

PHENOTYPIC ANALYSIS OF GLOBIN GENE EXPRESSION:
THE THALASSAEMIAS

R.A. Flavell, N.K. Moschonas, E. deBoer,
G.C. Grosveld, M. Busslinger, H.H.M. Dahl
and F.G. Grosveld

National Institute for Medical Research,
Mill Hill, London, NW7 1AA, U.K.

Introduction

Structural analysis of cloned globin genes and indirect restriction mapping performed by genomic blotting experiments has permitted the detailed elucidation of globin gene organization. (See e.g. Maniatis et al., 1980) for a recent review). A number of generalizations can be made from these observations.

1. THE RELATED GENES ARE CLUSTERED IN A SHORT REGION OF THE CHROMOSOME. They are present in the order $5' \epsilon G_{\gamma} A_{\gamma} \delta \beta 3'$ for the β -related human globin gene locus and $\zeta \alpha 1 \alpha 2$ for the human α locus. In this case the genes are present on the chromosome in the order that they are expressed, but this does not appear to be so for the chicken globin genes, where the two adult α genes are flanked on either side by an embryonic α -globin gene. Though the α - and β -related genes are normally on separate chromosomes, this is clearly not obligatory, since in XENOPUS the α - and β -globin genes are closely linked (Jeffreys et al., 1980). Within a given cluster the globin genes appear to be transcribed from the same DNA strand.
2. THE DISTANCES BETWEEN THE GLOBIN GENES BY FAR EXCEEDS THE SIZE OF THE GENES THEMSELVES. Thus, the distances are ϵ 13.5kb G_{γ} 3.5kb A_{γ} 13.5kb δ 5.5kb β ; each globin gene (including transcribed introns) is about 1.5kb in length.
3. REPETITIVE DNA WITHIN GENE CLUSTERS. Little is known about the structures present in the intergenic DNA. Two types of DNA sequences have been recognized to date. First,

numerous reiterated sequences are found at several sites within the β -related globin gene cluster. These have been mapped extensively in the case of the rabbit genes. Though the position of these repeats is well characterized, nothing is known of their function. The characterization of repeated sequences in the human globin gene system is less advanced than in the rabbit. The so-called Alu family of repeats, however, have been the subject of considerable attention. About 300,000 of these sequences, each approximately 200bp in length, are present in the human genome and they have a set of intriguing properties (Jelinek et al., 1980):

- they are homologous both to an abundant class of small nuclear RNAs and with the double-stranded regions of heterogeneous nuclear RNA. This shows that at least some of the repeats, though not necessarily all, are transcribed in vivo
- they serve as efficient transcription templates for RNA polymerase III in vitro
- they contain a sequence that is homologous to a sequence found near the replication origin of SV40, polyoma and BK viral DNAs.

From the last observation it is tempting to speculate that these sequences may represent cellular origins of replication. There is, however, no evidence to support this idea at present.

Genetic Analysis of Globin Gene Expression

Structural analysis alone clearly cannot explain the subtleties of gene expression. To understand gene expression we must employ genetics. Two options are available at present. The approach most frequently followed is to analyse the phenotype conferred by mutant DNA molecules which have been generated in vitro - the process which has been called reversed genetics. This powerful method has been particularly useful for the study of eukaryotic promoters (see for example the articles by Chambon and Birnstiel, this volume). The second, classical approach is to study natural mutants. Structural analysis of mutations which confer a specific and well defined phenotype would alone be very informative. Since, however, detailed genetic analysis of mammals is not feasible, it is commonly not possible to map mutations precisely, nor to determine an exact phenotype in vivo. The best strategy available therefore is to combine classical

genetics with reversed genetics. The approach in this case is to clone a gene from an individual with a given phenotype. The structure of the mutant gene is determined and the phenotype conferred by this mutant gene is analysed in detail, using reversed genetics in a given assay system (see below). In this article, we will illustrate the use of this approach in the study of human genetic diseases. Before considering the experimental details we should consider the reasons why the globin gene system is suitable for these studies.

The Thalassaemias as a Model for a Human Genetic Defect

The success of the globin gene model has several underlying features:

1. The gene products have been characterized extensively; the primary sequence of most of the human globins has been determined; higher order structures have been resolved both for normal and mutant haemoglobins.

2. The primary sequence of these mRNAs and their genes (summarized in Efstratiadis *et al.*, 1980) have been available for some time.

3. The changes in globin gene expression during human embryonic development have been worked out. The zeta chain is a primitive α -like globin present in early embryos and is probably present in association with ϵ -chains, the first β -like globin polypeptide to appear in human embryos. The zeta-chains are probably rapidly replaced by α -globin chains, which remain present during the remainder of foetal and adult life. The ϵ -chains are in turn replaced by the γ -globins which are the major foetal β -like globin chains (HbF $\approx \alpha_2\gamma_2$). Finally, in the last stages of foetal development, the γ -chains are replaced by the β - and, to a minor extent, δ -globin chains.

4. Numerous well-defined cases of defects in the functioning of the globin genes have been described. These fall into two classes:

a) abnormal globin proteins are produced, usually because of amino acid substitutions which result from point mutations in the structural gene; less commonly, deletions or fusions cause other types of abnormal haemoglobins.

b) The level of otherwise normal haemoglobin can be reduced or even zero in the diseases known as the thalassaemias. In this article, we shall restrict the discussion to the β -thalassaemias and related diseases. The levels of mRNA have

been determined both in erythroid cells from patients with β -thalassaemia and from normal individuals using complementary DNA (cDNA) hybridization. The introduction of Southern-blotting techniques to the study of the thalassaemias (Flavell *et al.*, 1979; Fritsch *et al.*, 1979; Bernards *et al.*, 1979; Ottolenghi *et al.*, 1979 and van der Ploeg *et al.*, 1980) has provided much valuable information on gross alterations in the structure of the human β -globin locus which accompany some of these diseases. Fig. 1 shows our current understanding of the structure of the human β -globin gene cluster and the extent of the deletion of DNA sequences in a number of these diseases. The significance of these data has been considered in detail elsewhere (Fritsch *et al.*, 1979; Bernards *et al.*, 1979; van der Ploeg *et al.*, 1980; Bernards and Flavell, 1980) and will not be discussed here.

