MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 5, No. 1, February, pp. 18–32, 1996
ARTICLE NO. 0004

Evolutionary Strategies for the Elucidation of *cis* and *trans* Factors That Regulate the Developmental Switching Programs of the β-like Globin Genes

DEBORAH L. GUMUCIO,* DAVID A. SHELTON,* WEI ZHU,* DAVID MILLINOFF,* TODD GRAY,†
JEFFREY H. BOCK,‡ JERRY L. SLIGHTOM,‡ AND MORRIS GOODMAN§

**Department of Anatomy and Cell Biology, University of Michigan Medical School, Ann Arbor, Michigan 48109-0616; †Biochemical Research Building 739, Case Western Reserve University, Cleveland, Ohio 44106-4955; ‡Molecular Biology Unit 7242, The Upjohn Company, Kalamazoo Michigan 49007; §Department of Anatomy and Cell Biology, Wayne State School of Medicine, Detroit, Michigan 48201*

Received July 12, 1995

We describe three strategies for the identification of specific *cis* and *trans* factors that regulate globin gene expression, all three of which are based on the evolution of the globin genes and their expression patterns. The first approach, phylogenetic footprinting, relies on a search for sequence similarities and is designed to elucidate the factors that control those expression patterns which are shared by orthologous globin genes of all eutherian mammals (e.g., the expression of the ε globin genes in the embryonic yolk sac and its repression in fetal and adult hematopoietic tissues). The second approach, differential phylogenetic footprinting, relies on a search for sequence differences. This approach may be of value in identifying the mechanisms underlying the generation of novel expression patterns in specific lineages (e.g., the expression of γ as a fetal gene in the simian primates in contrast with the embryonic expression of γ in all other mammals). Finally, motif-based phylogenetic analysis takes into consideration the fact that many transcription factors are quite flexible in the recognition of their cognate sites. The approach allows the detection of functionally conserved binding sites despite their sequence variation.

INTRODUCTION

For decades, hemoglobin has been a popular focus of scientific attention. An abundant protein comprising fully one third of the protein mass of a red cell, its easy isolation has made it a favorite model system for the investigation of protein biochemistry, protein structure, protein sequence, and protein evolution. Upon the advent of molecular cloning techniques, the globins were among the first cDNAs and genes to be isolated and sequenced. In the early 1980s, the sequence of 73 kb of the human β cluster, including its five active β-like globin genes were determined (Collins and Weissman, 1984). Sequences of the β globin clusters in additional mammals (rabbit, mouse, goat, and galago) followed, providing the basis for a molecular understanding of the structural evolution of the cluster. As sequences from additional species were gathered, the β-like globin locus became a valuable tool for the establishment of phylogenetic relationships.

Interest in another aspect of the globin genes, the regulation of their developmental stage-specific expression patterns, began with descriptions of the changing composition of the hemoglobin molecule during development as assessed by electrophoretic separation of the various globin chains (Bunn and Forget, 1986). It soon became clear that each globin gene is expressed in a restricted developmental stage and that these expression patterns are tightly regulated. Coalescence of expression studies on the RNA and protein level with the structural studies of the β-like globin locus soon established that the five active genes of the human β globin cluster are arranged on the chromosome in the order of their developmental expression (Fig. 1A). The most 5′ gene in the human cluster, ε, is expressed in early embryonic life when erythropoiesis is centered in the yolk sac. The movement of the site of erythropoiesis from the yolk sac to the fetal liver is accompanied by the repression of ε gene expression and the simultaneous activation of the paired and highly homologous γ genes. Expression of the γ-genes in the human fetal liver continues until shortly after birth, when a final movement in the site of erythropoiesis from the fetal liver to the bone marrow coincides with down regulation of the γ genes and activation of δ and β gene transcription. Because of the dramatic "switches" in gene activity that accompany development, the process by which the patterns of globin gene expression change during ontogeny have become known as "hemoglobin switching."

1055-7903/96 $18.00

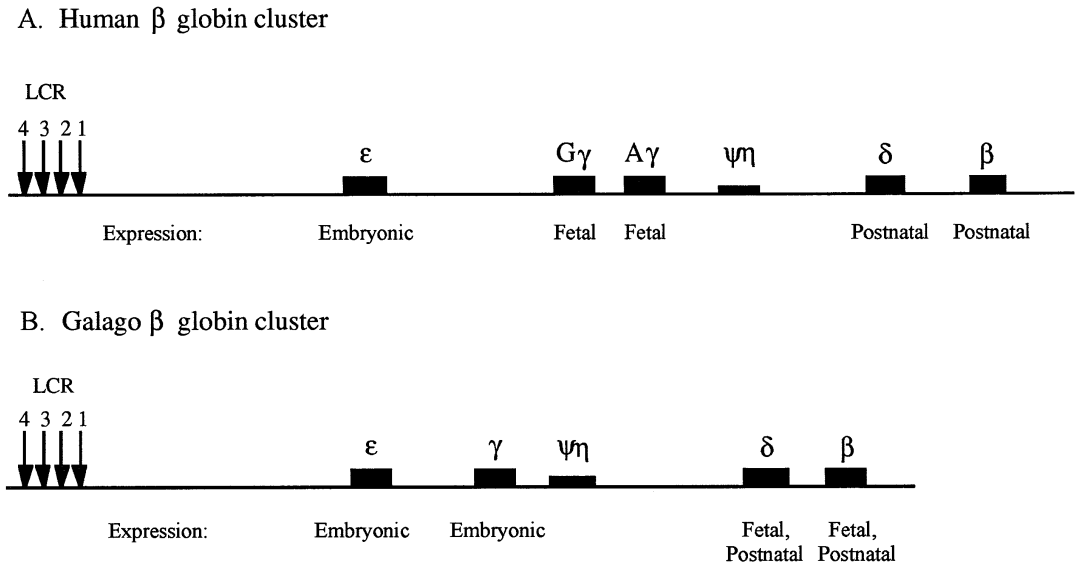

FIG. 1. Schematic representation of the human and galago β globin clusters. LCR, locus control region. Arrows refer to positions of four DNaseI hypersensitive sites. The time period in which each gene is expressed is indicated.

In the clinical arena, the molecular bases for severe hemoglobinopathies such as β thalassemia and sickle cell anemia have been uncovered as deletions or point mutations in the adult β globin gene. These genetic defects cause absence or dysfunction of the β gene's protein product. Interestingly, these anemias are less severe when they occur in *cis* or *trans* to other mutations in the γ gene promoters (Miller *et al.*, 1987; Safaya *et al.*, 1989). These γ mutations cause overexpression of the γ genes in adult life. That is, the mutations somehow interfere with the normal mechanisms that silence γ. This finding has given additional impetus to attempts to further understand the process of hemoglobin switching, particularly the factors that regulate the silencing of the γ and ε genes. Reactivation of these developmentally suppressed genes in patients with defective β genes could potentially cure these anemias.

Below, we summarize the structural evolution of the mammalian β globin locus, using this as a basis on which to trace the evolution of the developmental stage-specific expression patterns that characterize each of the genes of this locus. Attention will then be focused on the molecular mechanisms that regulate these stage-specific expression patterns in eutherian mammals. Three approaches that are of value for identification of the *cis* and *trans* factors that regulate hemoglobin switching will be presented. All three are based on the evolution of the globin genes and their expression patterns. Through the use of these strategies, the complex regulatory machinery that underlies hemoglobin switching can be more thoroughly probed. The knowledge gained may eventually lead to a therapeutic approach to the reactivation of the γ or ε genes in patients with severe hemoglobinopathies.

The mammalian β-like globin locus: structure and evolution. The first β-like globin gene arose by duplication of an ancestral globin gene during early vertebrate evolution approximately 450 million years ago (Czeluzniak *et al.*, 1982). This duplication produced the α and β globin genes. (The subsequent evolution of the α gene cluster will not be discussed here). Further duplication of the β gene about 180 MYA in the early mammals gave rise to an upstream locus (proto ε) and a downstream locus (proto β) (Goodman *et al.*, 1987; Koop and Goodman 1988). By the time of the last common ancestor of Metatheria (marsupials) and Eutheria (placental mammals), about 125 MYA, the proto ε and proto β loci differed from each other in both their coding and promoter sequences and also in their times of developmental expression (Koop and Goodman, 1988). The expression of proto ε was restricted to embryonic life while proto β was repressed until the end of embryonic life, when the developmental switch from proto ε to proto β expression occurred. This two-gene cluster has persisted in marsupials as demonstrated by findings on the American opossum (Koop and Goodman, 1988) and an Australian dasyruid marsupial (Cooper and Hope, 1993). It should be noted here that the avian β globin cluster also evolved via a series of gene duplications. However, the duplications that generated the β clusters of mammals and birds occurred independently in these two lineages (Goodman *et al.*, 1987; Hardison, 1981). Despite this, studies of globin switching in chicken and mammals seem to indicate that at least a part of the molecular machinery that regulates the switch from embryonic to postembryonic β-like globin expression had already evolved prior to the divergence of avian and mammalian lineages (300 MYA). This is discussed further below.

About 100 MYA, in the early eutherian mammals, tandem duplications of proto ε produced the ε, γ, and η loci, while a tandem duplication of proto-β led to the δ- and β loci (Goodman *et al.,* 1984; Hardison, 1984; Harris *et al.,* 1984). The extant globin clusters of all mammals bear some resemblance to this five-member ancestral cluster. However, in some lineages, certain of the globin genes have become dispensable: γ was deleted in artiodactyls; η was inactivated in early primates and deleted in rodents and lagomorphs; and δ became a pseudogene in rodents, lagomorphs, and artiodactyls (Goodman *et al.,* 1984; Harris *et al.,* 1984). Also in some lineages, further duplications have acted on individual genes (the γ gene was duplicated in primates; both the ε and the β genes were duplicated in mice) or on the entire locus (the locus is triplicated in the goat and duplicated in the cow).

A particularly interesting evolutionary history is exhibited by the γ gene. In artiodactyls, it was deleted (Goodman *et al.,* 1984). In other mammals, with the exception of the anthropoid primates, γ continued to function, like ε, as an embryonic gene (Tagle *et al.,* 1988; Hardison, 1981; Hill *et al.,* 1984). However, in the anthropoid primates, γ was recruited to a fetal expression pattern (Tagle *et al.,* 1988). Among other mammals, a specific fetal globin gene is only seen in the artiodactyls; in this lineage, however, a duplicated β gene and not the γ gene was coopted for fetal expression (Goodman *et al.,* 1984).

