Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 6018–6022, July 1993
Evolution

Phylogenetic footprinting reveals unexpected complexity in trans factor binding upstream from the ε-globin gene

DEBORAH L. GUMUCIO*†, DAVID A. SHELTON*, WENDY J. BAILEY‡, JERRY L. SLIGHTOM‡§, AND MORRIS GOODMAN‡

*Department of Anatomy and Cell Biology, University of Michigan Medical School, Ann Arbor, MI 48109-0616; ‡Department of Anatomy and Cell Biology, Wayne State School of Medicine, Detroit, MI 48201; and §Molecular Biology Unit 7242, The Upjohn Company, Kalamazoo, MI 49007

Communicated by Roy J. Britten, March 29, 1993

ABSTRACT **The human ε-globin gene undergoes dramatic changes in transcriptional activity during development, but the molecular factors that control its high expression in the embryo and its complete repression at 6–8 weeks of gestation are unknown. Although a putative silencer has been identified, the action of this silencer appears to be necessary but not sufficient for complete repression of ε gene expression, suggesting that multiple control elements may be required. Phylogenetic footprinting is a strategy that uses evolution to aid in the elucidation of these multiple control points. The strategy is based on the observation that the characteristic developmental expression pattern of the ε gene is conserved in all placental mammals. By aligning ε genomic sequences (from −2.0 kb upstream to the ε polyadenylylation signal), conserved sequence elements that are likely binding sites for trans factors can be identified against the background of neutral DNA. Twenty-one such conserved elements (phylogenetic footprints) were found upstream of the ε gene. Oligonucleotides spanning these conserved elements were used in a gel-shift assay to reveal 47 nuclear binding sites. Among these were 8 binding sites for YY1 (yin and yang 1), a protein with dual (activator or repressor) activity; 5 binding sites for the putative stage selector protein, SSP; and 7 binding sites for an as yet unidentified protein. The large number of high-affinity interactions detected in this analysis further supports the notion that the ε gene is regulated by multiple redundant elements.**

The human ε-globin gene is expressed at extremely high levels during early embryonic life but is silenced completely at 6–8 weeks of gestation. Elucidation of the factors that control this repression is important since the reactivation of this gene in individuals with sickle cell anemia and β-thalassemia could potentially cure these anemias. Studies in transgenic mice have shown that the cis sequences that direct the developmental silencing of ε are located near the gene (1, 2). In fact, candidate silencer sequences have been identified within the ε promoter. In transient transfection assays, the region between bp −177 and −392 represses reporter gene activity by 3-fold in erythroid cells and by 10-fold in nonerythroid cells (3). In transgenic mice, deletion of these sequences results in persistence of human ε transgene expression in definitive mouse erythrocytes (4). However, the deleted construct is still subject to considerable down-regulation. Therefore, although this silencer region is indeed necessary for complete stage-specific repression of ε, additional silencer elements must also exist. Additional silencers have not been detected in the deletional studies that have been carried out (3).

An alternative approach to the identification of these silencers is the use of evolutionary clues. Mammals of such different eutherian orders as Primates (e.g., human), Lagomorpha (e.g., rabbit), Rodentia (e.g., mouse), and Artiodactyla (e.g., cow and goat) possess ε-globin genes that were derived from the same ancestral proto-ε gene (5). In all of these lineages, the characteristic pattern of ε gene expression during development is nearly identical. Thus, the controlling factors may be evolutionarily conserved. In support of this hypothesis is the observation that the human ε-globin gene is properly regulated during development in the transgenic mouse background (1, 2). To pinpoint conserved sequence motifs that might be important in ε gene regulation, we have used a strategy called "phylogenetic footprinting." In an earlier study, this technique was applied to the γ-globin genes (6, 7). In alignments of orthologous sequences from several mammalian species, sequence motifs that showed 100% conservation in all species over a region of ≥6 contiguous base pairs were defined as "phylogenetic footprints" (6–8). Of the 13 phylogenetic footprints that were identified upstream from the γ gene, gel-shift analysis showed that 12 bound nuclear proteins whereas only 2 of 9 nonconserved regions bound proteins. These findings validated the phylogenetic footprinting approach as an efficient means to identify potentially important cis sequences and the nuclear proteins that bind to them.

In this study, we have aligned ε sequences from six mammalian species: human, orangutan, gibbon, capuchin monkey, galago, and rabbit. The total number of evolutionary years included in such an alignment is additive. Since this data base encompasses >270 million years of evolution (9), each nucleotide has had adequate time to accumulate changes if such changes are tolerated. Stretches of sequence that are invariant among all species are likely to be truly conserved and possibly functionally important. Twenty-one conserved elements were identified in the 2 kb of sequence immediately upstream from ε, nearly double the number identified in this region of the γ sequence (6, 7). This may be a direct reflection of the fact that all of these ε genes are expressed in the same developmental stage (embryonic), whereas the γ genes studied were heterogeneous in their expression pattern (some fetal and some embryonic).