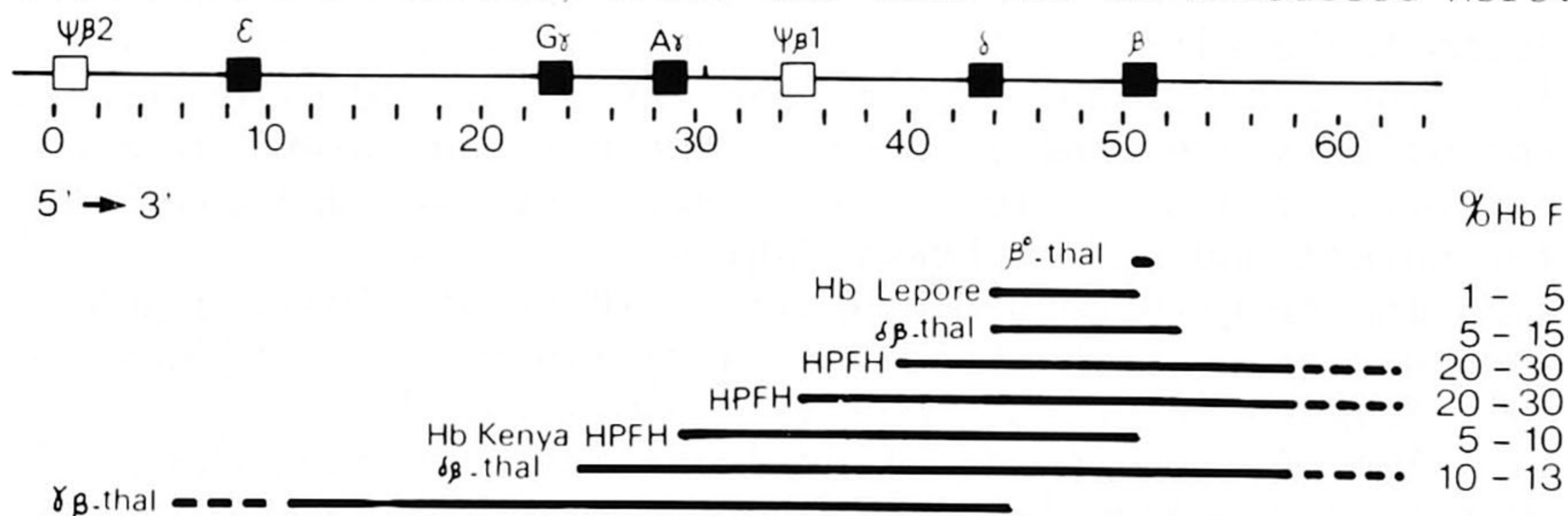


Fig. 1

The structure of the $\epsilon\gamma\delta\beta$ globin region in normal DNA and in a number of hereditary diseases. The DNA deleted in the hereditary diseases indicated is depicted by the horizontal line. A dashed line indicates that the precise location of the end point of the deletion in that area is not known. The deletion in Hb Kenya has not been established at the DNA level. See also Maniatis *et al.*, 1981, and Flavell *et al.*, 1981.

The Molecular Defects in β -thalassaemia

Recently we have turned our attention to the study of β -thalassaemia. β -thalassaemia is generally divided into two classes. In the first, β^+ thalassaemia, a low but detectable level of β -globin protein is found; this low level of β -globin protein reflects a low level of β -globin mRNA. It seems likely, *a priori*, that this disease would be the result of a defect in the transcription or processing of β -globin mRNA precursors. In β^0 thalassaemia no β -globin protein is

Phenotypic Analysis of Globin Gene Expression / 29

detected. In this case, therefore, a number of causes of the genetic lesion can be envisaged. In some cases, globin mRNA is readily detectable (Temple *et al.*, 1977), whereas in others, β -globin mRNA could not be detected using cDNA hybridization techniques (e.g. Ottolenghi *et al.*, 1975). It is likely, therefore, that there are several different molecular forms of β^0 thalassaemia.

Unlike the diseases discussed above, in β -thalassaemia, gross alterations of DNA structure are not usually observed (Flavell *et al.*, 1979; Orkin *et al.*, 1979). Fig. 2 shows a representative blot of the β - and δ -globin genes from patients with β -thalassaemia. In all but one case, no difference can be seen between the normal and thalassaemic globin genes (Flavell *et al.*, 1979). In the exceptional form of β -thalassaemia which does show a difference in blotting pattern (lane 11 of Fig. 2) we showed that this abnormal globin gene had a deletion of about 600 bp which included the third coding segment of the β -globin gene (Flavell *et al.*, 1979). The detailed structure of this thalassaemic gene has been elucidated by cloning (Orkin *et al.*, 1980).

