

# The $\beta$ -globin dominant control region: hypersensitive site 2

Sjaak Philipsen, Dale Talbot, Peter Fraser and Frank Grosveld

Laboratory of Gene Structure and Expression, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

Communicated by F. Grosveld

**The Dominant Control Region (DCR) of the human  $\beta$ -globin gene locus consists of four strong hypersensitive sites (HSS) upstream of the  $\epsilon$ -globin gene. Addition of these sites confers copy number dependent expression on the human  $\beta$ -globin gene in murine erythroleukaemia cells and transgenic mice, at levels comparable with the endogenous mouse globin genes. We have shown previously that a 1.9 kb fragment comprising HSS 2 accounts for 40–50% of the full effect of the DCR. In this paper we describe a deletional analysis of HSS 2. We show that a 225 bp fragment is sufficient to direct high levels of expression of the human  $\beta$ -globin gene which is copy number dependent and integration site independent. This 225 bp fragment overlaps the major region that is hypersensitive 'in vivo'. DNase I footprinting shows the presence of four binding sites for the erythroid specific protein NF-E1; the three other footprinted regions display a remarkable redundancy of the sequence GGTGG and bind a number of proteins including Sp1 and the CACC box protein. The significance of these results for the regulation of globin gene expression is discussed.**

**Key words:** Dominant Control Region/erythroid factors/ $\beta$ -globin

## Introduction

The human  $\beta$ -globin gene locus consists in the 5' to 3' direction of the  $\epsilon$ -globin gene, which is expressed in embryonic stages, the G $\gamma$  and A $\gamma$  genes, which are expressed during fetal development and the  $\delta$ - and  $\beta$ -globin genes which are expressed during adult life. The gene cluster comprises ~60 kb (for review, see Collins and Weissman, 1984) and expression is restricted to cells of the erythroid lineage. To understand the mechanism underlying the stage and tissue specific expression of the genes in the human  $\beta$ -globin gene cluster, DNA fragments containing the individual genes were used to generate transgenic mice (Magram *et al.*, 1985; Chada *et al.*, 1985; Kollias *et al.*, 1987; Behringer *et al.*, 1987). It was shown that the  $\gamma$ -globin genes behaved like the mouse embryonic  $\beta$ H1 gene, while the human  $\beta$ -globin gene followed the expression pattern of the mouse  $\beta$ major gene. However, expression levels were very low and dependent on the integration site in the mouse genomic DNA.

Recently, we have shown that the addition of a region containing four very strong hypersensitive sites (HSS) found upstream of the  $\epsilon$ -gene (Tuan *et al.*, 1985; Forrester *et al.*,

1987) to a  $\beta$ -globin gene fragment containing all its known local regulatory elements overcomes the dominant action of position effects (Grosveld *et al.*, 1987). Moreover, expression of the transgene was found to match that of the endogenous globin genes and was dependent on the copy number of the integrated construct. For this reason, the construct was termed the 'minilocus' and the HSS were called the Dominant Control Region (DCR) of the human  $\beta$ -globin gene cluster. The four 5' HSS were contained in a 20 kb fragment in the original minilocus; we subsequently linked the HSS together as a 6.5 kb fragment and showed that this configuration, designated the microlocus, was comparable with the minilocus as defined in murine erythroleukaemia (MEL) cells and transgenic mice (Blom van Assendelft *et al.*, 1989; Talbot *et al.*, 1989). The analysis of a similar construction has recently been reported by Forrester *et al.* (1989).

The discovery of DCR sequences in the human  $\beta$ -globin gene locus has opened the way for a realistic approach towards gene therapy and to the development of animal models for human haemoglobinopathies. The feasibility of the latter has been demonstrated in our laboratory by the development of transgenic mice that carry the human allele for sickle haemoglobin under the control of the DCR (Greaves *et al.*, 1990). The erythrocytes of a mouse producing more  $\beta^s$  than endogenous mouse  $\beta$ -globin sickled both *in vivo* and *in vitro*, providing a mouse model to study anti-sickling drugs and gene therapy protocols.

For somatic gene therapy, introduction of the transgene via retroviral insertion in stem cells appears to be the most realistic approach at the moment. For this, the development of retroviral constructs passaging with high titres is essential to obtain a high frequency of infection in the targeted cells. When single HSS of the DCR were tested in MEL cells, it was shown that HSS 1 and 4 gave a level of ~10% of the microlocus, while both HSS 2 and 3 conferred ~50% to a linked  $\beta$ -globin gene (Collis *et al.*, 1990). These results have been confirmed in transgenic mice (Fraser *et al.*, 1990).

In this paper we describe a detailed analysis of HSS 2. Fine mapping shows that the HSS 2 is, in fact, a small hypersensitive region rather than a defined site and functional analysis of MEL cells and transgenic mice shows that a core 225 bp fragment, coinciding with the HSS, allows high level of position independent expression.

## Results

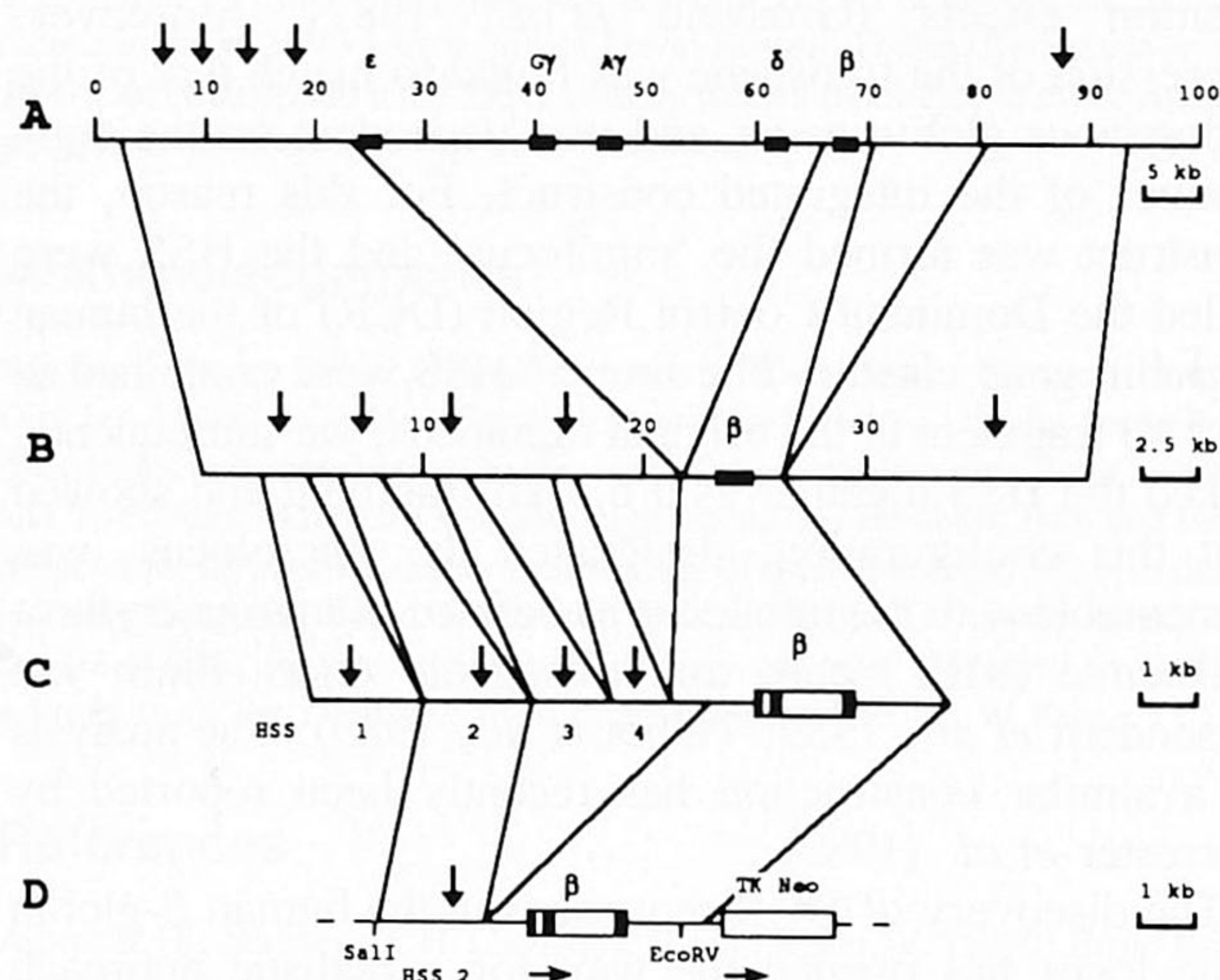
### Fine mapping of HSS 2 in MEL cells

The locations of the HSS in the  $\beta$ -globin DCR were previously mapped on large restriction fragments (Tuan *et al.*, 1985; Forrester *et al.*, 1987; Grosveld *et al.*, 1987). This allowed us to construct a smaller fully functional DCR containing the individual HSS on 1–2 kb fragments (Talbot



