CHAPTER 3

The Normal Structure and Regulation of Human Globin Gene Clusters

Bernard G. Forget and Ross C. Hardison

The genes encoding the different globin chains of hemoglobin are members of an ancient gene family. In this chapter we will review the structural features of the globin genes, with particular attention to the sequences needed for proper regulation of gene expression. Some of these have been well- conserved during mammalian evolution and therefore are likely to provide a common function in many mammals. Others are only found in higher primates, and may play roles in lineage-specific regulation. We will first describe the structural characteristics of the human globin genes and then provide a comparative analysis of the genomic contexts, regulatory regions and evolutionary conservation of features present in the globin gene clusters.

NUMBER AND CHROMOSOMAL LOCALIZATION OF HUMAN GLOBIN GENES

Hemoglobin is a heterotetramer that contains two polypeptide subunits related to the α-globin gene subfamily (referred to here as α-like globins) and two polypeptide subunits related to the β-globin gene subfamily (β-like globins). Globin polypeptides bind heme, which in turn allows the hemoglobin in erythrocytes to bind oxygen reversibly and transport it from the lungs to respiring tissues. In humans, as in all vertebrate species studied, different α-like and β-like globin chains are synthesized at

progressive stages of development to produce hemoglobins characteristic of primitive (embryonic) and definitive (fetal and adult) erythroid cells (Figure 3.1).

Before precise knowledge of globin gene organization was gained by gene mapping and molecular cloning, a general picture of the number and arrangement of the human globin genes emerged from the genetic analysis of normal and abnormal hemoglobins and their pattern of inheritance. The number and subunit composition of the different normal human hemoglobins (Figure 3.1) suggested that there must exist at least one globin gene for each of the different globin chains: α, β, γ, δ, ε, and ζ. Evidence from the study of hemoglobin variants and the biochemical heterogeneity of the chains in fetal hemoglobin (HbF) showed that the α- and γ-globin genes were duplicated. Persons were identified whose red cells contained more than two structurally different α-globin chains that could be best explained by duplication of the α-globin gene locus, and the characterization of the structurally different $^{G}\gamma$ and $^{A}\gamma$ globin chains of HbF imposed a requirement for duplication of the γ-globin gene locus.

Studies of the pattern of inheritance of hemoglobin variants from persons carrying both an α chain and a β chain variant revealed that the α- and β-globin genes are on different chromosomes (or very widely separated if on the same chromosome). Variants of α-globin and β-globin chains were always observed to segregate independently in offspring of doubly affected parents (reviewed in 1). Linkage of the various β-like globin genes to one another was established from the study of interesting hemoglobin variants that contained fused globin chains, presumably resulting from nonhomologous crossover between different β-like globin genes. Characterization of Hb Lepore (2), with its δβ fusion chain, established that the δ-globin gene was linked to and located on the 5' (or N-terminal) side of the β-globin gene. Analysis of Hb Kenya (3), with its $^{A}\gamma\beta$ fusion chain, provided evidence for linkage of the $^{A}\gamma$ gene, and presumably the $^{G}\gamma$ gene as well, to the 5'-side of the δ- and β-globin genes.

Thus, the general arrangement of the globin genes that emerged from these various genetic analyses can be represented as illustrated in Figure 3.1. It was also assumed, but unsupported by genetic evidence, that the embryonic α-like (ζ) and β-like (ε) globin genes were likely to be linked to the loci encoding their adult counterparts.

By using rodent-human somatic hybrid cells containing only one or a few human chromosomes, Deisseroth and colleagues (4, 5) clearly established that the human α- and β-globin genes resided on different chromosomes. The α-like globin genes are located on chromosome 16, whereas the β-like globin genes are on chromosome 11. The latter results were obtained by hybridizing a solution of total cellular DNA from the various somatic hybrid cells to radioactive cDNAs, synthesized from α- and β-globin mRNAs by reverse transcriptase. These results were later confirmed and extended by various groups using the gene mapping procedure of Southern blot analysis with DNA from various hybrid cell lines containing different translocations or deletions of the involved chromosomes.

These studies also localized the globin gene loci to specific regions on their respective chromosomes: the β-globin gene cluster to the short arm of chromosome 11, and the α-globin gene cluster to the short arm of chromosome 16 (Figure 3.1). These chromosomal assignments were further confirmed and refined by *in situ* hybridization of radioactive cloned globin gene probes to metaphase chromosomes and by fluorescence-based *in situ* hybridization (FISH). Thus, the β-globin gene cluster was assigned to 11p15.5 and the α-globin gene cluster to 16p13.3. Subsequent DNA sequencing of entire human chromosomes and alignment with maps of chromosome bands places the β-globin gene cluster in 11p15.4. The α-globin gene cluster is only about 150 kilobase pairs (kb) from the telomere of the short arm of chromosome 16.

GLOBIN GENE STRUCTURE: INTRONS AND THEIR REMOVAL

The coding region of each globin gene in humans and other vertebrates is interrupted at two positions by stretches of noncoding DNA called intervening sequences (IVSs) or introns (6). In the β-like globin genes, the introns interrupt the sequence between codons 30 and 31 and between codons 104 and 105; in the α-globin gene family, the intervening sequences interrupt the coding sequence between codons 31 and 32 and between codons 99 and 100 (Figure 3.2.A). Although the precise codon position numbers at which the interruption occurs differ between the α- and β-like globin genes, the introns occur at precisely the same position in the aligned primary sequence of the α-

and β-globin chains. Thus, given the likely possibility that the α- and β-globin gene families originally evolved from a single ancestral globin gene (7), these gene sequences are homologous, and we infer that the presence of the introns at these positions predates the separation of α-globin and β-globin genes about 500 million years ago (in an ancestral jawed vertebrate). The first intervening sequence (IVS-1) is shorter than the second intervening sequence (IVS-2) in both α- and β-globin genes, but IVS-2 of the human β-globin gene is much larger than that of the α-globin gene (Figure 3.2.A).

The pattern of intron sizes of the ζ-like globin genes differs from that of the other α-like globin genes. Whereas the introns in the α and ψα genes are small, e.g. fewer than 150 base pairs (bp), those of the ζ and ψζ genes are larger (8). Furthermore, the first introns of the ζ and ψζ genes are much larger than their second introns; in fact they are eight to ten times larger than the first introns of any other globin gene.

The presence of intervening sequences that interrupt the coding sequences of structural genes imposes a requirement for some cellular process to remove these sequences in the mature mRNA. As illustrated in Figure 3.2.B, intervening sequences are transcribed into globin (and other) precursor mRNA molecules (9), but they are subsequently excised and the proper ends of the coding sequences joined to yield the mature mRNA (10). This posttranscriptional processing of mRNA precursors to remove introns has been termed splicing. A crucial prerequisite for the proper splicing of globin (and other) precursor mRNA molecules is the presence of specific nucleotide sequences at the junctions between coding sequences (exons) and intervening sequences (introns). Comparison of these sequences in many different genes has permitted the derivation of two different consensus sequences, which are almost universally found at the 5' (donor) and 3' (acceptor) splice sites of introns (11, 12). The consensus sequences thus derived are shown in Figure 3.2A, along with the consensus surrounding the branch point A involved in the initiation of splicing. The dinucleotides GT and AG shown in boldface, at the 5' and 3' ends, respectively, of the intron, are essentially invariant and are thought to be absolutely required for proper splicing. This is the so-called GT-AG rule. Rare examples have been described in which GC instead of GT is found at the donor splice site junction.

The importance of these consensus sequences is underscored by the fact that mutations that either alter them or create similar consensus sequences at new sites in a globin gene can lead to abnormal processing of globin mRNA precursors; these constitute the molecular basis for many types of thalassemia (Chapters 13 and 16). Throughout this chapter we will refer to human mutations that affect some aspect of the pathway for gene expression. Readers desiring more information may want to use databases such as HbVar (http://www.bx.psu.edu) (13) or the Locus Variants track on the UCSC Genome Browser (http://genome.ucsc.edu) (14) to find positions, genotypes and phenotypes for the greater than 1000 known globin gene variants.