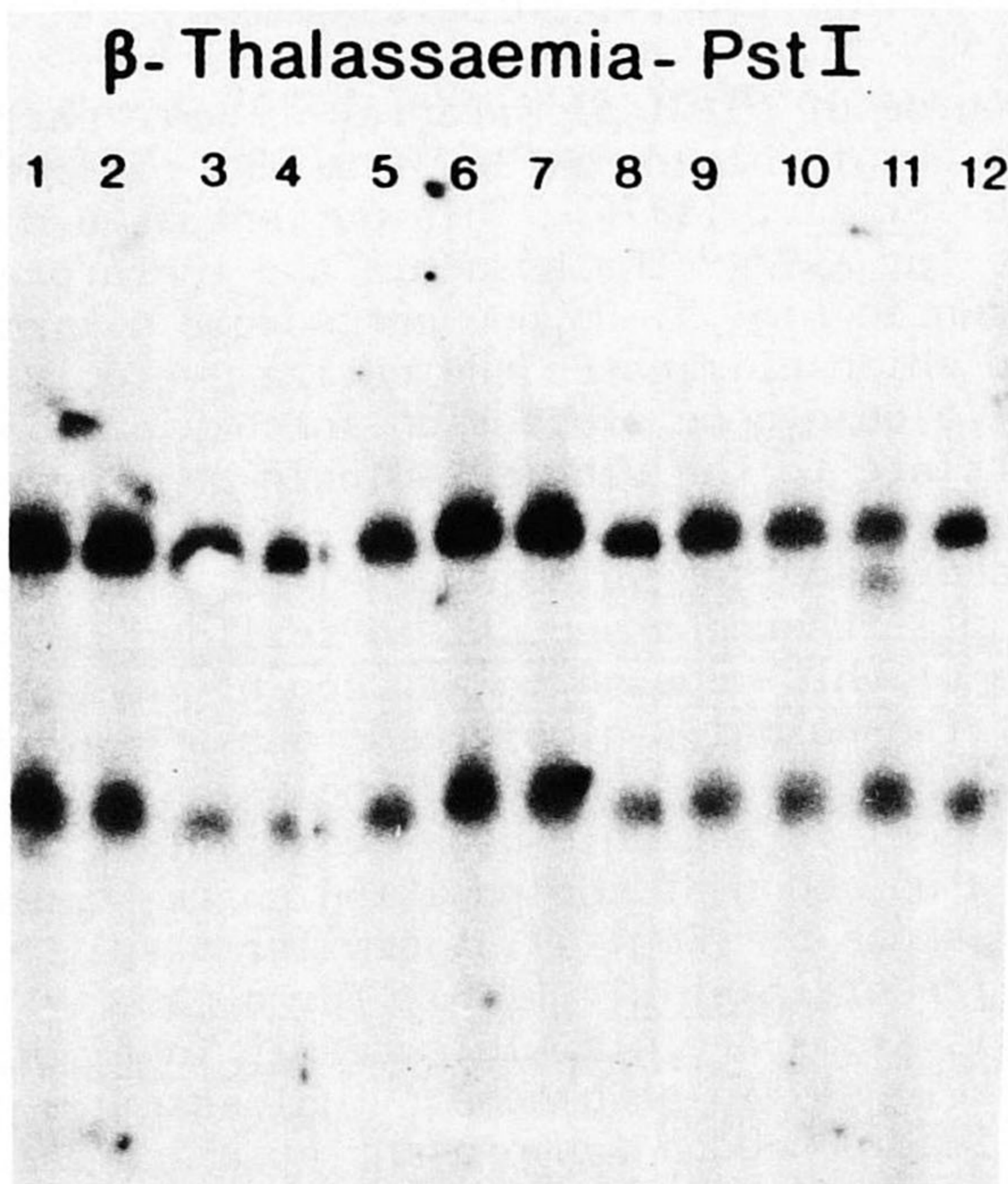


Fig. 2

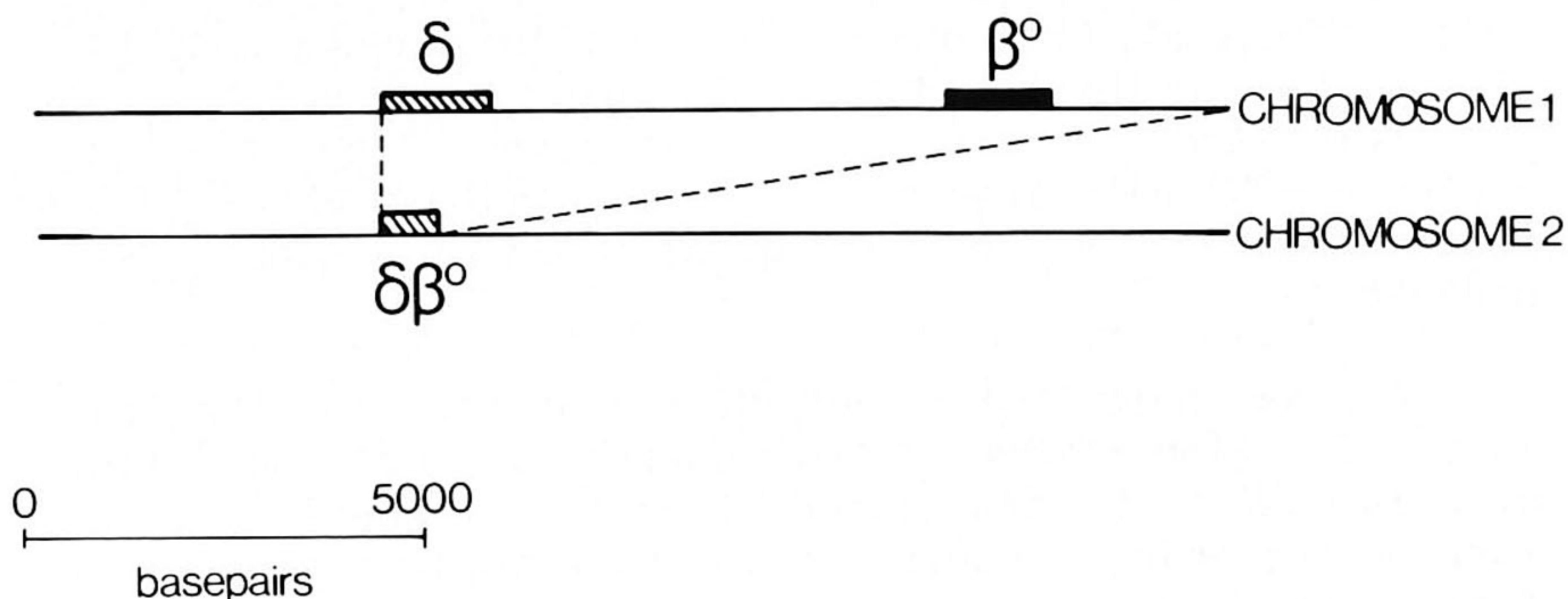
Fig. 2

β - and δ -globin gene fragments in digests of DNA from patients homozygous for β -thalassaemia. DNA was isolated from either blood or spleen from patients with homozygous β -thalassaemia as described in Jeffreys and Flavell (1977) and digested to completion with PstI. The samples were analysed for the β - δ -globin genes as described (Flavell *et al.*, 1978). Patient 5 is described by Comi *et al.* (1977), patient 11 in Tolstoshew *et al.* (1976). Patients 2,3 and 10 have been diagnosed as β^0 thalassaemia (Southern Italian; Ottolenghi, personal communication) and the remainder as homozygous β^+ thalassaemics. 1, 4, 8 and 12 show control normal DNA (derived from a Dutch placenta).

