Use of Long Sequence Alignments to Study the Evolution and Regulation of Mammalian Globin Gene Clusters

Ross Hardison* and Webb Miller†

*Department of Molecular and Cell Biology, and †Department of Computer Science and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University

The determination of long segments of DNA sequences encompassing the β- and α-globin gene clusters has provided an unprecedented data base for analysis of genome evolution and regulation of gene clusters. A newly developed computer tool kit generates local alignments between such long sequences in a space-efficient manner, helps the user analyze the alignments effectively, and finds consistently aligning blocks of sequences in multiple pairwise comparisons. Such sequence analyses among the β-like globin gene clusters of human, galago, rabbit, and mouse have revealed the general patterns of evolution of this gene cluster. Alignments in the flanking regions are very useful in assigning orthologous relationships. Investigation of such matches between the mouse and human β-like globin gene clusters has led to a reassessment of some orthologous assignments in mouse and to a revision of the proposed pathway for evolution of this gene cluster. In general, the interspersed repetitive elements have inserted independently, presumably via a retrotransposition mechanism, in the different mammalian lineages. However, some examples of ancient L1 repeats are found, including one between the ε- and γ-globin genes that appears to have been in the ancestral eutherian gene cluster. Prominent matching sequences are found in a long region 5' to the ε-globin gene, the locus control region (LCR) that is a positive regulator of the entire gene cluster. Three-way alignments among the human, goat, and rabbit sequences can extend for >3 kb in part of the LCR (DNase hypersensitive site 3), indicating that the cis-acting components of this complex regulatory region cover a long segment of DNA. In contrast to the β-like globin gene clusters, the α-like globin gene clusters of many mammals occur in very G+C-rich isochores and contain prominent CpG islands. The regions between the α-like globin genes are evolving faster than the intergenic regions of the β-like globin gene clusters. The contrasts between the two gene clusters can be attributed to differences in DNA metabolism in the isochores. The proximal control elements of the rabbit α-globin gene are located both 5' to and within the gene. All of this region is part of a prominent CpG island that may be acting as an extended, enhancer-independent promoter. One can hypothesize that the analogue to the LCR in the α-globin gene cluster may interface with the distinctive α-globin promoter in ways different from the interaction between the β LCR and the promoters of β-like globin genes. The net result, however, is balanced and coordinated expression of these different genes in erythroid cells.

1. Key words: β-globin genes, α-globin genes, hemoglobin, mammals, locus control regions, local alignments, multiple alignments.

Address for correspondence and reprints: Ross Hardison, 206 Althouse Laboratory, Department of Molecular and Cell Biology, The Pennsylvania State University, University Park, Pennsylvania 16802.
Introduction

The hemoglobin tetramer consists of four globin polypeptides, two α-like and two β-like, each with its associated heme. The amino acid sequences of the α-like chains match those of the β-like chains at ~50% of the positions. The α-like and β-like globins fold into similar but complementary shapes, so that the heme can snuggle into a pocket formed by three α helices and bind O₂ reversibly. The complementary interactions between the α- and β-globin subunits allow O₂ to bind cooperatively (Dickerson and Geis 1983, pp. 20-63). Hence, the production of functional hemoglobin requires the balanced synthesis of both α-like and β-like globins in erythroblasts and reticulocytes, and indeed mutations leading to inadequate production of either α- or β-globin cause the inherited anemias called “α-thalassemias” or “β-thalassemias” (Bunn and Forget 1986, pp. 223-379).

All species that use hemoglobins to carry O₂ produce different forms of hemoglobin at different developmental stages. Mammals produce different hemoglobins in embryonic and adult life, with some mammals (in particular, simian primates and ruminants) synthesizing a fetal specific form, HbF, or α₂γ₂. In most other mammals, such as the galago, rabbit, and mouse to be discussed here, the embryonic form is replaced by the “adult” hemoglobin in the fetal liver. This latter hemoglobin is produced through the rest of the animal’s life (Bunn and Forget 1986, pp. 126-167; Tagle et al. 1988).

This phenomenon of hemoglobin switching has provided a rationale for attempting to cure or ameliorate the effects of hemoglobinopathies by regulating the fetal to adult switch in humans. In those cases in which the adult β-globin gene is defective, such as in sickle cell anemia or β-thalassemia, the fetal γ-globin gene can substitute for the defective gene, as exemplified by the phenotype of “hereditary persistence of fetal hemoglobin” (Bunn and Forget 1986, pp. 223-379). Some of these patients have suffered a complete deletion of the δ- and β-globin genes, but the continued production of γ-globin in adult life leads to an essentially asymptomatic condition. Thus, intervention in the normal switching process, either pharmacologically or by genetic manipulation, offers a possible avenue to therapy for these common genetic hemoglobinopathies. Such a hope, as well as the intrinsic interest in genome evolution and regulation, has led many investigators to study both globin gene structure and globin gene function in many different organisms. Comparative studies among the different species should provide important new insights into these problems, and such studies will be reviewed here for the mammalian globin gene clusters.

The regulation of globin genes is complex, requiring at least three potentially interacting modes. The genes are expressed only in erythroid cells with a balanced production of α-like and β-like globins, and different forms of hemoglobin are made at progressive stages of development. Clues regarding this tissue specificity, coordination, and developmental regulation can come from many sources. Many gene regulatory elements, including promoters, enhancers, and silencers, reside in regions of chromatin that are readily cleaved by DNase (Gross and Garrard 1988). Deletions or other mutations that alter the level or pattern of gene expression also reveal candidates for control sequences. In this review the utility of analyzing multiple alignments of large segments of gene clusters will be emphasized as an additional source of candidates for control elements. These candidates can then be tested in functional assays for their effect on gene regulation.
Evolution of Mammalian Globin Gene Clusters

A phylogenetic analysis indicates that α- and β-globin genes were formed by duplication of a common ancestor ~450 Mya, early in the evolution of jawed vertebrates (Czelusniak et al. 1982). This ancestral gene had three exons separated by two introns. After the initial duplication, the α- and β-globin genes continued the process of duplication and divergence to generate the current gene clusters. The β-like globin genes are those descended from the ancestral β and include ε, γ, η, δ, and θ. Likewise, the α-like globin genes are those descended from the ancestral α and include ζ, α, and 0. The duplications of the α-like globin genes were fixed prior to the divergence of birds and mammals, so that the chicken embryonic π gene is orthologous to the mammalian embryonic ζ, and the chicken and mammalian α-globin genes are orthologous. However, the duplications that generated the β-like globin gene clusters have occurred independently in the avian and mammalian lineages (reviewed in Goodman et al. 1987; Hardison 1991).

The process of gene duplication and divergence in eutherian mammals has generated the gene clusters shown in figure 1, which illustrates the arrangement of globin genes in representatives of five orders of mammals, i.e., artiodactyls (goat, Capra hircus), lagomorphs (rabbit, Oryctolagus cuniculus), primates (the prosimian, Galago crassicaudatus, and human, Homo sapiens), rodents (mouse, Mus domesticus), and perissodactyls (horse, Equus caballus). The number and positions of these genes differ in each case. The distance between the globin genes is variable, in many cases as a result of multiple insertions of retroposon-type repeats.