DETAILED CHROMOSOMAL ORGANIZATION OF THE HUMAN GLOBIN GENES

A precise picture of the chromosomal organization of the α- and β-like human globin gene clusters, with respect to the number of structural loci and intergenic distances, was obtained by a number of different techniques: (1) restriction endonuclease mapping of genomic DNA (e.g. 15, 16) using the gel blotting procedure of Southern (17) and, (2) gene isolation and sequencing using recombinant DNA technology (e.g., 18). Sets of overlapping genomic DNA fragments spanning the entire α- and β-globin gene clusters were obtained by gene cloning, initially in bacteriophage λ and larger fragments in cosmid vectors. Detailed analysis of these recombinant DNA clones and complete DNA sequencing led to the determination of the gene organization illustrated in Figure 3.3. Some results were expected, such as the finding of single δ- and β-globin gene loci and duplication of the α- and γ-globin gene loci. In addition, single loci for the embryonic ζ and ε globin chains were found linked to the α- and β-globin gene clusters, respectively. It is noteworthy that the genes in each cluster are in the same transcriptional orientation and are arranged, in a 5' to 3' direction, in the same order as their expression during development.

An unexpected finding was the presence in the globin gene clusters of additional gene-like structures with sequence homology and an exon-intron structure similar to the actively expressed globin genes. These DNA segments have been called pseudogenes (19). One, called ψβ1, is in the β-like globin gene cluster between the γ- and δ-globin genes. At least two (and possibly four) are in the α-like globin gene cluster. The two clear examples are ψζ1 and ψα1, located between the active ζ-globin and α-globin genes (Fig. 3.3). All three (ψβ1, ψζ1, and ψα1) are characterized by the presence of one or more mutations that render them incapable of encoding a functional globin chain. This inability to encode a functional globin polypeptide does not necessarily render the pseudogenes inactive for transcription. The pseudogene ψβ1 is transcribed and spliced, as shown by several spliced ESTs (expressed sequence tags), whereas no evidence has been provided that ψα1 is transcribed. These pseudogenes appear to have arisen by gene duplication events within the globin gene clusters followed by mutation and inactivation of the duplicated gene and subsequent accumulation of additional mutations through loss of selective pressure.

Two other α-like globin genes have been identified and characterized in the α-globin gene cluster, but their roles, if any, in encoding globin polypeptides are still uncertain. The θ-globin gene located to the 3' or C-terminal side of the duplicated α-globin genes (20). It is more closely related to the α-globin genes than to the ζ-globin genes and is expressed at low levels in erythroid cells (21, 22). Clear homologs to the θ-globin gene are found in the homologous position in other mammalian α-like globin gene clusters. The μ-globin gene is located just 3' of the ψζ1-globin pseudogene (23, 24); it was initially called ψα2 (25) but with more accurate sequencing it is clear that this gene does not contain mutations that would render it inactive. It is a distant relative, being equally divergent from both α-globin and ζ-globin genes. Its closest relatives are the α^D-globin genes, which are actively expressed in red cells of reptiles and birds (24, 26). DNA sequences similar to that of the human μ-globin gene are found in other mammals, but in some species, such as mouse, the sequence has diverged so much that no obvious gene structure is found. Thus the presence of the θ-globin gene is conserved in all mammals examined but the μ-globin gene has been lost in some but not all lineages. Both

the θ-globin gene and the μ-globin gene are transcribed and spliced in erythroid cells, albeit at much lower levels than the α-globin gene. Curiously, no hemoglobin containing the θ-globin chain or the μ-globin chain has been identified, even by sensitive mass spectrometry (23). Furthermore, the predicted structure (translated amino acid sequence) of the θ-globin chain suggests that it would be unlikely to function normally as a hemoglobin subunit(27). Thus these genes remain a puzzle. They tend to be retained over mammalian evolution, hence indicating constraint for some function. They are expressed at the RNA level but do not appear to be translated into a polypeptide. Perhaps they or their RNA transcripts play some role that has yet to be discovered.

GENOMIC CONTEXT OF THE α-GLOBIN AND β-GLOBIN GENE CLUSTERS

The separation of α- and β-globin gene clusters to different chromosomes has allowed them to diverge into strikingly different genomic contexts, with paradoxical consequences for our understanding of their regulation. Given that all contemporary vertebrates have developmentally regulated hemoglobin genes encoding proteins used for oxygen transport in erythrocytes, it would have been reasonable to expect that the molecular mechanisms of globin gene regulation would be conserved in vertebrates. Certainly, the coordinated and balanced expression of α- and β-globin genes to produce the heterotypic tetramer $\alpha_2\beta_2$ in erythrocytes should be a particularly easy aspect of regulation to explain. Because the two genes would have been identical after the initial duplication in the ancestral vertebrate, with identical regulatory elements, it is parsimonious to expect selection to keep the regulatory elements very similar.

However, much has changed between the α- and β-like globin gene clusters since their duplication. Not only are they now on separate chromosomes in birds and mammals, but in mammals they are in radically different genomic contexts (28). A major determinant of the genomic environment is the G+C content. A G+C rich DNA segment has a high mole fraction of the nucleotides guanidylic acid (G) and cytidylic acid (C), whereas an A+T rich DNA segment has a high mole fraction of the nucleotides adenylic acid (A) and thymidylic acid (T). The G+C content for the human genome on average is

low (about 41%) but some segments can be much lower or higher, ranging from 30% to 65% in 20kb windows (29). Regions that are G+C rich tend to be enriched in genes, and those genes tend to be expressed in a broad range of tissues. They also tend to have islands with an abundance of the dinucleotide CpG (30). This is in stark contrast to the bulk of the genome, which has very few CpGs because these are the sites for DNA methylation, and substitution of CpG to TpG or CpA is very rapid on an evolutionary timescale (as much as 10 times faster than the rates of other substitutions). The CpG islands are thus short regions (a few hundred bp) in which the CpG dinucleotides are not methylated; these have been associated with important functions such as promoters for transcription.

The β-globin gene clusters in humans and other mammals are A+T rich, with no CpG islands (31), whereas the α-like globin gene clusters are highly G+C rich, with multiple CpG islands (32). This correlates with several important differences in the structure and regulation of the two gene clusters. Tissue-specific gene expression of the β-like globin genes is correlated with an increased accessibility of the chromatin only in expressing cells (33), and hence “opening” of a chromatin domain is a key step in activation of these genes. In contrast, the α-like globin genes, which are in constitutively open chromatin (28). The β-globin gene cluster is subject to tissue-specific DNA methylation(34), but, in keeping with the presence of CpG islands, the α-globin gene cluster is not methylated in any cell types (35). The β-globin gene clusters are replicated early in S phase only in cells expressing them, whereas the human α-globin genes are replicated early in all cells (36-38). Thus the mammalian α-globin genes have several characteristics associated with constitutively expressed "housekeeping" genes. The strikingly different genomic contexts of the two gene clusters affect several aspects of DNA and chromatin metabolism, including timing of replication, extent of methylation, and the type of chromatin into which the loci are packaged. Rather than selecting for similarities to insure coordinate and balanced expression, the processes of evolution at these two loci have made them quite different. The full implications of these differences may not yet be known. For instance, the two “healthy” genes with no known function in

the α-like globin gene cluster, θ and μ, are themselves CpG islands. Could this be a clue to a role for these genes outside the conventional one of coding for proteins?

The types of genes that surround the α-like and β-like globin gene clusters are quite different (Figure 3.3). The β-like globin gene cluster is surrounded by olfactory receptor (*OR*) genes, which encode G-protein coupled receptors expressed in olfactory epithelium (39). Several *OR* gene clusters containing about a thousand genes and pseudogenes are found in the human genome. The *OR* gene cluster surrounding the β-like globin genes is a particularly large one, with about a hundred genes extending almost 1 million bp (Mb) past *HBB* (the β-globin gene) and over 3 Mb toward the centromere from *HBE1* (the ε-globin gene). This arrangement is found in homologous regions in mammals and in chickens. Thus the erythroid-specific regulation of the β-like globin gene cluster is exerted in a chromosomal environment that is largely devoted to olfactory-specific expression. Perhaps this has had an impact on selection for a particularly powerful enhancer, to override the olfactory-specific regulation. As shown in Figure 3.3.A, several deletions causing β-thalassemia or Hereditary Persistence of Fetal Hemoglobin (HPFH) not only remove β-like globin genes, but they also fuse the remaining genes with sequences close to an *OR* gene. The phenotype of patients carrying such deletions may be explained in part by bringing positive or negative regulatory elements normally associated with *OR* genes into proximity of the β-like globin genes (40, 41).