A major form of β^0 thalassaemia found in Southern Italy differs from the above type in that no gross deletion of DNA could be detected; it is characterized by levels of β -globin that were lower than the detection levels of cDNA hybridization (e.g. Ottolenghi *et al.*, 1975; Tolstoshev *et al.*, 1976). We decided to analyse the molecular defect in more detail in this type of β^0 thalassaemia.

As a source of clinical material we were particularly fortunate to obtain blood samples from the patient described in Ottolenghi *et al.*, (1975). This patient is a double heterozygote for $\delta\beta^0/\beta^0$ thalassaemia and therefore has the genotype shown in Fig. 3. On one chromosome a large deletion has occurred which eliminates the entire β -globin gene. Studies of β -globin gene expression in this patient, therefore, only relate to the single β -globin gene present.

We have cloned the β -globin gene from this patient as a 7.5 kb HindIII fragment and have determined the entire DNA sequence of this gene. A comparison of this sequence with that of the normal β -globin gene shows that the only difference in the entire gene (and including both intervening sequences) is a single base substitution that generates a stop codon at the position corresponding to amino acid residue 39 (Fig. 4). A similar result has been obtained by Trecartin *et al.* (1981) for a case of β^0 thalassaemia from Sardinia. This result is similar to that described previously in a Chinese β^0 thalassaemic; in this case, a termination codon was present at amino acid 17 (Chang and Kan, 1979).



Schematic Genotype of the Thalassaemic Patient

Fig. 3

Schematic genotype of the $\delta\beta^0/\beta^0$ thalassaemic patient at the δ and β -globin locus. Superfluous $\delta\beta^0$ refers to the partially deleted $\delta\beta^0$ thalassaemic globin gene and β^0 to the thalassaemic β -globin gene.

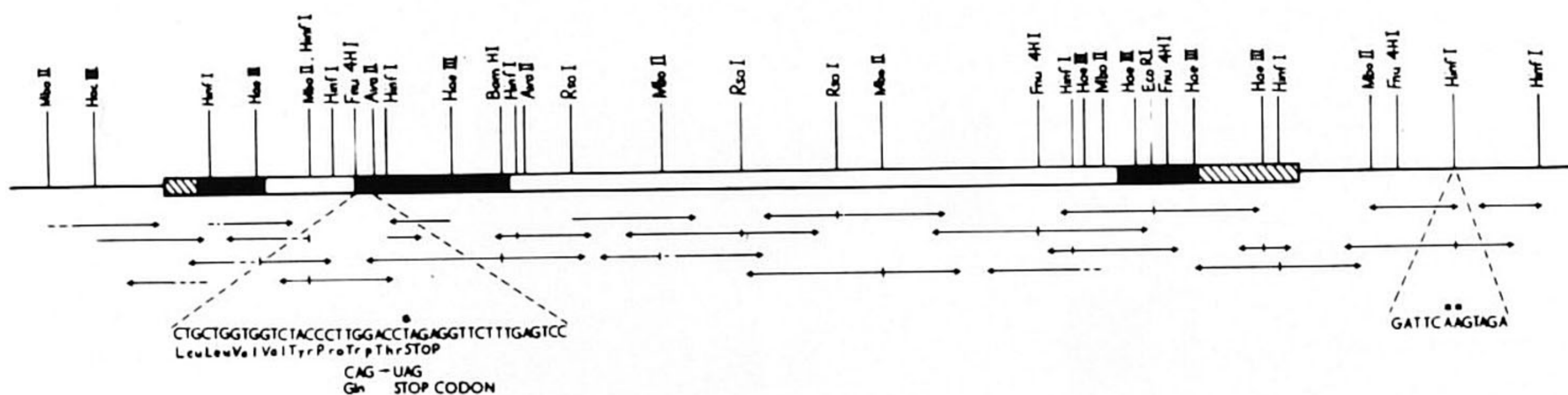


Fig. 4

The sequencing strategy for the β^0 -thalassaemic globin gene and the sequence differences observed. The β -globin gene is indicated as protein coding (filled boxes) 5' or 3' untranslated sequences (hatched boxes) or intron (open boxes).

This result clearly explains why this β -globin gene is thalassaemic. Why then does this patient not produce detectable levels of β -globin mRNA? To re-examine this point, we have probed for α - β - and γ -globin mRNA using sensitive S_1 -nuclease mapping procedures (Fig. 5). This shows that β -globin mRNA is in fact detectable using these methods, but that this level is very low. A comparison of the level of β -globin mRNA with that of the α - or γ -globin mRNAs suggest that only about 2-5% of the normal level of β -globin mRNA was present.

A reasonable, but untested, explanation for the low level of β -globin mRNA in this patient is that the failure of this mRNA to be translated effectively causes it to be more susceptible to nucleases in the cytoplasm and therefore to be degraded rapidly after synthesis.