*et al.*, 1989; Figure 1C). In this microlocus construct, HSS 2 is present on a 1.9 kb *Hind*III fragment. To determine the position of HSS 2 more accurately, we took advantage of a MEL cell line containing four copies of the minilocus construct (clone C, Blom van Assendelft *et al.*, 1989). Nuclei were isolated (Gorski *et al.*, 1986) and treated with different amounts of DNase I for 5 min on ice. As a non-erythroid control, mouse L-cells containing the same construct were used (Blom van Assendelft *et al.*, 1989). Southern blots of *Hind*III digested DNA were probed for

HSS 2 via indirect end-labelling. With a probe specific for the 5' end of the 1.9 kb *Hind*III fragment, we reproducibly found two hypersensitive regions in MEL cells, the 5' region being the weakest (Figure 2, panel B). With the 3' probe, only the stronger 3' region is seen (Figure 2, panel A). We were unable to detect hypersensitivity in L-cells (Figure 2, panels A and B). From these results we infer that the 5' hypersensitive site is located between nucleotides 950 and 1150, and the 3' site between 1250 and 1550 (Figure 3, top line).

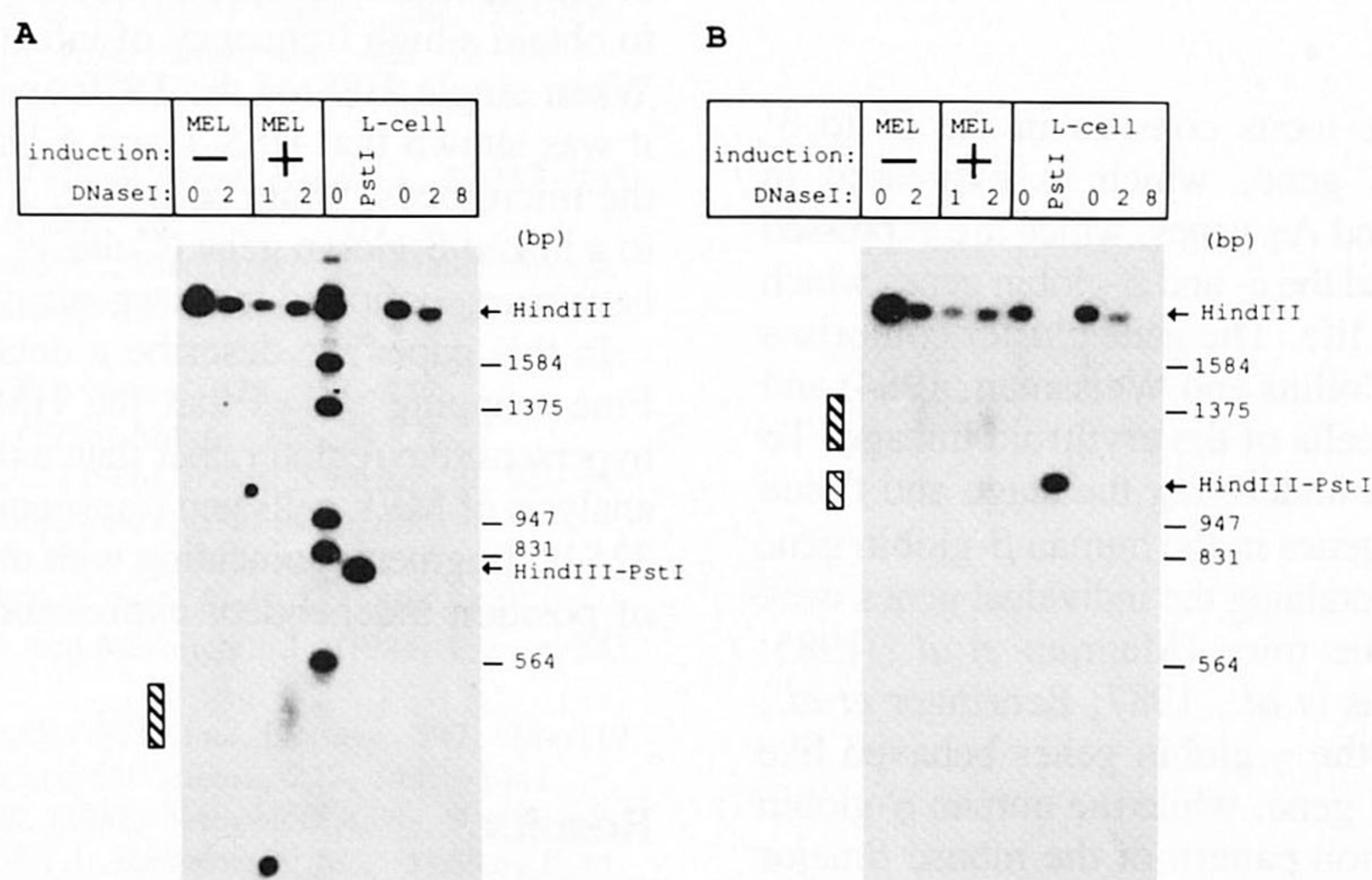


**Fig. 1.** The Dominant Control Region of the human  $\beta$ -globin gene cluster. (A) The human  $\beta$ -globin gene cluster on the short arm of chromosome 11. The DCR, characterized by four hypersensitive sites 5' to the  $\epsilon$ -globin, and one hypersensitive site 3' to the  $\beta$ -globin gene is indicated by vertical arrows: (B) The minilocus described by Grosfeld *et al.* (1987), combining the DCR and the 3' hypersensitive site with the human  $\beta$ -globin gene. (C) The microlocus constructed by Talbot *et al.* (1989), containing the DCR region as four restriction fragments of 1–2 kb. (D) The plasmid constructs used in this study. All the fragments tested are cloned in the *Hpa*I site at 800 bp in front of the cap site of the human  $\beta$ -globin gene. The *Sal*I and *Eco*RV sites used to isolate the DNA containing the gene plus HSS 2 for transgenic mice are indicated.

### Functional analysis of HSS 2 deletions

A series of HSS 2 deletions was made (Figure 3) and cloned in the *Hpa*I site at position –800 in front of the human  $\beta$ -globin gene as shown in Figure 1D. Except when indicated otherwise, the natural sense orientation was used. Plasmids were linearized with *Pvu*I and transfected into MEL cells by electroporation. After selection in G418 containing medium, populations were induced by the presence of 2% DMSO for 4 days and expression of the construct was measured by quantitative S1 analysis using expression of the endogenous mouse  $\alpha$ -globin genes as a control (Figure 4).

Construct 1 is the 1.9 kb *Hind*III fragment, which serves as a reference for the full activity of HSS 2 on its own. Construct 2 is a doublet of this fragment. Interestingly, expression of the test gene is remarkably higher with this doublet, indicating that this tandem array allows cooperativity between the two HSS 2 fragments (Table I). Constructs 3–8 are various 5' and 3' fragments of HSS 2; analysis of their expression patterns allows us to draw the conclusion that only the fragments containing the major 3' HSS (Figure 3, Table I) give the full expression observed with the original HSS 2 fragment. Constructs 6 and 7 are particularly instructive, since construct 6 contains the 5' part of the 1.9 kb *Hind*III fragment, including the weak 5' HSS, while construct 7 contains the remaining 3' part comprising the 3' HSS. The 5' border of construct 7 is just inside the 5' border of the strong HSS as mapped in MEL cells (Figures 2 and 3). Only construct 7 gives a level of



**Fig. 2.** Fine mapping of HSS 2 in MEL and L-cells. Nuclei were isolated and treated with DNase I (see Materials and methods). DNA was purified and restricted with *Hind*III and Southern blotted. MEL DNA was from uninduced (–) cells or from cells induced for 2 days with 2% DMSO (+). The amount of DNase I ( $\mu$ g/mg DNA) is indicated. The major and minor hypersensitive regions are indicated by dashed blocks. (A) Hybridized with the 3' *Bst*NI–*Hind*III fragment. (B) Hybridized with the 5' *Hind*III–*Bam*HI fragment. The position of the major and the minor hypersensitive regions within the 1.9 kb fragment are indicated in Figure 3.



expression comparable with the full site; the level of construct 6 is essentially the same as that of the vector without any DCR fragment (Table I).