In contrast, the α-like globin genes are surrounded by a variety of genes (Figure 3.3.B), many of which are widely expressed and carry out fundamental roles in cellular metabolism and physiology, such as *MPG* (encoding the DNA repair enzyme methyl purine glycosylase) and *POLR3K* (encoding a subunit of RNA polymerase III) (42). Although the α-like globin gene cluster and surrounding DNA is in constitutively open chromatin, histones are hyperacetylated (another mark of active loci) in erythroid cells in a more restricted region encompassing the globin genes and their regulatory sequences (43). Although the regions homologous to that surrounding the α-like globin gene cluster have undergone inter- and intra-chromosomal rearrangements in various vertebrates lineages, the genes from *POLR3K* through *HBQ1* have remained together in all species

examined from fish to mammals (44). This suggests that this region encompasses all the sequences needed in *cis* for appropriate regulation of the α-like globin genes.

Despite these many differences between α-like and β-like globin gene clusters in mammals, the appropriate genes are still expressed coordinately between the two loci, resulting in balanced production of α-like and β-like globins needed for the synthesis of normal hemoglobins. The mechanisms that accomplish this task still elude our understanding.

One important aspect that is common to the genomic contexts of both gene clusters is the presence of distal strong enhancers. The discovery of these was aided by mapping of deletions that result in β-thalassemia or α-thalassemia, which are inherited deficiencies in the amount of β-globin or α-globin, respectively (see Chapters 13 and 16). Some of these deletions removed distal sequences but retained all the globin genes, e.g. the deletions associated with Hispanic $(\varepsilon\gamma\delta\beta)^0$ thalassemia and the Ti~ α^0 thalassemia (Figure 3.3). Within the deleted intervals are critical long-range enhancers needed for high level expression of any gene in the linked globin gene clusters. These are the locus control region (LCR) for the β-globin gene cluster and HS-40 or major regulatory element (MRE) for the α-globin gene cluster. Thus regulation of expression of globin genes involves DNA sequences both close to the genes (proximal) and as much as 70 kilobase pairs (kb) away from the genes (distal). These will be examined in more detail in the next section.

EVOLUTIONARY INSIGHTS INTO REGULATION OF GLOBIN GENE CLUSTERS

Motivation

One avenue for improving the conditions of patients with hemoglobinopathies could involve regulation of expression of the globin genes. This hope is based on the normal human variation in phenotypes presented for a given mutant genotype. For

example, patients with naturally higher concentrations of HbF ($\alpha_2\gamma_2$) in their erythrocytes tend to have milder symptoms of either sickle cell disease or thalassemia (Chapters 17 & 19). The α-globin gene status can affect the severity of β-thalassemia, with more balanced production of α-globin and β-globin associated with milder disease. Thus considerable effort has gone into studying the stage-specific expression of the globin genes, with a long-term goal of enhancing or restoring production of embryonic or fetal hemoglobins in adult life or reducing expression of deleterious alleles. Although no current treatment by gene therapy is in practice as of this writing, much effort continues in this area. The use of hydroxyurea in treatment of sickle cell disease is an outgrowth of studies on mechanisms of regulation of globin genes. Current studies aim to discover more sophisticated and directed pharmacological methods for enhancing production of embryonic and fetal hemoglobins.

Studies over the past three decades have revealed much about the regulation of the human globin genes. In this section, we will summarize some of the information about DNA sequences needed in *cis* (i.e. on the same chromosome) for regulation of the globin genes. Chapter 4 will cover the proteins interacting with these regulatory DNA sequences.

Common versus lineage-specific regulation

Comparison of noncoding genomic DNA sequences among related species is a powerful approach to identifying and better understanding *cis*-regulatory modules (CRMs). However, it is important to distinguish what is similar and what is distinctive about the patterns of regulated expression of the genes in the species being compared. If one is searching for CRMs that carry out a function common to most or all mammals, then conservation across all mammals and evidence of strong constraint in noncoding DNA will provide good candidates for further experimental tests (e.g. 45, 46, 47). Such constrained noncoding sequences can have within them short, almost invariant regions that frequently correspond to transcription factor binding sites. These have been called phylogenetic footprints (48). However, if one is studying a type of regulation that only

occurs in higher primates, then searching for sequences conserved in other mammalian orders will be futile. Instead, the search should focus on sequences conserved in the species with a common mode of regulation but which differ from the homologous regions in species with a different regulation. These have been called differential phylogenetic footprints (49).

Regulatory features of globin genes common to many vertebrate species include tissue specificity and some aspects of developmental specificity. Expression of the α-like and β-like globin genes in all vertebrate species examined is restricted to the erythroid lineage. Thus some determinants of tissue specificity should be common to all these genes. One example is binding by the transcription factor GATA-1. As will be detailed in the following sections, either the promoter, enhancers or both for all globin genes have binding sites for GATA-1. Another feature common to all mammals is the expression of the ε-globin and ζ-globin genes exclusively in primitive erythroid cells, which are produced during embryonic life. Thus one might expect determinants of embryonic expression to be conserved in many species. Indeed, conservation of the upstream promoter regions of these genes in eutherian mammals is more extensive than is seen for other promoters in their globin gene clusters (50).

An example of lineage-specific regulation is the recruitment of the γ-globin genes for expression in fetal erythroid cells. In most eutherian mammals, the γ-globin genes are expressed in primitive erythroid cells, similar to the ε-globin gene, and the β-globin gene is expressed in definitive erythroid cells both during fetal and adult life. However, simian primates, including humans, express the γ-globin genes during fetal erythropoiesis, and the expression of the β-globin gene is delayed. The extent of delay varies in different primate clades, but in humans it is largely delayed until just before birth. Thus when examining interspecies alignments of the regulatory regions of the β-globin gene (*HBB*) and the γ-globin genes (*HBG1* and *HBG2*), one will be seeing a combination of CRMs used in common (e.g. for adult erythroid expression of *HBB*) and in a lineage-specific manner (e.g. fetal expression of *HBG1*).

Quantitative analysis of sequence alignments

Alignments of genomic DNA sequences reveal the segments that are similar between species, and often these reflect homology (descent from a common ancestor). These sequence matches tend to have highest similarity in the protein-coding exons, but significant stretches of noncoding sequences also align between mammalian species (for globin gene complexes, see (51-53). Further analysis is required to discern which sequence matches simply reflect common ancestry (aligned neutral DNA) versus those in sequences that are under constraint (sequences with a common function) (54, 55).

Several bioinformatic tools have been developed to help interpret the alignments of multiple sequences. Results from two of these, each analyzing alignments of several mammals (human, chimpanzee, rhesus macaque, mouse, rat, dog, cow, and sometimes additional ones), are shown in Figure 3.3. The Conservation track plots the phastCons score at each position of the human sequence. This score is an estimate of the posterior probability that a given nucleotide is in the most strongly constrained (i.e. most slowly changing) portion of the genome (56). Higher scores are associated with a greater likelihood that a position or region is under strong purifying selection. Sequences that are needed for a feature that is common to these several placental mammals would be expected to have a high Conservation score.

A discriminatory analysis of the multiple alignments was used to generate a Regulatory Potential (RP) score (57). This machine-learning approach estimates the likelihood that a given aligning segment is a CRM, given the frequency of patterns in the alignments that are distinctive for CRMs as opposed to neutral DNA. The patterns are strings of alignment columns, and their discriminatory power is determined by the frequency of the patterns in training sets of alignments in CRMs versus alignments in neutral DNA. Although the RP score is influenced by features in addition to constraint, it is designed for finding CRMs that are common among species.