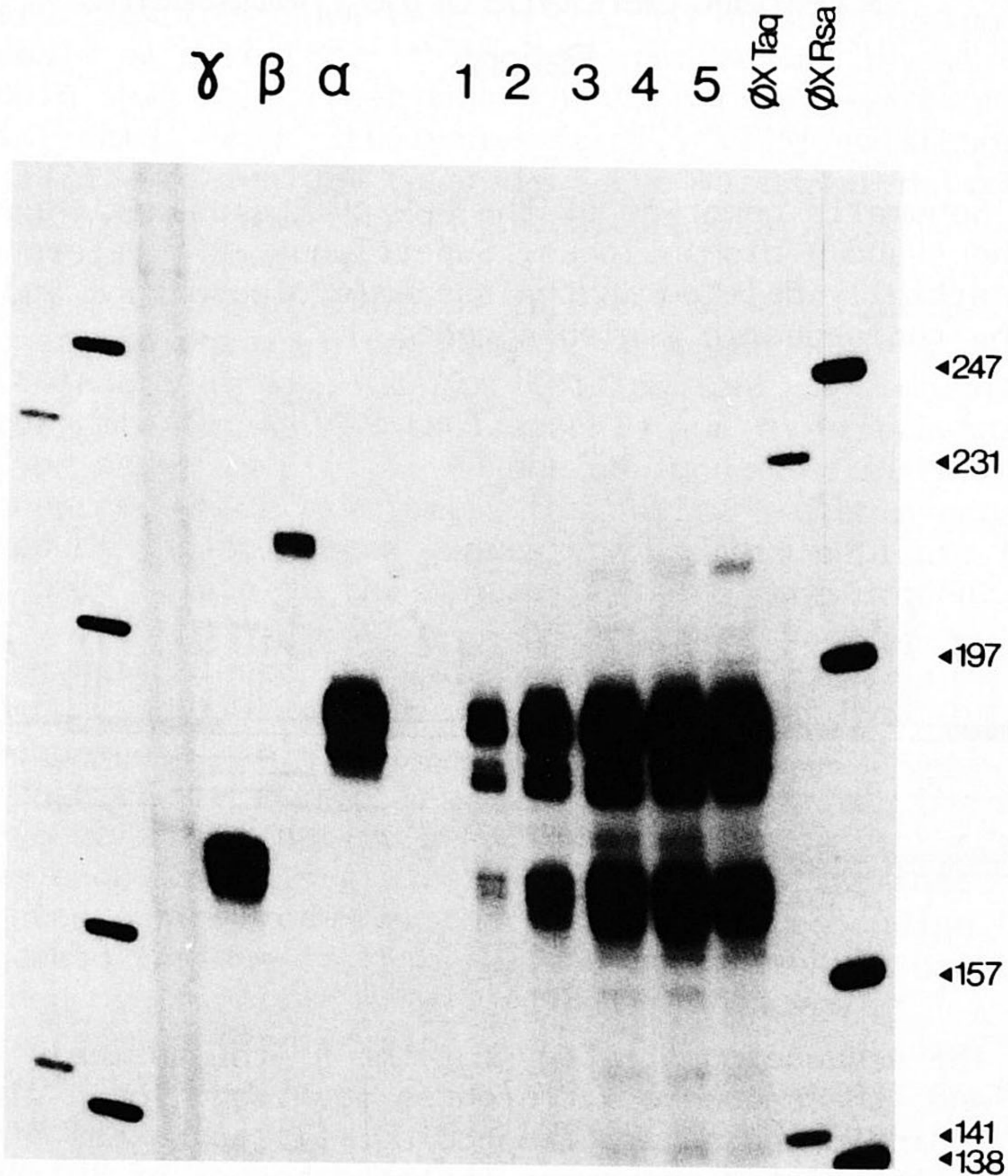


Fig. 5

Fig. 5

Determination of the α -, β - and γ -globin mRNA levels in whole blood RNA from this patient. The RNA levels were determined by hybridization to 3' end labelled probes for the β (2×10^4 cpm/p mol 3' ends) and γ -globin genes (1.72×10^4 cpm/p mol 3' ends) and to a 5' labelled probe for the α -globin gene (1.5×10^4 cpm/p mol 5' ends). On the left side of the figure the probes have been hybridized separately to an excess of total blood RNA in a volume of 10 μ l. On the right side the three probes were mixed and hybridized to 4 ng, 20 ng, 100 ng, 400 ng and 2 μ g respectively (1 to 5), of total RNA. The markers are 5' labelled fragments of ϕ X 174 DNA cut with *TaqI* or *RsaI*.

Phenotypic Assay of Inherited Defects in Human Genes

The structural analysis of the type of β^0 thalassaemia discussed above obviously gives a clear answer. A stop codon present in the β -globin gene causes the β^0 thalassaemic phenotype. Yet it is likely that molecular lesions will exist in which the molecular defect is not so easy to discern, either because the mutation detected by structural analysis is not clearly interpretable, or because a large number of polymorphic differences are present which mask the true mutation which causes the disease. To solve this problem, it is essential to correlate the structural data with phenotypic expression.

We have recently used the techniques of reversed genetics to study β -thalassaemia. The approach followed is shown in Fig. 6 and derives from the work of Banerji, Rusconi and Schaffner (personal communication). A globin gene is linked to a segment of SV40 which contains the DNA sequences around the origin cloned in a bacterial plasmid. This triple recombinant is then introduced into Hela cells by calcium phosphate coprecipitation. A high proportion of the cells take up the DNA and up to 30% of those cells express the SV40 early region as estimated by immunofluorescence. In Fig. 7, we can see that under these conditions the rabbit β -globin gene is expressed specifically from its own promoter by two criteria. First, the expression of the β -globin gene is independent of the orientation of the β -globin gene with respect to the SV40 DNA sequences. Second, the 5' ends of the β -globin RNA map at the natural cap site of the β -globin gene. It can be seen from Fig. 7 that the β -globin RNA produced is identical to natural rabbit β -globin mRNA.