Despite this diversity, some common features are found. First, the β-like globin genes occur in clusters separate from the α-like globin genes; in fact, these gene clusters are located on different chromosomes in all birds and mammals investigated (Collins and Weissman 1984). Hence the balance in globin production requires coordinate gene regulation between dispersed gene clusters. Second, each gene cluster contains a set of developmentally regulated genes that are arranged roughly in their temporal order of expression. In the β-like globin gene clusters, the earlier-expressed genes (embryonic ε and γ) typically are located at the 5’ end of each gene cluster, and the later-expressed genes (fetal and adult β) are located at the 3’ end of each gene cluster (Collins and Weissman 1984; Hardison 1991). The duplicated γ-globin genes expressed in the fetal livers of higher primates are orthologous to the γ-globin genes expressed in the embryonic life of rabbits (Rohrbaugh and Hardison 1983), mouse (β1; Farace et al. 1984), and galago (Tagle et al. 1988). Hence, the anthropoid γ-globin genes were most likely recruited for fetal expression in an ancestral simian primate (Tagle et al. 1988). The genes orthologous to the δ-globin gene, located in the middle of the gene cluster, range from nonexpressed pseudogenes in mice and rabbits to a low-expressed adult gene in humans to a highly expressed gene in galagos. In the α-globin gene clusters, the earliest expressed gene, the embryonic ζ, is also located at the 5’ end of the cluster, and α-globin genes are located on the 3’ side of the ζ-globin genes. However, the θ-globin gene at the 3’ end of the α-globin gene cluster is a pseudogene in rabbits (Cheng et al. 1986) and horse (Clegg 1987). The human θ-globin gene also is likely to be a pseudogene (Clegg 1987), even though it is transcribed at a low level (Leung et al. 1987; Hsu et al. 1988). Although the α-globin genes located between ζ and θ are expressed abundantly in fetal and adult life, they are also expressed at low levels in embryonic life in humans, rabbits, and perhaps other mammals (Rohrbaugh
Goat $\beta$

Galago $\beta$

Human $\beta$

Rabbit $\beta$

Mouse $\beta$

Goat $\alpha$

Horse $\alpha$

Human $\alpha$

Rabbit $\alpha$

Mouse $\alpha$

FIG. 1.—Maps of representative mammalian globin gene clusters. Each gene (including exons plus introns) is shaded according to its proposed orthologous relationships; all genes are oriented 5' to 3', from left to right. The horse $\psi\alpha$-globin gene is orthologous to the $\theta$-globin gene of humans and rabbits; it is not known whether a $\theta$-globin gene is present in goat or mouse. Both $\theta1$- and $\theta2$-globin genes in rabbits are pseudogenes, and it is likely that all the duplicated copies are also pseudogenes (Cheng et al. 1986). References for the maps are as follows: goat $\beta$, Townes et al. (1984); galago $\beta$, Tagle et al. (1992); human $\beta$, Efstratiadis et al. (1980) and Collins and Weissman (1984); rabbit $\beta$, Margot et al. (1989); mouse $\beta$, Shehee et al. (1989); goat $\alpha$, Schon et al. (1982) and Wernke and Lingrel (1986); horse $\alpha$, Flint et al. (1988); human $\alpha$, Nicholls et al. (1987); rabbit $\alpha$, Cheng and Hardison (1988) and Cheng et al. (1988); and mouse $\alpha$, Leder et al. (1981).

and Hardison 1983; Peschle et al. 1985). Thus the most consistent feature is the early expression of the 5'-most genes in each gene cluster.

Two exceptions to this general pattern of gene organization result from block duplications of sets of genes. In the first case, the goat $\beta$-like globin gene cluster (fig. 1) has triplicated an $\varepsilon$-$\eta$-$\delta$-$\beta$ globin gene quartet (Townes et al. 1984). The first two
Evolution of Mammalian Globin Gene Clusters

The 3'-most genes in each quartet, each a descendant of the β-globin gene, are expressed at different developmental stages, either in fetal (β'), juvenile (βc), or adult (βa) life (Townes et al. 1984). Thus fetal specific expression is not limited to γ-globin genes, and the recruitment to fetal expression has occurred in independent events in different orders of mammals. The remaining genes in the goat β-globin gene cluster appear to be pseudogenes (Lingrel et al. 1985). Likewise, the rabbit α-globin gene cluster was formed by the duplication of a ζ-α-θ quartet, followed by deletion of the α gene from one set and then multiple duplications of the resultant ζ-α-θ trio (Cheng et al. 1986, 1987, 1988). In this case most of the genes are not active, leaving one active α-globin gene and at least one active ζ-globin gene (on the basis of the evidence for ζ-globin in the hemoglobins in embryonic rabbit erythrocytes; Melderis et al. 1974). After such block duplications, the actual number of gene sets is highly polymorphic in the population (Cheng and Hardison 1988; Garner and Lingrel 1988), indicating continued recombinations between these homology blocks to generate variable numbers of pseudogenes. But even in these more complicated cases in mammals, the 5'-most gene is expressed early in development, and the adult genes are located on the 3' side of the embryonic genes.

Structural Comparisons among Mammalian Globin Gene Clusters

In constructing evolutionary histories of multigene families, it is necessary to discriminate between paralogous relationships (genes that diverged as a result of gene duplications, either prior to or after speciation events) and orthologous relationships (genes that diverged as a result of speciation events). However, the many recombinations and gene conversions that have occurred in the gene clusters have greatly complicated the analysis (Maeda and Smithies 1986). Fortunately, large segments of intronic and flanking sequences have been determined, and these have proved invaluable in assigning orthologous relationships (Hardies et al. 1984; Hardison 1984; Hutchison et al. 1984). Even at an essentially neutral rate of evolution (e.g., silent-site substitution rate), the intergenic regions still matched at a detectable level, though the separation of mammalian orders occurred ~80 Mya. Therefore, it was possible to deduce orthology on the basis of alignments in the introns and flanking regions. The information in the alignments has allowed the construction of plausible histories for the evolution of the gene clusters, as will be illustrated here for the four β-globin gene clusters in which all the known genes and intergenic regions have been sequenced. Complete sequences are now available from the β-globin gene clusters of human (73,360 bp; compiled in Collins and Weissman 1984), rabbit (44,594 bp; Margot et al. 1989), mouse (57,232 bp; Shehee et al. 1989; Moon and Ley 1990), and galago (41,101 bp; Tagle et al. 1992).

The initial long-range comparisons of these gene clusters used dot-matrix analysis, or dot plots (Konkel et al. 1979; Zweig 1984), in which the segments at any position in the two sequences that fulfilled a particular match criterion (e.g., 23 matches in a window of 30 nucleotides) were identified as a dot on a two-dimensional plot. Although many of the general features of the sequence comparisons were evident as prominent diagonals of dots (Margot et al. 1989; Shehee et al. 1989), the background of irrelevant matches was very high, likely obscuring some of the more subtle features. Another disadvantage of the dot-plot approach was that no alignments were generated and one had to use other programs to generate alignments. Consequently, only small portions of the gene clusters were subjected to this laborious analysis.