Basal promoters

Promoters are DNA sequences needed for accurate initiation of transcription. For some promoters including the globin gene promoters, one DNA segment interacts with RNA polymerase II and its accessory factors (such as TFIID and TFIIB) to determine the start site of transcription; this is the basal promoter (58). Five motifs have been associated with basal promoters, and these are found in the promoters of human globin genes (Figure 3.4.A). They include the familiar TATA box to which TBP binds, along with the BRE to which TFIIB binds and the Inr and DPE motifs to which components of TFIID binds (58).

Early studies revealed the presence of the ATAAA motif about 25-30 bp 5' to the start site of transcription of the globin genes (59), and this is by far the most restricted in its consensus, i.e. this motif appears to be under evolutionary constraint in globin genes. Recent studies on other promoters are revealing the roles of additional motifs close to the start site of transcription, but on both sides. Matches to these motifs can be found readily at the appropriate positions in the human globin genes (Figure 3.4.A). The motifs other than TATA do not have well-defined consensus sequences, either for genes in general or for the human globin genes, and thus their presence alone may not signify function. Also, only the TATA box, Inr and DPE show evidence of constraint in homologs in other mammalian species (Figure 3.5.A, conservation track). However, each of the motifs except BRE has been implicated in function by finding a mutation in at least one case of β-thalassemia. Every base in the TATA box has been altered in one or another thalassemia, and mutations in Inr, MTE and DPE also are associated with thalassemia (Figure 3.5.A, Compilation of Human Disease Variants and Other Mutations). The BRE overlaps with the β-direct repeat element (βDRE), which is a *cis*-regulatory element bound by βDRF and demonstrated to function in regulation of the β-globin gene by mutagenesis and expression in transfected cells (60). Thus, the mutagenesis data (natural and directed) indicate that all five motifs are important for appropriate expression of the β-globin gene. The presence of similar motifs in the basal promoters for other human globin genes suggests that they are active in these genes as well.

Although it is common to describe promoters recognized by RNA polymerase II by the motifs shown in Figure 3.4.A, it is important to realize that this is true for only a minority of human genes. Globin gene promoters fall into the category of promoters with well-defined TATA boxes at a restricted location and one major start site for transcription. Recent studies show that these comprise a small minority of promoters, perhaps only 10 to 20%. Most promoters are CpG islands with no obvious TATA box, and in some cases they have a broad distribution of start sites (61).

Upstream regulatory sequences

Adjacent to the basal promoter is the upstream regulatory region (58) , which in globin genes runs from about positions -40 to -250 (Figure 3.4.B). Only one motif in this region is found in all the highly expressed globin genes: the CCAAT box. Proteins such as NF-Y and CP1 bind to this motif (62, 63), and it has been implicated in promoter function because of its presence in many promoters and the results of mutagenesis and binding studies (59). It is missing from the δ-globin gene (*HBD*) promoter, but this gene is expressed at a low level (about 1-2% of *HBB*).

Two motifs are found in many but not all promoters. One is the CACC box, which is bound by transcription factors in the Krüppel-like zinc finger class (KLF). The first erythroid KLF discovered was EKLF, which binds to the CACC box in the *HBB* promoter and is needed for erythropoiesis (64, 65). The CACC boxes in globin promoters tend to be highly conserved in other mammals, albeit not as constrained as the CCAAT box (Figure 5.5.B). Mutations in almost every position in the proximal CACC box have been associated with β-thalassemia (Figure 5.5.B). Thus many lines of evidence point to the importance of this motif. Other KLFs may bind to the CACC boxes in other globin gene promoters, such as FKLF or KLF13 (66) for the *HBG1* and *HBG2* promoters.

The other motif occurring frequently in upstream regulatory regions is WGATAR, the binding site for GATA-1 and related proteins (Figure 3.4.B). GATA-1 plays a critical role in erythroid-specific gene activation and repression (67-69), and the binding sites in these upstream regions have been implicated in positive regulation of the

respective genes (70, 71). The GATA-1 binding sites upstream of *HBE1*, *HBG1*, *HBG2* and *HBZ2* are conserved in most mammals, but the ones upstream of *HBB* are not. GATA-1 binds to the promoter regions of β-globin genes in both human (63) and mouse (72), but the binding site motif occurs in different places in the two promoters (73). This is an example of alterations in the binding site being associated with changes in the pattern of regulation, e.g. the delay in onset of expression in humans.

A different set of binding sites is distinctive to each type of gene. For instance, βDRF (60) and BB1-binding protein (72, 74) have been implicated in the regulation of the β-globin gene but not other globin genes (Figure 3.4.B). Both binding sites are conserved in many placental mammals (Figure 3.5.B and 73). Likewise, binding of OCT1 and γPE has been shown for the upstream regions of g-globin genes but not others (75).

The *cis*-elements close to the γ-globin genes are key determinants of fetal versus embryonic expression. One of the clearest demonstrations of this is from transgenic mouse experiments using a construct containing an LCR to enhance expression of globin genes. The γ-globin gene of prosimians, e.g. the bush-baby galago, is expressed embryonically, and when it is included in the test construct in transgenic mice, the transgene is also expressed embryonically. In contrast, a human γ-globin gene, normally expressed during fetal life in humans, is expressed fetally when transferred into transgenic mice in an otherwise identical construct (76). Thus one would expect to find alterations in the regulatory regions of anthropoid (monkey, ape and humans γ-globin genes that are associated with this change in stage-specificity (i.e., sequences that are conserved in anthropoid primates but are different in prosimians and nonprimate mammals). Examination of aligned sequences for differential phylogenetic footprints (49) led to the identification of a stage selector element (SSE) in the human γ-globin gene promoter (Fig. 3.4.B). The SSE is a binding site for a factor called the stage-selector protein, or SSP, which has been implicated in the differential expression of γ- and β-globin genes (77). Additional DNA sequence that binds several proteins implicated in fetal silencing of the γ-globin gene (49). Parallel protein-binding and mutagenesis studies led to the discovery of a novel protein that binds to an element called the γPE the

upstream regulatory region of the γ-globin genes, which has also been implicated in regulation of this gene (75).

The most distinctive globin gene promoters are those of the α-globin genes (*HBA1* and *HBA2*). These promoters are CpG islands, and among the hemoglobin genes, only those encoding α-globin have this feature. (The θ-globin and μ-globin genes also have promoters in CpG islands, but as discussed above, it is not clear that they encode components of hemoglobin.) While the majority of mammalian promoters are CpG islands (61), most of the associated genes are expressed in multiple tissues and few if any are expressed at such a high level as the α-globin gene. Thus the presence of a CpG island in the promoter for a globin gene is curious, and it leads to several unanswered questions about the α-globin gene promoters. What prevents their expression in nonerythroid tissues? What sequences in addition to the CpG island lead to very high level expression in erythroid cells? No GATA-1 binding site is found in the α-globin gene promoters of most placental mammals (the mouse α-globin genes is a notable exception), so sequence-directed binding of this protein to the proximal sequences is not the answer. Several studies have shown that the CpG island is a key component of the *cis*-regulatory elements for the α-globin gene of humans and rabbits, possibly through its effects on chromatin structure (78, 79).

The differences in the arrays of proteins functioning at ε-, γ-, β- and α-globin genes indicate that a distinct battery of proteins functions in the promoter for each type of gene. Indeed, this is consistent with the observation that *cis*-acting sequences needed for stage-specific regulation of expression map close to the genes (80).

Proximal Enhancers

Enhancers are DNA sequences that increase the activity of promoters; they can be located on either side of a gene or internal to it, and they can act at considerable distances from genes (81). Two enhancers have been found close to genes in the β-globin gene cluster, one that is 3' to *HBB* and one that is 3' to *HBG1* (Figure 3.3.A). In both cases the enhancers are less than 1 kb downstream of the polyA additional signal for the respective genes. The *HBB* enhancer was discovered by its effect on developmental timing of

expression of globin transgenes when introduced into mice. High level expression of human γ- or β-globin transgene constructs in fetal erythroid cells (the normal onset of expression of mouse β-globin genes) is dependent on the presence of the enhancer (74, 82-84). The *HBG1* enhancer was discovered as the only DNA segment in a 22kb region surrounding the γ-globin genes for DNA segments that boosted expression of a reporter gene driven by a γ-globin gene promoter in transfected erythroid cells (85). Deletion of this enhancer from a large construct containing the human LCR and β-like globin genes had no effect on expression levels in transgenic mice (86), which could mean that it actually has no function, or that other sequences compensate for its loss, or that its function is not apparent in mice.