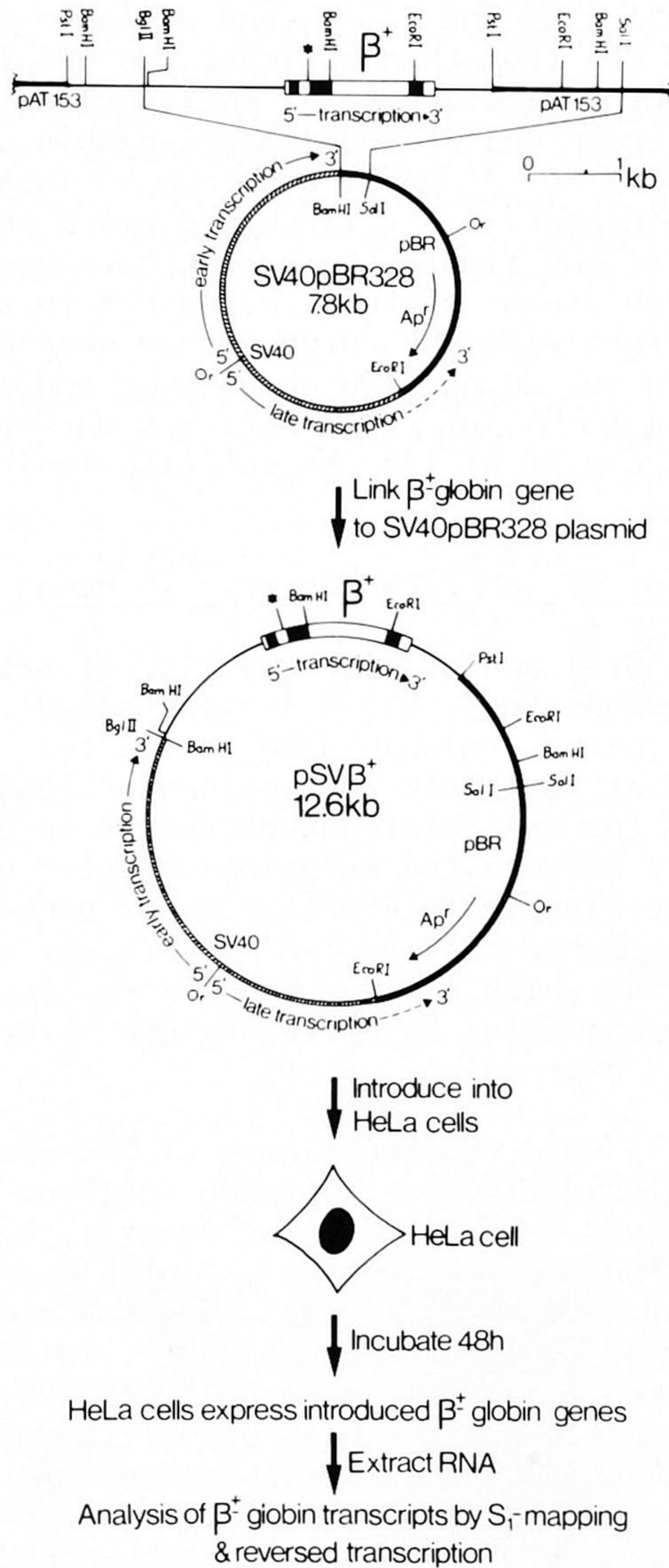


Fig. 6

A strategy for the analysis of globin gene expression *in vivo* in HeLa cells. See text for details. In this case a β^+ thalassaemic globin gene is the subject of investigation.