The exact sequence of events that led to the fetal recruitment of the γ-gene in the anthropoid primates is not known, but phylogenetic reconstructions suggest a possible scenario. A duplication of the γ gene occurred in the early simians about 45 MYA. This time frame places the duplication event after the separation of the prosimian primates (represented by the extant galago and lemur lineages) from the simian primates (Old World and New World monkeys, apes, and man), but prior to the divergence of platyrrhine and catarrhine primates. The γ duplication was mediated in misaligned chromatids by an unequal crossover between repetitive LINE elements that flanked the γ locus (Fitch *et al.,* 1991). It has been proposed that due to the increased redundancy of the γ genes, at least one of these genes (possibly the one most distant from the ε locus) collected changes that altered its function and developmental timing of expression from the embryonic to the fetal time period (Fitch *et al.,* 1991; Hayasaka *et al.,* 1992; Meireles *et al.,* 1995). Presumably, such changes were beneficial and were not only selected for, but transferred to the other γ gene by gene conversion.

It should be noted that this is only one possible scenario. The timing of this fetal recruitment event has not been precisely determined, but the fact that all simian species studied (both platyrrhine and catarrhine) express γ in fetal life suggests that fetal recruitment, like the γ duplication, occurred in the stem of the anthropoid primates. It is not known whether the fetal recruitment event preceded or postdated the γ duplication. It is clear, however, that in the altered genetic program for hemoglobin switching that emerged in simians, the γ genes were fully expressed in fetal life and repressed in postnatal life. In contrast, in prosimian primates such as the galago, the γ gene remained embryonic. It is of interest that the galago δ and β genes are activated early in fetal life, concomitant with the silencing of γ in that species. But in simians, the δ and β genes are delayed in their activation until late fetal or postnatal life, again in accord with γ silencing. This suggests a connection between γ silencing and β activation, a proposal that has now been experimentally verified (discussed below).

For purposes of further investigating the molecular mechanisms that underlie the fetal recruitment phenomenon, it is useful to identify representatives of the two most closely related primate lineages that exhibit differences in γ expression. We have chosen the human as a model simian primate (fetal γ) and the galago as a model prosimian (embryonic γ). The structure of the human and galago globin clusters and a summary of the expression patterns of each of the globin genes within both clusters is presented in Fig. 1.

Although the actual alterations in γ regulatory regions that were responsible for its fetal recruitment are not yet known, there is good molecular evidence for a burst of nucleotide substitutions occurring at the same time as the fetal transition. Rates of nonsynonymous and promoter substitutions in the γ genes accelerated approximately 40–45 MYA, then decelerated after the fetal pattern was well established (Goodman *et al.,* 1987; Fitch *et al.,* 1991; Hayasaka *et al.,* 1992), suggesting that selection favored the emergence of a distinct fetal hemoglobin in the simian primates. This acceleration in nucleotide changes is unique to the γ gene (the ε and β genes did not undergo a similar process) and unique to this particular window of evolutionary time. One or more of these anthropoid-specific promoter substitutions may have been responsible for the regulatory change that converted γ from an embryonic to a fetal expression pattern. Other nonsynonymous substitutions in the simian ancestral γ gene specified amino acid changes in regions of the protein that function to bind DPG. By drastically reducing the DPG-binding capacity of fetal hemoglobin, these changes ensured a favorable balance in the transport of oxygen from mother to fetus, making possible the prolonged intrauterine fetal life and extensive prenatal brain development which characterizes simian primates.

The regulation of hemoglobin switching. Despite years of study, many of the mechanistic details underlying the process of hemoglobin switching remain unclear. Early studies of human globin gene expression in transgenic mice revealed that *cis* elements located

near the γ and β genes provide some degree of tissue and stage-specific regulation of expression (Townes *et al.*, 1985; Chada *et al.*, 1986; Kollias *et al.*, 1986). Thus, human constructs containing these genes and only a few kilobases of 5′ flanking sequences are expressed in the proper tissue and in the proper developmental time period in the murine background. However, such constructs are expressed at extremely low levels compared to the endogenous mouse globin genes. In fact, constructs containing human ε genes and their immediate flanking sequences are not expressed at all in the mouse (Shih *et al.*, 1990; Raich *et al.*, 1990). It was later realized that, for proper high level expression of these human constructs, sequences located several kilobases upstream from ε are required. These unusual sequences, collectively referred to as the locus control region (LCR, Fig. 1), were first identified by virtue of their extreme sensitivity to DNase I (Tuan *et al.*, 1985; Forrester *et al.*, 1986). Recently, it has been shown that the same sequences contain powerful enhancers and binding sites for ubiquitous as well as erythroid-restricted factors (Ryan *et al.*, 1989; Curtin *et al.*, 1989; Fraser *et al.*, 1990; Ney *et al.*, 1990).

The LCR elements are important for the expression of each of the β-like globin genes. Studies of individuals with deletion forms of $\gamma\delta\beta^0$-thalassemia have shown that deletion of the LCR region results in failure to express any of the downstream globin genes even though all of these genes themselves are intact (Driscoll *et al.*, 1989). This finding confirms the enhancer function of the LCR. But further studies in transgenic mice suggest that the LCR effect is even more complicated. Constructs containing the LCR linked directly to the human β gene are expressed at high levels in the mouse, but proper developmental regulation is lost (Behringer *et al.*, 1990; Enver *et al.*, 1990). The powerful enhancers of the LCR appear to be dominant over the silencing mechanisms that normally repress the β gene in embryonic life. Likewise, linking the LCR directly to the human γ gene results in high level expression. In the case of the γ gene, however, some developmental regulation is retained (Behringer *et al.*, 1990; Enver *et al.*, 1990; Dillon and Grosveld, 1991). The variable degree to which the gene is stage-specifically silenced appears to be dependent upon the construct used as different laboratories report different results. Nevertheless, when constructs containing LCR sequences linked to both γ and β are tested, each gene is expressed only in its proper developmental time period (Behringer *et al.*, 1990; Enver *et al.*, 1990; Dillon and Grosveld, 1991). On this basis, it has been proposed that the γ promoter and the β promoter compete in *cis* for interaction with LCR sequences. In humans, a region of the γ promoter has been identified that may be important in the maintenance of preferential LCR contact with the γ gene during fetal life, thus ensuring β gene silence (Jane *et al.*, 1992). This region of the γ gene (−50) binds a protein named SSP (stage selector protein). The SSP protein has been recently cloned (Jane *et al.*, 1995) and shown to be homologous to a protein which mediates a similar *cis* competition in the chicken (NF-E4). SSP is therefore likely to be an important part of the switching machinery.

On the surface, SSP-mediated *cis* competition between γ and β seems to explain the observation that the β gene is activated earlier in those mammalian species in which the γ ortholog is silenced earlier. However, it is of interest to note that the galago γ gene does not have a binding site for SSP at the −50 region of its promoter (Jane *et al.*, 1992). Recently, numerous binding sites for SSP were detected in conserved regions of the ε gene (Gumucio *et al.*, 1993). While none of these sites has been functionally tested as yet, it is possible that in the nonanthropoids, it is the ε gene and not the γ gene which outcompetes the β gene for LCR contact in early ontogeny and which therefore controls the timing of β activation.

Thus, at least three types of regulation seem to be important in the stage-specific silencing process for the γ and β genes: repressors/activators which bind near the genes, enhancers/chromatin modifying elements in the LCR, and competition effects. Other studies have revealed that gene order or distance from the LCR may also affect the expression of these genes (Hanscombe *et al.*, 1991). With respect to the β gene, then, any of the following models may apply: (a) the β gene does not need an active silencing mechanism because it cannot engage the LCR in early developmental time periods since it is outcompeted by the γ gene; (b) in addition to the competitive mechanism, the β gene may require active silencing in the stages in which it is not expressed; (c) an as yet unidentified stage-specific transactivator may be important in the initiation of β gene expression after birth. The absence (or inactivity) of such a transactivator in earlier stages may account for the inactivity of the β gene in those stages.

For the ε gene, developmental control appears to be more straightforward. Constructs containing LCR sequences linked to the ε gene exhibit ε expression at high levels in transgenic mice; silencing of the human ε gene at the end of the embryonic period in the developing mouse is complete and does not require the presence of the γ and/or β gene (Shih *et al.*, 1990; Raich *et al.*, 1990). This suggests that *cis* sequences located near the ε gene direct its stage-specific down regulation. Furthermore, these silencers are dominant over the enhancing effect of the LCR. Candidate silencer elements have been detected in the ε promoter by transfection studies. A region between −177 and −392 bp (with respect to the ε cap site) represses reporter gene activity by 3-fold in K562 cells and by 10-fold in HeLa cells (Cao *et al.*, 1989). In transgenic mice, deletion of these sequences results in persistence of human ε transgene expression in definitive mouse erythrocytes (Raich *et al.*, 1992).

However, since a considerable degree of downregulation of this mutant ε transgene still occurs, the silencer region must be necessary but not sufficient for complete stage-specific repression of ε. Therefore, although ε regulation would appear to present a more simple problem than control of γ or β expression, multiple *cis* and *trans* factors appear to be responsible for its autonomous suppression.

The γ genes are intermediate to the ε and β genes, with respect to position on the chromosome, stage of expression, and mechanism of silencing. Experiments in transgenic mice have shown that competition with the β gene may occur, and depending upon how the test construct is designed, the γ gene may be fully dependent upon the β gene for proper silencing (that is, inappropriately expressed in adult stages in the absence of the β gene; see Enver *et al.*, 1990; Behringer *et al.*, 1990) or completely autonomous (that is, silenced efficiently by repressors that bind in this vicinity of the γ gene as for the ε gene; see Dillon and Grosveld, 1991). Several regions of the γ gene have been shown to be important in determing the outcome of γ silencing, including the proximal promoter (Miller *et al.*, 1987; Safaya *et al.*, 1989; Chada *et al.*, 1986; Perez-Stable and Costantini, 1990), distal upstream sequences (Stamatoyannopoulos *et al.*, 1993), regions 3′ to the gene (Lloyd *et al.*, 1992; Stamatoyannopoulos *et al.*, 1993; Behringer *et al.*, 1990; Enver *et al.*, 1990; Dillon and Grosveld, 1991), and portions of the LCR (Fraser *et al.*, 1993; Li and Stamatoyannopoulos, 1994). It therefore seems likely that silencing the γ gene requires several control points.