Probes spanning each phylogenetic footprint bound proteins in the gel-shift assay, revealing an unexpectedly complex pattern of redundant binding motifs. Among the 47 binding interactions characterized were 8 sites for YY1 (yin and yang 1) (10). A single binding site for this protein, earlier called CSBP-1 (conserved sequence binding protein 1), was found (6) upstream from each γ gene. This unusual protein has been shown to have the properties of a "switch" protein (10) in that it can act as a repressor (6, 10–12) or as an activator (10, 13, 14). YY1 binding sites were found both within and upstream of the putative ε silencer region. These upstream sites could represent the additional silencer sites

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

Abbreviations: YY1, yin and yang 1; SSP, stage selector protein; CSBP, conserved sequence binding protein.

†To whom reprint requests should be addressed.

mentioned above. In addition to YY1, several binding sites were found for the stage selector protein (SSP) (15, 16) and for another protein that we have named CSBP-2.

MATERIALS AND METHODS

The alignment used here was generated as described (10) and covered 2.0 kb upstream from the ε gene, the entire ε gene, and 300 bp 3′ to the termination codon. The sequences used are available through GenBank; the alignment itself is available through EMBL (File server, ALIGN:DS8841.DAT) or by request from the authors. Selected portions of this alignment are presented in Fig. 1. Sequence elements displaying 100% conservation for ≥6 bp were defined as phylogenetic footprints. Nineteen double-stranded oligonucleotide probes (Table 1) were made spanning the 21 phylogenetic footprints found in this alignment (probes −835 and −1735 spanned two phylogenetic footprints). Oligonucleotides were synthesized on Applied Biosystems synthesizer model 380B at Upjohn. The 5′ end of the antisense strand contained a GATC overhang that facilitated labeling of the annealed probes by using the Klenow fragment of DNA polymerase I. Labeling of probes, preparation of nuclear extracts, and performance of gel-shift assays were as described (17). Binding studies were performed using erythroid (K562, HEL, and MEL) and nonerythroid (HeLa and HepG2) nuclear extracts.

```
                -2039
HSA  ATATAATAAATAACAAAGTCAGGTTATAGAAGAGAGAAACGCTCTTAGTAAACTTGGAAT
PPY  AT  A  A A AA  AG  CA    T G AG     ACGC T TA TA  CT G A T
HLA  AC  A  A A AA  AG  CA    T G AG     ATGC C TA TA  CT G T T
CAL  AC  A  T A AA  AG  GA    C G AG     TTAA C TA GG  CT G A T
GCR  AG  A  A A GC  AG  AG    A G TA     GAGT C CA TA  TA G A T
OCU  TA  C  A G AC  GC  AG    T A CA     GAGT C GG TA  CT A A C

                                        -1735
HSA  AATTCCTAGATCTGGT*GGGGCAAGGGGAGCCATAG*GAGAAAGAAATGGTAGAAATGGA
PPY    T C C GA C GGT*G G  A GGGG  CCATA *GA      T A      TGGA
HLA    T T C GA C GGT*G G  A GGGG  CCATA *GA      T A      TGGA
CAL    T C C GA C GTC*A G  A GGGA  ACATT *GG      T A      ****
GCR    C C C GG C GGAAG G  C ATAG  CACCA AGA     A A      ****
OCU    T T C TG A AGCAA A  A AGGA  AACTA *AG     G A      ****

          -835
HSA  CCCTTCCAGTGAGAAGTATAAGCAGGACAGACAGGCAAGCAAGAAGAGCCCCAGGCAA
PPY  CC  TCCCAG G G      A C AGA     G  AAGC  GAA  GAGCCCCA  C
HLA  CC  TCCCAG G G      T C GGA     G  AAGC  GAA  GAGCCCCA  C
CAL  GC  TTCCGA G G      G C GCA     G  AAGC  GAA  GAG*CCCA  C
GCR  CT  CACTGA G G      G C GCA     G  A**C  A**  TAGCTCAG  A
OCU  GC  CATGGA T *      A T GAG     A  CAAA  AGA  G****TCA  A

                        -698
HSA  ATTCCCTGGAAGCACTGGATGTAATCTT*TTCTGTCTGTCCTCTCTAGGGAATCACCCCA
PPY  ATTCCCTGG     A T  ATGG  T   *TT  G       C  TGG G  T  CCC A
HLA  ATTTCCTGG     A T  ATGT  T   *TT  G       C  TAG G  T  CCC A
CAL  AGTCCCTGA     C C  GCGT  T   *TT  G       C  TAG G  T  TCC A
GCR  ATTCCCTGG     A C  GTGT  T   GTT  T       C  CC* G  C  GGT T
OCU  *********     A A  AGAT  C   G**  G       G  TAG A  T  CTC T

       -48
HSA  CAGGGGGCCAGAACTTCGGCAGTAAAGAATAAAAGGCCAGACAGAGAGGCAGCAGCACAT
PPY  C G    CC   A T  G     GAA A          CACAGAG GG    A   C
HLA  C G    CC   A T  G     AAA A          CACAGAG GG    A   C
CAL  T G    GC   * T  A     AAA A          CGCAGAG AG    C   C
GCR  C A    CT   A T  A     GGA G          CATAGAA AT    A   T
OCU  C A    CC   C A  A     GGG A          AGCCTTG AG    A   C

                    -1095
HSA  CAACTTCCAGTTTCAGCTCTACCAAGTAAAGAGCTAGCAAGTCATCAAAATGGGGACATA
PPY  CAAC TCC    T CAGC CT C  AGT     GC  C A      C C ATGGGGACA *
HLA  CAAC TCC    T CAGC CT C  GGT     GC  G A      C C ATGGGGACA *
CAL  CAAC TCT    C CAGT AT C  GGT     AC  G A      C C ACAGGGACA *
GCR  CAGC *CC    T TAGC TC A  TAT     GC  A A      T C A******CA A
OCU  GGAT TAC    T TGTC CC A  TGC     AT  G C      C C TCAGGGGGC T
```