To obtain optimized local alignments of very long sequences, Huang et al. (1990)...
developed a program (*sim*) that applies dynamic programming techniques while using computer space efficiently. Subsequently, an interactive user interface for viewing and analyzing the alignments (*lax*) and methods for plotting the coordinates of the alignments along with sequence features (*laps*) were developed (Schwartz et al. 1991). An additional program (*pab*) will identify blocks of sequences that align consistently in a set of separate pairwise comparisons (Boguski et al. 1992; Miller, accepted). The results of applying these programs to the four sequences of the mammalian β-like globin gene clusters are presented here.

The aligning segments plotted in the top panel of figure 2 show the positions of sequences that still match in the β-like globin gene clusters 55 Myr after the divergence between humans and galagos (Tagle et al. 1992). Virtually the entire galago gene cluster aligns with the human sequence, with the exception of the insertions of Ll repeats between ε and γ in human and between Ψη and δ in galago. Some of these long alignments are interrupted by additional insertions of shorter repeated sequences. These long blocks of matching sequences, extending for many kilobases around the genes, are convincing evidence that this chromosomal arrangement has been conserved and was present in the last common ancestor; hence the genes within the long alignments are indeed orthologous. In addition, the exons of each β-like globin gene from one species match the exons of all β-like globin genes in the other species. For example, this can be seen as a set of short diagonals including exons 1 and 2 and a separate diagonal for exon 3 under human ε-, γ-, Ψη-, and δ-globin genes when compared with the galago β-globin gene. Such exonic matches off the main diagonals denote paralogous relationships between genes that had duplicated prior to the split between the two species being compared. The galago δ-globin gene shows a longer alignment with human β than with the human δ-globin gene; this galago δ-globin gene underwent a gene conversion event with β, such that much of its sequence is now like a β-globin gene and is expressed at an unusually high level in adult tissues (Tagle et al. 1991). As expected for a relatively recent duplication, both the human γ-globin genes show extensive matches with the galago γ and its flanking region. In most cases, the matches between repetitive elements are far off the main diagonal; e.g., the L1Gc repeats located between galago Ψη and δ-globin genes match with the L1Hs repeat located 3' to the human β-globin gene, but the alignments fall at the far right of the plot in the top panel of figure 2. This is expected for matches between repeated sequences that have inserted independently in the two lineages. In contrast, the left end (3' end in the conventional orientation) of L1HS3, located between the human ε and γ-globin gene, matches with the 3' end of an L1Gc repeat in the same position in galago, extending into the flanking single-copy DNA (Tagle et al. 1992; fig. 2). Thus, an L1 most likely was in this position before the separation between simian and prosimian primates.

The next comparison spans a greater evolutionary time of ~80 Myr since the ancestors to humans and rabbits diverged. However, the overall pattern seen in the middle panel of figure 2 is very similar to that observed for human versus galago. Again, the characteristic long alignments extend for several kilobases around orthologous genes, and the paralogous pairs are evident as shorter matches off the main diagonal, confined largely to the coding regions. In both comparisons (fig. 2, *top* and *middle*), the duplication units of the γ-globin genes are bounded by truncated L1 retroposon repeats, consistent with the proposed mechanism of duplication via unequal crossing-over, with the L1 repeats being the targets of the recombination (Maeda and Smithies 1986; Fitch et al. 1991). The plot in the middle panel of figure 2 also shows the deletion of the η-globin gene from rabbits, and the matches at the far 5' end of
The rabbit \( \psi \delta \)-globin gene forms a long alignment with the human \( \beta \) globin gene (fig. 2, middle), which is explained by a gene conversion with the \( \beta \) globin gene at its 5' end (Hardison and Margot 1984). The \( \delta \) globin gene also was converted by the \( \beta \) globin gene in a separate event in higher primates (Martin et al. 1983) and the prosimian galago (Tagle et al. 1991). In rabbits, the conversion was not sufficient to rescue the expression of this pseudogene, but in humans the \( \delta \) globin gene was resuscitated to a low level of expression. The \( \delta \) globin gene locus has been involved in many independent recombinations and conversions in different mammalian lineages (reviewed in Hardison and Margot 1984); hence it appears to be a recombinogenic locus. It is interesting that a region between the human \( \delta \) - and \( \beta \)-globin genes is a hot spot for recombination both in human (Orkin et al. 1982) and in yeast (Treco et al. 1985).

In addition to the foregoing, the alignment plots presented here revealed several features that were obscured by the high background of previous dot plots. In particular, two ancient L1 repeats were missed, one between the rabbit \( \epsilon \) - and \( \gamma \)-globin genes (orthologous to the L1 in this position in galago and human) and one 5' to the human \( \delta \)-globin gene (Schwartz et al. 1991). The \( \epsilon \)-\( \gamma \) intergenic region of rabbits is peppered with multiple C repeats (analogous to human \( Alu \) repeats; Krane et al. 1991), but it also has several short segments that match around the 3' end of L1Hs3, the long L1 in the human \( \epsilon \)-\( \gamma \) intergenic region (fig. 2, middle). Analysis of these alignments solely on the basis of the pairwise comparisons is ambiguous. The highest-scoring aligning segment favors separate integration events of the highly truncated L1Oc10 repeat in this region (Huang et al. 1990). However, two other DNA segments that align between rabbit and human also are found in the galago-human alignments; these segments extend from the 3' end of the L1 into the immediate flanking region (fig. 3; Tagle et al. 1992), suggesting that they resulted from a common integration event in the ancestral species.

The L1 repeats are retroposons (Dombrowski et al. 1991) and have a similar structure in all mammals, with two long open reading frames (ORFs) bounded by 5' and 3' untranslated regions. All of ORF2 and the 3' portion of ORF1 are quite similar in L1's from different mammalian species, but the 5' and 3' ends are distinctive (Demers et al. 1989; Hutchison et al. 1989). Many individual L1's are truncated, but this is almost always from the 5' end. In contrast, the common L1 repeats found in the \( \epsilon \)-\( \gamma \) intergenic region in humans, galagos, and rabbits are missing the 3' untranslated region characteristic of most L1's. The presence of these L1's in the same position in the \( \beta \)-like globin gene cluster is, currently, the only example of orthologous repeats in different mammalian orders. Thus, these repeats are members of a very old class of L1 repeats (Rogan et al. 1987) that apparently existed prior to the eutherian radiation (Tagle et al. 1992). They may be examples of the parents of the contemporary L1 repeats. Certainly these ancient L1 repeats predate the acquisition of the distinctive 3' untranslated regions characteristic of more recent L1's. The L1 located 5' to the human \( \delta \) globin gene, L1Hs9, also contains a remnant of the 3' end of ORF2 with no 3' untranslated region (fig. 3). Quantitative comparisons also support the conclusion that these ancient L1 repeats predate the eutherian radiation, in that they are as similar to L1's from other species as to their own (\( \sim 60\%-65\% \) identity; Schwartz et al. 1991), whereas more recently transposing L1's are much more similar within species than between species (Hutchison et al. 1989).