The complexity and apparent redundancy of the regulatory machinery for hemoglobin switching has impeded the elucidation of key control elements. However, since the globin genes of all eutherian mammals undergo switches in gene activity during development and since the globin clusters of all mammals were derived from the same five-member cluster, it is possible that significant clues as to the *cis* elements important to hemoglobin switching could be derived by evolutionary approaches. Identification of these *cis* elements would immediately facilitate the elucidation of their cognate binding factors. Here, we describe three strategies for the elucidation of key regulatory motifs, all of which are based on evolutionary considerations: phylogenetic footprinting, differential phylogenetic footprinting, and motif-based phylogenetic analysis. These three approaches represent valuable additions to the collection of strategic tools that may eventually help to unlock the secrets of hemoglobin switching.

MATERIALS AND METHODS

Phylogenetic footprinting analysis. This strategy is designed to identify conserved *cis* elements that could be important for the regulation of those expression patterns that are shared among orthologous globin genes of all mammalian species. Alignments of orthologous regions of the globin cluster using sequences from several mammalian species are produced as described (Bailey *et al.*, 1992). Within the alignments, regions are sought in which the sequence is identical in all species used in the alignment over six contiguous base pairs; these regions are called phylogenetic footprints. Although arbitrary, the use of the 6-bp criterion has been shown to successfully identify *trans* factor binding sites over 95% of the time (Gumucio *et al.*, 1992, 1993). Double-stranded oligonuceotide probes are synthesized spanning each phylogenetic footprint. The probes (30-mers) contain a *Bam*HI overhang on the 5′ end of the antisense strand to permit labeling by the Klenow reaction. Methods for the production of nuclear extracts and the performance of gel mobility shifts have been described (Gumucio *et al.*, 1988). The identity of the factors binding to each of the probes is established by several of the following criteria: (1) cell type specificity of binding, (2) comigration of the complex with a known protein complex, (3) successful competition of the complex with probes containing binding sites for known factors and not with the mutant counterparts of such probes, (4) limited protease digestion patterns of the complex compared to similarly digested known complexes, (5) antibody supershifts.

Differential phylogenetic footprinting. This analysis is based on the observation that the galago γ gene, like the γ genes of most other mammals, is expressed in the embryonic time period. In contrast, the human γ gene, as well as the γ genes of the rest of the anthropoid primates, is expressed in fetal life. The differential phylogenetic strategy starts with the assumption that this fetal recruitment of γ results from one or more *cis* changes that altered the *trans* factor binding pattern and that this, in turn, altered the γ expression pattern. The analysis is carried out by aligning orthologous sequences of the galago and the human γ globin genes and searching for sequence differences. Double-stranded oligonucleotide probes (30-mers) are synthesized which span the sequence differences. Two probes are made, one corresponding to the human sequence and one corresponding to the galago sequence. Gel mobility shifts are used to detect differences in the pattern of proteins binding to the two probes. Where differences are found, their functional consequences are tested in expression assays. The protocols for extract preparation, gel shift analysis, and transfection assays have been described (Gumucio *et al.*, 1988).

Motif-based phylogenetic analysis. It is clear that most transcription factors tolerate some degree of sequence variation in their binding sites. For example, Sp1 can bind with high affinity to GGCGGGG as well as to GGTGGGG (Letovsky and Dynan, 1989). Such sequence variation can destroy a phylogenetic footprint

without affecting functional binding capability. Motif-based phylogenetic analysis was developed as a means to detect conserved binding sites which show sequence variation. The strategy begins with the use of PCR-assisted binding site selection to characterize the sequence variations tolerated by a given binding protein. We have previously described the application of this method to the YY1 protein (Yant *et al.,* 1995). In this report, we have also applied the technique to the GATA-1 protein, an erythroid-specific protein which is known to be important in the expression of the globin genes.

A GATA-1–GST fusion protein was produced by ligation of a full-length mouse GATA-1 cDNA (Tsai *et al.*, 1989), kindly provided by Dr. David Martin (Fred Hutchinson Cancer Center, Seattle, WA) into a GST expression vector, pGSTag (Ron and Dressler, 1992). The fusion protein was expressed in *Escherichia coli*, strain DH5α, and harvested after IPTG induction by freeze–thaw lysis of the cells. Cleared lysate was incubated with glutathione–Sepharose resin (Pharmacia, Piscataway, NJ) and the derivatized, washed resin was used directly for binding.

A library of single-stranded oligonucleotides was synthesized containing 18 bp of known sequence flanking 24 random positions: 5′AACGGTCCCTGGCTAAA $C(N)_{24}$CAGTGTGTGGACTATTAG. To produce a double-stranded population, a primer complementary to the most 3′ 18 nucleotides was annealed and extended exactly as described previously (Yant *et al.*, 1995). Derivatized resin was incubated with 100 pmoles of the oligonucleotide library, and binding was carried out with continual rotation for 1 h at room temperature. The GATA-1–resin:DNA complexes were washed with binding buffer, the resin was pelleted, the supernatant was discarded, and the resin resuspended in amplification mix [2.5 m*M* $MgCl_2$, 200 μ*M* each dNTA, 1 μ*M* each primer, 1.25 U *Taq* polymerase in 1× *Taq* buffer] and overlaid with mineral oil. Amplification was performed as 94°C for 1 min, 60°C for 30 s, repeated for 30 cycles. The resultant PCR products were used in another round of binding and amplification. Four rounds of GATA-1 selection and PCR amplification were carried out before the resulting oligonuceotides were cloned. Individual clones were tested in a gel shift competition assay and only those that competed efficiently for binding of GATA-1 to a well-characterized GATA-1 binding site were sequenced.

It should be noted here that the PCR-assisted binding site optimization procedure was applied previously to the GATA-1 protein by three different groups (Ko and Engel, 1993; Merika and Orkin, 1993; Whyatt *et al.*, 1993). However, in none of those studies were the clones screened for GATA-1 binding prior to sequencing. A few were, however, tested for GATA-1 binding subsequent to sequencing. Where it was possible to differentiate, only those sequences that were shown to bind GATA-1 with high affinity were used in the analysis described here.

Once the resulting sequences were aligned and a consensus was derived, this consensus was used to search the human globin cluster for motifs that are likely to represent high-affinity GATA-1 binding sites. Each of these putative GATA-1 motifs was then examined at its orthologous position in the galago sequence. Motifs in the galago sequence that fit the defined GATA-1 binding consensus established by the PCR analysis were designated "conserved" even if the sequence of the galago motif did not exactly match the sequence of the corresponding human motif.

RESULTS AND DISCUSSION

The search for sequence identity: phylogenetic footprinting as a tool to identify cis *and* trans *factors important for anciently conserved regulatory properties.* The existence of a large amount of sequence information for the globin genes of a number of mammalian species has provided a unique opportunity to utilize sequence alignments in order to identify highly conserved sequence motifs that could control conserved regulatory properties. Within these multispecies alignments, motifs that show 100% conservation over six or more contiguous bases are tagged as phylogenetic footprints. In order to reveal highly conserved motifs, some of the species represented in the alignment should be from different primate suborders (i.e., have a last common ancestor 55–65 MYA) and at least one species should be from a different eutherian order (e.g., Lagomorpha). All together, the branches of the phylogenetic tree uniting the species to their ancestral nodes should represent more than 200 MY of evolutionary time.

The phylogenetic footprinting strategy has been applied to the ε upstream region from −1 to −2000 bp (Gumucio *et al.*, 1993) and to the entire intergenic region between ε and γ (Gumucio *et al.*, 1992). The results of these analyses are depicted in Fig. 2A (for ε) and Fig. 2B (for γ). In the case of the ε upstream region, 21 phylogenetic footprints were found. Double-stranded oligonucleotide probes spanning these 21 footprints were used in gel-shift assays to identify the *trans* factors that bind within and nearby these conserved motifs. Multiple binding interactions were identified, including redundant sites for: the erythroid-specific activator GATA-1; the ubiquitously distributed activator Oct-1; the stage selector protein SSP; and a protein with demonstrated repressor activity, YY1 (discussed further below). In addition, multiple sites for a novel protein, CSBP-2 (conserved sequence binding protein-2) were detected. The role of this latter protein in globin gene expression is still under investigation.

As summarized in Fig. 2, the γ gene yielded far fewer phylogenetic footprints (12) than the ε gene (21) despite the fact that a larger sequence region was analyzed.