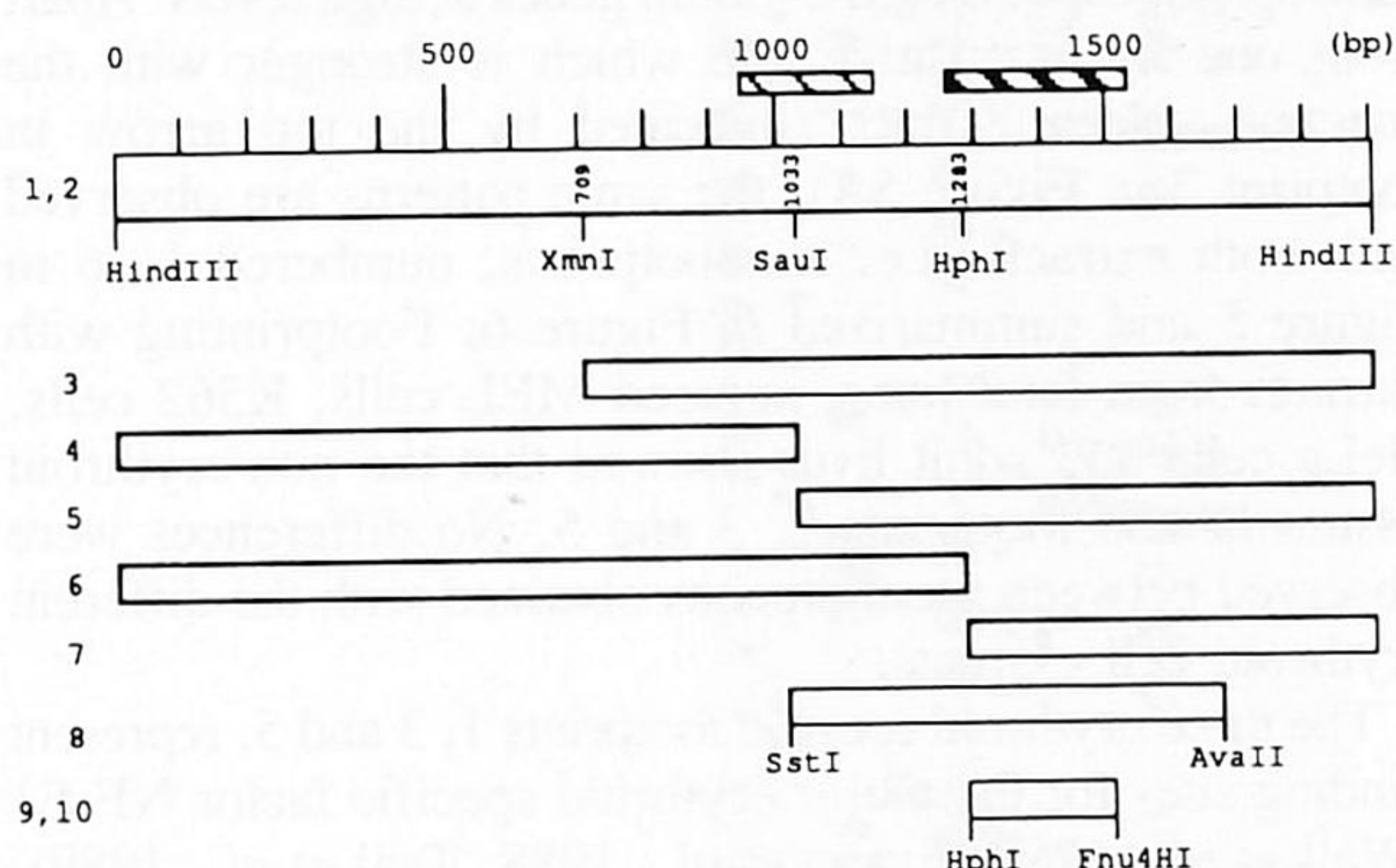
These results prompted us to test a 225 bp *HphI*–*Fnu4HI* fragment which is slightly smaller than the major HSS. This fragment was tested in both the sense (construct 9) and the antisense (construct 10) orientation. The results are shown in Figure 4, panels A and B and Table I. Even this small restriction fragment provides nearly full expression when compared with constructs 1,3,5,7 and 8, independent of orientation. This predicts that this small fragment is capable of providing position independent high levels of expression. However, analyses using MEL cell populations are rather limiting in assaying position independent expression. First

of all, cell transformants would have to be cloned and analysed for expression and transgene integrity. Second, the test construct must be integrated into active chromatin regions since the transfected cells are selected in G418 for expression of the linked marker *tk-neo* gene, thus biasing the result. Finally, only a limited copy number range can be achieved in these MEL cells and transcription levels per gene are decreased with increasing copy numbers (>4–5) with similar type small constructs (Talbot *et al.*, 1990).

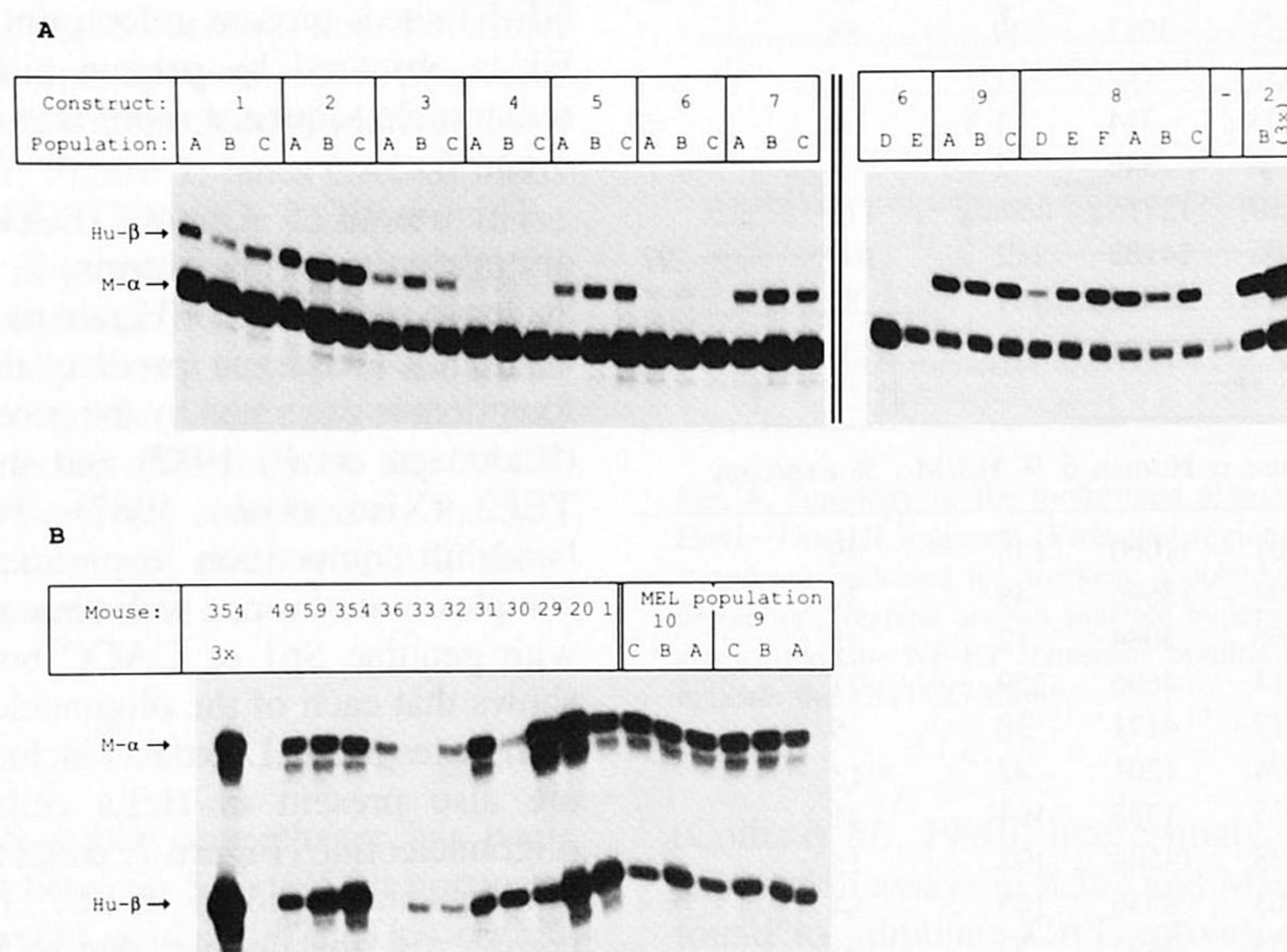
### Transgenic mice

To test for position independent levels of expression we used the *SalI*–*EcoRV* fragment of construct 9 (Figures 1 and 3) for microinjection into fertilized mouse eggs. Transgenic 13.5 day fetuses were collected and expression was compared with 13.5 day fetuses containing the 1.9 kb *HindIII* fragment driving the human  $\beta$ -globin gene (Fraser *et al.*, 1990). To determine copy numbers of seven transgenic fetuses, the initial blots of placenta DNA were probed with a human  $\beta$ -globin probe and an endogenous mouse *Thy-1* probe as a loading control. To screen for mosaicism of the transgene, DNA from body, head and yolk sac was then also analysed, using the same probes. This showed that mice 31 and 32 were mosaic, since different tissues had a different ratio of human  $\beta$ -globin to *Thy-1* signal on the blots (not shown).