FIG. 1. Selected portions of the alignment used in this study. Sequences are from human (HSA), orangutan (PPY), gibbon (HLA), capuchin monkey (CAL), galago (GCR), and rabbit (OCU). Asterisks denote gaps added to maximize alignment; blank regions correspond to 100% homology among all aligned sequences. Positions of the oligonucleotide probes used are shown as overlines and identified with a number that indicates the position of the 5′ base of the probe with respect to the aligned sequences. This number corresponds to the probe numbers used in the text.

Table 1. Oligonucleotide probes used in this study

Probe	Sequence
(−48/−48)	CAGAACTTCGGCAGTAAAGAATAAAAGGC-CAGACAGAGAG
(−95/−95)	GGTCAGCCTTGACCAATGACTTTTAAGTAC
(−125/−124)	GGACCTGACTCCACCCCTGAGGACACAGG
(−147/−146)	CATCCATCACTGCTGACCCTCTCCGGACCT
(−177/−176)	TCCAGCACACATTATCACAAACTTAGTGTC
(−325/−303)	TTTCTTGGAAAAGGAGAATGGGAGAGAT-GG
(−667/−605)	CCAAGGTACTGTACTTTGGGATTAAGGCTT
(−698/−638)	ATCTTTTCTGTCTGTCCTCTCTAGGGAATC
(−835/−774)	TCCCAGTGAGAAGTATAAGCAGGACAGAC-AGGCAAGCAAG
(−883/−820)	ATTCTGGCTTTAAATAATTTTAGGATTTT
(−906/−843)	ACTTTTTTCTCTGTTTGATGACAAATTCTG
(−1095/−959)	CAGCTCTACCAAGTAAAGAGCTAGCAAGT-CATCAAAATGG
(−1213/−1073)	TATGAGGATAATGACAATGGTATTATAAGG
(−1257/−1115)	TAGTTAAATAATTTCTGTGAATTTATTCCT
(−1327/−1185)	ATTAACAATGCTGGAATTTGTGGAACTCT
(−1735/−1589)	TAGGAGAAAGAAATGGTAGAAATGGATG-GA
(−1792/−1638)	CATAGGAAATTGTAGGAACAGAAATTCC-TA
(−1939/−1772)	ATAACAAAGTCAGGTTATAGAAGAGAGAA-ACGCTCTTAGT
(−2093/−1925)	AATTCAGTGGCCTGGAATAATAACAATTTG

Sequences of the sense strand are given. The two numbers in parentheses indicate the location of the 5′ end of the probe in the aligned sequence data base and the location of the 5′ end of the probe with respect to the ε cap site. In the text and figures, the first number is used to identify the probe. The underlined bases represent phylogenetic footprints (defined as ≥6 contiguous bp of conservation among all species tested). In some cases, additional conserved regions of 4 or 5 contiguous base pairs were also seen in the region spanned by the oligonucleotides. These additional conserved regions are also underlined in the sequences above.