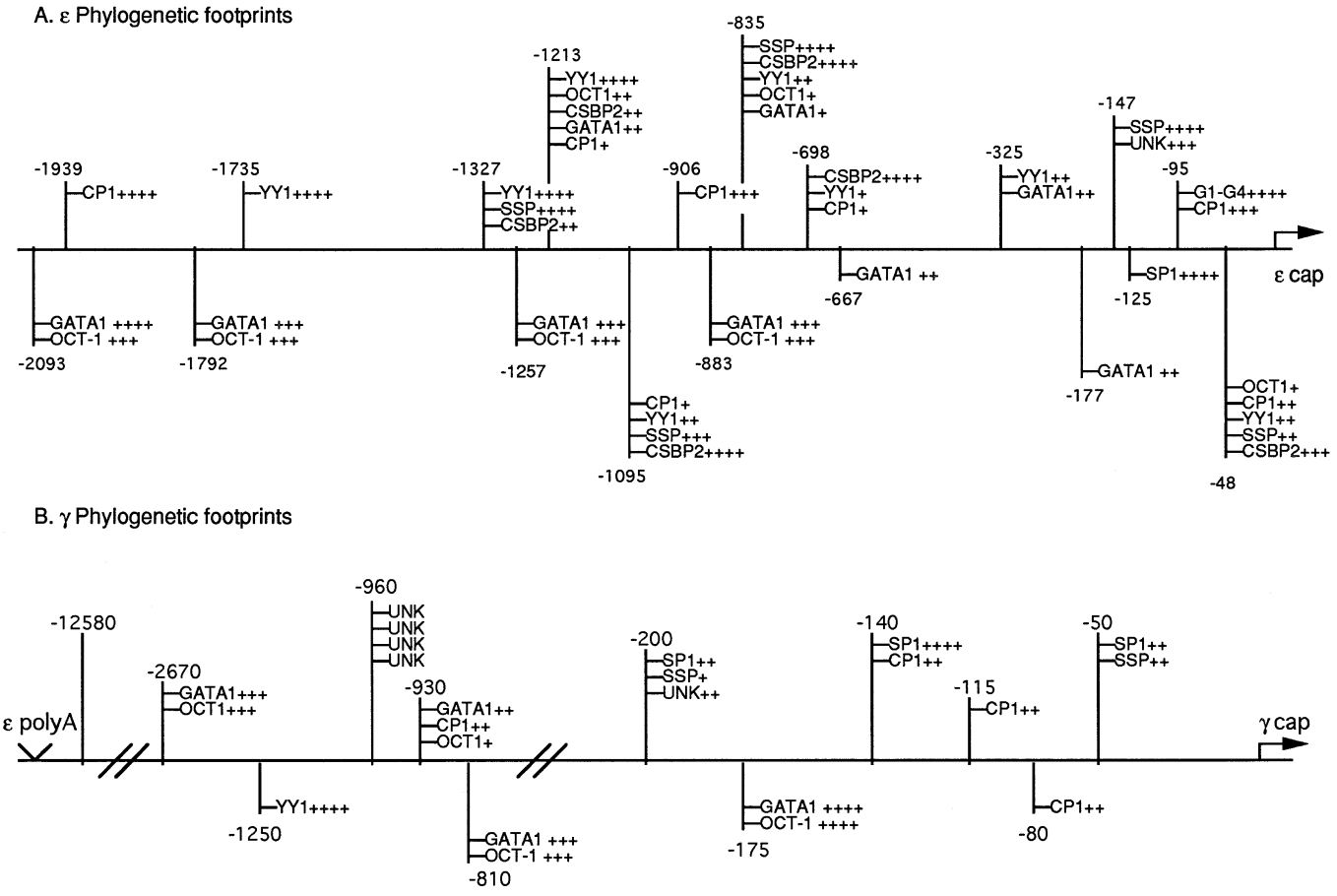

FIG. 2. Summary of phylogenetic footprinting results for the ε and γ globin genes. The relative affinity of binding of each factor is given as + (low affinity) to ++++ (high affinity). The positions of the probes used in both regions are drawn approximately to scale except for two regions of the γ promoter at which breaks in the scale are indicated by double diagonal lines. Additional descriptions for the factors are given in the original references (Gumucio *et al.*, 1992, 1993).

This is consistent with the observation that the pattern of regulation of ε genes among all mammalian species is similar (expressed in the early embryo and silenced in early or mid fetal life) while the pattern of γ gene regulation varies (γ is embryonic in nonprimates and in nonanthropoid primates, but fetal in anthropoid primates). Although several of the specific proteins identified in the analysis of the γ gene were similar to those found in the analysis of the ε gene, several significant differences were noted. In particular, no CSBP-2 sites were identified in γ gene phylogenetic footprints; the ε region had more YY1 binding sites (10 as opposed to 1 for γ) and more SSP binding sites 5 vs 1 for γ), while the γ region had more Sp1 binding sites (3 as opposed to 1 for ε).

An exciting outcome of the phylogenetic footprinting studies applied to date is the finding of multiple YY1 binding sites upstream from the ε and γ genes. A total of 10 binding sites for this factor were identified within the 2 kb upstream from the ε gene, and an additional high-affinity YY1 binding site was located at −1086 upstream from each of the γ genes. Recent preliminary results indicate that several more high-affinity YY1 binding sites are located in phylogenetic footprints found within the cores of hypersensitive sites 2 and 3 of the LCR (D.L.G., unpublished observations). Counting these LCR sites, a total of 23 YY1 binding sites have now been detected at evolutionarily conserved regions of the globin cluster. The large number of YY1 sites is intriguing in light of the function of the YY1 protein in other contexts. YY1 was cloned simultaneously in four different laboratories and given four different names: YY1 (Shi *et al.*, 1991) NF-E1 (Park and Atchison, 1991), δ (Hariharan *et al.*, 1991), and UCRBP (Flanagan *et al.*, 1992). The name YY1, for yin and yang 1, has been most commonly adopted since it reflects the dual nature of the protein. This ubiquitously distributed transcription factor has been shown to function in various systems either as an activator or as a repressor. Particularly pertinent to the globin regulation problem is the fact that YY1 can mediate a switch between activation and repression. Such a switch has been seen in three

separate systems and always depends upon the interaction of YY1 with a second protein. Indeed, YY1 has been shown to interact directly with Sp1 (Seto *et al.*, 1993; Lee *et al.*, 1993), E1A (Shi *et al.*, 1991), and c-*myc* (Shivastava *et al.*, 1994). In addition, YY1 has been implicated in stage-specific enhancer activity in the Igκ gene locus (Park and Atchison, 1991). Finally, YY1 can bend DNA (Natesan and Gilman, 1993). This could be extremely important in the globin cluster where it has been postulated that the LCR is folded into a type of holodomain that is capable of interacting via different faces with the promoters of each of the active genes during development (Fraser *et al.*, 1993). All of the diverse properties of YY1 make this protein an extremely interesting factor in the context of globin gene regulation.

Recent functional studies have implicated YY1 in the silencing of both the ε and γ genes. A single high-affinity YY1 site has been detected at −1086 upstream from each of the γ genes. This site, which is conserved in anthropoid and nonanthropoid primates and in rabbit, mediates repression of reporter gene activity in transient transfection assays (Gumucio *et al.*, 1992). For ε, a silencer region that mediates partial, but not complete, suppression of ε activity was detected in transfection assays (Cao *et al.*, 1989). Binding studies done in our laboratory (Gumucio *et al.*, 1992, 1993) and in the Noguchi laboratory (Peters *et al.*, 1993) have identified several binding sites for the YY1 protein within this silencer region. One of the four YY1 binding sites that have been identified in this region is evolutionarily conserved. Deletion of the ε silencer was shown to partially interfere with normal stage-specific ε silencing in transgenic mice (Raich *et al.*, 1992). The failure of this deletion to completely reverse ε silencing may be due to the fact that the deletion also removed activators that are important for full ε promoter strength (several GATA-1 sites are found within this region). Alternatively, this could indicate that the silencer carries only a portion of the silencing signal.

In a more recent report, point mutations in both the YY1 and GATA-1 motifs implicate both of these proteins in ε silencing in transgenic mice (Raich *et al.*, 1995). Again, however, the data suggest that additional silencing regions must exist because significant downregulation of ε is still observed. In accord with this interpretation is the fact that several additional high-affinity YY1 binding sites were detected by phylogenetic footprinting at conserved motifs upstream from the ε silencer (see Fig. 2A). The possibility that all of these sites (and/or additional conserved YY1 binding motifs recently detected by the motif-based phylogenetic footprinting approach; see below) are together responsible for ε silencing can now be directly tested.

Data from the phylogenetic footprinting studies completed to date suggest that globin gene regulation is the result of the redundant use of a relatively restricted set of transcription factors. It is likely that both gene activation and repression involve many points of control that coordinately result in the appropriate expression level. Indeed, recent careful dissection of the ε upstream region by deletion mutagenesis and analysis in transfection assays indicates that several regions exhibit different functions: activators and repressors are interspersed over the 900 bp analyzed (Trepicchio *et al.*, 1993). If the concentration of nuclear factor binding sites shown in Fig. 2 is any indication of the true distribution of important regulatory elements, then it is highly likely that the deletion mutagenesis approach has underestimated the number of functional motifs. For the γ gene, deletion analysis has not identified a single *cis* sequence that has a dominant effect on γ gene silencing (Perez-Stable and Costantini, 1990; Stamatoyannopolous *et al.*, 1993). The phylogenetic footprinting approach, in contrast, provides a rich source of binding data on which to base functional hypotheses. The involvement of the conserved binding interactions detected by phylogenetic footprinting in globin gene regulation *in vivo* will be best tested by specific mutagenesis of several *trans* factor binding sites in the context of intact constructs that are known to be correctly regulated during development in transgenic mice.

The search for sequence differences: differential phylogenetic footprinting as a means to identify cis *and* trans *factors that account for the variations in globin gene regulation characteristic of specific mammalian species.* While the phylogenetic footprinting approach provides a tool to detect anciently conserved *cis* elements, differential phylogenetic footprinting provides a strategy to identify sequence differences that are associated with changes in the pattern of developmental expression of individual genes within the globin cluster. Differential phylogenetic footprinting stems from our discovery that the human and galago γ genes, which are separated by approximately 65 MY of evolution, exhibit different developmental expression patterns (Fig. 1). To determine which *cis* changes might account for this expression change, human and galago γ gene promoter sequences were aligned and sequence differences were noted. Binding and expression assays were then carried out to determine the functional consequences of these sequence differences (Gumucio *et al.*, 1994).

Sequence changes that caused alterations in the patterns of proteins binding to human and galago promoter sequences were found at −50 (a GC-rich region), −88 (proximal CCAAT box), −140 (CACCC region), and −175 (GATA motif). Most interesting was the proximal CCAAT box region. Here, comparison of the binding profiles revealed several proteins that bind avidly to the galago sequence but poorly if at all to the corresponding human sequence. In expression assays carried out in human erythroleukemia cells that express a predominantly fetal program, the binding of these proteins (G1, G2, G3, and G4) was associated with re-

pression of promoter strength. To further assess whether the binding of these proteins was correlated with an embryonic expression pattern, two additional analyses were performed. First, binding of these proteins to the rabbit and mouse γ orthologues as well as an additional prosimian primate was confirmed (all of these species also exhibit an embryonic γ expression pattern). Second, phylogenetic reconstructions were used to estimate the sequence of the primate ancestor prior to the simian/prosimian split as well as the simian ancestor prior to the catarrhine/platyrrine split. When these reconstructed sequences were tested in binding assays, it was clear that reduced affinity of G1 and G2 binding correlated well with the emergence of fetal γ expression: the reconstructed primate ancestral sequence (as well as the extant galago, lemur, rabbit, and mouse) was bound by the G proteins with high to moderate affinity while the simian ancestral sequence (as well as extant simian primate sequences) showed severely reduced binding (Fig. 3).

Although the G proteins do not bind to the human γ gene, binding sites for these proteins have been found in other potentially important regions of the human β globin cluster: several evolutionarily conserved motifs near hypersensitive site 3 of the LCR bind G1 and G2; and binding sites for all four proteins are located near the CCAAT box of the human ε gene (Gumucio *et al.*, 1994). These findings, coupled with the apparent repressor function of the γ proteins, suggest that these proteins may play a role in the regulation of the extant human globin genes. In addition, evolutionary loss of the binding sites for these proteins in the γ gene may have contributed to the recruitment of the γ gene to the fetal time period.

Motif-based phylogenetic footprinting: detecting conserved binding motifs in the context of variable sequence. A final evolutionary strategy for the elucidation of key regulators of hemoglobin switching relies on the detection of conserved binding motifs by virtue of their binding function rather than their sequence conservation per se. This strategy was developed because it was recognized that the phylogenetic footprint criterion, requiring six contiguous base pairs of conserved sequence, was extremely strict considering the well known ability of transcription factors to recognize their cognate sites with a certain degree of flexibility. A modification of the phylogenetic footprinting criterion, e.g., to accept one variable nucleotide in every six contiguous positions, might help to recover some of the conserved motifs overlooked by the stricter definition, but would also result in a much larger group of phylogenetic footprints to examine. The strategy outlined here can successfully identify conserved binding sites regardless of the level of sequence conservation.