Figure 4B and Table I show the results of a quantitative RNA analysis in fetal livers of the seven transgenic fetuses. As expected for mosaic mice, 31 and 32 show low levels of expression per copy of the gene. The other mice show expression levels per gene of 40–50% when compared with the construct containing the entire HSS 2 region (nos. 49 and 59). The exceptions are the mice with extremely high copy numbers (nos. 30 and 354), showing lower levels of



**Fig. 3.** Deletions of HSS 2 tested in stably transformed MEL cells. All fragments are cloned in the *HpaI* site of the vector GSE 1273 in the sense orientation, with the exception of construct 10, which is the antisense orientation of construct 9. Construct 2 contains the 1.9 kb *HindIII* fragment duplicated in a tandem array. Hatched boxes indicate hypersensitive sites.



**Fig. 4.** S1 analysis of the hypersensitive site 2 deletions. In **panel A** a probe for the 3' end of the human  $\beta$ -globin mRNA was used, in **panel B** a probe for the 5' end. The probe for the endogenous mouse  $\alpha$ -globin mRNA was for the 5' end in all cases. The combination of a human 3'  $\beta$ -globin probe and the mouse 5'  $\alpha$ -globin probe results in a background  $\beta$ -signal which was subtracted from all  $\beta$  signals to obtain the values in Table I. Constructs are numbered as in Figure 3. Transgenic mice 1, 20 and 29–33 were made with the *SalI*–*EcoRV* fragment of construct 9 (Figures 1 and 3). Mouse 36 is the non-transgenic control. Transgenic mice 49, 59 and 354 contain the 1.9 kb *HindIII* fragment 3' to the 4.8 kb *BglII* fragment of the human  $\beta$ -globin gene (Fraser *et al.*, 1990). Specific activities: panel A, left:  $\beta:\alpha = 1:5.3$ ; panel A, right:  $\beta:\alpha = 2.7:1$ ; panel B,  $\beta:\alpha = 1.5:1$ .



**Table I.** Expression of hypersensitive site 2 constructs in MEL cells and transgenic mice

MEL populations						
Construct	Copy no.	Mouse $\alpha$	Human $\beta$	% H $\beta$ /M $\alpha$	% exp/copy	Av.
1A	>8	13282	3509	24	—	
1B	1	14785	1968	13	52	44
1C	2	17882	3127	18	36	
2A	2	8687	4636	53	106	
2B	2.5	17599	10770	61	98	106
2C	2	13672	7767	57	114	
3A	2	8733	2177	25	50	
3B	3	7075	3345	47	63	59
3C	2	7174	2266	32	64	
4A	1	11304	182	2	8	
4B	1	15832	134	1	4	5
4C	1	9121	47	0.5	2	
5A	3	6597	2822	43	57	
5B	2.5	10904	3711	34	54	52
5C	2	17914	4113	23	46	
6A	1	17603	59	0.3	1	
6B	0.5	16339	81	0.5	2	
6C	0.5	10372	44	0.4	3	2
6D	3	32389	122	0.4	0.53	
6E	dl	nd	nd	—	—	
7A	3	10866	3411	31	42	
7B	2	20820	4171	20	40	41
7C	2	15761	3145	20	40	
8A	4.5	8769	5046	58	52	
8B	3	7134	2285	32	43	
8C	4	8234	4060	49	49	50
8D	dl	11367	1537	13	—	
8E	3	10617	3257	30	40	
8F	3	9190	4555	50	67	
9A	5	9222	4634	50	40	
9B	2	11403	3631	32	64	48
9C	3.5	12922	4543	35	40	
10A	1.5	9025	2369	26	69	
10B	7.5	9688	5436	56	30	52
10C	2.0	10754	3052	29	58	
1273A no DCR	1	13951	117	1	4	
1273B no DCR	1.5	20703	314	1.5	6	5
1273C no DCR	2	17178	346	2	4	
1401A $\mu$ l	3.5	13210	12716	96	110	
1401B $\mu$ l	4.5	12606	14182	112	100	97
1401C $\mu$ l	5	12999	13088	101	80	
Transgenic mice						
Mouse no.	Copy no.	Mouse $\alpha$	Human $\beta$	% H $\beta$ /M $\alpha$	% exp/copy	
1	11	8101	11060	136	49	
20	5	27494	13696	49	39	
29	1	24666	3084	12	48	
30	50	1814	4696	259	21	
31 mosaic	30	11017	4171	38	5	
32 mosaic	13	2874	1201	42	13	
33	19	715	1388	194	41	
354	50	7538	14504	192	15	
59	5	8103	8536	105	84	
49	2	6655	3492	53	106	
36 non transgenic	0	3228	43	1	—	

Bands were cut out of the gel and Cerenkov counted. A similar sized gel fragment just above the band of interest was also counted for background correction. The data given are corrected for the relative specific activities of the probes used. Copy numbers were determined from Southern blots as described (Talbot *et al.*, 1990). The microlocus controls were those used by Talbot *et al.*, (1989). dl, deletion;  $\mu$ l, microlocus; nd, not determined.

expression, a phenomenon previously observed for both HSS 2 and 3 (Fraser *et al.*, 1990). This indicates that although the small *HphI*–*Fnu4HI* fragment shows a reduced activity when compared with the large *HindIII* fragment in fetal liver versus MEL cells, this core fragment has retained the capacity to provide copy number dependent, integration site independent expression on the  $\beta$ -globin gene.