The comparison between human and mouse \( \beta \)-like globin gene clusters allows a
look farther back in time, given that mouse diverged from the ancestor to human and rabbit earlier than the latter two species diverged (Easteal 1988; Li et al. 1990). At the more stringent scoring parameters used in the comparisons of human with galago and rabbit, most of the aligning segments between mouse and human sequences are confined to the regions close to the genes, with only limited matches in the intergenic regions (Shehee et al. 1989). To test for more divergent alignments in intergenic regions, sim alignments were generated with more relaxed criteria. Indeed, many informative matches were found (fig. 2, bottom), and one can trace alignments through much of the gene cluster. The mouse \( y \) gene, encoding an \( \varepsilon \)-globin, shows a strong alignment with the human \( \varepsilon \)-globin gene, as do the regions around HS1 and HS2 of the LCR. The 3' flank of mouse bh0 matches with that of the human \( \varepsilon \)-globin gene (circled in fig. 2, bottom); thus, it appears that mouse bh0 is orthologous to the human \( \varepsilon \)-globin. This contrasts with the previous conclusion that it is orthologous to the human \( \gamma \)-globin gene, a conclusion based on comparisons of the gene regions (Hill et al. 1984). The 5' flank of mouse bh1 matches with that of the human \( G^\gamma \)-globin gene, confirming that these two genes are orthologous (Hill et al. 1984). Farther through the mouse gene cluster, the 3' flank of mouse bh2 matches with that of the human \( \delta \)-globin gene, and the 5' flanks of both bh1 and bh2 in mouse match with the 5' flank of the human \( \beta \)-globin gene. These observations also fit with previous conclusions that mouse bh2 and bh3 are \( \delta \) orthologues (Hardies et al. 1984; Hutchison et al. 1984) and that mouse bh1 and bh2 are \( \beta \) orthologues (Hardies et al. 1984). But although the mouse bh3 gene itself matches best with the human \( \delta \)-globin gene, its 5' flank matches with that of the human \( G^\gamma \)-globin gene (circled in fig. 2, bottom; Shehee et al. 1989).

The current model for evolution of the mouse \( \beta \)-like globin gene cluster (Hardies et al. 1984; Hill et al. 1984; fig. 4A) proposes a duplication of the DNA segment containing the ancestral \( \delta \)- and \( \beta \)-globin genes, followed by unequal crossing-over to fuse \( \beta \)- and \( \delta \)-globin genes to form the ancestor to the bh3 gene. By postulating that the \( \beta \)-globin gene parent to this fusion gene was inactivated early in the rodent lineage, the highly divergent 5' end of the bh3 gene can be explained (Hardies et al. 1984). On the basis of the strong similarity between mouse genes bh0 and bh1, Hill et al.

---

**Fig. 2.**—Plots of aligning DNA sequences in pairwise comparisons among \( \beta \)-like globin gene clusters. Top, Comparison between galago and human. The program *sim* (Huang et al. 1990) was used to find local alignments between the DNA sequences of the galago (Tagle et al. 1992) and human (Collins and Weissman 1984; Li et al. 1985) \( \beta \)-like globin gene clusters, by using scoring parameters of match = 1, mismatch = -1.2, gap open penalty = 6, and gap extension penalty = 0.2 per position in the gap. Alignments are displayed only if their score exceeds a threshold value of \( \tau \), chosen so that the probability is 0.05 that random sequences matching the given sequences in length and nucleotide composition have a gap-free alignment scoring at least \( \tau \) (Karlin and Altschul 1990). The program *laps* (Schwartz et al. 1991) was used to plot the positions of the aligning segments, along with notable sequence features along the axes. Genes are shown as boxes, with exons black filled; short interspersed repeats are shown as black-filled triangles, pointing to the 3' end of the repeat; long interspersed L1 repeats are shown as open arrows, again pointing to the 3' end; and DNase hypersensitive sites (HS1–HS4) in the LCR are displayed as unfilled boxes. Middle, Comparison between rabbit (Margot et al. 1989) and human. Bottom, Comparison between mouse (Shehee et al. 1989; Moon and Ley 1990) and human. The scoring parameters were changed to less severe mismatch and gap open penalties (match = 1, mismatch = -1, gap open penalty = 4, and gap extension penalty = 0.2 per position in the gap), thereby allowing more alignments in the intergenic regions. As in the top and middle panels, the cutoff score \( \tau \) was still chosen so that the probability is 0.05 that random sequences matching these two sequences in length and nucleotide composition will form a gap-free alignment scoring at least \( \tau \), by using these different scoring parameters. The circled regions are discussed in the text.
(1984) proposed that the ancestral γ-globin gene was duplicated, followed by successive gene conversions, first with the ε-globin gene and later with the γ-like parent to bh1 (fig. 4A). However, this model does not explain the alignments between the 3' flanks of the mouse bh0- and the human ε-globin genes or between the 5' flanks of the mouse bh3- and the human γ-globin genes.

These two aspects can be explained by the model in figure 4B, which proposes that the unexpected match between the 5' flanks of the human γ-globin and mouse bh3 genes arose by an unequal crossing-over involving the δ-β intergenic region on one chromosome and the ε-γ intergenic region on the other. This may have been a nonhomologous recombination, or it could have involved an interspersed repetitive element. After deletion of the 3' γ-globin gene, two δ-globin genes would remain with the 5' flank of a γ-globin gene between them, as is retained in the contemporary gene cluster (fig. 4B). To account for the structure of the bh0 gene, the model proposes that the ε-globin gene was duplicated, followed by a relatively recent conversion so that the gene became more like bh1 but still leaving the 3' flanking region similar to
FIG. 4.—Models for the evolution of the mouse β-globin gene cluster. Two different pathways are presented for the generation of the gene cluster in contemporary mice from that of the eutherian ancestor [evidence for the ancestral arrangement is reviewed by Collins and Weissman (1984) and Hardison (1991)]. The pathway in panel A is from Hardies et al. (1984) and Hill et al. (1984). The pathway in panel B incorporates more of the intergenic matches seen in the bottom panel of fig. 2. The segment 5' to the human αγ globin gene that matches with the 5' flank of the mouse bh3 gene is shown as a small box with dark diagonal fill. The timing of the ε gene duplication could be early, as shown in panel B, or it could be later. Although no defects are apparent in the bh0 gene from the Hbbε haplotype (Hill et al. 1984), much of it is deleted in the Hbbε haplotype (Holdener-Kenny and Weaver 1986), showing that it is an expendable gene, if not a pseudogene.
that of the orthologous ε-globin gene. The model in figure 4B does not address the hybrid nature of the bh3 pseudogene, and, indeed, an early inactivation of a β-like globin gene followed by a fusion with a δ-globin gene (fig. 4A) could be added to it. Certainly the evolutionary history of the mouse β-like globin gene cluster is complex, and this discussion has not ventured into the added complications of variations in repeated DNA elements and in gene structures in different haplotypes (Burton et al. 1985; Shyman and Weaver 1985; Holdener-Kenny and Weaver 1986). However, any model should show the bh0 gene as orthologous to ε-globin genes and should account for the remnant of the 5' flank of γ-globin genes found 5' to bh3.