Motif-based phylogenetic footprinting begins with definition of the sequence variation tolerated by a given transcription factor within the context of high-affinity binding. This is often done using PCR-assisted binding site optimization. In this procedure, the binding protein is used to select, from a family of random oligonucleotides, those sequences which represent high-affinity binding sites. An important further requirement of the procedure, as applied to the motif-based phylogenetic footprinting technique, is that each of the selected sequences be tested to ensure that it indeed contains a high-affinity binding site for the protein used in the selection. This is necessary because, despite repeated selection and amplification, a portion of the final cloned sequences in fact bind weakly or not at all (Merika and Orkin, 1993; Whyatt *et al.*, 1993). Affinity testing ensures that the resultant consensus derived from the analysis is built from avid binding sites and therefore enhances the ability of the consensus to predict high-affinity sites in any target sequence. Once the consensus is derived, it is used to search the human globin cluster, resulting in the identification of a large number of putative high-affinity binding sites. Each of these potential binding motifs is then examined in the context of the galago β globin cluster. Those galago motifs that still fit the consensus established by the PCR selection are designated "conserved."

In a recent study, motif-based phylogenetic analysis was applied to the YY1 transcription factor, a protein with possible repressor activity (Yant *et al.*, 1995). Here, we describe the application of this strategy to GATA-1. The importance of this factor to the establishment of definitive erythropoiesis and to the transcription of the individual globin genes has been well established. However, while a few of the individual GATA-1 binding sites within promoter regions or within the LCR have been functionally defined, our understanding of the key GATA-1 binding sites within the β globin cluster is far from complete. This approach provides a strategy for the identification of potentially important sites at any position within the globin cluster that can be aligned between human and galago.

The sequence characteristics of GATA-1 binding sites were previously determined in three separate studies (Ko and Engel, 1993; Merika and Orkin, 1993; Whyatt *et al.*, 1993). Merika and Orkin sequenced 53 selected oligonucleotides and established the consensus, SVNGATDGBB (S = G or C; V = A, C, or G; N = any base; D = A, G, or T; B = C, G, or T). However, only 7 of these were subsequently tested in binding assays; 1 of the 7 was bound very weakly by GATA-1. Ko and Engel sequenced 47 selected oligonucleotides; in this study, the central GAT motif was fixed with random sequence on both sides. Only 4 of the selected oligonucleotides were tested for GATA-1 binding; all were found to bind with high affinity. The consensus established in this study was AGATAA. Finally, Whyatt *et al.*, sequenced 83 oligonucleotides and subsequently tested 37 of these in gel-shift assays. Of those tested,

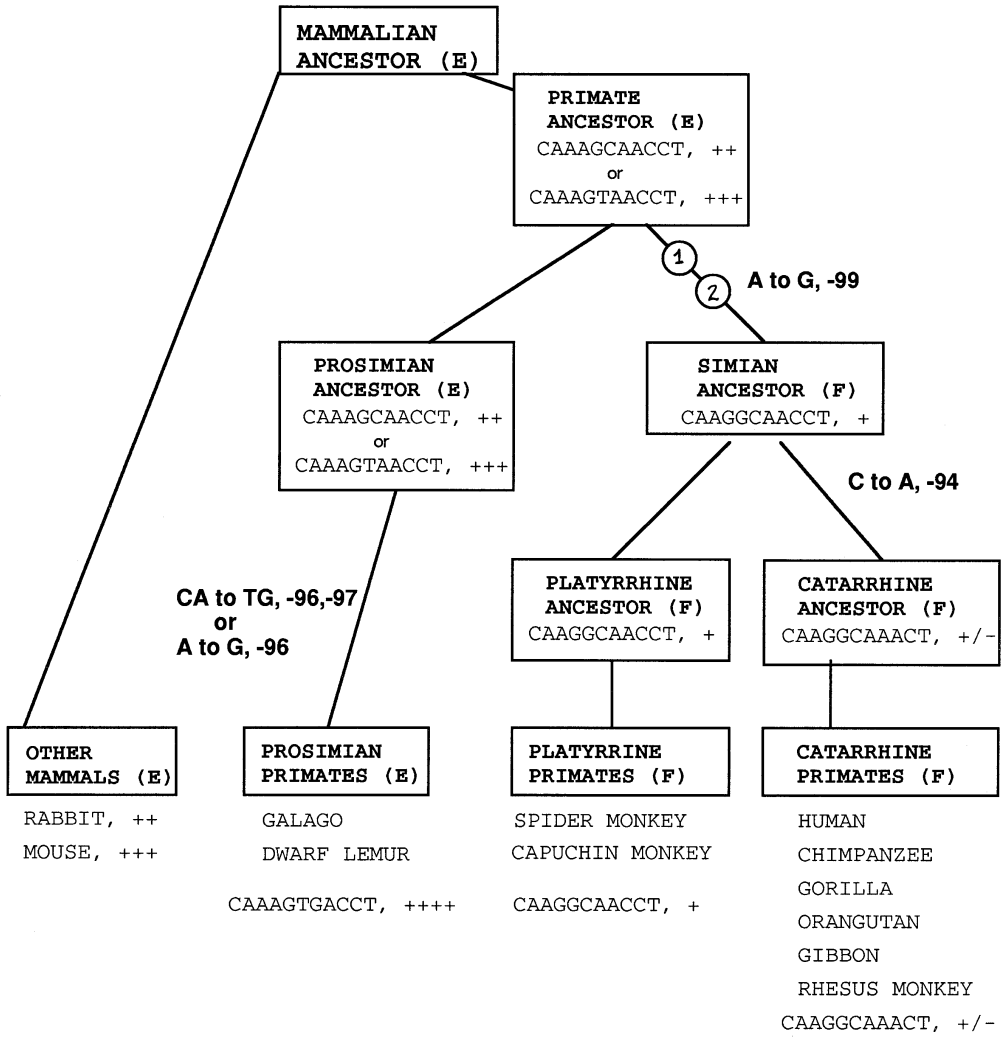

FIG. 3. Evolution of G protein binding. Sequence changes that occurred in the −88 region during primate evolution are correlated with affinity of G protein binding to extant sequences and reconstructed ancestral sequences. The reconstructed primate ancestral sequence has two equally possible solutions due to an ambiguity at position −97 (galago and dwarf lemur γ genes share T in this position, while brown lemur carries a C). Since this ambiguity cannot be resolved without additional sequence data from other nonprimates and prosimians, both possible solutions were tested. Both reconstructed primate ancestral sequences bind G1 and G2 with moderate affinity. Evolution of the prosimian ancestor proceeded with no additional changes in this sequence. Subsequent evolution of the extant prosimians was characterized by one to two base changes (CA to TG or simply A to G, depending upon which of the primate ancestral sequences is correct). These changes increased the binding affinity of all four G proteins to the extant galago and dwarf lemur γ genes. Simian evolution was initially characterized by a duplication of the γ gene (indicated by the circle labeled 1), an event known to occur early in the stem anthropoids. The recruitment of the γ gene to a fetal expression pattern is indicated by the circle labeled 2. However, as discussed in the text, it is not presently clear whether fetal recruitment actually preceded or postdated the γ duplication. Among the burst of anthropoid-specific sequence changes that are known to have occurred prior to the platyrrhine/catarrhine split, one (G to A at −99) is located within the cluster of four bases known to be necessary and sufficient for G protein binding (Gumucio *et al.*, 1994). This change may be primarily responsible for the four- to sixfold decrease in G1/G2 binding affinity to the simian ancestral sequence compared to the primate ancestral sequence. A second base change, C to A at −94, characterizes catarrhine evolution. This change further decreased G protein binding affinity since the human sequence, which is identical to that of the simian ancestor except for this base, binds the G proteins less avidly than does the simian ancestral sequence. The expression stage of the γ genes shown is indicated by F (fetal) or E (embryonic). This figure was reproduced with permission from an earlier publication (Gumucio *et al.*, 1994).

14 (38%) bound GATA-1 with high affinity, 6 (16%) bound very weakly, and 8 (22%) did not bind GATA-1. Only a portion of the sequenced oligonucleotides were used to produce compiled consensus. Two consensus sequences were derived: one surrounding a canonical GATA core (compiled from 31 sequences), AGATAG GGG; and one surrounding a GATT core (compiled from 12 sequences), ANGNNGATTWNNG (W = A or T).

In Fig. 4 (top), we have compiled the selected sequences from the three studies above. In the case of the Whyatt *et al.*, study, only the 14 sequences that were proven to bind GATA-1 with high affinity were used.

Merica,et al.	Merica,cont.	Ko, et al.	Ko cont.	This report
GT GAT AG	TG CAT CG	-A GAT AA	TA GAT AA	AA GAT AG
GA GAT AG	GC GAT GG	AC GAT AA	TT GAT GT	GA GAT AA
AT GAT AG	-C GAT CC	CA GAT AA	GC GAT AA	AA GAT AT
GA GAT AA	GC GAT TG	CT GAT AA	TA GAT AA	AA GAT AA
GG GAT AG	CG GAT GC	TT GAT AA	AG GAT TA	GA GAT AA
CC GAT AC	CG GAT GT	AA GAT AA	GA GAT AA	AC GAT AG
GG GAT AC	GA GAT CT	CA GAT AA	TA GAT AA	AA GAT AG
GC GAT AG	CG GAT GG	GT GAT AA	CA GAT AA	TA GAT AG
CC GAT AC	GT GAC AG	CC GAT AA	TA GAT CT	TA GAT AC
AT GAT AC	AA GTT AA	AT GAT AA	CT GAT CT	AG GAT AA
GC GAT AG	AT GAT TT	AA GAT AA	AT GAT TA	AA GAT AA
CA GAT AC	CA GAT GC	CA GAT AA	AA GAT AA	CC GAT AC
CA GAT TT	-A GAT GT	CA GAT AT	AA GAT AA	GA GAT AA
CT GAT GG	AT GTT AA	CA GAT AA		GC GAT AA
GT GAT TG	AG GAT TG	TA GAT CT		AG GAT AA
GA GAT GG	CT GAT GT	CA GAT AT	Whyatt, et al.	AA GAT AA
GT GAT CG	TG GAT TT	AA GAT AA	TG GAT AG	AT GAT AA
GT GAT TG	-T GCT AA	CT GAT AA	GT GAT AA	AT GAT AG
GT GAT GA	GG GAT CT	GA GAT AA	AT GAT AG	CA GAT AA
CT GAG GA		TT GAT AA	AA GAT AA	TA GAT AA
AA GAT GG		CA GAT AA	AT GAT AG	AC GAT AA
GG GAT TT		CT GAT GT	CA GAT AC	CC GAT AT
AC GAT CG		CA GAT CT	AC GAT AT	AC GAT AA
GC GAT TG		CT GAT AA	AA GAT AA	
-A GAT GC		AA GAT TT	GA GAT TA	
-A GAT GC		GA GAT AA	TG GAT TG	
-C GAT TC		CA GAT CT	AA GAT TA	
TT GAT TC		CC GAT AA	AC GAT TC	
AC GAT GG		AT GAT TA	CT GAT AG	
TG GAT GG		CA GAT AA		
TT GAT GC		AA GAT AA		
GC GAT GT		TA GAT CA		
AT GAG AG		TT GAT CT		