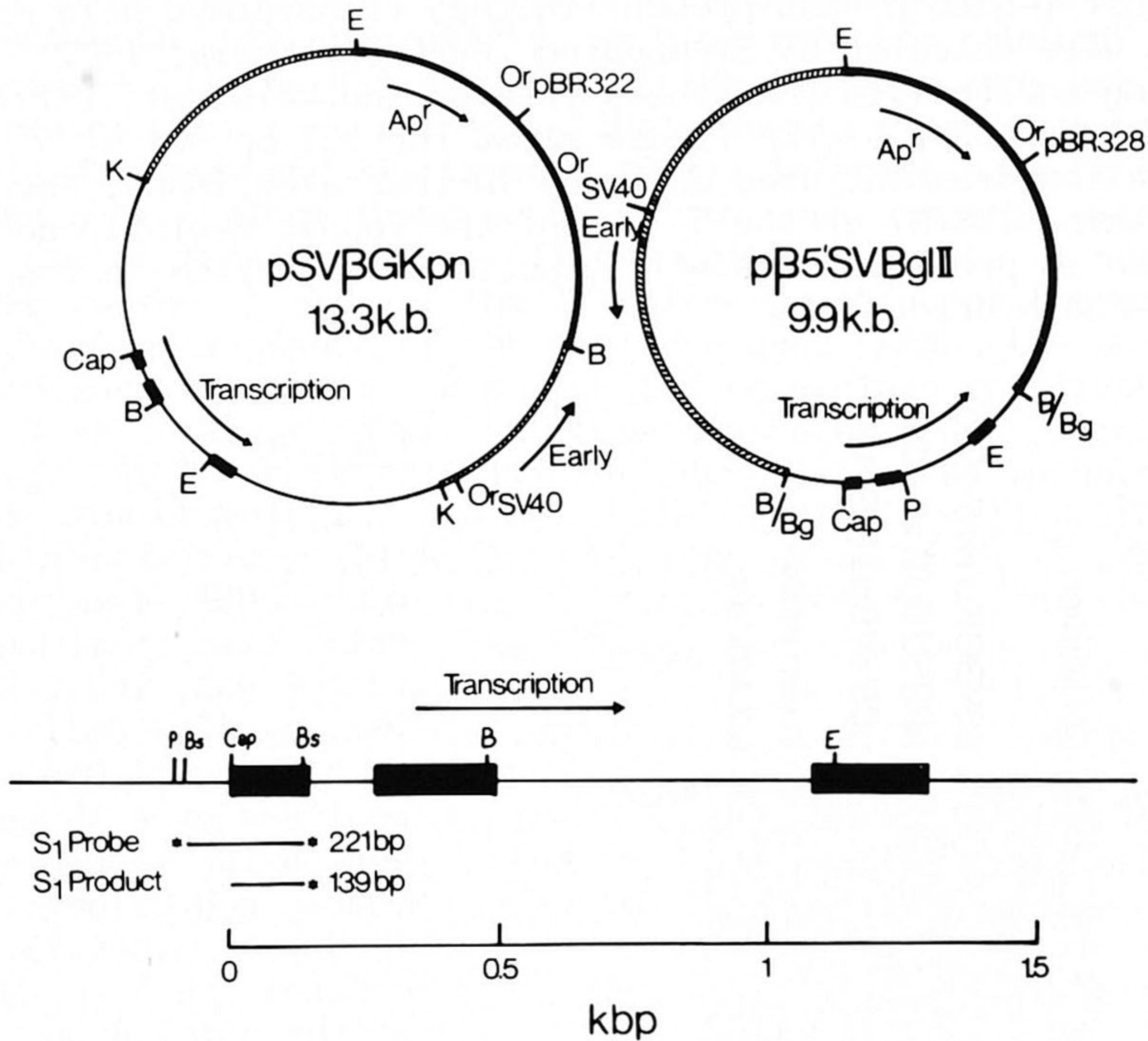


Fig. 7a

Transient expression of the rabbit β -globin gene linked to SV40 sequences in Hela cells. 15 μ g DNA from either pSV β G Kpn, p β 5' SV BglII or p β 3' SV BglII (which has the same structure but the β -globin gene segment is in the opposite orientation with respect to the SV40 sequences) was mixed with 25 μ g of salmon sperm DNA and used to transform a half confluent 100mm dish of Hela cells using the calcium phosphate method (Wigler *et al.*, 1979). Three dishes were used for each DNA sample. After 10 hours the medium (H21, 10% newborn calf serum) was changed, the cells were grown for another 24 hours, harvested and the RNA was isolated by the LiCl urea-method (Auffray and Rougeon, 1980). The figure shows schematic restriction maps of pSV β G Kpn and p β 3' SV BglII DNA. pSV β G Kpn contains the 3.9 kb Bam-Eco fragment of pBR322, the 4.4kb Bam-Eco fragment of SV40 and the 4.8 kb KpnI rabbit β -globin fragment ligated into the KpnI site of SV40 DNA. p β 3' SV Bgl and p β 5' SV Bgl DNA contain the 2kb rabbit β -globin BglII partial fragment in two orientations with respect to the SV40 sequences ligated

into the BamHI site of the SV40-pBR328 DNA junction. The rabbit β -globin RNAs present in the transformed Hela cell RNA were detected by S_1 -mapping (Berk and Sharp, 1977) using endlabelled DNA probes (Weaver and Weissman, 1979). The lower part of the figure shows the 221 bp Bst NI 5' labelled fragment used as probe in this experiment. This contains 139 NT of the 5' end of the rabbit β -globin gene, which is protected against S_1 -nuclease digestion in the DNA-RNA hybrid.

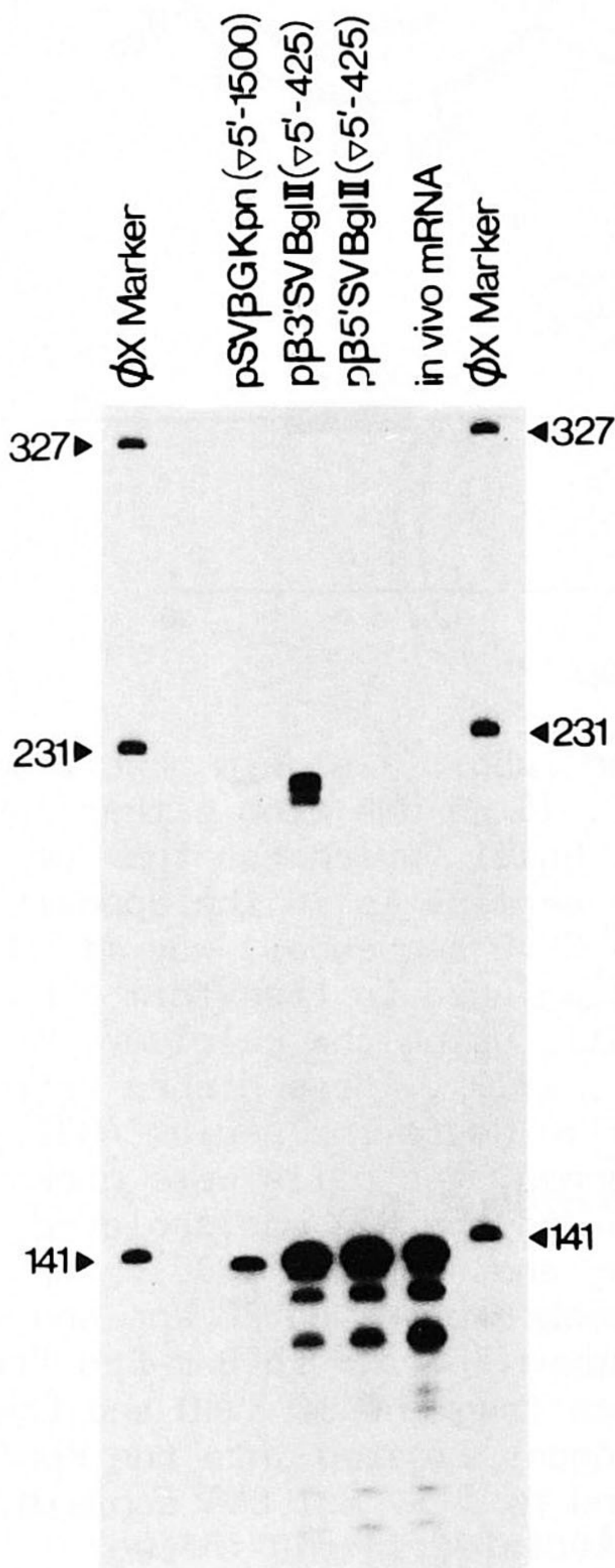


Fig. 7b

This figure shows the S_1 -products of globin RNAs from HeLa cells transformed with vector DNA, pSV β G Kpn DNA, p β 3'SV BglIII or p β 5'SV BglIII DNA analysed on a 5% polyacrylamide sequence gel (Maxam and Gilbert, 1980). As a positive control the S_1 -product of the BstHI probe with *in vivo* rabbit β -globin mRNA is shown. The labelled marker is ØX 174 x HaeIII. The numbers on the side are the lengths of the marker fragments in nucleotides. Matched areas represent SV40 DNA: solid lines plasmid sequences: thin lines and filled boxes represent the rabbit β -globin exons.