The pattern of aligning sequences among rabbit, human, and mouse β-like globin gene clusters is summarized in figure 5. This is the pattern that would be expected if the contemporary gene clusters were descended from a linked set of ε-γ-η-δ-β genes in the eutherian ancestor. Substantial portions of the intergenic regions still match in the pairwise alignments, frequently interrupted by retroposon repeats.

**Regulatory Regions of β-like Globin Gene Clusters**

The matching regions in pairwise comparisons described above are excellent guides to the evolutionary history of the mammalian β-like globin gene clusters, but to what extent are these sequences actually conserved in different mammals? The percent divergence of some of the flanking regions in the rabbit-human and galago-human comparisons are consistent with their divergence at a rate comparable to a pseudogene rate of 4–5 × 10⁻⁹ substitutions per site per year (Margot et al. 1989; Tagle et al. 1992). However, this is a very rough approximation, given that no theory is in place for handling the large numbers of deletions and insertions that occur in intergenic regions. Although some of the pairwise matches in flanking regions represent the remnants of the ancestral chromosome that still form detectable alignments in two contemporary species, these would be much less likely to align consistently in more than two species. Thus, the program pat was used to search the pairwise alignments among the sequenced mammalian β-like globin gene clusters (human, rabbit, galago, and mouse) for blocks of sequence that consistently align. The results are:

![Fig. 5](https://example.com/figure5.png)

**Fig. 5.** Summary of aligning segments in the mouse, human, and rabbit β-like globin gene clusters. The alignments displayed in fig. 2 are summarized here as filled regions between the maps of the gene cluster in each species. Lighter fill is between a single gene in one species and the second of two duplicated genes in another species. The DNase hypersensitive sites in the LCR are shown as numbered boxes. Other conventions are as in figs. 1 and 2; each gene is shown as a single box without division into exons and introns. The sequences between DNase hypersensitive sites in the rabbit LCR have not been determined completely; further study may reveal more extensive matches.
plotted in figure 6, using the coordinates of the rabbit-versus-human comparison. Each diagonal marks the position of an alignment recovered in all the pairwise comparisons—and hence, one that is a strong candidate for a conserved sequence, i.e., one subject to purifying selection because changes in it place the organism at a selective disadvantage. As expected, the coding regions are displayed prominently, but, what is more interesting, the alignments between orthologous genes extend for a considerable distance from the actual genes (fig. 6). These matches in the flanking regions are not as long as those seen in the comparisons between human and galago or rabbit (fig. 2, top and middle), and those pairwise alignments lost in the multiple alignments are likely to be those that simply retained matches after neutral evolution. The multiple alignments among the ε-globin gene orthologues start ~2 kb on the 5′ side of the gene, run through the gene including the introns, and end in a region about ~1 kb on the 3′ side of the gene (fig. 6). A similar pattern is seen for orthologues to both the γ- and the β-globin genes.

These observations suggest that a substantial region both flanking and internal to the genes contains proximal, cis-acting regulatory sequences, and indeed this is frequently the case (reviewed by Evans et al. 1990). Not only are the proximal promoters located in the first 100 bp on the 5′ side of the gene (Dierks et al. 1983), but enhancers have been recognized on the 3′ side of both the β- and γ-globin genes (Behringer et al. 1987; Bodine and Ley 1987; Kollias et al. 1987; Trudel and Costantini 1987). Important regulatory functions are found in the introns of the β- and γ-globin genes (Donovan-Peluso et al. 1987; Collis et al. 1990), and both positive and negative cis-acting sequences have been mapped on the 5′ side of the ε-globin gene (Cao et al. 1989). In an extensive study of the 5′ flank of the γ-globin gene, conserved sequences have been observed for >1,000 bp on the 5′ side of the cap site, and nuclear factors were found to bind specifically to most of these conserved sites (Gumucio et al. 1991). These proximal, cis-acting sequences are sufficient to direct expression of human γ- and β-globin genes in erythroid cells, at the same time of development as their endogenous mouse orthologues, i.e., embryonic expression of the human γ-globin transgene (Chada et al. 1986) and fetal and adult expression for the human β-globin transgene (Magram et al. 1985). However, expression levels were low, and many transgenic progeny did not express the transgene, presumably as a result of negative position effects at the site of integration in the mouse genome. The consistently aligning blocks shown in figure 6 correlate very well with the extensive regulatory sequences being identified in many laboratories and indicate that regulatory sequences can occupy almost as much DNA as do the exon and intron regions for these relatively small genes.

Additional regulatory sequences are located far upstream from the ε-globin gene, in the LCR (reviewed by Townes and Behringer 1990; Evans et al. 1990; Orkin 1990). Three lines of evidence show that this region is required for high-level, erythroid expression of β-like globin genes. Deletion of the LCR in patients with Hispanic, English, or Dutch γδβ-thalassemia prevented expression of the still-intact, distal β-globin gene, (Townes and Behringer 1990); four prominent developmentally stable DNase hypersensitive sites (HS1–HS4) map in this region in nuclei from erythroid cells (Tuan et al. 1985; Forrester et al. 1986); and addition of the LCR sequences to human globin gene constructs allowed their expression in erythroid cells of transgenic mice, at levels approaching those of the endogenous mouse genes, independently of the site of integration of the transgene in the mouse genome (Grosveld et al. 1987). Thus, at least two functions map within the LCR—enhancement (Tuan et al. 1989) and position independence (Grosveld et al. 1987). The latter function could be quite
Fig. 6.—Plot of consistently aligning sequences among the human, galago, rabbit, and mouse β-like globin gene clusters. The program pab was used to search for consistently aligning blocks of sequences among the pairwise comparisons of human with rabbit, rabbit with galago, galago with human, and human with mouse. The positions of those blocks that align in all four pairwise comparisons are plotted along the coordinates of the rabbit and human sequences. In other words, this plot shows the rabbit-human alignments that remain after filtering them by searching for sequence blocks that also align with goat and mouse; it is a first approach to a four-way alignment among these sequences. It should be noted that the pab results are sensitive to the choice of pairwise alignments used to make the four-way alignment; other groups of pairwise alignments (e.g., human-rabbit, rabbit-galago, galago-mouse, and mouse-human) show less extended matches. The pairwise alignments for this figure were chosen to best illustrate the long segments that do form multiple alignments around the ε-, γ-, and β-globin genes.
complex, involving either a dominant effect of opening a chromosomal domain, a
neutral insulation from negative effects of surrounding sequences (Stief et al. 1989;
Kellum and Schedl 1991), or a combination of both the dominant and neutral effects.
In addition, the LCR doubtless plays an important role in developmental switching,
by interacting with gene-proximal sequences that determine developmental timing
(Enver et al. 1990; Dillon and Grosveld 1991).