RESULTS AND DISCUSSION

Identification of 21 ε Phylogenetic Footprints. The alignment covered a contiguous stretch of genomic DNA from 2.0 kb upstream of ε to 300 bp 3′ of the polyadenylylation site. Within this region, 21 phylogenetic footprints were detected; all were found upstream from the gene. None were found in introns or in the short region 3′ to the gene. Exons were excluded from the analysis. All 19 of the probes shown in Table 1 were found to bind nuclear proteins; most bound multiple proteins. Fig. 2*A* presents gel-shift patterns for most of the probes tested in this study. The identity of each of the binding proteins was determined by competition assays as illustrated in Fig. 2 *C* and *D*. The binding data are summarized schematically in Fig. 3.

Previous binding studies carried out in other laboratories had already established that the transcription factor GATA-1 binds to the −165 region, Sp1 binds to the CACCC sequence at the −111 position, and CP1 binds to the CCAAT box at the −81 position (18, 19). All three of these regions were identified as phylogenetic footprints and these binding results were confirmed in our analysis. However, we found that in addition to CP1, the CCAAT box is bound by two or three additional proteins (number varies with extract source) that have not yet been fully characterized. The same proteins bind to the proximal CCAAT box from the galago γ-globin gene but not to the proximal CCAAT box of the human γ gene (7). This is of interest because the galago γ gene, like the human ε gene, is expressed in embryonic life (8) whereas the human γ gene is silent in the embryo, activated in fetal life, and silenced again in the adult. Thus, the binding of these