Phenotypic Analysis of Globin Gene Expression / 37

Clearly this assay can be used to study the expression of a defective globin gene from a thalassaemic patient. To investigate this possibility, we have recently studied β thalassaemia. The DNA sequence of the β -globin gene has been determined for two cases of β^+ thalassaemia (Spritz et al., 1981; Westaway and Williamson, 1981). Interestingly, both genes show a single base substitution when compared with the normal β -globin gene - TTGG -> TTAG at a site 21 nucleotides to the 5' side of the 3' splice junction of the small intervening sequence of the β -globin gene. Among the possible phenotypes that could be postulated to explain this result, the most interesting is that this generates a new splice acceptor site in the small intervening sequence. We have recently established this using the SV40-Hela cell system (Busslinger et al., 1981). S₁ mapping of the β -globin RNA produced in Hela cells from a β^+ thalassaemic β -globin gene shows that about 95% of the β -globin mRNA produced is spliced incorrectly, that is, to this abnormal splice site. This generates a β -globin mRNA with an additional 19 nucleotides which derive from the small intervening sequence; since this is not a multiple of 3, this insertion results in the alteration of the translational reading frame and in fact generates an in-phase stop codon 5 amino acids after the splice.

Conclusions

From the above, it should be clear that a combination of structural and functional analyses will enable us in the future to elucidate the molecular nature of defects in a large number of human diseases. In the examples provided here, we show that defects at the level of protein synthesis and RNA processing can be elucidated using these systems. Since the globin mRNA produced in these cells is spliced, polyadenylated and exported to the cytoplasm (Busslinger et al., 1981), it is likely that defects at the level of RNA transport can also be studied. Elsewhere, we have described how this system can be used to study transcriptional defects (Grosveld et al., 1981). It is to be anticipated that the combination of structural and functional analyses in the next few years will bring some exciting discoveries in the field of the molecular biology of human diseases.

Acknowledgements

We are grateful to Cora O'Carroll for the preparation of this manuscript. N.M. and H.H.M.D. are the grateful recipients of EMBO fellowships and M.B. of a Swiss National Research Council fellowship. We gratefully acknowledge the helpful advice offered to us by W. Schaffner on the SV40 system. This work was supported by the British Medical Council.

References

- Auffray C, Rougeon F (1980) *Eur J Biochem* 107: 303.
Berk AJ, Sharp PA (1977) *Cell* 12: 721.
Bernards R, Flavell RA (1980) *Nucl Acids Res* 8: 1521.
Bernards R, Kooter JM, Flavell RA (1979) *Gene* 6: 265.
Busslinger M, Moschonas N, Flavell RA (1981) *Cell* (in press).
Chang JC, Kan YW (1979) *Proc Natl Acad Sci (USA)* 76: 2886.
Comi P, Giglioni B, Barvarano L, Ottolenghi S, Williamson R, Novakova M, Maser G (1977) *Eur J Biochem* 79: 617.
Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel, JK, Forget B, Weissmann SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ (1980) *Cell* 21: 653.
Flavell RA, Bernards R, Kooter JM, deBoer E, Little PFR, Annison G, Williamson R (1979) *Nucl Acids Res* 6: 2749.
Flavell RA, Bud H, Bullman H, Dahl H, deBoer E, deLange T, Groffen J, Grosveld F, Grosveld G, Kioussis D, Moschonas N, Shewmaker C (1981) In Stamatoyannopoulos G, Nienhuis AW (eds): "Organization and Expression of Globin Genes," New York: Alan R. Liss, p 119.
Flavell RA, Kooter JM, deBoer E, Little PFR, Williamson R (1978) *Cell* 15: 25.
Fritsch EF, Lawn RM, Maniatis T (1979) *Nature* 279: 598.
Grosveld GC, deBoer E, Shewmaker CK, Flavell RA (1981) *Nature* (submitted).
Jeffreys AJ, Flavell RA (1977) *Cell* 12: 429.
Jeffreys AJ, Wilson V, Wood D, Simons JP, Kay RM and Williams JG (1980) *Cell* 21: 555.
Maniatis T, Fritsch EF, Lauer J, Lawn RM (1981) *Ann Rev Gen* 14: 145.
Jelinek WR, Toomey TR, Leinwand L, Duncan CH, Biro PA, Choudary PN, Weissman SM, Rubin CM, Houck CM, Deininger PL, Schmid CW (1980) *Proc Natl Acad Sci (USA)* 77: 1398.

Phenotypic Analysis of Globin Gene Expression / 39

- Maxam AM, Gilbert W (1980) In Grossman L, Moldare K (eds): "Methods in Enzymology," New York: Academic Press, p 499.
- Ottolenghi S, Giglioni B, Comi P, Gianni AM, Polli E, Acquaye CTA, Oldham JH, Maserà G (1979) Nature 278: 654.
- Ottolenghi S, Lanyon WG, Williamson R, Weatherall DJ, Clegg JB, Pitcher CS (1975) Proc Natl Acad Sci (US) 72: 2294.
- Orkin SH, Kolodner R, Michelson A, Husson R (1980) Proc Natl Acad Sci (USA) 77: 3558.
- Orkin SH, Old JM, Weatherall DJ and Nathan DG (1979) Proc Natl Acad Sci (USA) 76: 2400.
- Spritz RA, Jagadeeswaran P, Choudary PV, Biro PA, Elder JT, deRiel JK, Manley JL, Gefter ML, Forget BG, Weissman SM (1981) Proc Natl Acad Sci (USA) (in press).
- Temple GF, Chang JC, Kan YW (1977) Proc Natl Acad Sci (USA) 74: 3047.
- Tolstoshev P, Mitchell J, Lanyon G, Williamson R, Ottolenghi S, Comi P, Giglioni B, Basera G, Modell B, Weatherall DJ, Clegg JB (1976) Nature 259: 95.
- van der Ploeg LHT, Konings A, Oort M, Roos D, Bernini LF, Flavell RA (1980) Nature 283: 637.
- Westaway D, Williamson R (1981) Nucl Acids Res 9: 1777.
- Weaver RF, Weissmann C (1979) Nucl Acids Res 6: 1175.
- Wigler M, Sweet R, Sim GK, Wold B, Pellicer A, Lacy E, Maniatis T, Silverstein S, Axel R (1979) Cell 16: 777.