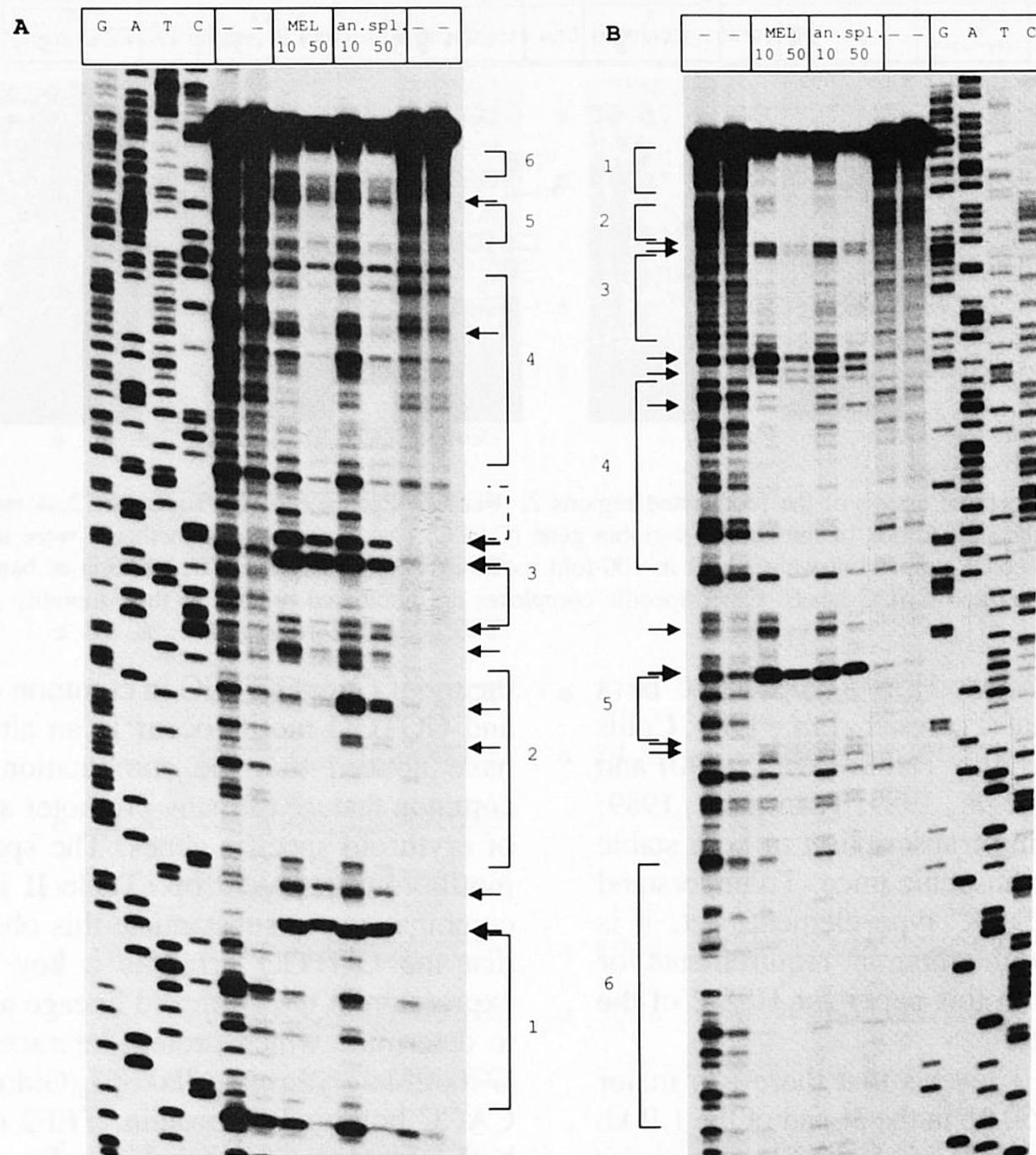
#### DNase I footprinting

Based on the expression data in transgenic mice and MEL cells we analysed protein–DNA interactions in the 225 bp *HphI*–*Fnu4HI* fragment by *in vitro* DNase I footprinting. Figure 5 shows the results obtained with nuclear extracts (Gorski *et al.*, 1986) from uninduced MEL cells, representing the non-expressing adult erythroid stage and anaemic adult spleen expressing the globin genes at high levels. Apart from one hypersensitive site which is stronger with the anaemic spleen extract (indicated by the top arrow in footprint 2 in Figure 5A), the same patterns are observed with both extracts, i.e. six footprints, numbered 1–6 in Figure 5 and summarized in Figure 6. Footprinting with extracts from fetal liver, induced MEL cells, K562 cells, HeLa cells and adult liver showed that the non-erythroid tissues lacked footprints 1, 3 and 5. No differences were observed between the footprints obtained with the different erythroid cell extracts.

The three erythroid specific footprints 1, 3 and 5, represent binding sites for the major erythroid specific factor NF-E1 (Wall *et al.*, 1988; Evans *et al.*, 1988; Tsai *et al.*, 1989). As a hallmark of NF-E1 binding very strong hypersensitive sites are observed immediately upstream of the binding site (indicated by horizontal arrows in Figure 5 and vertical arrows in Figure 6). Competition experiments with NF-E1 specific oligonucleotides of the human  $\beta$ -globin gene 3' enhancer (Wall *et al.*, 1988) also confirm that footprints 1, 3 and 5 are NF-E1 sites (not shown). Another consensus NF-E1 site is present in footprint 4, but the presence of NF-E1 is obscured by protein binding to the neighbouring remarkable sequence motif that is repeated in footprints 2, 4 and 6.

This motif 5'-GnnnGGTGG-3' occurs in the same orientation twice in footprint 2, three times in footprint 4 and once in footprint 6 (Figure 6). The presence of additional Gs in the sequence predicts that at least part of these footprints is generated by the general transcription factor Sp1 (Kadonaga *et al.*, 1987) and the CACC binding protein TEF2 (Xiao *et al.*, 1987). We therefore carried out bandshift/competition experiments to determine which complexes are formed with these regions and compared these with genuine Sp1 or CACC box binding sites. Figure 7 shows that each of the oligonucleotides forms a number of complexes in MEL extracts including Sp1 and TEF2, which are also present in HeLa cells (not shown). The Sp1 oligonucleotide (Figure 7, panel Sp1) contains a dimer Sp1 binding site (Gidoni *et al.*, 1985) which forms four complexes, the doublet due to binding of the 95 kd and 105 kd forms of Sp1 (Jackson and Tjian, 1988) and a slower mobility complex [labelled (2 $\times$ )] which is the result of two Sp1 binding sites on the same oligonucleotide. The nature of the faster mobility complex (also labelled Sp1) which is observed in all of our extracts is, at present, not clear. It could be a degradation product of Sp1 (Gustafson and Kedes, 1989), but could also be a different protein (Xiao *et al.*,



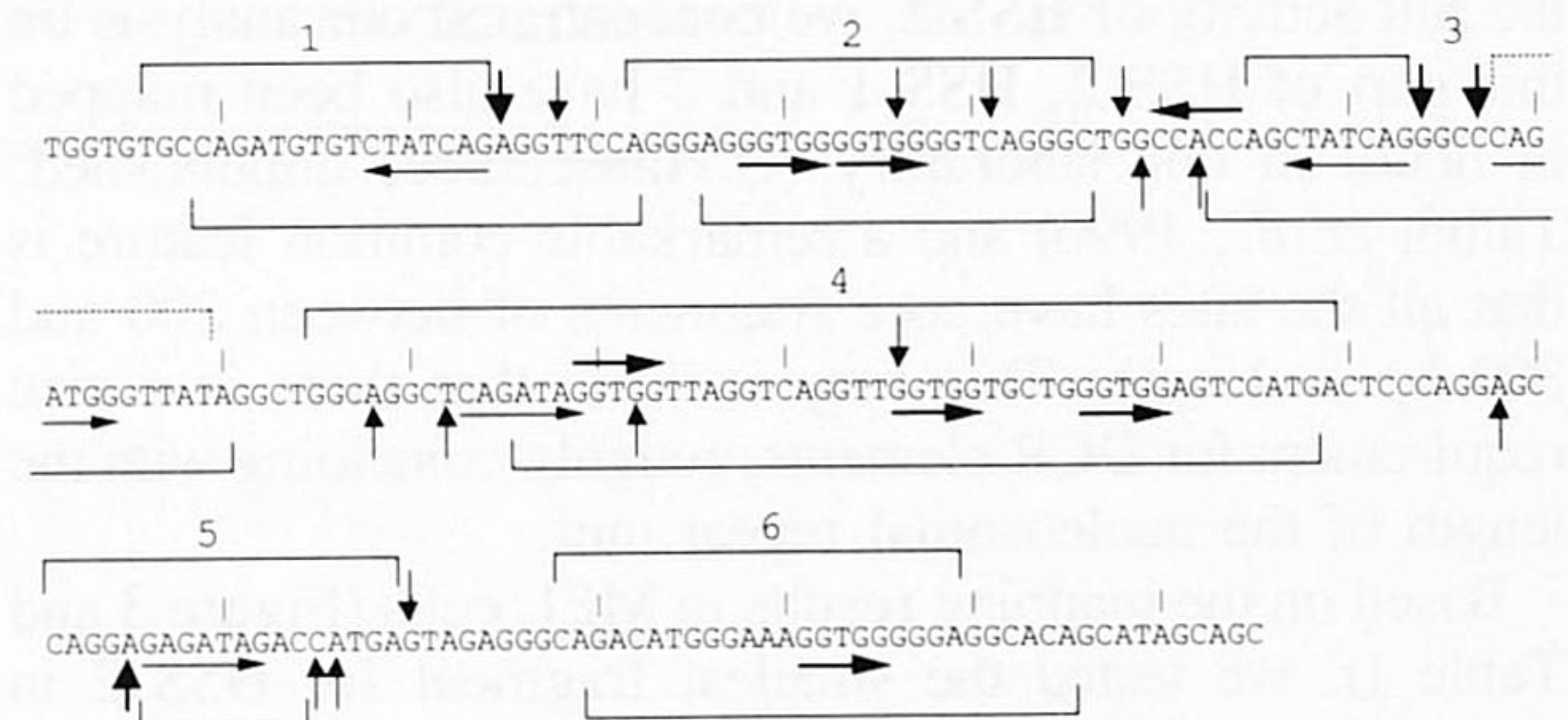


**Fig. 5.** DNase I footprinting of the 225 bp *HphI*–*Fnu4HI* fragment. The DNA in the ‘no protein’ (–) lanes was treated with 1, 0.5, 0.25 and 0.125  $\mu$ g DNase I from left to right in each panel. DNA pre-incubated with 10  $\mu$ g nuclear extract from MEL cells or anaemic spleens (an.spl.) was treated with 1  $\mu$ g DNase I, DNA pre-incubated with 50  $\mu$ g nuclear extract was treated with 2  $\mu$ g DNase I. **Panel A** shows the sense strand, **panel B** the antisense strand. Footprinted regions are indicated by numbered brackets, a weak footprint by a dotted bracket. Arrows indicate hypersensitive sites.

1987). In support of the latter possibility, is the fact that an antibody specific for Sp1 (gift of S.Jackson) does not affect the mobility of this band (E.Spanopoulou and F.Grosveld, unpublished). The CACC box oligonucleotide (derived from the  $\beta$ -globin gene promoter, Figure 7, lanes CACC) binds the CACC box protein (labelled CACC) in addition to those proteins bound by the Sp1 oligonucleotide. It also binds a number of fast mobility complexes. The site 2 probes 2, 4 and 6 (see Figure 6) specifically form a number of additional complexes (labelled 1–10). Of the additional complexes, only number 1 can be competed efficiently by the CACC box oligonucleotide, while none are competed by the Sp1 oligonucleotide. It is at present not clear which of these proteins is functionally important.