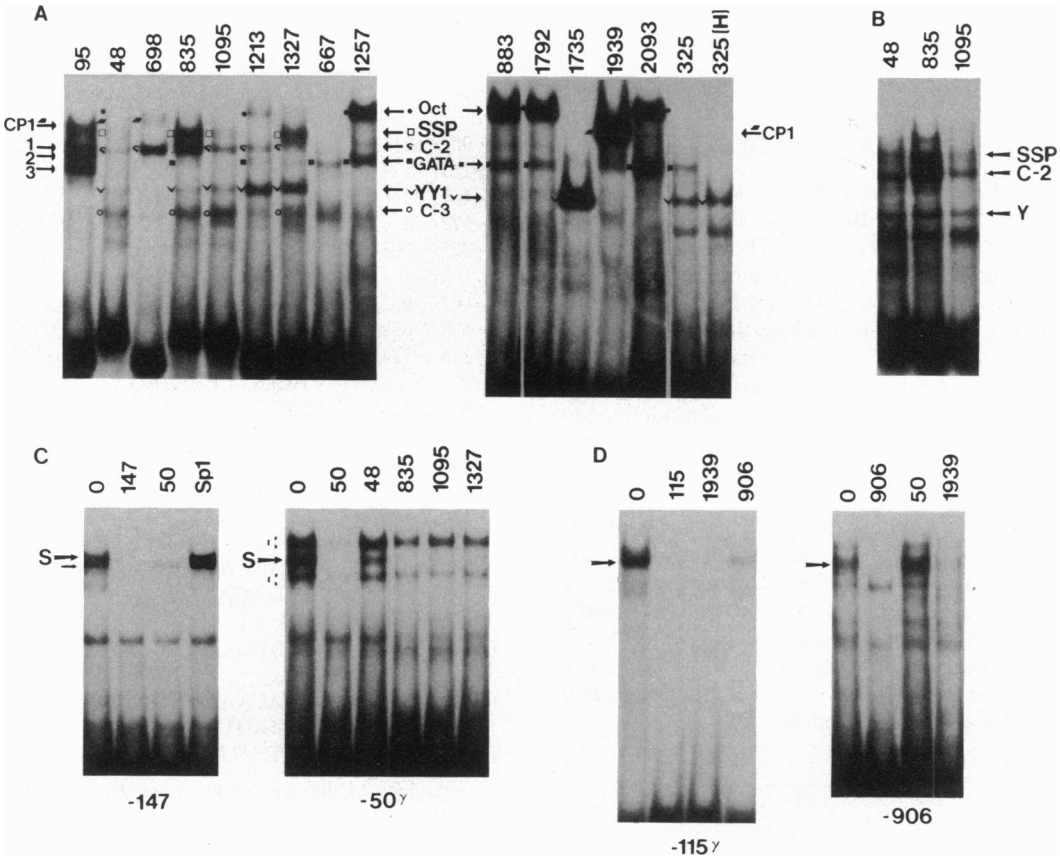

FIG. 2. (*A*) Binding patterns seen with 15 of the 19 probes tested in this study. Binding to the positions −125 and −177 has been described elsewhere (18, 19); binding to probes −147 and −906 is shown in *C* and *D*, respectively. K562 cell nuclear extracts were used except in the lane marked [H], in which a HeLa cell extract was used. Complexes are identified in the margins. The numbers 1, 2, and 3 (to the left) refer to three uncharacterized complexes that are detected with the probe that spans the CCAAT box (probe −95). Y, C-2, and C-3 (in the center) refer to YY1, CSBP-2, and CSBP-3, respectively. CSBP-3 has not yet been fully characterized and may represent a proteolytic product of CSBP-2. (*B*) Similar binding patterns are seen with probes −48, −835, and −1095; each probe binds SSP, YY1 (Y), and CSBP-2 (C-2). (*C*) Five ε oligonucleotides bind SSP. In the four lanes to the left, probe −147 is used; this region binds SSP (S) and an uncharacterized protein (−147B, small arrow). The left lane contains no added competitor oligonucleotide; the three lanes to its right contain 100 ng of the unlabeled oligonucleotide indicated above the lane. The Sp1 oligonucleotide (6) binds Sp1 with high affinity but not SSP; oligonucleotide −50γ(50, derived from the human γ globin −50 region, ref. 6) binds Sp1 and SSP. In the six lanes to the right, the probe is oligonucleotide −50γ from the γ promoter (6). Complexes correspond to Sp1 (open arrows) and SSP (S). Competitor oligonucleotides (100 ng) are shown above the lanes. The competition profiles indicate that oligonucleotide −48 binds SSP weakly, whereas oligonucleotides −835, −1095, and −1327 bind SSP strongly. (*D*) Binding of CP1 to oligonucleotides −1939 and −906. In the four lanes to the left, the probe is −115γ (derived from the distal CCAAT box region of the γ promoter, ref. 17); this probe binds CP1 (arrow). Competition with 100 ng of the oligonucleotide shown above the lane indicates that this complex can be effectively competed by self-competition (−115) and by oligonucleotide −1939. Weak competition is also seen with oligonucleotide −906. In the four lanes to the right, the probe is oligonucleotide −906. The competition assay illustrates efficient self-competition (probe −906), no competition by an unrelated probe (probe −50γ), and complete competition by probe −1939. Thus, CP1 binds oligonucleotide −906 weakly and oligonucleotide −1939 strongly, despite the fact that neither of these oligonucleotides contains a consensus CCAAT motif.

uncharacterized proteins correlates with expression in the embryonic time period.

Multiple Binding Sites for a Stage-Selector Protein. Among the 21 phylogenetic footprints, five binding sites were detected for a putative stage selector protein that has been called SSP (15, 16). This protein was first detected as a shifted band in a gel-retardation assay when the −50 region of the γ promoter was used as a probe (probe −50γ) (20). Transfection studies in K562 cells showed that this protein seems to provide the γ promoter with a selective advantage over a cis-linked β promoter when the two are in competition for interaction with enhancers from the upstream locus control region (15). The detection of five binding sites for this protein in the ε promoter (documented in Fig. 2*C*) suggests that this protein could also play a role in the expression of the ε gene. This remains to be directly tested.

Multiple Binding Sites for the YY1 Protein. The ubiquitously distributed YY1 protein (10) has been detected simultaneously in several laboratories and has been called NF-E1 (12), UCRBP (11), δ (13), CF-1 (14), and CSBP-1 (6). The name yin and yang 1 (YY1) best describes the ability of this protein to act as a repressor (6, 10–12) or as an activator (10, 13, 14). In an earlier study, we detected a single strong binding site for this protein upstream of each γ gene and three sites within the ε silencer region (6). In this report, seven additional YY1 binding sites are documented, all of which lie upstream from the silencer. Six of these are shown in Fig. 2*A* (band Y). The location of these potential repressor sites is of interest considering the fact that, in transgenic mice, deletion of the silencer region attenuates but does not abolish the stage-specific silencing of ε (4).

We have now detected 11 binding sites for YY1 upstream from γ and ε. To establish an affinity hierarchy among the sites, competition studies were done using the gel-shift assay (data not shown). In these experiments, each of the sites was used as a probe and the efficiency of cross-competition of YY1 by each of the other sites was titrated. The results can be summarized as follows: −1735, −260 > −1250γ, −1327, −1213 > −325 > −835, −1095 > −48 > −667, −250. The sequences of the probes containing the six highest-affinity