Indeed, comparative sequence analysis of these proximal enhancers strongly supports the conclusion that both play roles in higher primates but not in other species. As illustrated in Figure 3.4.C, both enhancers contain binding sites for GATA-1 (87, 88), and the *HBG1* enhancer also binds to the γPE protein (75). However, the DNA homologous to the *HBB* enhancer in other mammals is not strongly conserved, even in the GATA motifs. Furthermore, two of the GATA1-binding sites in the *HBG1* enhancer were introduced via an LTR-type transposable element that is present only in higher primates (Figure 3.6.A). Thus the presence of the *HBG1* proximal enhancer correlates with the fetal recruitment of γ-globin gene expression in anthropoids, and its function may not be observed in transgenic mice. Likewise, the presence of GATA-1 bindings sites only in higher primates suggests that the function of the *HBB* proximal enhancer may also be lineage-specific, perhaps related to the delay in expression of *HBB* in higher primates. In this case, an effect on developmental timing is readily demonstrable in transgenic mice, but because of the differences in timing of *HBB* expression in humans (the source of the transgene) and mouse (the host species), it is difficult to fully understand this function.

Distal Enhancers

In addition to the proximal promoters and enhancers, both the α-like and β-like globin gene clusters are regulated by distal control regions. The β-like globin cluster is

regulated by the distal LCR (reviewed in 89, 90), and the α-like globin gene cluster is regulated by HS-40 (91). In both cases, deletion of the distal control region is associated with thalassemia (Figure 3.3). Addition of the distal control regions has profound effects on expression of linked genes in transgenic mice. Without the LCR, erythroid expression of a β-globin transgene is not seen in all mouse lines (92), presumably because of integration in a repressive region of a chromosome (a position effect). With the LCR, the β-globin transgene is expressed at a high level in erythroid cells in almost all mouse lines, indicating strong enhancement and a reduction in position effects (93). HS-40 of the α-globin gene complex is a strong enhancer of globin gene expression, both in transgenic mice (91, 94) and in transfected cells (95).

The β-globin LCR is a very large regulatory region, containing at least five DNase hypersensitive sites in humans spread over about 17 kb (96-98) between *HBE1* and an *OR* gene (Figure 3.3.A). This region is highly conserved in mammals, with highly similar sequences indicative of constraint found both in the hypersensitive sites and between them (50, 90). This can be seen in Figure 3.3.A as the string of peaks of conservation and RP in this region.

The distal enhancer for the α-globin gene, HS-40, is much smaller than the LCR. It is about 250 bp in length (99), located in a widely expressed gene called *C16orf35* (Figure 3.3.B). Additional erythroid DNase hypersensitive sites are present in this large gene, but none have been shown to play a role in regulation of globin genes (26). HS-40 is sufficient for strong enhancement and high activity in erythroid cells of transgenic mice, especially during embryonic and fetal development (91). It is very strongly conserved in mammals, with obvious matches to species as distant as opossum (Figures 3.3.B and 3.6.B). Functional tests have shown that the homologous regions of chicken and fish also have enhancer activity, despite considerable divergence outside the protein-binding sites (44).

Regulatory activities in addition to tissue-specific enhancement have been attributed to the β-globin LCR, but they are not seen consistently in multiple lines of investigation (100). Examination of chromatin structure after deletion of the LCR led to the inference that the LCR is needed for tissue-specific chromosomal domain opening

(101). Chromosome 11 from a patient with the Hispanic $(\varepsilon\gamma\delta\beta)^0$ thalassemia (missing most of the LCR and some adjacent sequences) was transferred through multiple somatic cells to generate a hybrid murine erythroleukemia cell line containing the mutant human chromosome. The β-globin gene cluster in this hybrid cell line is inactive and is insensitive to DNase, indicating that the LCR is needed for opening a chromosomal domain (101). However, an engineered mouse line carrying a deletion of the mouse β-globin LCR and the sequences homologous to those lost in the Hispanic deletion retains an open chromatin conformation (accessible to DNase) in the mouse β-globin gene (102). Although expression of the mouse β-globin genes is reduced substantially, the locus is not silenced. Thus the repressive heterochromatin seen in the hybrid murine erythroleukemia cells carrying human chromosome 11 with the Hispanic deletion may have been produced during the chromosome transfers between cell lines. Currently, the DNA sequence determinants of chromatin opening have still not been discovered. The β-globin LCR has also been implicated in overcoming position effects in transgenic mice (103), in keeping with the inferred effect on opening a chromatin domain. However, transgene constructs containing the β-globin can still show position effect variegation (104). Both the β-globin LCR and the α-globin HS-40 are very strong, erythroid-specific enhancers needed for the expression of any of the linked globin genes. They also can overcome some but not all repressive effects after integration at a variety of chromosomal locations. This could be a consequence of the strong enhancement.

Three transcription factor-binding motifs are present in almost all DNase hypersensitive sites that have a strong function in the distal enhancers (Figure 3.4.D). All contain Maf-response elements, or MAREs, to which transcriptional activator proteins of the basic leucine zipper class can bind (105). A subfamily of proteins related to AP1, such as NF-E2, LCRF1/Nrf1, and Bach1, bind to this element (reviewed in 106, 107). All are heterodimers containing a Maf protein as one subunit, which is the basis for the name of the response element. All the hypersensitive sites have GATA motifs, to which GATA-1 and its relatives bind (108). The third common motif is CACC, to which a family of Zn-finger proteins including erythroid Krüppel-like factor (EKLF) can bind (64). At HS3 in the β-globin LCR, there is evidence that motifs related to CACC are

bound by additional Krüppel-like factors, such as Sp1 (109). HS2 of the β-globin LCR also has three E-boxes , which are the binding sites for TAL-1 and its heterodimeric partners (47). This protein has been implicated in regulation of hematopoiesis, and it appears to also play a role in enhancement by HS2.

Initial studies of protein binding at these and other CRMs used various *in vitro* methods and *in vivo* footprinting (99, 110-112). Recent experiments using chromatin immunoprecipitation have demonstrated occupancy of the CRMs by several of these proteins in erythroid cells (e.g., 113, 114-116). Many of the sites have implicated directly in activity by mutagenesis and gene transfer (e.g., 47, e.g., 117, 118, 119).

The protein binding sites in the distal positive regulators show some common patterns (Figure 3.4.D). A MARE plus two GATA motifs is present in most of the CRMs, and this arrangement has been shown to be needed for formation of a hypersensitive site at HS4 (120). The strongest enhancers (as assayed by gene transfer in somatic cells) are HS2 and HS-40. Both of these have two MAREs, and mutation of those MAREs removes much of the enhancing activity (117, 119). Thus the MAREs and proteins binding to them are critical for high-level enhancement, but the other binding sites contribute to function as well.

The CRMs marked by these hypersensitive sites in the distal positive regulators are conserved across almost all mammals (26, 90). The portion of the alignments for HS-40 shown in Figure 3.6.B indicates the very strong constraint seen in the known binding sites and additional short segments both for this enhancer and for HS2. Most of the binding sites in HS3 are also highly conserved, but some are not, likely reflecting both common and lineage-specific functions. HS4, with the MARE and two GATA motifs, is conserved across a wide span of placental mammals, but this DNA sequence is part of an LTR-type repeat, a member of the ERV1 repeat family. This appears to be an old transposable element (predating most of the mammalian radiation), but one that continues to provide a regulatory function.

