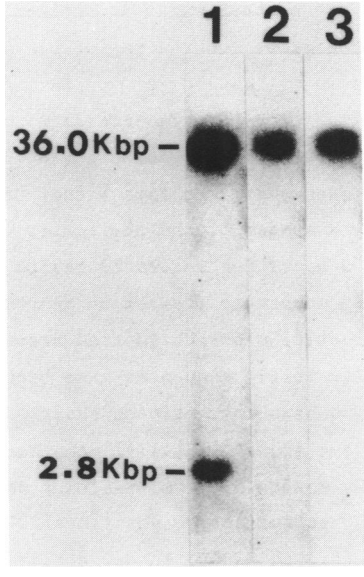

Figure 4. The effect of poly(rA) on the hybridization of the $^{G}\gamma$ globin cDNA plasmid pJW151 (8) to DNA covering the human $\psi\beta_2$ region. The autoradiographs are of Southern blots of agarose electrophoresis gels.

1)An EcoRI/Hind III digest of $\psi\beta_2$-containing bacteriophage IX.10 DNA hybridized to ^{32}P-labelled pJW151 in the absence of poly(rA).

2)Same as 1) but with poly(rA) in the hybridization mixture.

3)The same digest as 1) but with the plasmid vector, pMB9, as the radioactive probe.

The 2.8 kbp band contains the $\psi\beta_2$-region. The 36.0 kbp band is from the adhered arms of the bacteriophage vector, Charon 28, and a part of inserted DNA in clone IX.10.

same digest but using ^{32}P nick-translated plasmid vector, pMB9, as the probe. A 2.8 kbp band containing the $\psi\beta_2$ region hybridizes to pJW151 in the absence (strip 1) of added poly(rA), but fails to hybridize in its presence (strip 2). The hybridizing band at about 36 kbp seen in control strip 3 and the other two strips is due to cross-hybridization between the two vectors, Charon 28 and pMB9, used in constructing IX.10 and pJW151.

Dr. Ed Fritsch kindly repeated his hybridization tests (1) on our Hind III subclone, p5-36, and on the fragment he had originally studied; he confirms our observations that poly(rA) inhibits the hybridization of pJW151 to the $\psi\beta_2$ region.

Since the addition of poly(rA) blocks the hybridization of the probe to the $\psi\beta_2$ region (the 2.8 kbp band in Figure 4) we conclude that this

hybridization is largely, and perhaps entirely, the consequence of poly (dA-dT) tracts common to the probe and the region.

Some significant sequences in $\psi\beta_2$ region

The foregoing analysis of the DNA sequence data and the hybridization results demonstrate that there are no detectable globin-related genes in the $\psi\beta_2$ region. Nevertheless, there are some significant sequences in this region. One of them, a member of the Alu family of repeated DNA (9), occurs within the first 500 bp of the sequenced region (overlined in Figure 2). This Alu family member is flanked by short direct repeats of 15 nucleotides with 1 nucleotide mismatch (dotted arrows in Figure 2). Similar repeats flanking Alu family sequences have been noted previously by several investigators who have commented on their possible significance in relation to transposable elements (16, 17). Another interesting pair of sequences within the $\psi\beta_2$ region are two stretches of repeats of the dinucleotide AC which occur at nucleotide positions 894 through 907, and nucleotide positions 2811 through 2834. In our earlier papers (2, 5) we suggested (5) that the sequence $(TG)_n$, or its complement $(AC)_n$, in which n is greater than or equal to 4 might be a recognition signal involved in recombination. This repeat family is observed in many eukaryotic species, and in humans may have a copy number between 30,000 and 60,000 (18).

Although our present data indicated that there are no globin-related genes in the sequenced DNA, we were still interested in asking whether there might be other protein-coding sequences in the region. In searching for such possible sequences we note that in the 5' flanking region of genes there are several specific sequences, such as the CAT box (12), TATA box (19) and cap site (20), that intervening sequences start with a GT and end with an AG dinucleotide (21), and that the hexanucleotide AATAAA commonly precedes the poly (A) addition site (22). Taking these several sequences and their variations into account, together with the consensus sequences for donor and acceptor sites for splicing described by Mount (23), we have located one possible protein-coding region in the newly sequenced DNA. A possible CAT box (CCAAT) occurs at nucleotide positions 1095 to 1099. The corresponding TATA box is located at nucleotide positions 1135 to 1140 (ATATAT). A cap site sequence occurs from nucleotide position 1162 to 1167 (CTTCTG). An initiator codon occurs at nucleotide position 1240. A possible intervening sequence could occur between nucleotide positions 1300 (GT) and 1474 (AG). A second exon could lie between positions 1475 and 1564 with a second intron between

nucleotide positions 1565 (GT) and 3046 (AG). The putative open reading frame might terminate at the nucleotide position 3155 (TGA) with a poly(A) addition sequence AATAAA located at position 3174. Whether this possible open reading frame, equivalent to a polypeptide of 85 amino acids, corresponds to any real gene is being tested experimentally. Preliminary tests utilizing a cloned segment corresponding to the open coding sequence to blot poly(A)$^+$ RNA from a human erythroleukemic cell line, K562 (24), failed to reveal any transcript, although transcripts corresponding to the ϵ and γ globin genes were readily seen with appropriate probes (data not shown).