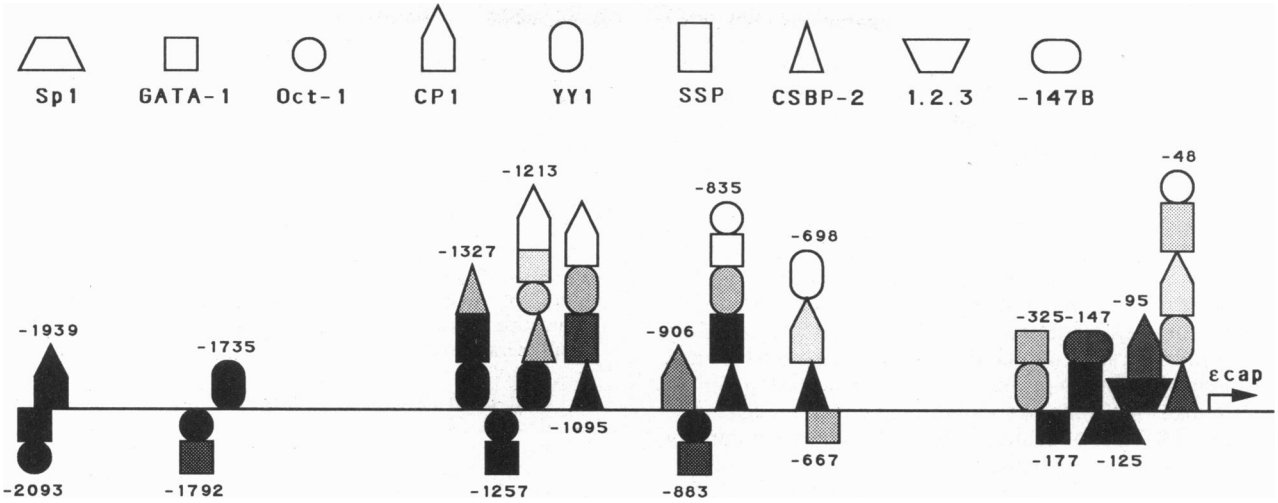

FIG. 3. Schematic representation of the 47 proteins that bind to the 19 probes tested in this study. The proteins are differentiated by shape as indicated by the key above. Shading indicates the relative affinity of the binding interaction: solid symbols represent high-affinity interactions and open symbols indicate very weak interactions. Shaded symbols represent the gradient of affinity between these two extremes. The three uncharacterized proteins that bind the CCAAT region probe are indicated by a single symbol. CP1 also binds to this probe.

globin YY1 sites are compared with other published high-affinity sites for this protein in Fig. 4*A*.

Although the human YY1 protein is ubiquitously distributed, its possible dual nature (activator vs. repressor), its ability to interact with other proteins (10), and its involvement in the stage-specific silencing of another tissue-specific gene (12) suggest that this protein could also play a role in

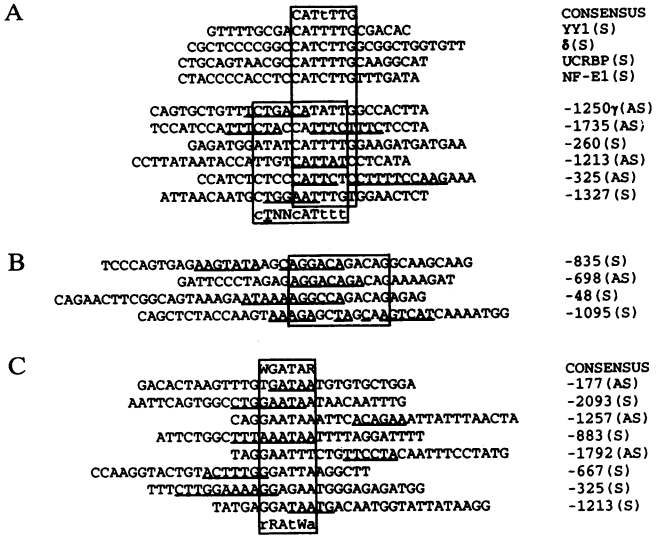

FIG. 4. (*A*) Alignment of the oligonucleotide probes that bind YY1 with high affinity and comparison with NF-E1, YY1, UCRBP, and δ binding sites (10–13). Letters in parentheses indicate whether the sense (S) or antisense (AS) strand was used in the alignment. The −1250γ and −260 sites have been described (6). Evolutionarily conserved bases are underlined. The boxes indicate the two slightly different binding motifs that are observed if (*i*) the four sites described in other systems are aligned, CATTTTG, or (*ii*) only the high-affinity globin sites are considered, cTNNcATttt [lowercase type indicates the most common base where >1 base is seen and the underlined T appears in all of these high-affinity globin sites as well as in four of the five lower-affinity globin sites (data not shown)]. (*B*) Comparison of the probes that bind CSBP-2 with high affinity. Evolutionarily conserved bases are underlined. (*C*) Comparison of the nine probes that bound GATA-1. Probes are aligned according to strength of GATA binding as determined by gel-shift competition assays. The sequences that are most closely related to the canonical GATA site, (A/T)GATA(A/G), are boxed.

globin gene regulation. In this regard, it is of interest that all of the high-affinity globin YY1 binding sites (Fig. 4*A*) and four of the five low-affinity globin sites (data not shown) include a thymidine located 3 bp upstream from the general YY1 consensus sequence (CATTTTG). Although footprinting data are not yet available for all of these sites, this thymidine is within the protected region of the −1250γ probe, as determined by methylation interference (7). It will be of interest to determine whether this thymidine is important for YY1 function in erythroid cells.