CONCLUDING REMARKS

Molecular clones containing mammalian globin gene clusters were isolated about 30 years ago. Intense study since then has revealed much about their structure, evolution and regulation. However, understanding sufficient to lead to clinical applications continues to elude us. The myriad levels of regulation and function that operate within these gene clusters certainly confound attempts to find simplifying conclusions. Despite these challenges, studies of the globin gene clusters have consistently provided new insights into function, regulation and evolution. The lessons being learned as we try to integrate information from classical molecular biology and genetics, new high through-put biochemical assays, and extensive interspecies sequence comparisons are paving the way for applying these approaches genome-wide. The globin gene clusters illustrate the need to distinguish common from lineage-specific regulation. Although simple generalizations are rare, the extensive information that one needs for interpreting data in the context of comparative genomics is readily accessible. Throughout this chapter, we have illustrated points using output from the UCSC Genome Browser (http://genome.ucsc.edu), with special emphasis on the tracks showing Conservation, Regulatory Potential, and Locus Variants. Deeper information on the variants associated with disorders of the hemoglobins can be obtained from HbVar (http://www.bx.psu.edu). We hope that the examples presented here will be helpful in guiding interpretation of the multitude of data available to the readers now and in the future.

ACKNOWLEDGEMENTS

RH was supported by NIH grant R01 DK065806 and BGF was supported by NIH grants R01 DK19482 and P01 HL63357.

REFERENCES

1. Weatherall DJ and Clegg JB. Thalassemia Syndromes. 3rd. ed. 1981, Oxford: Blackwell Scientific Publications.
2. Baglioni C. The fusion of two peptide chains in hemoglobin Lepore and its interpretation as a genetic deletion. Proc Natl Acad Sci U S A 1962;48:1880-6.
3. Kendall AG, Ojwang PJ, Schroeder WA, and Huisman TH. Hemoglobin Kenya, the product of a gamma-beta fusion gene: studies of the family. Am J Hum Genet 1973;25:548-63.
4. Deisseroth A, Nienhuis A, Turner P, et al. Localization of the human alpha globin structural gene to chromosome 16 in somatic cell hybrids by molecular hybridization assay. Cell 1977;12:205-18.
5. Deisseroth A, Nienhuis AW, Lawrence J, Giles RE, Turner P, and Ruddle FH. Chromosomal localization of the human beta globin gene to human chromosome 11 in somatic cell hybrids. Proc Nat Acad Sci, USA 1978;75:1456-60.
6. Tilghman SM, Tiemeier DC, Seidman JG, et al. Intervening sequence of DNA identified in the structural portion of a mouse beta-globin gene. Proc Natl Acad Sci USA 1978;75:725-9.
7. Goodman M, Czelusniak J, Koop B, Tagle D, and Slightom J. Globins: A case study in molecular phylogeny. Cold Spring Harbor Symp Quant Biol 1987;52:875-90.
8. Proudfoot NJ, Gil A, and Maniatis T. The structure of the human zeta-globin gene and a closely linked, nearly identical pseudogene. Cell 1982;31:553-63.
9. Tilghman SM, Curtis PJ, Tiemeier DC, Leder P, and Weissmann C. The intervening sequence of a mouse beta-globin gene is transcribed within the 15S beta-globin mRNA precursor. Proc Natl Acad Sci USA 1978;75:1309-13.
10. Krainer AR, Maniatis T, Ruskin B, and Green MR. Normal and mutant human beta-globin pre-mRNAs are faithfully and efficiently spliced in vitro. Cell 1984;36:993-1005.
11. Mount SM. A catalogue of splice junction sequences. Nucleic Acids Res 1982;10:459-72.

12. Padgett RA, Grabowski PJ, Konarska MM, Seiler S, and Sharp PA. Splicing of messenger RNA precursors. Annu Rev Biochem 1986;55:1119-50.
13. Patrinos GP, Giardine B, Riemer C, et al. Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. Nucleic Acids Res 2004;32 Database issue:D537-D41.
14. Giardine B, Riemer C, Hefferon T, et al. PhenCode: connecting ENCODE data with mutations and phenotype. Hum Mutat 2007;28:554-62.
15. Jeffreys AJ and Flavell RA. The rabbit beta-globin gene contains a large large insert in the coding sequence. Cell 1977;12:1097-108.
16. Tuan D, Biro PA, deRiel JK, Lazarus H, and Forget BG. Restriction endonuclease mapping of the human gamma globin gene loci. Nucleic Acids Res 1979;6:2519-44.
17. Southern EM. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 1975;98:503-17.
18. Fritsch E, Lawn R, and Maniatis T. Molecular cloning and characterization of the human beta-like globin gene cluster. Cell 1980;19:959-72.
19. Zhang Z and Gerstein M. Large-scale analysis of pseudogenes in the human genome. Curr Opin Genet Dev 2004;14:328-35.
20. Hsu S, Marks J, Shaw J, et al. Structure and expression of the human theta 1 globin gene. Nature 1988;331:94-6.
21. Ley TJ, Maloney KA, Gordon JI, and Schwartz AL. Globin gene expression in erythroid human fetal liver cells. J Clin Invest 1989;83:1032-8.
22. Albitar M, Peschle C, and Liebhaber SA. Theta, zeta and epsilon globin messenger RNA are expressed in adults. Blood 1989;74:629-37.
23. Goh SH, Lee YT, Bhanu NV, et al. A newly discovered human alpha-globin gene. Blood 2005;106:1466-72.
24. Cooper SJ, Wheeler D, De Leo A, et al. The mammalian alphaD-globin gene lineage and a new model for the molecular evolution of alpha-globin gene clusters at the stem of the mammalian radiation. Mol Phylogenet Evol 2006;38:439-48.

25. Hardison RC, Sawada I, Cheng J-F, Shen C-KJ, and Schmid CW. A previously undetected pseudogene in the human alpha globin gene cluster. Nucleics Acids Research 1986;14:1903-11.
26. Hughes JR, Cheng JF, Ventress N, et al. Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. Proc Natl Acad Sci USA 2005;102:9830-5.
27. Clegg JB. Can the product of the theta gene be a real globin? Nature 1987;329:465-6.
28. Craddock CF, Vyas P, Sharpe JA, Ayyub H, Wood WG, and Higgs DR. Contrasting effects of alpha and beta globin regulatory elements on chromatin structure may be related to their different chromosomal environments. EMBO J 1995;14:1718-26.
29. Lander ES,Linton LM,Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.
30. Bird AP. CpG-rich islands and the function of DNA methylation. Nature 1986;321:209-13.
31. Collins FS and Weissman SM. The molecular genetics of human hemoglobin. Prog Nucl Acids Res & Mol Biol 1984;31:315-462.
32. Fischel-Ghodsian N, Nicholls RD, and Higgs DR. Unusual features of CpG-rich (HTF) islands in the human α-globin complex: association with nonfunctional pseudogenes and presence within the 3' portion of the ζ genes. Nucl Acids Res 1987;15:9215-25.
33. Groudine M, Kohwi-Shigematsu T, Gelinas R, Stamatoyannopoylos G, and Papyannopoulou T. Human fetal to adult hemoglobin switching: Changes in chromatin structure of the β-globin gene locus. Proc Natl Acad Sci, USA 1983;80:7551-5.
34. van der Ploeg LHT and Flavell RA. DNA methylation in the human γ-δ-β globin locus in erythroid and nonerythroid tissues. Cell 1980;19:947-58.

35. Bird A, Taggart M, Nicholls R, and Higgs D. Non-methylated CpG-rich islands at the human α-globin locus: implications for evolution of the α-globin pseudogene. EMBO J 1987;6:999-1004.
36. Epner E, Rifkind RA, and Marks PA. Replication of alpha and beta globin DNA sequences occurs during early S phase in murine erythroleukemia cells. Proc Natl Acad Sci USA 1981;78:3058-62.
37. Goldman MA, Holmquist GP, Gray MC, Caston LA, and Nag A. Replication timing of genes and middle repetitive sequences. Science 1984;224:686-92.
38. Dhar V, Mager D, Iqbal A, and Schildkraut CL. The co-ordinate replication of the human β-globin gene domain reflects its transcriptional activity and nuclease hypersensitivity. Mol Cell Biol 1988;8:4958-65.
39. Bulger M, Bender MA, von Doorninck JH, et al. Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse β-globin gene clusters. Proc Natl Acad Sci, USA 2000;97:14560-5.
40. Feingold EA and Forget BG. The breakpoint of a large deletion causing hereditary persistence of fetal hemoglobin occurs within an erythroid DNA domain remote from the beta-globin gene cluster. Blood 1989;74:2178-86.
41. Anagnou NP, Perez-Stable C, Gelinas R, et al. Sequences located 3' to the breakpoint of the hereditary persistence of fetal hemoglobin-3 deletion exhibit enhancer activity and can modify the developmental expression of the human fetal A gamma-globin gene in transgenic mice. J Biol Chem 1995;270:10256-63.
42. Flint J, Thomas K, Micklem G, et al. The relationship between chromosome structure and function at a human telomeric region. Nature Genetics 1997;15:252-7.
43. Anguita E, Johnson CA, Wood WG, Turner BM, and Higgs DR. Identification of a conserved erythroid specific domain of histone acetylation across the alpha-globin gene cluster. Proc Natl Acad Sci USA 2001;98:12114-9.
44. Flint J, Tufarelli C, Peden J, et al. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. Hum Mol Genet 2001;10:371-82.