DISCUSSION

Our nucleotide sequence data have failed to reveal any globin-related sequences in the region originally postulated by Fritsch et al. (1) to contain a human globin-related sequence, $\psi\beta_2$. The results of our hybridization experiments suggest that these investigators may have been misled by the occurrence of hybridization between the poly (dA-dT) tails of their probe and stretches of poly (dA-dT) which we here show occur in the $\psi\beta_2$ region. Accordingly we conclude that $\psi\beta_2$ is not a pseudogene.

When looking for distantly related sequences by Southern blotting experiments, a critical balance must be sought between choosing overly stringent conditions which lead to missing important relationships and choosing insufficiently stringent conditions which lead to the detection of sequences having non-significant relationships. Unfortunately there is no certain formula which strikes this balance, although the addition of appropriate complex and simple DNA or RNA competitors can be of considerable help. DNA sequencing experiments will usually resolve uncertainties, but they are time consuming and are not lightly undertaken if the region to be sequenced is likely to have little intrinsic interest. In the present instance, Fritsch et al. detected a sequence $\psi\beta_1$, which is now known to be a globin pseudogene, when they used conventional hybridization conditions. The sequence $\psi\beta_2$, detected with low stringency conditions in the absence of any simple sequence competitors, has been here shown by DNA sequencing not to be globin related. However, the region in question is so close to an important gene cluster that knowledge of its sequence is happily of interest in its own right.

ACKNOWLEDGEMENTS

We thank Pat Powers for her gift of the clone IX.10 used to isolate

the $\psi\beta_2$ region, John Devereux and the University of Wisconsin Genetics Computer Group for computer analyses and Dr. Ed Fritsch for hybridization tests on p5-36. This work was supported by NIH grants GM20069 and AM20120. This is paper No. 2600 from the Laboratory of Genetics, University of Wisconsin-Madison.

REFERENCES

1. Fritsch, E. F., Lawn, R. N. and Maniatis, T. (1980) Cell 19, 959-972.
2. Slightom, J. L., Blechl, A. E. and Smithies, O. (1980) Cell 21, 627-638.
3. Rimm, D. L., Horness, D., Kucera, J. and Blattner, F. R. (1980) Gene 12, 301-309.
4. Maxam, A. M. and Gilbert, W. (1977) Proc. Nat. Acad. Sci. USA 74, 560-564.
5. Shen, S. H., Slightom, J. L. and Smithies, O. (1981) Cell 26, 191-203.
6. Southern, E. M. (1975) J. Mol. Biol. 98, 503-517.
7. Blattner, F. R., Blechl, A. E., Denniston-Thompson, K., Faber, H. E., Richards, J. E., Slightom, J. L., Tucker, P. W. and Smithies, O. (1978) Science 202, 1279-1284.
8. Wilson, J. T., Wilson, L. B. , DeRiel, J. K. , VillaKomaroff, L., Efstratiadia, A., Forget, B. G. and Weissman, S. M. (1978) Nucl. Acids Res. 5, 563-581.
9. Jelinek, W. R., Toomey, T. P., Leinwand, L., Duncan, C. H., Biro, P. A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L. and Schmid, C. W. (1980) Proc. Nat. Acad. Sci. 77, 1398-1402.
10. Maizel, J. B., Jr. and Lenk, R. P. (1981) Proc. Natl. Acad. Sci. 78, 7665-7669.
11. Baralle, F. E., Shoulders, C. C. and Proudfoot, N. J. (1980) Cell 21, 621-626.
12. 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. and Proudfoot, N. J. (1980) Cell 21, 653-668.
13. Proudfoot, N. J. and Maniatis, T. (1980) Cell 21, 537-544.
14. Lawn, R. M., Efstratiadis, A., O'Connell, C. and Maniatis, T. (1980) Cell 21, 647-651.
15. Spritz, R. A., DeRiel, J. K., Forget, B. G. and Weissman, S. M. (1980) Cell 21, 639-646.
16. Jagadeeswaran, P., Forget, B. G. and Weissman, S. M. (1981) Cell 26, 141-142.
17. Grimaldi, G. and Singer, M. F. (1982) Proc. Natl. Acad. Sci. 263, 211-214.
18. Miesfeld, R., Krystal, M. and Arneheim, N. (1981) Nuc. Acids Res. 9, 5931-5947.
19. Goldberg, M. (1979) Ph. D. thesis, Stanford University, Stanford, California.
20. Baralle, F. E. and Brownlee, G. G. (1978) Nature 274, 84-87.
21. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978) Proc. Natl. Acad. Sci. USA 75, 4853-4857.
22. Proudfoot, N. J. and Brownlee, G. G. (1978) Nature 263, 211-214.
23. Mount, S. M. (1982) Nuc. Acids Res. 10, 459-472.
24. Lozzio, C. B. and Lozzio, B. B. (1975) Blood 45, 321-334.
25. Fritsch, E. F., Shen, C. K. J., Lawn, R. M. and Maniatis, T. (1981) Cold Spring Harb. Symp. Quant. Biol. 95, 761.