Multiple Binding Sites for an Uncharacterized Protein. Analysis of these phylogenetic footprints upstream from ε has revealed seven binding sites for an uncharacterized protein that we have named CSBP-2; six of these sites are shown in Fig. 2*A*. No binding sites for this protein were detected in our earlier analysis of the γ promoter. Oligonucleotides with binding sites for transcription factors Sp1, CP1, Oct-1, GATA-1, AP1, NF-E2, ets, CSBP-1, and CREB did not compete for binding of the CSBP-2 complex (data not shown). The two strongest binding sites for CSBP-2 as determined by cross-competition assays are found in probes −698 and −835. These two probes share an identical 11-bp sequence, AGGACAGACAG (Fig. 4*B*). A third probe that binds CSBP-2 with high affinity spans the ATA box (probe −48) and contains a sequence that matches this motif at 10 of 11 bp. A fourth weaker binding probe, −1095, also contains a related sequence (Fig. 4*B*). For each of these four probes, this sequence element overlaps the phylogenetic footprint.

Binding of GATA-1 and CP1 to Noncanonical Sites. The fact that nine binding sites for the erythroid-specific GATA-1 protein were detected in this analysis is perhaps not surprising given the apparent role of this protein in all stages of erthropoiesis (21). However, examination of the sequences of the probes that bind GATA-1 reveals that relatively high-affinity binding is observed with probes that do not contain good matches for the canonical GATA site. In this study, shifted complexes were identified as containing GATA-1 if formation of the complex was erythroid-specific, if the shifted band comigrated with the GATA-1 complex formed on the −175 region of the γ promoter (a well-characterized GATA-1 binding site, ref. 22), and if the shifted band was effectively competed away by oligonucleotide −175γ. In Fig. 4*C*, the probes that bound GATA-1 in this study are listed according to their relative affinity of binding. The sequences within each probe that are most closely related to the

canonical site are aligned. It should be emphasized that all of these probes also bound additional proteins. Thus, the weak binding of GATA-1 may be stabilized in these *in vitro* assays by cobinding of other proteins. The actual bases in any of these probes that are contacted by GATA-1 will need to be confirmed by footprinting studies with purified GATA-1 protein and functional assays will be required to assess their role in ε gene regulation.

Fig. 2 *A* and *D* demonstrates that the probe spanning the phylogenetic footprint at position −1939 binds CP1, the CCAAT box binding protein, with very high affinity. The binding complex is competed only by probes that contain CP1 sites, the complex comigrates with a well-characterized CP1 complex, and the same subbands are produced when CP1 and the complex binding to probe −1939 are subjected to limited protease digestion (data not shown). However, as discussed above for GATA-1, oligonucleotide −1939 does not contain a CCAAT motif or an element resembling a CP1 consensus binding site. To confirm that no errors had occurred during oligonucleotide synthesis, this oligonucleotide was cloned and three independent clones were sequenced; no errors were found (data not shown). As indicated by our experience with CP1 and GATA-1, the phylogenetic-footprinting approach not only facilitates the detection of uncharacterized proteins such as CSBP-2 but also may reveal information about the binding interactions of well-established transcription factors.

Patterns of Binding Interactions. This analysis also revealed several interesting binding patterns. Probes spanning phylogenetic footprints at −48 (the ATA box), −835, and −1095 each bound YY1, CSBP-2, and SSP (Fig. 2*B*). Furthermore, six of the seven probes that were bound by CSBP-2 were also bound by YY1. Since YY1 has been shown to interact with other proteins (10), it will be of interest to test whether these binding patterns are functionally important in ε gene expression. Finally, seven of nine probes binding GATA-1 also bound Oct-1. Several examples of this codetection of Oct-1 and GATA-1 were observed in an earlier analysis of the γ promoter (6).

Utility of the Phylogenetic Footprinting Technique. The complex problem of identifying all of the control elements that regulate the tissue-specific and developmental-stage-specific expression of individual genes is further amplified in a coordinately regulated gene family such as the globins. For the ε gene, functional studies using deletional approaches have provided valuable information regarding the location of a silencer element (3, 4). However, while this silencer seems to play a role in the repression of ε expression at the end of embryonic life, complete down-regulation appears to require additional silencers (4) that are not detected by deletional studies. The cis sequences identified by phylogenetic footprinting represent possible control elements that have been identified by an alternative criterion, their evolutionary conservation. The unexpected number of binding sites observed not only demonstrates that this approach provides a sensitive and efficient strategy to detect these potentially important cis elements but also supports the suggestion that regulation of the ε gene may be controlled by multiple redundant elements. This redundancy may account for the inability of the deletional approach to clearly identify the control points. It is possible, with the information we now have, to design direct functional tests to analyze the contribution of individual binding proteins, to probe possible interactions among binding proteins, or to determine the relevance of binding sites with sequences that deviate from the consensus sequence. As a prime example, site-directed mutagenesis of the multiple YY1 sites individually and in combinations and analysis of these mutant constructs in transgenic mice will allow a careful assessment of the role of this protein in ε gene silencing.