45. Gumucio DL, Heilstedt-Williamson H, Gray TA, et al. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human γ and ε globin genes. Mol Cell Biol 1992;12:4919-29.
46. Gumucio D, Shelton D, Zhu W, et al. Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the beta-like globin genes. Mol Phylog and Evol 1996;5:18-32.
47. Elnitski L. Conserved E boxes in the locus control region contribute to enhanced expression of beta-globin genes via TAL1 and other basic helix-loop-helix proteins. 1998, The Pennsylvania State University.
48. Tagle DA, Koop BF, Goodman M, Slightom J, Hess DL, and Jones RT. Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*): Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 1988;203:7469-80.
49. Gumucio DL, Shelton DA, Blanchard-McQuate K, et al. Differential phylogenetic footprinting as a means to identify base changes responsible for recruitment of the anthropoid γ gene to a fetal expression pattern. J Biol Chem 1994;269:15371-80.
50. Hardison R and Miller W. Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. Mol Biol Evol 1993;10:73-102.
51. Margot JB, Demers GW, and Hardison RC. Complete nucleotide sequence of the rabbit beta-like globin gene cluster: Analysis of intergenic sequences and comparison with the human beta-like globin gene cluster. J Mol Biol 1989;205:15-40.
52. Shehee R, Loeb DD, Adey NB, et al. Nucleotide sequence of the BALB/c mouse β-globin complex. J Mol Biol 1989;205:41-62.
53. Hardison R, Krane D, Vandenbergh D, et al. Sequence and comparative analysis of the rabbit alpha-like globin gene cluster reveals a rapid mode of evolution in a G + C-rich region of mammalian genomes. J Mol Biol 1991;222:233-49.
54. Hardison RC. The nucleotide sequence of the rabbit embryonic globin gene β4. J Biol Chem 1983;258:8739-44.

55. Cooper GM, Brudno M, Stone EA, Dubchak I, Batzoglou S, and Sidow A. Characterization of evolutionary rates and constraints in three Mammalian genomes. Genome Res 2004;14:539-48.
56. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005;15:1034-50.
57. Taylor J, Tyekucheva S, King DC, Hardison RC, Miller W, and Chiaromonte F. ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements. Genome Res 2006;16:1596-604.
58. Maston GA, Evans SK, and Green MR. Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genomics Hum Genet 2006;7:29-59.
59. Efstratiadis A, Posakony JW, Maniatis T, et al. The structure and evolution of the human β-globin gene family. Cell 1980;21:653-68.
60. Stuve LL and Myers RM. A directly repeated sequence in the β-globin promoter regulates transcription in murine erythroleukemia cells. Mol Cell Biol 1990;10:972-81.
61. Carninci P, Sandelin A, Lenhard B, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 2006;38:626-35.
62. Cohen RB, Sheffery M, and Kim CG. Partial purification of a nuclear protein that binds to the CCAAT box of the mouse α1-globin gene. Mol Cell Biol 1986;6:821-32.
63. deBoer E, Antoniou M, Mignotte V, Wall L, and Grosveld F. The human β-globin promoter; nuclear protein factors and erythroid specific induction of transcription. EMBO J 1988;7:4203-12.
64. Miller IJ and Bieker JJ. A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the *Kruppel* family of nuclear factors. Mol Cell Biol 1993;13:2776-86.
65. Perkins AC, Sharpe AH, and Orkin SH. Lethal β-thalassaemia in mice lacking the erythroid CACCC-transcription factor EKLF. Nature 1995;375:318-22.

66. Asano H, Li XS, and Stamatoyannopoulos G. FKLF, a novel Kruppel-like factor that activates human embryonic and fetal beta-like globin genes. Mol Cell Biol 1999;19:3571-9.
67. Pevny L, Simon MC, Robertson E, et al. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. Nature 1991;349:257-60.
68. Simon MC, Pevny L, Wiles MV, Keller G, Costantini F, and Orkin SH. Rescue of erythroid development in gene targeted GATA-1- mouse embryonic stem cells. Nat Genet 1992;1:92-8.
69. Welch JJ, Watts JA, Vakoc CR, et al. Global regulation of erythroid gene expression by transcription factor GATA-1. Blood 2004;104:3136-47.
70. Martin D and Orkin S. Transcriptional activation and DNA binding by the erythroid factor GF-1/NF-E1/Eryf 1. Genes & Dev 1990;4:1886-98.
71. Gong Q-H and Dean A. Enhancer-dependent transcripion of the ε-globin promoter requires promoter-bound GATA-1 and enhancer-bound AP-1/NF-E2. Mol Cell Biol 1993;13:911-7.
72. Macleod K and Plumb M. Derepression of mouse β-major-globin gene transcription during erythroid differentiation. Mol Cell Biol 1991;11:4324-32.
73. Hardison R, Chao K-M, Schwartz S, Stojanovic N, Ganetsky M, and Miller W. Globin gene server: A prototype E-mail database server featuring extensive multiple alignments and data compilation. Genomics 1994;21:344-53.
74. Antoniou M, deBoer E, Habets G, and Grosveld F. The human β-globin gene contains multiple regulatory regions: Identification of one promoter and two downstream enhancers. EMBO J 1988;7:377-84.
75. Lloyd JA, Case SS, Ponce E, and Lingrel JB. Positive transcriptional regulation of the human γ-globin gene: γPE is a novel nuclear factor with multiple binding sites near the gene. J Biol Chem 1994;269:26-34.
76. TomHon C, Zhu W, Millinoff D, et al. Evolution of a fetal expression pattern via *cis*-changes near the γ-globin gene. J Biol Chem 1997;272:14062-6.

77. Jane SM, Ney PA, Vanin EF, Gumucio DL, and Nienhuis AW. Identification of a stage selector element in the human γ-globin gene promoter that fosters preferential interaction with the 5' HS2 enhancer when in competition with the β-promoter. EMBO J 1992;11:2961-9.
78. Pondel M, Murphy S, Pearson L, Craddock C, and Proudfoot N. Sp1 functions in a chromatin-dependent manner to augment human alpha-globin promoter activity. Proc Natl Acad Sci USA 1995;92:7237-41.
79. Shewchuk BM and Hardison RC. CpG islands from the α-globin gene cluster increase gene expression in an integration-dependent manner. Mol Cell Biol 1997;17:5856-66.
80. Trudel M, Magram J, Bruckner L, and Costantini F. Upstream G gamma-globin and downstream beta-globin sequences required for stage-specific expression in transgenic mice. Mol Cell Biol 1987;7:4024-9.
81. Tjian R and Maniatis T. Transcriptional activation: A complex puzzle with few easy pieces. Cell 1994;77:5-8.
82. Trudel M and Costantini F. A 3' enhancer contributes to the stage-specific expression of the human β-globin gene. Genes & Devel 1987;1:954-61.
83. Behringer RR, Hammer RE, Brinster RL, Palmiter RD, and Townes TM. Two 3' sequences direct adult erythroid-specific expression of human beta-globin genes in transgenic mice. Proc Natl Acad Sci USA 1987;84:7056-60.
84. Liu Q, Bungert J, and Engel JD. Mutation of gene-proximal regulatory elements disrupts human epsilon-, gamma-, and beta-globin expression in yeast artificial chromosome transgenic mice. Proc Natl Acad Sci USA 1997;94:169-74.
85. Bodine D and Ley T. An enhancer element lies 3' to the human A gamma globin gene. EMBO J 1987;6:2997-3004.
86. Liu Q, Tanimoto K, Bungert J, and Engel JD. The A gamma-globin 3' element provides no unique function(s) for human beta-globin locus gene regulation. Proc Natl Acad Sci USA 1998;95:9944-9.