As might be expected, the hypersensitive-site regions of the LCR are quite well
conserved in different mammals. The absence of alignments in the LCR in figure 6
results simply from several of the sequence files not extending this far. However, a
large region 5′ to the ε′-globin gene of goats has been sequenced (Li et al. 1990, 1991)
and the positions of alignments with the human sequence (fig. 7) clearly show the
extensive matches in the HS3, HS2, and HS1 regions previously reported (Li et al.
1990). The aligning sequences around HS3 extend for ≳3.5 kb, with the alignment
breaks resulting from insertion of both an Alu repeat in humans and Nla repeats in
goat. The matches around HS2 are more circumscribed, extending for ~1 kb, and a
series of long matches extends through the HS1 region. As pointed out above, matches
in the HS2 and HS1 regions also are apparent in the pairwise comparisons between
human and mouse (Moon and Ley 1990) or rabbit HS regions (fig. 2, middle and
bottom). The locations of the DNase cleavage sites and protein-binding sites have
been mapped with considerable precision in both HS2 (Ney et al. 1990; Talbot et al.
1990) and HS3 (Philipsen et al. 1990) of humans (fig. 7, open boxes). The alignments
extend much farther than the mapped binding sites. Thus, although these minimal,
or core, regions of the hypersensitive sites are sufficient for some aspects of the en-
hancement and position independence conferred by the LCR, the more extended
regions may also play some functional role.

To test whether these extended matches were consistently found in several pairwise
comparisons among different species, the HS3 region from rabbits (authors’ unpub-
lished data) was compared with the human and goat sequence. As has been reported
elsewhere (Boguski et al. 1992), figure 8 illustrates the very extensive matches found
in all three species. The consistently aligning blocks were tested against a small library
of recognition sites for DNA-binding proteins. These are shown in figure 8 as short
lines of negative slope, perpendicular to the main diagonal mapping the alignments.
They are found primarily in the region where protein-binding sites have been mapped
in vitro (Philipsen et al. 1990), with additional matches to both an NFE-2 site im-
mediately 5′ to this region and a CCAAT-binding site located 3′ to a human Alu
repeat. The aligning regions clearly extend beyond these sites. The fact that these long
regions match in all three species argues against the possibility that the matches result
simply from orthologous regions still aligning after neutral evolution. In a rough com-
parison of the LCR versus intergenic regions, in alignments between rabbit and human,
1,946 (64%) of 3,042 bp match in the HS3 region, whereas only 1,245 (41%) of 3,042
bp match in the δ-β intergenic region. Given the inability to apply the considerable
literature on substitution rates for protein-coding regions (Nei 1987, pp. 64–110) to
sequences with large numbers of insertions and deletions, this comparison does not
give a good estimate of the degree of conservation in the hypersensitive-site regions
in the LCR. However, the long, consistently aligning blocks in three species, covering
∼3 kb, argue that these blocks are conserved. A similar four-way analysis of HS2,
using sequences from human, goat, mouse, and rabbit (authors’ unpublished data),
supports a similar conclusion, although the consistently aligning blocks extend for a
shorter region.
Fig. 7.—Plot of aligning sequences between the LCRs of the human and goat β-like globin gene clusters. The human sequence is from Li et al. (1985), and the goat sequence is from Li et al. (1990, 1991). The scoring parameters, analysis, and conventions are as in the top panel of fig. 2; the D repeats in goat are shown as boxes with light fill.
FIG. 8.—Plot of consistently aligning sequence blocks from HS3 of the LCRs of human, rabbit, and goat. The pairwise alignments between human and rabbit (unpublished data) sequences are plotted as the thin lines, and those segments that also align in rabbit-goat and goat-human pairwise comparisons are shown as the thick lines (Boguski et al. 1992). Sequences that match the binding sites for notable DNA-binding proteins are shown as short lines of negative slope perpendicular to the main diagonal. Four of these correspond to protein-binding sites mapped in HS3 by Philipsen et al. (1990), marked by the open box on the horizontal axis. Matches to binding sites for NFE-2 and a CCAAT-binding protein are also found at ~4500 and ~5900, respectively, in the human sequence. The filled triangle represents an Alu repeat.
The diagram in figure 9 summarizes two general features of the patterns of conservation in the β-globin LCRs. First, the hypersensitive sites are in very similar positions relative to the ε-globin gene in rabbit, human, and goat, despite both the many insertions of retroposon repeats and several deletions. This could indicate that an appropriate spacing between the hypersensitive sites is important for full function of the LCR. Second, the conserved sequences extend for a considerable distance beyond the known protein-binding sites for all four hypersensitive sites. The functional dissection of this large, complex regulatory locus is a challenging prospect, and the ability to cull out strong candidates for conserved sequences illustrated here should be a helpful guide in this project.

Comparisons among Mammalian α-Globin Gene Clusters

Since the α- and β-globin genes descended from a common ancestral gene, albeit ~450 Mya, one would hypothesize that coordinate regulation and balanced production were maintained simply by keeping the genes under the same regulatory elements after their duplication and divergence. This is definitely not the case for most mammalian α-like globin genes. Instead, it appears that the α-like genes have settled into a genomic environment radically different from that of the β-like globin genes and have evolved a rather different regulatory mechanism. In fact, α-globin gene regulation may be more similar to that of housekeeping genes than to that of tissue-specific genes, of which the β-globin gene is a paradigm. The surprising differences between α-like and β-like globin genes form an illustrative example of the division of vertebrate genomes into different isochores (Bernardi et al. 1985), very long segments of homogeneous base composition, each with distinct properties.

Sometime prior to the divergence of birds and mammals from their reptilian ancestors, the α-like and β-like globin gene clusters moved to separate chromosomes.
although these genes are still linked in amphibians (reviewed in Collins and Weissman 1984). The β-like globin gene clusters are part of a well-conserved syntenic group, including the genes for Ha-ras and parathyroid hormone in human, mouse, cat, bovine, and rabbit chromosomes (O’Brien et al. 1988; Xu and Hardison 1989). In contrast, the genes linked to the α-globin gene cluster in humans are not linked in mouse (Searle et al. 1987). The human α-globin gene cluster is located close to the telomere of chromosome 16 (Buckle et al. 1988), whereas the mouse α-like globin gene cluster is located internally on chromosome 11 (Davisson et al. 1990). The rabbit α-like globin gene cluster has been mapped close to the end of the long arm of chromosome 6 (Xu and Hardison 1991), in a region predicted to be similar to that of human chromosome 16 (Dutrillaux et al. 1980). The chromosome mapping data suggest that the mammalian β-like globin gene clusters are located in a chromosomal segment that has been relatively stable since the eutherian radiation. This is essentially a large-scale view of the stability evident at the nucleotide level, on the basis of the sequence comparisons presented above. In contrast, the α-globin gene clusters seem to be in a more unstable chromosomal segment, one that is quite different between human and mouse but that is similar between human and rabbit.