Compiled Base Frequency

G	31	16	134	0	1	20	34
A	42	57	0	132	0	83	61
T	21	36	0	2	133	19	23
C	34	24	1	1	1	13	17
Total	128	133	135	135	135	135	135

Percent Representation

G	**24**	12	**99**	0	1	15	**25**
A	**33**	**43**	0	**98**	0	**61**	**45**
T	16	**27**	0	1	**98**	14	17
C	**27**	18	1	1	1	10	13
Consenus:	**V**	**W**	**G**	**A**	**T**	**A**	**R**

FIG. 4. Compilation of GATA-1 motifs identified by PCR-assisted binding site selection. The data are from four sources as indicated (Merika and Orkin, 1993; Ko and Engel, 1993; Whyatt *et al.*, 1993; this report). The frequency with which each nucleotide occupies each position is tallied at bottom. The percentage representation of each nucleotide at each position is also given. The nuceotides seen most often are indicated in bold. The consensus was determined as bases occupying a given position >60% of the time.

We selected several additional sequences using bacterially expressed mouse GATA-1–GST fusion protein and carrying out the selection directly on the glutathione resin (see Materials and Methods). The selected sequences were first tested for GATA-1 binding before sequencing.

Figure 4 (bottom) represents the consensus derived from compilation of all 135 of the sequences shown in Fig. 4 (top), VWGATAR (R = A or G). This consensus was used to search the human globin cluster, yielding 163 matches. Each of these was then examined in the galago sequence. The galago was chosen for this analysis because the sequence of the entire galago β globin cluster including the LCR region has been determined (J.L.S. and J.H.B., unpublished data: Tagle *et al.*, 1992). Moreover, the galago and human lineages diverged 55–65 MYA, long enough to allow significant sequence variation at unconstrained regions, but recently enough to allow alignment of major portions of the cluster, as illustrated in the matrix plot in Fig. 5. The matrix plot also reveals that some portions of the human and galago clusters cannot be aligned due to insertions or deletions of repetitive DNA elements subsequent to the divergence of these two lineages.

Of the 168 potential GATA-1 sites detected in the human globin cluster, 98 could be aligned in the galago cluster, and 22 of these were found to be conserved. Figure 6 shows the location of the conserved sites; most (11/23) were found in the LCR, 3 were found near ε, 4 near Gγ, 3 near Aγ, and 1 downstream from δ. The detection of the majority of sites in the LCR supports the hypothesis that this protein is involved in opening chromatin in the globin cluster during the early steps of erythroid maturation (Stamatoyannopoulos *et al.*, 1995). Interestingly, only 1 of the conserved GATA-1 sites was found near the δ and β genes. This finding is consistent with the possibility that the β gene does not require a specific local complement of GATA-1 activators to establish its adult expression pattern.

The fact that motif-based phylogenetic analysis could be of considerable value in the identification of key regulatory elements is supported by the finding that 7 of the 23 conserved sites for GATA-1 (30%) have already been shown to be functionally important in other studies: sites 4624, 4654, and 4717 are located in the core of HS3, within regions that are footprinted *in vitro* (Philipsen *et al.*, 1990) and *in vivo* (Strauss and Orkin, 1992). Site 8736 is immediately downstream from the NF–E2 dimer in the HS2 core. Although the major enhancing effect of the HS2 core is derived from the NF–E2 interaction, the GATA-1 site has been shown to augment this enhancing activity (Ellis *et al.*, 1993). Site 19345 is located −159 bp upstream from ε, and its functional importance for enhancer-dependent transcription was demonstrated by Gong and Dean (1993). Finally, sites 34312 and 39248 are located at −184 bp upstream from Gγ and Aγ, respectively, and represent the upstream halves of a bipartate GATA-1 site that has presumptive importance in γ regulation (Martin *et al.*, 1989). Thus, the other GATA-1 sites in Fig. 6 represent apt starting places for future functional analyses. With the knowledge of the sequence characteristics of these important *trans* factor binding sites, it will be possible to use site-directed mutagenesis to destroy several binding sites simultaneously. Thus, the problem of redundancy in *trans* factor binding, which has

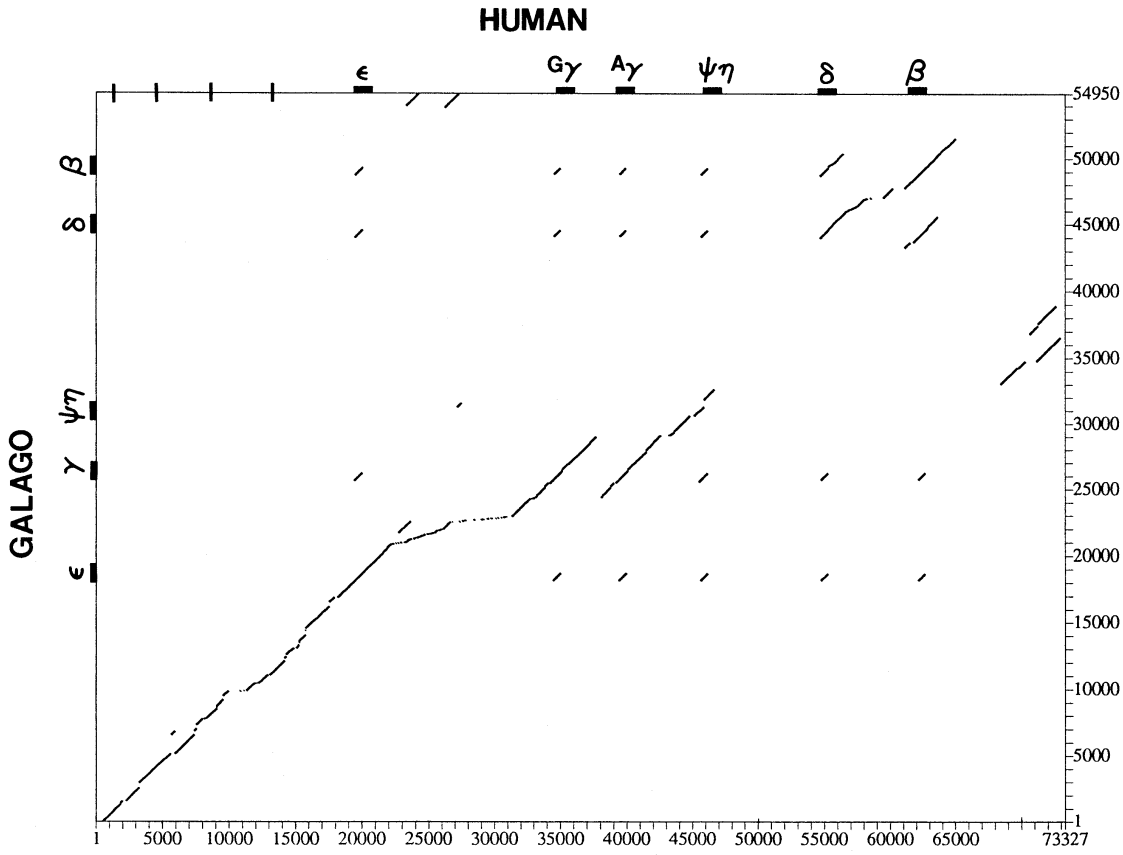

FIG. 5. Dot-matrix similarity analysis of the galago and human β globin clusters. The positions of the individual globin genes are indicated along the top and left margins. Areas covered by diagonals can be aligned between these two clusters. Some regions (e.g., the area surrounding the pseudogene, ψη) have diverged since the separation of simians and prosimians by insertion or loss of repetitive transposable elements and can no longer be aligned.

provided an impediment to interpretation of deletional assays, can be circumvented.

Future approaches. Further exploitation of the functional motif approach to evolution will be greatly facilitated by the recent introduction of a Globin Gene Server by Hardison and Miller (Hardison *et al.*, 1994). This electronically accessible facility is already in operation and affords a constantly updated resource for sequence alignments as well as integrated functional data. This facility will support the search for anciently conserved sequences or binding motifs that are important for regulatory aspects such as the embryonic expression of ε, the embryonic repression of β, or the adult

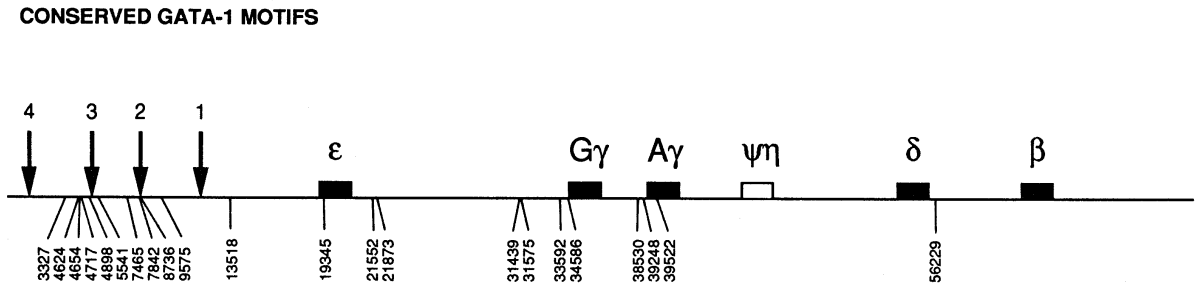

FIG. 6. Outcome of motif-based phylogenetic analysis as applied to the GATA-1 protein. The location binding sites that are conserved between human and galago are indicated using a number that denotes the position of the motif within the human globin cluster as given in GenBank. The position of each of the five active genes is shown on the top line as boxes, and the locations of the four hypersensitive sites within the LCR are indicated by arrows.

expression of β in all mammalian species. In addition, it will facilitate the search for differences in sequences or binding motifs that are responsible for alterations in expression patterns. This latter type of analysis is best carried out within sequence alignments composed of many closely related species such as the primates. The inclusion of other mammalian species in the alignment could give misleading results if there is coevolution of the *trans* factors themselves in these other species.