We thank Ms. Debora Riley for oligonucleotide synthesis and Drs. Francis Collins, Kenji Hayasaka, Miriam Meisler, Juanita Merchant, Diane Robins, and Linda Samuelson for helpful review of the manuscript. Support is gratefully acknowledged from National Institutes of Health Grant HL33940, from the Michigan Phoenix Foundation (D.L.G.), and from the University of Michigan Rackham Faculty Grant Program (D.L.G.).

1. Shih, D. M., Wall, R. J. & Shapiro, S. T. (1990) *Nucleic Acids Res.* **18,** 5465–5472.
2. Raich, N., Enver, T., Nakamoto, B., Josephson, B., Papayannopoulou, T. & Stamatoyannopoulos, G. (1990) *Science* **250,** 1147–1149.
3. Cao, S. H., Gutman, P. D., Dave, H. P. G. & Schlechter, A. N. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 5306–5309.
4. Raich, N., Papayannopoulou, T., Stamatoyannopoulos, G. & Enver, T. (1992) *Blood* **79,** 861–864.
5. Goodman, M., Koop, B. F., Czelusniak, J., Weiss, M. L. & Slightom, J. L. (1984) *J. Mol. Biol.* **180,** 803–823.
6. Gumucio, D. L., Heilstedt-Williamson, H., Gray, T. A., Tarle, S. A., Shelton, D. A., Tagle, D. A., Slightom, J. L., Goodman, M. & Collins, F. S. (1992) *Mol. Cell. Biol.* **12,** 4919–4929.
7. Gumucio, D. L., Blanchard-McQuate, K. L., Heilstedt-Williamson, H., Tagle, D. A., Gray, T. A., Tarle, S. A., Gragowski, L., Goodman, M., Slightom, J. L. & Collins, F. S. (1991) in *Proceedings of the 7th Conference on Hemoglobin Switching*, eds. Stamatoyannopoulos, G. & Nienhuis, A. (Johns Hopkins Univ. Press, Baltimore), pp. 277–289.
8. Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. & Jones, R. T. (1988) *J. Mol. Biol.* **203,** 439–455.
9. Bailey, W. J., Slightom, J. L. & Goodman, M. (1992) *Science* **256,** 86–89.
10. Shi, Y., Seto, E., Chang, L.-S. & Shenk, T. (1991) *Cell* **67,** 377–388.
11. Flanagan, J. R., Becker, K. D., Ennist, D. L., Gleason, S. L., Driggers, P. H., Levi, B.-Z., Appella, E. & Ozato, K. (1992) *Mol. Cell. Biol.* **12,** 38–44.
12. Park, K. & Atchison, M. L. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 9804–9808.
13. Hariharan, N., Kelley, D. E. & Perry, R. P. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 9799–9803.
14. Riggs, K. J., Merrell, K. T., Wilson, G. & Calame, K. (1991) *Mol. Cell. Biol.* **11,** 1765–1769.
15. Jane, S. M., Ney, P. A., Vanin, E. F., Gumucio, D. L. & Nienhuis, A. (1992) *EMBO J.* **11,** 2961–2968.
16. Jane, S. M., Gumucio, D. L., Ney, P. A., Cunningham, J. M. & Nienhuis, A. W. (1993) *Mol. Cell. Biol.*, in press.
17. Gumucio, D. L., Rood, K. L., Gray, T. A., Riordan, M. F., Sartor, C. I. & Collins, F. S. (1988) *Mol. Cell. Biol.* **8,** 5310–5322.
18. Gong, Q.-H., Stern, J. & Dean, A. (1991) *Mol. Cell. Biol.* **11,** 2558–2566.
19. Yu, C.-Y., Motamed, K., Chen, J., Bailey, A. D. & Shen, C.-K. (1991) *J. Biol. Chem.* **266,** 8907–8915.
20. Gumucio, D. L., Rood, K. L., Blanchard-McQuate, K. L., Gray, T. A., Saulino, A. & Collins, F. S. (1991) *Blood* **78,** 1853–1863.
21. Pevny, L., Simon, M. C., Robertson, E., Klein, W. H., Tsai, S.-F., D'Agati, V. D., Orkin, S. H. & Costantini, F. (1991) *Nature (London)* **349,** 257–260.
22. Martin, D. I. K., Tsai, S.-F. & Orkin, S. H. (1989) *Nature (London)* **338,** 435–438.