The rabbit α-like globin genes form a complex array (fig. 10), generated by a series of block duplications of gene sets consisting initially of ζ-ζ-α-θ (Cheng et al. 1987). A characteristic J sequence is found at the junctions of the duplicated blocks (Cheng et al. 1987). The J sequence is a curious hybrid, beginning with a retroposon C repeat but also containing internal and 3′ flanking sequences related to those of α-globin genes (Hardison et al. 1991). The J sequence has been proposed as a hot spot for recombination involved in the deletion of an α-globin gene, the duplication of ζ-globin genes, and the expansion and contraction of the rabbit α-like globin gene cluster by changing the numbers of ζ-ζ-θ blocks. In support of this hypothesis, part of the J-sequence matches (Hardison et al. 1991) with a sequence that is on the 3′ side of human α-globin genes and that can enhance recombination in cultured mammalian cells (Hu and Shen 1987). Figure 10 also shows the many C repeats that have inserted into this gene cluster, often recursively with younger retroposons inserting into the older repeats (Krane et al. 1991).

The human and rabbit α-like globin gene clusters are unusually G+C rich and contain several prominent CpG islands (Bird et al. 1987; Fischel-Ghodsian et al. 1987; Hardison et al. 1991). The G+C content hovers around 65%–70% for most of the
rabbit α-like globin gene cluster, whereas the average G+C content of the rabbit β-like globin gene cluster is ~45% (Margot et al. 1989) and CpG islands are not present (Hardison et al. 1991). The rabbit α-like globin gene cluster has two CpG islands, one that extends from ~300 bp on the 5' side of the cap site to the second intron of the α-globin gene and another that extends from the 5' flank through much of the θ-globin pseudogene (fig. 11). One long CpG island in the human α-globin gene cluster covers parts of two pseudogenes, ζ1 and ψα2, and other CpG islands encompass each of the active α-globin genes, α2 and α1, as well as the θ1-globin gene (Fischel-Ghodsian et al. 1987; fig. 11). In contrast, the mouse α-globin gene does not have the properties of a CpG island (Whitelaw et al. 1989), resembling more the β-like globin genes in being relatively A+T rich and showing a lower than expected frequency of CpG dinucleotides.

Comparison of a block of sequence containing the rabbit ζ1-, α-, and θ-globin genes with the sequence of the human α-like globin gene cluster reveals a pattern of alignments quite different from that of the β-like globin gene clusters (fig. 11). Each of the α-like globin genes matches with all the others, with the longest matches involving the respective ζ-, α-, and θ-globin genes from each cluster. (Note the smaller scale, compared with plots of comparisons between β-like globin genes.) These latter matches extend for short distances into the flanking regions of the genes, whereas matches between, e.g., rabbit α- and human ζ1-globin genes are limited mainly to the coding regions. A quantitative analysis of the number of matches in all pairwise comparisons among the genes shows that the rabbit α-, θ-, and ζ-globin genes are each most like the human α-, θ-, and ζ-globin genes, respectively. On the basis of these considerations and the matches in the proximal flanks of the genes, it is likely that the rabbit α- and human ζ1-globin genes are orthologous, as are the rabbit and human α-globin genes and the rabbit and human θ-globin genes (Hardison and Gelinas 1986; Hardison et al. 1991).

Additional matches are seen in figure 11, between each of the CpG islands; these appear as a set of adjacent diagonals, indicating tandemly repeating sequence motifs (Proudfoot ct al. 1982; Schon et al. 1982; Cheng et al. 1986). Also, the rabbit J sequence matches with intron 1 of human ψα1, the 3' untranslated region, and the 3' flank of the α-globin genes and a segment of the X homology block in humans, leading to the proposal that the rabbit J sequence was derived in part from the ancestor to the human ψα1-globin gene (Hardison et al. 1991). Matches in the flanking sequences indicate that the human ψα1 gene is orthologous to the goat α- and horse α2-globin genes (fig. 1; Hardison and Gelinas 1986; Sawada and Schmid 1986; Flint et al. 1988), but no full-length homologue to human ψα1 is found in rabbit. Perhaps, after extensive deletions and fusions in rabbits, it evolved into the J sequence.

Other than the aforementioned regions, almost no matching sequences are found in the intergenic regions separating the orthologous genes (fig. 11). This contrasts with the long matches seen in the intergenic comparisons of mammalian β-like globin gene clusters (fig. 2). Given that the gene clusters are being compared in the same two species and that the gene products are interacting subunits of the hemoglobin tetramer (and hence subject to comparable, if not identical, selective pressure), one has to conclude that the intergenic regions are evolving at a faster rate for mammalian α-globin gene clusters than for the β-like globin gene clusters. This faster rate applies to the intergenic regions, not to the coding regions, which show amounts of divergence similar to those of the β-like globin genes (Efstratiadis et al. 1980).

The human and rabbit α-like globin gene clusters differ from mammalian β-like
FIG. 11.—Plots of aligning segments between the rabbit and human α-globin gene clusters. Local alignments were generated with the program sim using the parameters match = 1, mismatch = -1, gap penalty = 4, and gap extension penalty = 0.4 per position in the gap. The cutoff value of τ was chosen so that the probability was 0.8 that random sequences matching the given sequences in length and nucleotide composition would form a gap-free alignment scoring at least τ (i.e., with random sequences, the plot would contain one or two species). Note that this is a more relaxed cutoff than was used for the β-like globin gene comparisons and that much shorter sequences are being compared (one-quarter to one-half the size of the smaller β-like globin gene clusters). Local alignments scoring above the threshold τ were plotted using the program laps. X, Y, and Z homology blocks mark the duplication region for the α2- and α1-globin genes in humans (reviewed in Collins and Weissman 1984). Positions of CpG islands are shown as open boxes. Short interspersed repeats (human Alu and rabbit C) are shown as filled triangles. I sequences are shown as open arrows. This figure is from Hardison et al. (1991) and is reprinted with permission.
globin gene clusters in many other ways. The α-like globin gene clusters have also sustained insertions only by Alu-like short interspersed repeats with no L1 repeats (Nicholls et al. 1987; Krane et al. 1991). The α-globin genes are not methylated in any tissues examined (Bird et al. 1987), whereas the β-like globin genes are more methylated in nonexpressing tissues (Shen and Maniatis 1980; van der Ploeg and Flavell 1980). The α-like globin genes are replicated early in S phase in both expressing and nonexpressing cell lines (Goldman et al. 1984), whereas the β-globin genes, like most tissue-specific genes, are replicated early in S phase in expressing tissues but late in S phase in nonexpressing tissues (Epner et al. 1988). These are many of the functional differences between sequences in different isochores (Bernardi et al. 1985; Bernardi 1989), and, in fact, the human and rabbit globin gene clusters map to different isochores. The α-like globin gene cluster is in the A+T-rich L1 isochore, containing the bulk of mammalian genomic DNA, whereas the α-like globin gene cluster is in the very dense, highly G+C-rich H3 isochore that constitutes a small part of the genome but is rich in genes (Bernardi et al. 1985). This H3 isochore maps to the T bands found at the telomeres of many mammalian chromosomes (Saccone et al. 1992), just as the human and rabbit α-globin genes map close to the telomeres of their chromosomes. It is interesting that the mouse genome has no H3 isochore component (Salinas et al. 1986). Since the mouse is the only known example of a mammal whose α-globin genes lack a CpG island, one may tentatively propose that the presence of α-globin genes in the telomeric, G+C-rich isochores is the ancestral arrangement in mammals.