The evolutionary approach to hemoglobin switching has provided new paradigms that will result in more complete understanding of the evolution of expression patterns in clustered multigene families. Two new assay systems have recently become available that will be important tools in the analysis of the emerging data. First, it has been demonstrated that murine embryonic stem cells can be induced to differentiate *in vitro* into embryoid bodies with blood islands (Keller *et al.*, 1993). The induction of both primitive and definitive erythroblast populations is possible with differential cytokine and growth factor treatment. This system may afford an important alternative to transgenic mice for the analysis of those elements important in early embryonic and fetal globin gene expression. Second, appropriate stage-specific expression of chicken globin genes has been achieved in an *in vitro* system utilizing artificial nuclei (Barton and Emerson, 1994). These nuclei, produced by incubation of DNA expression templates with histones membrane fractions derived from xenopus oocytes and transcription factors isolated from erythroid cells, undergo a single round of semi-conservative DNA replication after assembly (Newport, 1987). Following replication, enhancer-dependent transcriptional activation of the adult chicken β gene can be detected (Barton and Emerson, 1994). This system needs to be further tested to prove that it can support stage-specific expression of mammalian globin genes. If this can be established, it will provide an important resource for functional analysis. Both of these systems will provide important new resources in the analysis of the molecular regulation of hemoglobin switching.

ACKNOWLEDGMENTS

These studies were facilitated by grant support from the Michigan Chapter of the American Cancer Society (D.L.G.) and from PHS Grants NIH-HL48802 (D.L.G.) and NIH-HL 33940 (M.G.). Portions of this study were carried out using computer facilities provided by the General Clinical Research Center at the University of Michigan, which is supported by a grant (M01RR00042) from the National Center for Research Resources, National Institutes of Health.

REFERENCES

Bailey, W. J., Hayasaka, K., Skinner, C. G., Kehoe, S., Sieu, L. C., Slightom, J. L., and Goodman, M. (1992). Reexamination of the African hominoid trichotomy with additional sequences from the primate β-globin gene cluster. *Mol. Phylogenet. Evol.* **1:** 97–135.

Barton, M. C., and Emerson, B. M. (1994). Regulated expression of the β-globin gene locus in synthetic nuclei. *Genes Dev.* **8:** 2453–2465.

Becker, K. G., Jedlicka, P., Templeton, N. S., Liotta, L., and Ozato, K. (1994). Characterization of hUCRBP (YY1, NF-E1, δ): A transcription factor that binds the regulatory regions of many viral and cellular genes. *Gene* **150:** 259–266.

Behringer, R. R., Ryan, T. M., Palmiter, R. D., Brinster, R. L., and Townes, T. M. (1990). Human γ- to β-globin gene switching in transgenic mice. *Genes Deve.* **4:** 380–389.

Bunn, H. F., and Forget, B. G. (1986). "Hemoglobin: Molecular, Genetic and Clinical Aspects," Saunders, Philadelphia.

Cao, S. H., Gutman, P. D., Dave, H. P. G., and Schlechter, A. N. (1989). Identification of a transcriptional silencer in the 5′-flanking region of the human ε-globin gene. *Proc. Natl. Acad. Sci. USA* **86:** 5306–5309.

Chada, K., Magram, J., and Costantini, F. (1986). An embryonic pattern of expression of a human fetal globin gene in transgenic mice. *Nature* **319:** 685–687.

Collins, F. S., and Weissman, S. M. (1984). The molecular genetics of human hemoglobin. *Prog. Nucleic Acid Res. Mol. Biol.* **31:** 315–462.

Cooper, S. J. B., and Hope, R. M. (1993). Evolution and expression of a β-like globin gene of the Australian marsupial *Sminthopsis crassicaudata*. *Proc. Natl. Acad. Sci. USA* **90:** 11777–11781.

Curtin, P. T., Liu, D., Liu, W., Change, J. C., and Kan, Y. W. (1989). Human β-globin gene expression in trangenic mice is enhanced by a distant DNase I hypersensitive site. *Proc. Natl. Acad. Sci. USA* **86:** 7082–7086.

Czeluzniak, J., Goodman, M., Hewett-Emmett, D., Weiss, M. L., Venta, P. J., and Tashian, R. E. (1982). Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. *Nature* **298:** 297–300.

Dillon, N., and Grosveld, F. (1991). Human γ-globin genes silenced independently of other genes in the β-globin locus. *Nature* **350:** 252–255.

Driscoll, M. C., Dobkin, C. S., and Alter, B. P. (1989). γκβ-thalassemia due to a *de novo* mutation deleting the 5′ β-globin gene activation-region hypersensitive sites. *Proc. Natl. Acad. Sci. USA* **86:** 7470–7474.

Ellis, J., Talbot, D., Dillon, N., and Grosveld, F. (1993). Synthetic human β-globin 5′HS2 constructs function as locus control region only in multicopy transgene concatamers. *EMBO J.* **12:** 127–134.

Enver, T., Raich, N., Ebens, A. J., Papayannopoulou, T., Costantini, F., and Stamatoyannopoulos, G. (1990). Developmental regulation of human fetal-to-adult globin gene switching in transgenic mice. *Nature* **344:** 309–313.

Fitch, D. H. A., Bailey, W. J., Tagle, D. A., Goodman, M., Sieu, L., and Slightom, J. L. (1991). Duplication of the γ-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. *Proc. Natl. Acad. Sci. USA* **88:** 7396–7400.

Flanagan, J. R., Becker, K. D., Ennist, D. L., Gleason, S. L., Driggers, P. H., Levi, B.-Z., Appella, E., and Ozato, K. (1992). Cloning of a negative transcription factor that binds to the upstream conserved region of moloney murine leukemia virus. *Mol. Cell. Biol.* **12:** 38–44.

Forrester, W. C., Thompson, C., Elder, J. T., and Groudine, M. A. (1986). Developmentally stable chromatin structure in the human β-globin gene cluster. *Proc. Natl. Acad. Sci. USA* **83:** 1359–1363.

Fraser, P., Hurst, J., Collis, P., and Grosveld, F. (1990). DNaseI hypersensitive sites 1, 2, and 3 of the human β-globin dominant control region direct position-independent expression. *Nucleic Acids Res.* **18:** 3503–3507.

Fraser, P., Pruzina, S., Antoniou, M., and Grosveld, F. (1993). Each hypersensitive site of the human β-globin locus control region con-

fers a different developmental expression pattern of expression on the globin genes. *Genes Dev.* **7:** 106–113.

Gong, Q., and Dean, A. (1993). Enhancer-dependent transcription of the ε-globin promoter requires promoter-bound GATA-1 and enhancer-bound AP-1/NF-E2. *Mol. Cell. Biol.* **13:** 911–917.

Goodman, M., Koop, B. F., Czelusniak, J., Weiss, M. L., and Slightom, J. L. (1984). The eta-globin gene: Its long evolutionary history in the beta-globin gene family of mammals. *Mol. Biol.* **180:** 803–823.

Goodman, M., Czelusniak, J., Koop, B. F., Tagle, D. A., and Slightom, J. L. (1987). Globins: A case study in molecular phylogeny. *Cold Spring Harbor Symp. Quant. Biol.* **52:** 875–890.

Gumucio, D. L., Rood, K. L., Gray, T. A., Riordan, M. F., Sartor, C. I., and Collins, F. S. (1988). Nuclear proteins that bind the human γ-globin gene promoter: Alterations in binding produced by point mutations associated with hereditary persistence of fetal hemoglobin. *Mol. Cell. Biol.* **8:** 5310–5322.

Gumucio, D. L., Heilstedt-Williamson, H., Gray, T. A., Tarle, S. A., Shelton, D. A., Tagle, D. A., Slightom, J. L., Goodman, M., and Collins, F. S. (1992). Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human γ and ε globin genes. *Mol. Cell. Biol.* **12:** 4919–4929.

Gumucio, D. L., Shelton, D. A., Bailey, W. L., Slightom, J. S., and Goodman, M. (1993). Phylogenetic footprinting reveals unexpected complexity in *trans* factor binding upstream from the ε globin gene. *Proc. Natl. Acad. Sci. USA* **90:** 6018–6022.

Gumucio, D. L., Shelton, D. A., Blanchard-McQuate, K., Tarle, S. A., Gray, T. A., Heilstedt-Williamson, H., Slightom, J. S., Collins, F. C., and Goodman, M. (1994). Evolutionary sequence changes near the proximal CCAAT box of the γ globin gene alter *trans* factor binding and promoter strength. *J. Biol. Chem.* **269:** 15371–15380.

Hanscombe, O., Whyatt, D., Fraser, P., Yannoutsos, N., Greaves, D., Dillon, N., and Grosveld, F. (1991). Importance of globin gene order for correct developmental expression. *Genes Dev.* **5:** 1387–1395.

Hardison, R. C. (1981). The nucleotide sequences of rabbit embryonic globin gene β 3. *J. Biol. Chem.* **256:** 11780–11786.

Hardison, R. C. (1984). Comparison of the β-like globin gene families of rabbits and humans indicates that the gene cluster 5′-ε-γ-δ-β-3′ predates the mammalian radiation. Mol Biol. Evol. **1:** 390–410.

Hardison, R., Chao, K., Schwartz, S., Stojanovic, N., Ganetsky, M., and Miller, W. (1994). Globin gene server: A prototype e-mail database server featuring extensive multiple alignments and data compilation for electronic genetic analysis. *Genomics* **21:** 344–353.

Hariharan, N., Kelley, D. E., and Perry, R. P. (1992). δ, A transcription factor that binds to downstream elements in several polymerase II promoters, is a functionally versatile zinc finger protein. *Proc. Natl. Acad. Sci. USA* **88:** 9799–9803.

Harris, S., Barrie, P. A., Weiss, M. L., and Jeffreys, A. J. (1984). The primate ψβ1 gene: An ancient β-globin pseudogene. *J. Mol. Biol.* **180:** 785–801.

Hayasaka, K., Fitch D. H. A., Slightom, J. L., and Goodman, M. (1992). Fetal recruitment of anthropoid γ-globin genes. Findings from phylogenetic analyses involving the 5′-flanking sequences of the ψγ globin gene of spider monkey *Ateles geoffroyi*. *J. Mol. Biol.* **224:** 875–881.