## Discussion

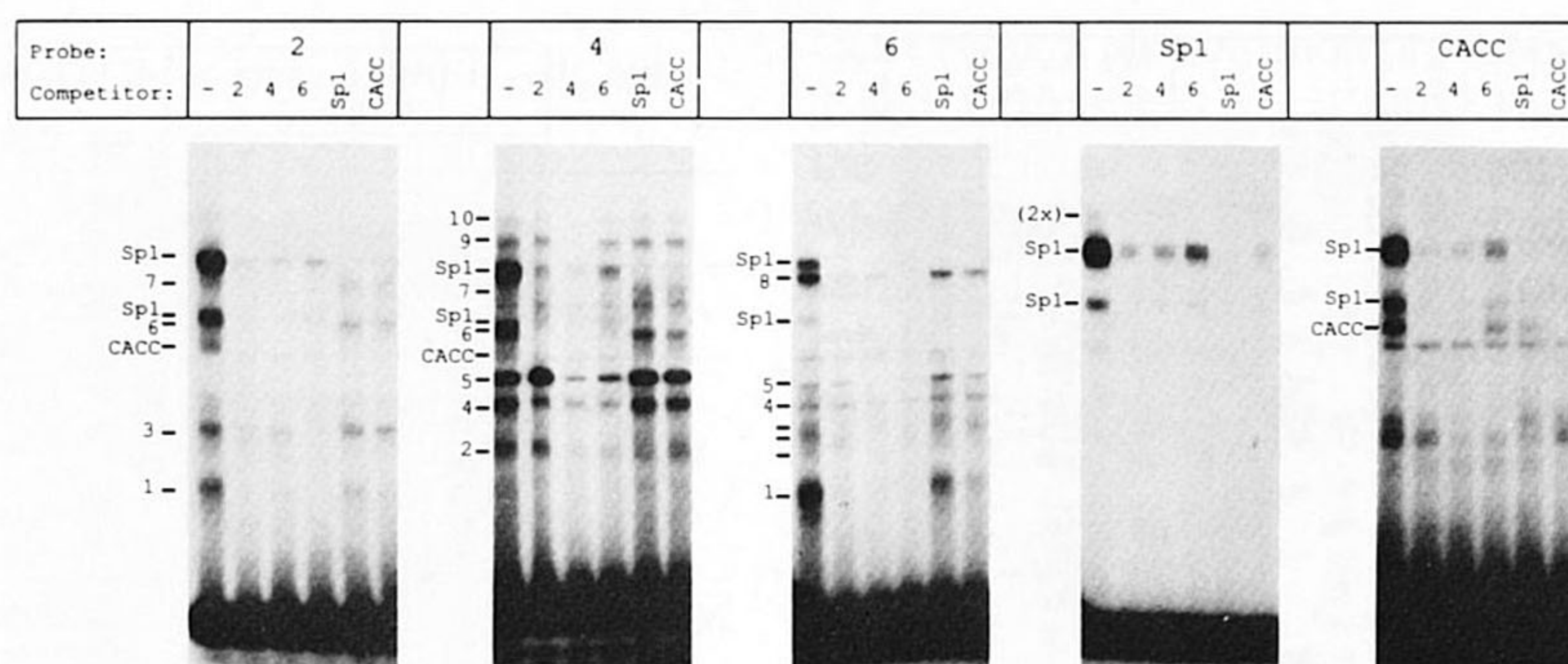
The DCR of the human  $\beta$ -globin gene cluster has been defined to the DNA region between 5 and 25 kb upstream of the  $\epsilon$ -globin gene (Grosveld *et al.*, 1987). This region contains four erythroid specific ‘super’ hypersensitive sites for DNase I (Tuan *et al.*, 1985; Forrester *et al.*, 1987), which were shown to be the functional elements of the DCR (Talbot *et al.*, 1989; Forrester *et al.*, 1989). Deletional analysis of the microlocus construct demonstrated that HSS 2 and 3, as single sites, conferred high levels of expression on the human  $\beta$ -globin test gene in a stable transfection assay



**Fig. 6.** Summary of the footprinted regions in the 225 bp *HphI*–*Fnu4HI* fragment. Protected regions on the sense and antisense strand are indicated by brackets, a dotted bracket indicates weak protection; vertical arrows indicate hypersensitive sites. Thin horizontal arrows indicate NF-E1 consensus binding sites; thick horizontal arrows indicate the GGTGG motif.

(Collis *et al.*, 1990). Interestingly, when tested in transient expression assays in K562 and MEL cells, only HSS 3 was found to stimulate CAT activity (Tuan *et al.*, 1989 and C.H.Chang and P.Dierks, submitted, respectively). This suggests that integration into chromatin is an important prerequisite for proper functioning of the DCR. It clearly distinguishes ‘DCR’ type elements from classical enhancers, which were originally defined in transient transcription assays (for review see Serfling *et al.*, 1985; Maniatis *et al.*, 1987; Dynan, 1989) and do not necessarily provide integration





**Fig. 7.** Gel mobility shift and competition assays of the footprinted regions 2, 4 and 6. Probes covering footprints 2, 4 and 6 (Figure 6) and probes for the SV40 GC-box (Sp1) and the CACC box of the human  $\beta$ -globin gene (CACC) (see Materials and methods) were used in a gel shift assay with nuclear extracts from MEL cells. Competitors were added in 100-fold molar excess as indicated. The position of bands specific for the Sp1 and CACC proteins are indicated by Sp1 and CACC labels. Other specific complexes are numbered relative to their mobility.

position independent expression. HSS 3 appears to be a combination of a DCR element (Forrester *et al.*, 1989; Collis *et al.*, 1990; Fraser *et al.*, 1990; Talbot *et al.*, 1990) and a powerful enhancer (Curtin *et al.*, 1989; Tuan *et al.*, 1989; Ryan *et al.*, 1989) stimulating transcription in both stable and transient assays and in transgenic mice. To understand the mechanism by which 'DCR' type elements act, it is therefore important to define the minimal requirements for this activity, as we describe in this paper for HSS 2 of the human  $\beta$ -globin gene cluster.

Precise mapping of HSS 2 reveals that there is a major hypersensitive region of  $\sim 250$  bp in the 3' end of the 1.9 kb *Hind*III fragment, and a minor site of 150 bp 5' to the major site. The minor site is found only when probing from the 5' end of the *Hind*III fragment and footprinting analysis (not shown) shows very few, if any, protein binding sites. Since deletional analysis in MEL cells shows that all the constructs containing the major site (1, 3, 5, 7, 8, 9 and 10) retain the full activity of HSS 2, we concentrated our analysis on this part of HSS 2. HSS 1 and 3 have also been mapped in detail in our laboratory (O.Hanscombe, unpublished; Talbot *et al.*, 1990) and a remarkable common feature is that all the sites have core fragments of between 200 and 300 bp in length. This may indicate that there is a size requirement for DCR elements, roughly coinciding with the length of the nucleosomal repeat unit.

Based on the mapping results in MEL cells (Figure 3 and Table I), we tested the smallest fragment for HSS 2 in transgenic mice. This fragment has retained the most important properties of the entire 1.9 kb HSS 2, i.e. orientation and integration position independent expression of the  $\beta$ -globin gene, albeit at a reduced level. The latter could be due to the fact that the smallest fragment has lost some enhancing sequences, e.g. a potential NF-E2 binding site (Mignotte *et al.*, 1989a,b), which is located just upstream of this fragment and could act as an enhancer (Talbot *et al.*, 1990). It also highlights the occurrence of differences between MEL cell assays and transgenic mice analysis, something we have previously observed with HSS 1 (Collis *et al.*, 1990; Fraser *et al.*, 1990). This could be due to stage specific differences between the two systems or some inherent limitation of the MEL cells.