The loss of the CpG island may have occurred as the mouse α-globin gene evolved into a different genomic context, perhaps concomitantly with its movement to the internal region of the chromosome and during the general loss of the H3 isochore from the mouse genome. This movement of the mouse α-globin gene would explain the absence of common syntenic loci close to this gene in mouse and human.

The differences between α-like and β-like globin gene clusters specifically and between G+C-rich and A+T-rich isochores in general are all related to various aspects of DNA metabolism—replication timing, DNA methylation, insertions of different classes of retroposon repeats, DNA repair, and recombination. One can postulate different compartments in the nucleus, either functionally or physically separated, wherein the G+C-rich isochores are acted on by enzymatic machineries different than those acting on the A+T-rich isochores. Hence, these differences between α-like and β-like globin gene clusters may not be related so much to gene function as to their radically different genomic contexts. These pronounced differences in the genomic context of the α-like and β-like globin genes are likely to have an important effect on the regulation of the genes. Early studies showed that these genes showed different properties when expressed transiently in transfected nonerythroid cell lines—the β-globin gene required a viral enhancer in cis for expression (Banerji et al. 1981), whereas the α-globin gene was expressed without an added enhancer (Mellon et al. 1981). When stably integrated into mouse erythroid cell lines, a DNA segment containing the human β-globin gene was inducible along with the endogenous mouse genes, in contrast to the human α-globin gene, which was constitutively expressed prior to induction (Charnay et al. 1984). In general, DNA fragments containing the human (Mellon et al. 1981) or rabbit (Cheng et al. 1986) α-globin genes are expressed in an essentially deregulated manner in many different cell types. In contrast, the mouse α-globin gene is not expressed promiscuously (Whitelaw et al. 1989), suggesting a correlation between the presence of a prominent CpG island and deregulated expression in transfected cells. Fusions between α- and β-globin genes have mapped the sequences...
responsible for this deregulated expression of the human α-globin gene 3' to the start
codon (Charney et al. 1984).

Recent efforts to map the sequences that allow the enhancer-independent expression
of α-globin genes in transfected cells have confirmed a role for internal regions.
Experiments with both the human (Brickner et al. 1991) and rabbit (Yost et al. 1991;
M. James-Pederson, S. E. Yost, and R. Hardison, unpublished data) α-globin genes
have shown that a DNA segment extending from the proximal 5' flank into exon 2 is
required for expression in transfected erythroid and nonerythroid cells. The internal
sequences of the α-globin gene do not serve as enhancers of β-globin gene promoters,
and it is plausible that this internal region is effective only in its natural position, i.e.,
downstream from the proximal 5' flank. These cis-acting control sequences are located
within the CpG island, and one can propose that the CpG island with appropriate
binding sites for transcription factors forms an extended (i.e., encompassing both 5'
flanking and internal sequences), enhancer-independent promoter that by itself will
transcribe promiscuously in a variety of cell types.

However, in the proper chromosomal context, this α-globin gene is expressed
only in erythroid cells (Deisseroth and Hendrick 1978), implicating more distal se-
quences for gene control. Evidence from deletional α-thalassemias (Wilkie et al. 1990),
mapping of DNase hypersensitive sites, and functional tests (Higgs et al. 1990; Jarman
et al. 1991) have shown that a major regulatory element, HS-40, is located ~40 kb
5' to the human ζ2-globin gene. This HS-40 sequence will confer tissue-specific, po-
sition-independent expression on linked human α-globin genes in transgenic mice
and in erythroid cell lines (Higgs et al. 1990). Thus, it has many of the properties of
an LCR, although it does not produce the copy-number-dependent expression that is
one criterion for LCR function (Grosveld et al. 1987). Previous attempts to express
the human α-globin gene in transgenic mice were unsuccessful, even when large seg-
ments of DNA (but not HS-40) from the α-globin gene cluster were used (Palmiter
and Brinster 1986). Thus it is possible that the CpG island containing the α-globin
gene was inactivated at a particular developmental stage in the transgenic mouse,
possibly by methylation, but that the presence of the HS-40 sequence prevents this
gene inactivation in erythroid cells. Apparently a comparable inactivation does not
occur when the α-globin genes are transfected into cell lines, resulting in constitutive,
deregulated expression (Charmay et al. 1984). But, in the presence of HS-40, the α-
globin gene is appropriately regulated in transfected mouse erythroid cells and is now
capable of high-level expression (Higgs et al. 1990). As was seen for the β-like globin
genes, the interaction between distal (HS-40) and the proximal control sequences (a
CpG-island that forms an extended promoter) is critical for proper regulation of the
α-globin genes. But, in this case, not only does the HS-40 sequence apparently rescue
the transgene from inactivation in transgenic mice; it modulates its otherwise dereg-
ulated expression in transfected cells. Hence, the distal regulatory sequences of the β-
like and α-like globin gene clusters have several features in common, but, at this early
stage in the analysis, some regulatory features appear to be distinctive for each.

Concluding Remarks

Analysis of long sequence alignments is a useful guide to understanding complex
pathways of genomic evolution, such as the evolution of multigene families from
repeated rounds of gene duplication and divergence, complicated by repeated gene
conversions. Comparison of sequences from more than two species is an avenue to
identifying noncoding sequences that are under selective pressure, as opposed to se-
quences that still match after neutral evolution. Such analysis should help in dissecting and understanding the several layers of regulation exerted by the distal locus control sequences. Evolutionary rates and modes can vary dramatically in different species and in different regions of the genome, and long sequence alignments have provided important information for understanding the very different evolutionary histories of mammalian α- and β-globin genes. Similarly, these sequence comparisons provide a helpful perspective on the paradox of different regulatory mechanisms of the α- and β-globin genes, whose expression must be balanced in erythroid cells.

Availability

The programs used in this paper are available by anonymous ftp from groucho.cs.psu.edu.

Acknowledgments

R.H. was supported by PHS grants DK 27635 and HL 44491, and W.M. was supported by National Library of Medicine grant RO1 LM05110. Comments from the reviewers helped clarify the presentation.

LITERATURE CITED


CHARNAY, P., R. TREISMAN, P. MELLON, M. CHAO, R. AXEL, and T. MANIATIS. 1984. Dif-


ENVER, T., N. RAICH, A. J. EBENS, T. PAPAYANNOPOULOU, F. COSTANTINI, and G. STAMA-


WERNKE, S. M., and J. B. LINGREL. 1986. Nucleotide sequence of the great embryonic \( \alpha \)-globin
gene (ζ) and linkage and evolutionary analysis of the complete α globin cluster. J. Mol. Biol. 192:457-471.


Masatoshi Nei, reviewing editor

Received June 30, 1992; revision received August 12, 1992

Accepted August 13, 1992