Hill, A., Hardies S. C., Phillips, S. J., Davis, M. G., Hutchinson, C. A., III, and Edgell, M. H. (1984). Two mouse early embryonic β-globin gene sequences: evolution of the nonadult β-globins. *J. Biol. Chem.* **259:** 3739–3747.

Jane, S. M., Ney, P. A., Vanin, E. F., Gumucio, D. L., and Nienhuis, A. W. (1992). Identification of a potential stage selector element in the human γ-globin gene promoter that fosters preferential interaction with the HS2 enhancer when in competition with the β-promoter. *EMBO J.* **11:** 2961–2969.

Jane, S. M., Nienhuis, A. W., and Cunningham, J. M. (1995). Hemoglobin switching in man and chicken is mediated by a heteromeric complex between the ubiquitous transcription factor CP2 and a developmentally specific protein. *EMBO J.* **14:** 97–105.

Keller, G., Kennedy, M., Papayannopoulou, T., and Wiles, M. V. (1993). Hematopoietic commitment during embryonic stem cell differentiation in culture. *Mol. Cell. Biol.* **13:** 473–486.

Ko, L. J., and Engle, J. D. (1993). DNA-binding specificities of the GATA transcription factor family. *Mol. Cell. Biol.* **13:** 4011–4022.

Kollias, G., Wrighton, N., Hurst, J., and Grosveld, F. (1986). Regulated expression of human Aγ-, β-, and hybrid γβ-globin genes in transgenic mice: Manipulation of the developmental expression patterns. *Cell* **46:** 89–94.

Koop, B. F., and Goodman, M. (1988). Evolutionary and developmental aspects of two hemoglobin β-chain genes (ε^M and β^M) of opossum. *Proc. Natl. Acad. Sci. USA* **85:** 3893–3897.

Lee, J.-S., Galvin, K. M., and Shi, Y. (1993). Evidence for physical interaction between the zinc-finger transcription factors YY1 and Sp1. *Proc. Natl. Acad. Sci. USA* **90:** 6145–6149.

Letovsky, J., and Dynan, W. S. (1989). Measurement of the binding of transcription factor Sp1 to a single GC box recognition sequence. *Nucleic Acids Res.* **17:** 2639–2649.

Li, Q., and Stamatoyannopoulos, J. (1994). Position independence and proper developmental control of γ-globin gene expression require both a 5′ locus control region and a downstream sequence element. *Mol. Cell. Biol.* **14:** 6087–6096.

Lloyd, J. A., Krakowsky, J. M., Crable, S. C., and Lingrel, J. B. (1992). Human γ to β globin switching using a mini-construct in transgenic mice. *Mol. Cell. Biol.* **12:** 1561–1567.

Martin, D. I. K., Tsai, S.-F., and Orkin, S. H. (1989). Increased γ-globin expression in a nondeletion HPFH mediated by an erythroid-specific DNA-binding factor. *Nature* **338:** 435–438.

Meireles, C. M. M., Schneider, M. P. C., Sampaio, M. I. C., Schneider, I. I., Slightom, J. L., Chiu, C-H., Neiswanger, K., Gumucio, D. L., Czelusniak, J., and Goodman, M. (1995). Fate of a redundant γ-globin gene in the atelid clade of New World monkeys: Implications concerning fetal globin gene expression. *Proc. Natl. Acad. Sci. USA* **92:** 2607–2611.

Merika, M., and Orkin, S. H. (1993). DNA-binding specificity of GATA family transcription factors. *Mol. Cell. Biol.* **13:** 3999–4010.

Miller, B. A., Olivieri, N., Salameh, M., Ahmed, M., Antognetti, Huisman, T. H. J., Nathan, D. G., and Orkin, S. H. (1987). Molecular analysis of the high-hemoglobin-F phenotype in Saudi Arabian sickle cell anemia. *New Engl. J. Med.* **316:** 244–250.

Natesan, S., and Gilman, M. Z. (1993). DNA bending and orientation-dependent function of YY1 in the c-*fos* promoter. *Genes Dev.* **7:** 2497–2509.

Newport, J. (1987). Nuclear reconstitution *in vitro:* stages of assembly around protein-free DNA. *Cell* **48:** 205–217.

Ney, P. A., Sorrentino, B. P., McDonough, K. T., and Nienhuis, A. W. (1990). Tandem AP-1-binding sites within the human β-globin dominant control region function as an inducible enhancer in erythroid cells. *Genes Dev.* **4:** 993–1006.

Ney, P. A., Sorrentino, B. P., Lowrey, C. H., and Nienhuis A. W. (1990). Inducibiltiy of the HS2 enhancer depends on binding of an erythroid specific nuclear protein. *Nucleic Acids Res.* **18:** 6011–6017.

Park, K., and Atchison, M. L. (1991). Isolation of a candidate repressor/activator, NF-E1 (YY-1, δ), that binds to the immunoglobulin κ 3′ enhancer and the immunoglobulin heavy-chain μE1 site. *Proc. Natl. Acad. Sci. USA* **88:** 9804–9808.

Perez-Stable, C., and Costantini, F. (1990). Roles of fetal Gγ-globin promoter elements and the adult β-globin 3′ enhancer in the stage-specific expression of globin genes. *Mol. Cell. Biol.* **10:** 1116–1125.

Peters, B., Merezhinskaya, N., Diffley, J. F., and Noguchi, C. T. (1993). Protein–DNA interactions in the epsilon-globin gene silencer. *J. Biol. Chem.* **268:** 3430–3437.

Philipsen, S., Talbot, D., Fraser, P., and Grosveld, F. (1990). The β-globin dominant control region hypersensitive site 2. *EMBO J.* **9:** 2159–2167.

Raich, N., Enver, T., Nakamoto, B., Josephson, B., Papayannopoulou, T., and Stamatoyannopoulos, G. (1990). Autonomous developmental control of human embryonic globin gene switching in transgenic mice. *Science* **250:** 1147–1149.

Raich, N., Papayannopoulou, T., Stamatoyannopoulos, G., and Enver, T. (1992). Demonstration of a human ε-globin gene silencer with studies in transgenic mice. *Blood* **79:** 861–864.

Raich, N., Clegg, D. H., Grofti, J., Romeo, P.-H., and Stamatoyannopoulos. (1995). GATA1 and YY1 are developmental repressors of the human ε-globin gene. *EMBO J.* **14:** 801–809.

Ron, D., and Dressler, H. (1992). pGSTag—A versitile bacterial expression plasmid for enzymatic labeling of recombinant proteins. *BioTechniques* **13:** 866–868.

Ryan, T. M., Behringer, R. R., Townes, T. M., Palmiter, R. D., Brinster, R. L. (1989). High-level erythroid expression of human α-globin genes in transgenic mice. *Proc. Natl. Acad. Sci. USA* **86:** 37–41.

Safaya, S., Rider, R. F., Dowling, C. E., Kazazian, H. H., Jr., and Adams, J. G., III. (1989). Homozygous β-thalassemia without anemia. *Blood* **73:** 324–328.

Seto, E., Lewis, B., and Shenk, T. (1993). Interaction between transcription factors Sp1 and YY1. *Nature* **365:** 462–464.

Shi, Y., Seto, E., Chang, L.-S., and Shenk, T. (1991). Transcriptional repression by YY1, a human GLI-Kruppel-related protein, and relief of repression by adneovirus E1A protein. *Cell* **67:** 377–388.

Shih, D. M., Wall, R. J., and Shapiro, S. T. (1990). Developmentally regulated and erythroid-specific expression of the human embryonic β-globin gene in transgenic mice. *Nucleic Acids Res.* **18:** 5465–5472.

Shrivastava, A., Saleque, S., Kalpana, G. V., Artandi, S., Goff, S. P., and Calame, K. (1994). Inhibition of transcriptional regulator Yin-Yang-1 by association with c-*myc.* *Science* **262:** 1889–1892.

Stamatoyannopoulos, G., Josephson, B., Zhang, J-W., and Li, Q. (1993). Developmental regulation of human γ-globin genes in transgenic mice. *Mol. Cell. Biol.* **13:** 7636–7644.

Stamatoyannopoulos, J. A., Goodwin, A., Joyce, T., and Lowrey, C. H. (1995). NF-E2 and GATA binding motifs are required for the formation of DNase I hypersensitive site 4 of the human beta-globin locus control region. *EMBO J.* **14:** 106–116.

Strauss, E. C., and Orkin, S. H. (1992). *In vivo* protein–DNA interactions at hypersensitive site 3 of the human β-globin locus control region. *Proc. Natl. Acad. Sci. USA* **89:** 5809–5813.

Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. and Jones, R. T. (1988). Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatis*): Nucleotide and amino acid sequences, developmental regulation, and phylogenetic footprints. *J. Mol. Biol.* **203:** 439–455.

Tagle, D. A., Stanhope, M. J., Siemieniak D. R., Benson, P., Goodman, M., and Slightom, J. L. (1992). The beta globin gene cluster of the prosimian primate (*Galago crassicadatis*): Nucleotide sequence determination of the 41 Kb cluster and comparative sequence analysis. *Genomics* **13:** 1–20.

Tsai, F. S., Martin, D. I., Zon, L. I., D'Andrea, A. D., Wong, G. G., and Orkin, S. H. (1989). Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. *Nature* **399:** 446–451.

Townes, T. M., Lingrel, J. B., Chen, H. Y., Brinster, R. L., and Palmiter, R. D. (1985). Erythroid-specific expression of human β-globin genes in transgenic mice. *EMBO J.* **4:** 1715–1723.

Trepicchio, W. L., Dyer, M. A., and Baron, M. H. (1993). Developmental regulation of the human embryonic beta-like globin gene is mediated by synergistic interactions among multiple tissue- and stage-specific elements. *Mol. Cell. Biol.* **13:** 7457–7468.

Tuan, D., Soloman, W., Li., Q., and London, I. M. (1985). The "β-like-globin" gene domain in human erythroid cells. *Proc. Natl. Acad. Sci. USA* **82:** 6384–6388.

Whyatt, D. J., deBoer, E., and Grosveld, F. (1993). The two zinc finger-like domains of GATA-1 have different DNA binding specificities. *EMBO J.* **12:** 4933–5005.

Yant, S. R., Zhu, W., Millinoff, D., Slightom, J., Goodman, M., and Gumucio, D. L. (1995). High affinity YY1 binding motifs: Identification of two core types (CCAT and ACAT) and distribution of potential binding sites within the human β globin cluster. *Nucleic Acids Res.* **23:** 4353–4362.