DNase I footprinting with erythroid nuclear extracts shows that there are six protected regions in the 225 bp fragment. These comprise at least three NF-E1 binding sites (Wall *et al.*, 1988; Tsai *et al.*, 1989) and three regions which have

the motif GnnnGGTGG in common (Figure 6). The NF-E1 and GGTGG motifs occur in an alternating order and we have noticed that the combination of these motifs is a common feature of many promoter and regulatory elements of erythroid specific genes. The spacing between the two motifs is often  $\sim 30$  bp. Table II lists a number of such combinations to substantiate this observation. We propose that the GGTGG array is a key determinant for gene expression in the erythroid lineage and it will be important to determine which protein interacts functionally with the G-motif in erythroid cells. Sp1 (Gidoni *et al.*, 1985) and the CACC box binding protein, TEF2 (Xiao *et al.*, 1987) are both abundant in erythroid cells and both bind to this sequence *in vitro* (Figure 7), but these proteins are not erythroid specific. It has been shown that both of these can interact with other distal transcription elements and their factors to mediate their effect to the transcriptional machinery (Schüle *et al.*, 1988). However, the GGTGG motif is different from the Sp1 and the CACC consensus and it is clear that a number of other ubiquitous proteins are bound to this region which may provide the main activating function *in vivo*; a clue to this is provided by the conserved G residues 5' and 3' to the GGTGG repeats in HSS 2 (Table II). Point mutations at all the positions should resolve which of the proteins is active in a functional analysis. It is conceivable that the combination of such a factor and a tissue specific factor (NF-E1) is sufficient to activate and direct high levels of erythroid specific gene expression. Although the NF-E1 protein has the ability to *trans*-activate (Tsai *et al.*, 1989; Evans and Felsenfeld, 1989), the presence of NF-E1 sites alone is not sufficient for activation *in vivo*, since each of the globin genes (without a DCR) contains multiple NF-E1 binding sites, but is only expressed at low levels and in a highly position dependent manner. Therefore, NF-E1 could be a major activator only in combination with a second factor, where one protein could be the first to bind, without activation of the genes, but enabling the second to bind, thus mediating a transcriptional effect.

Finally, the work described here contributes to the construction of a fully functional DCR of the human  $\beta$ -globin gene cluster containing at least three of the HSS on an  $\sim 1$  kb fragment. This should greatly facilitate the construction of retroviral vectors which passage with titres high enough to transduce cultured bone marrow stem cells efficiently. It would be a significant step forward if  $\beta$ -thalassaemia could be cured by grafting of autologous 'repaired' bone marrow



**Table II.** Comparison of different GGTGG arrays in erythroid promoters and regulatory elements

AACCTCTGATAGACAC	a 19 s	GGGGAGGGTGGGGTGGG	s 30 a	GGGCCCTGATAGCTGG	HSS 2, 1324
AACCTCTGATAGACAC	a 24 s	GGGTGGGGTGGGGTCAG	s 25 a	GGGCCCTGATAGCTGG	HSS 2, 1329
GGGCCCAGATGGGTTA	s 30 s	CAGATAGGTGGTTAGGT			HSS 2, 1395
GCTCTCAGATAGGTGG	s 22 s	CAGGTTGGTGGTGTCTGG	s 41 s	CAGGAGAGATAGACCA	HSS 2, 1412
GCTCTCAGATAGGTGG	s 32 s	GTGCTGGGTGGAGTCCA	s 31 s	CAGGAGAGATAGACCA	HSS 2, 1422
CAGGAGAGATAGACCA	s 34 s	GGGAAAGGTGGGGGAGG			HSS 2, 1487
ATCGTGAGATAGACGT	a 27 s	AGAAGGGGTGGACTCCA			HSS 1, 899
CAGGGCAGATGGCAAA	s 26 a	TAGTCAGGTGGTCAGCT			HSS 3, 1019
CAGGGCAGATGGCAAA	s 42 a	GTTTGAGGTGGAGTTTT			HSS 3, 1035
		TGCCATGGTGGTTTGCT	s 27 a	TAATGTAGATGACGGG	HSS 4, 247
		GTTGGGGGTGGGGGGCT	a 24 s	AGTGTGTGATGTTCCC	HSS 4, 247
CAGCAGTGATGGATGG	a 30 a	CACAGGGGTGGAGTCAG			H. $\epsilon$ , -165
GCATTGAGATAGTGTG	a 44 a	CCCATGGGTGGAGTTTA			H. G $\gamma$ , -143
GGCCTATGATAGGGTA	a 10 a	ATTTGGGGTGGGGCCTA	a 40 s	TGTTTAAGATTAGCAT	H. $\beta$ -gl.enh., +231
		TTGTGGGGTGGCGCGTG	a 30 a	GGCTCCAGATTGAGAG	H. $\alpha$ -gl., -713
CGAGCGGGATGGGCGG	s 23 s	GTGGCGGGTGGAGGGTG			H. $\alpha$ -gl., -202
CGAGCGGGATGGGCGG	s 30 s	GTGGAGGGTGGAGACGT			H. $\alpha$ -gl., -195
		GGAAGGTGGGCCTGG	s 12 s	GGCCTGGGATAACAGC	H. Gph. A, -51
		AGGAAGGGTGGGGCCTG	a 31 a	GTAAAGAGATAAGGCC	H. PBGD, -100
		GCTGGGTGTGCCC	s 32 s	CCTCAGATAAGACC	Rat PK, -51
		CAGCTGGGTGGGGGCAG	s 16 s	GGTTGCAGATAAACAT	Ch. $\beta^A$ enh., +1882
		CAGCTGGGTGGGGGCAG	s 31 a	AAGTCTTGATAGCAAA	Ch. $\beta^A$ enh., +1882

## GGTGG

G	8	6	14	6	7	14	21	21	0	21	20	9	15	4	7	6	10
A	3	5	1	6	6	5	0	0	0	0	0	7	0	3	1	3	4
T	3	7	3	5	6	2	0	0	21	0	1	4	2	8	4	6	6
C	5	1	3	4	2	0	0	0	0	0	0	1	4	6	9	5	0

Consensus: - - G - - G G G T G G r G y <sup>C</sup>G - r

The GGTGG and NF-E1 motifs are underlined. Frequencies of nucleotides in each position are given. Note that all the sequences are given in the orientation that allows alignment, as indicated by s (sense) and a (antisense). Distances are calculated from the central base in the GGTGG motif to the central base in the NF-E1 motif RAGATNR. Numbering given for the HSS of the DCR is from the 5' restriction sites used to construct the microlocus (Talbot *et al.*, 1989); others are relative to the cap site. H., human; gl., globin; enh., enhancer; Gph., glycophorin; PBGD, phosphobinogen deaminase; PK, pyruvate kinase; Ch., chicken.

stem cells, opening the road to the treatment of human haematopoietic diseases by somatic gene therapy protocols.

## Materials and methods

### Plasmid constructions

Plasmid GSE1273 contains the human  $\beta$ -globin gene as a 4.8 kb *Bgl*II fragment linked to a tk-*neo*<sup>r</sup> gene (Talbot *et al.*, 1989). All the fragments

tested were blunted and cloned in *Hpa*I digested GSE1273, replacing the 650 bp *Hpa*I fragment in the 5' flanking region of the human  $\beta$ -globin gene (see Figure 1).

### DNase I sensitivity

Nuclei were isolated according to Gorski *et al.* (1986) and resuspended at a DNA concentration of 1 mg/ml in 15 mM Tris pH 7.5, 60 mM KCl, 15 mM NaCl, 0.2 mM EDTA, 0.2 mM EGTA, 0.15 mM spermine, 0.5 mM spermidine. DNase I (Worthington) was added to concentrations



of 1–8 µg/ml, and the reaction was started by the addition of MgCl<sub>2</sub> to 10 mM and CaCl<sub>2</sub> to 1 mM final concentration. The reaction was allowed to proceed for 5 min on ice, and stopped by the addition of 1 vol 50 mM EDTA, 1% SDS. DNA was isolated, digested with appropriate restriction enzymes, and analysed by Southern blotting.

#### Tissue culture and cell transfections

The MEL cell clone β-C and the L-cell population 4, each containing 4–5 copies of the human β-globin minilocus construct, have been described previously (Blom van Assendelft *et al.*, 1989).

The MEL cell line C88 was maintained in standard αMEM plus 10% fetal calf serum. Plasmid constructs were linearized at the *PvuI* site and transfected by electroporation as follows: log phase MEL cells (2 × 10<sup>7</sup> cells per transfection) were washed and resuspended in 1.5 ml HEPES buffered saline containing 50 µg of the linearized plasmid. After incubation on ice for 10 min the cells were electroporated with three pulses from a Hoefer 'Pro-Genitor' apparatus set to deliver 250 V for 10 ms. After 5 min at room temperature, they were divided to generate three independent transfected populations (Antoniou *et al.*, 1988). MEL cells were induced to undergo erythroid differentiation by incubation in the presence of 2% (v/v) DMSO for 4 days.

#### Transgenic mice

Plasmids were cut with *EcoRV* and *SalI* and the fragment containing the human β-globin gene was separated from plasmid sequences by agarose gel electrophoresis and recovered from the gel via isotachophoresis (Öfverstedt *et al.*, 1984). DNA was dissolved at 1–2 µg/ml in microinjection buffer and oocytes were injected into the male pronucleus, as described (Kollias *et al.*, 1987). Fetuses were collected 13.5 days after transfer to pseudopregnant foster mothers and analysed as described (Grosveld *et al.*, 1987).

#### RNA analysis

RNA was isolated from transfected cell populations and 13.5 day fetal mouse liver by the method of Auffray and Rougeon (1980).

Quantitative S1 nuclease analysis using human β-globin and mouse α-globin DNA probes was performed as described (Kollias *et al.*, 1987; Antoniou *et al.*, 1988; Talbot *et al.*, 1989).

#### DNA analysis

Southern blotting was performed essentially as described by Southern (1975) using nylon membranes (Nytran, Schleicher and Schüll) and the hybridization conditions described by Church and Gilbert (1985). Copy numbers of the human β-globin gene were determined by laser densitometry using the signal of the endogenous mouse Thy-1 gene as an internal loading control. 13.5 day transgenic fetuses were screened for mosaicism of the transgene by comparing the human β-globin/mouse Thy-1 ratio in DNA from placenta, body, head and yolk sac.

#### DNase I footprinting

The 225 bp *HphI*–*Fnu4HI* fragment of HSS 2 was blunted and cloned in both orientations in the *SmaI* site of M13mp8. 1 µg of single-stranded template DNA was annealed to 5 pmol of kinased Universal 17 mer sequencing primer and extended using Klenow polymerase and all four cold dNTPs. The reaction products were phenol extracted and passed over a Sephadex G50 column. After overnight digestion with 10 U *EcoRI*, the DNA was phenol extracted, ethanol precipitated and dissolved in 200 µl of TE buffer. Each footprinting reaction contained 1 µl of the probe mentioned above (3000 c.p.m.), 2 µg poly(dI-dC) and 10–50 µg nuclear protein as described by Wall *et al.*, 1988, but was pre-incubated for 10 min on ice. The samples were analysed on 6% sequencing gels, using dideoxy sequencing reactions as markers.

Nuclear extracts were prepared according to Gorski *et al.* (1986), with modifications as described in Wall *et al.* (1988). HeLa nuclear extracts were prepared according to Dignam *et al.* (1983).

#### Gel mobility shift assays and competition studies

Gel mobility shift assays were performed as described previously (deBoer *et al.*, 1988), using 5 µg MEL nuclear extract per reaction. Competitions were done by adding 100-fold molar excess of the indicated double-stranded oligonucleotides before addition of the extract.

Nucleotide sequences of the oligonucleotides used in these studies were:

oligonucleotide 2 sense strand:

ATCACAGGTTCCAGGGAGGGTGGGGTGGGGTCAGGGCTGGCC-AC

oligonucleotide 4 sense strand:

GCTCAGATAGGTGGTTAGGTCAGGTTGGTGGTGGTGGGTTGGAG-TCCATGACTCCCAG

oligonucleotide 6 sense strand:

TAGAGGGGAGACATGGGAAGGTGGGGGAGGCACAGCATAG

Sp1 oligonucleotide late strand:

CGATGGGCGGAGTTAGGGGCGGGACTAT

CACC oligonucleotide sense strand:

CGATCCGTAGAGCCACACCCTAGGTAT

## Acknowledgements

We are grateful to E.deBoer for assistance and to Cora O'Carroll for preparation of the manuscript. S.P. was supported by an EMBO long-term fellowship, D.T. by an NSERC (Canada) and a Commonwealth scholarship and P.F. by an NIH postdoctoral fellowship. This work was supported by the Medical Research Council (UK).

## References

- Antoniou, M., deBoer, E., Habets, G. and Grosveld, F. (1988) *EMBO J.*, **7**, 377–384.
- Auffray, C. and Rougeon, F. (1980) *Eur. J. Biochem.*, **107**, 303–314.
- Behringer, R.R., Hammer, R.E., Brinster, R.L., Palmiter, R.D. and Townes, T.M. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 7056–7060.
- Blom van Assendelft, M., Hanscombe, O., Grosveld, F. and Greaves, D.R. (1989) *Cell*, **56**, 969–977.
- Chada, K., Magram, J. and Costantini, F. (1985) *Nature*, **319**, 685–688.
- Church, J.M. and Gilbert, W. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 1991–1995.
- Collins, F.S. and Weissman, S.M. (1984) *Prog. Nucleic Acid Res. Mol. Biol.*, **31**, 315–462.
- Collis, P., Antoniou, M. and Grosveld, F. (1990) *EMBO J.*, **9**, 233–240.
- Curtin, P.T., Liu, D., Liu, W., Chang, J.C. and Kan, Y.W. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 7082–7086.
- deBoer, E., Antoniou, M., Mignotte, V., Wall, L. and Grosveld, F. (1988) *EMBO J.*, **7**, 4203–4212.
- Dignam, J., Lebowitz, R. and Roeder, R. (1983) *Nucleic Acids Res.*, **11**, 1475–1489.
- Dynan, W. (1989) *Cell*, **58**, 1–7.
- Evans, T. and Felsenfeld, G. (1989) *Cell*, **58**, 877–885.
- Evans, T., Reitman, M. and Felsenfeld, G. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 5976–5980.
- Forrester, W.C., Takegawa, S., Papayannopoulou, T., Stamatoyannopoulos, G. and Groudine, M. (1987) *Nucleic Acids Res.*, **15**, 10159–10177.
- Forrester, W.C., Novak, U., Gelinas, R. and Groudine, M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 5439–5443.
- Fraser, P., Hurst, J., Collis, P. and Grosveld, F. (1990) *Nucleic Acids Res.*, in press.
- Gidoni, D., Kadonaga, J.T., Barrera-Saldana, H., Takahashi, K., Chambon, P. and Tjian, R. (1985) *Science*, **230**, 511–514.
- Gorski, K., Carneiro, M. and Schibler, U. (1986) *Cell*, **47**, 767–776.
- Greaves, D.R., Fraser, P., Vidal, M., Hedges, M., Ropers, D., Luzzatto, L. and Grosveld, F. (1990) *Nature*, **343**, 183–185.
- Grosveld, F., Blom van Assendelft, G., Greaves, D.R. and Kollias, G. (1987) *Cell*, **51**, 975–985.
- Gustafson, T.A. and Kedes, L. (1989) *Mol. Cell. Biol.*, **9**, 3269–3283.
- Jackson, S. and Tjian, R. (1988) *Cell*, **55**, 125–133.
- Kadonaga, J.T., Carner, K.R., Masiara, F.R. and Tjian, R. (1987) *Cell*, **51**, 1079–1080.
- Kollias, G., Hurst, J., deBoer, E. and Grosveld, F. (1987) *Nucleic Acids Res.*, **15**, 5739–5747.
- Magram, J., Chada, K. and Costantini, F. (1985) *Nature*, **315**, 338–340.
- Maniatis, T., Goodbourn, S. and Fischer, J.A. (1987) *Science*, **236**, 1237–1245.
- Mignotte, V., Eleouet, J.F., Raich, N. and Romeo, P.-H. (1989a) *Proc. Natl. Acad. Sci. USA*, **86**, 6545–6552.
- Mignotte, V., Wall, L., deBoer, E., Grosveld, F. and Romeo, P.-H. (1989b) *Nucleic Acids Res.*, **17**, 37–54.
- Öfverstedt, L.-G., Hammarshom, K., Balgobin, N., Kjerten, S., Petterson, U. and Chattopadhyaya, J. (1984) *Biochim. Biophys. Acta*, **732**, 120–126.
- Ryan, T.M., Behringer, R.R., Townes, T.M., Palmiter, R.D. and Brinster, R.L. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 37–41.



- Schüle, R., Müller, M., Kaltschmidt, C. and Renkawitz, R. (1988) *Science*, **242**, 1418–1420.
- Serfling, E., Jasin, M. and Schaffner, W. (1985) *Trends Genet.*, **1**, 224–230.
- Southern, E.M. (1975) *J. Mol. Biol.*, **98**, 503–517.
- Talbot, D., Collis, P., Antoniou, M., Vidal, M., Grosveld, F. and Greaves, D.R. (1989) *Nature*, **338**, 352–355.
- Talbot, D., Philipsen, S., Fraser, P. and Grosveld, F. (1990) *EMBO J.*, **9**, 2169–2178.
- Tsai, S.-F., Martin, D.I.K., Zon, L.I., D'Andrea, A.D., Wong, G.G. and Orkin, S.H. (1989) *Nature*, **339**, 446–451.
- Tuan, D., Solomon, W., Qiliang, L.S. and Irving, M.L. (1985) *Proc. Natl. Acad. Sci. USA*, **32**, 6384–6388.
- Tuan, D.Y.H., Solomon, W.B., London, I.M. and Lee, D.P. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 2554–2558.
- Wall, L., deBoer, E. and Grosveld, F. (1988) *Genes Dev.*, **2**, 1089–1100.
- Xiao, J.-H., Davidson, J., Macchi, M., Rosales, K., Vigneron, M., Staub, A. and Chambon, P. (1987) *Genes Dev.*, **1**, 794–807.

Received on January 29, 1990; revised on March 27, 1990