87. Wall L, deBoer E, and Grosveld F. The human β-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein. Genes & Devel 1988;2:1089-100.

88. Puruker M, Bodine D, Lin H, McDonagh K, and Nienhuis AW. Structure and function of the enhancer 3' to the human Aγ-globin gene. Nucleic Acids Res 1990;18:407-7415.

89. Grosveld F, Antoniou M, Berry M, et al. The regulation of human globin gene switching. Philos Trans R Soc Lond 1993;339:183-91.

90. Hardison R, Slightom JL, Gumucio DL, Goodman M, Stojanovic N, and Miller W. Locus control regions of mammalian β-globin gene clusters: Combining phylogenetic analyses and experimental results to gain functional insights. Gene 1997;205:73-94.

91. Higgs D, Wood W, Jarman A, et al. A major positive regulatory region located far upstream of the human α-globin gene locus. Genes & Devel 1990;4:1588-601.

92. Chada K, Magram J, and Costantini F. Tissue- and stage-specific expression of a cloned adult beta globin gene in transgenic mice. Prog Clin Biol Res 1985;191:305-19.

93. Grosveld F, van Assendelft GB, Greaves D, and Kollias G. Position-independent, high-level expression of the human β-globin gene in transgenic mice. Cell 1987;51:975-85.

94. Sharpe JA, Chan-Thomas PS, Lida J, Ayyub H, Wood WG, and Higgs DR. Analysis of the human α-globin upstream regulatory element (HS-40) in transgenic mice. EMBO J 1992;11:4565-72.

95. Ren S, Luo X-n, and Atweh G. The major regulatory element upstream of the α-globin gene has classical and inducible enhancer activity. Blood 1993;81:1058-66.

96. Tuan D, Abelovich A, Lee-Oldham M, and Lee D. Identification of regulatory elements of human β-like globin genes, In: Stamatoyannopoulos G and Nienhuis AW, Editors Developmental Control of Globin Gene Expression. 1987, A. R. Liss, Inc.: New York. 211-20.

97. Forrester W, Takegawa S, Papayannopoulou T, Stamatoyannopoulos G, and Groudine M. Evidence for a locus activating region: The formation of developmentally stable hypersensitive sites in globin-expressing hybrids. Nucl Acids Res 1987;15:10159-77.
98. Dhar V, Nandi A, Schildkraut CL, and Skoultchi AI. Erythroid-specific nuclease-hypersensitive sites flanking the human β-globin gene cluster. Mol Cell Biol 1990;10:4324-33.
99. Jarman A, Wood W, Sharpe J, Gourdon G, Ayyub H, and Higgs D. Characterization of the major regulatory element upstream of the human α-globin gene cluster. Mol Cell Biol 1991;11:4679-89.
100. Higgs DR. Do LCRs open chromatin domains? Cell 1998;95:299-302.
101. Forrester WC, Epner E, Driscoll MC, et al. A deletion of the human β-globin locus activation region causes a major alteration in chromatin structure and replication across the entire β-globin locus. Genes & Devel 1990;4:1637-49.
102. Bender MA, Byron R, Ragoczy T, Telling A, Bulger M, and Groudine M. Flanking HS-62.5 and 3' HS1, and regions upstream of the LCR, are not required for beta-globin transcription. Blood 2006;108:1395-401.
103. Fraser P, Hurst J, Collis P, and Grosveld F. DNase I hypersensitive sites 1, 2 and 3 of the human β-globin dominant control region direct position-independent expression. Nucleic Acids Res 1990;18:3503-8.
104. Alami R, Greally JM, Tanimoto K, et al. beta-globin YAC transgenes exhibit uniform expression levels but position effect variegation in mice. Hum Mol Genet 2000;9:631-6.
105. Motohashi H, Shavit JA, Igarashi K, Yamamoto M, and Engel JD. The world according to Maf. Nucleic Acids Res 1997;25:2953-9.
106. Orkin S. Regulation of globin gene expression in erythroid cells. Eur J Biochem 1995;231:271-81.
107. Baron MH. Transcriptional control of globin gene switching during vertebrate development. Biochim Biophys Acta 1997;1351:51-72.

108. Evans T, Felsenfeld G, and Reitman M. Control of globin gene transcription. Annu Rev Cell Biol 1990;6:95-124.
109. Shelton DA, Stegman L, Hardison R, et al. Phylogenetic footprinting of hypersensitive site 3 of the β-globin locus control region. Blood 1997;89:3457-69.
110. Talbot D, Philipsen S, Fraser P, and Grosveld F. Detailed analysis of the site 3 region of the human β-globin dominant control region. EMBO J 1990;9:2169-78.
111. Strauss EC, Andrews NC, Higgs DR, and Orkin SH. In vivo footprinting of the human α-globin locus upstream regulatory element by guanine and adenine ligation-mediated polymerase chain reaction. Mol Cell Biol 1992;12:2135-42.
112. Reddy PMS, Stamatoyannopoulos G, Papayannopoulou T, and Shen C-KJ. Genomic footprinting and sequencing of human β-globin locus: Tissue specificity and cell line artifact. J Biol Chem 1994;269:8287-95.
113. Forsberg EC, Downs KM, and Bresnick EH. Direct interaction of NF-E2 with hypersensitive site 2 of the beta-globin locus control region in living cells. Blood 2000;96:334-9.
114. Sawado T, Igarashi K, and Groudine M. Activation of beta-major globin gene transcription is associated with recruitment of NF-E2 to the beta-globin LCR and gene promoter. Proc Natl Acad Sci USA 2001;98:10226-31.
115. Letting DL, Rakowski C, Weiss MJ, and Blobel GA. Formation of a tissue-specific histone acetylation pattern by the hematopoietic transcription factor GATA-1. Mol Cell Biol 2003;23:1334-40.
116. Anguita E, Hughes J, Heyworth C, Blobel GA, Wood WG, and Higgs DR. Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. Embo J 2004;23:2841-52.
117. Ney P, Sorrentino B, McDonagh K, and Nienhuis A. Tandem AP-1-binding sites within the human β-globin dominant control region function as an inducible enhancer in erythroid cells. Genes & Devel 1990;4:993-1006.
118. Caterina JJ, Ciavatta DJ, Donze D, Behringer RR, and Townes TM. Multiple elements in human β-globin locus control region 5' HS2 are involved in enhancer

activity and position-independent transgene expression. Nucl Acids Res 1994;22:1006-11.

119. Gong Q, McDowell JC, and Dean A. Essential role of NF-E2 in remodeling of chromatin structure and transcriptional activation of the ε-globin gene in vivo by 5' hypersensitive site 2 of the β-globin locus control region. Mol Cell Biol 1996;16:6055-64.

120. Stamatoyannopoulos JA, Goodwin A, Joyce T, and Lowrey CH. NFE2 and GATA binding motifs are required for the formation of DNase I hypersensitive site 4 of the human β-globin locus control region. EMBO J 1995;14:106-16.

121. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Res 2002;12:996-1006.

122. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, and Haussler D. The UCSC Known Genes. Bioinformatics 2006;22:1036-46.

123. Montgomery SB, Griffith OL, Sleumer MC, et al. ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. Bioinformatics 2006;22:637-40.

FIGURE LEGENDS

Figure 3.1. Basic organization of human globin gene complexes. The locations of the alpha-globin gene complex very close to the telomere of the short arm of chromosome 16 and the beta-globin gene complex on the short arm of chromosome 11 are shown at the top. The genes are shown as boxes on the second line, named according to the globin polypeptide that is encoded. In both diagrams, the 5'-3' transcriptional orientation is from left to right. Note that the orientations with respect to the centromere (CEN) and telomere (TEL) are opposite; the alpha-like globin genes are transcribed toward CEN whereas the beta-like globin genes are transcribed toward TEL. The composition of hemoglobins produced at progressive developmental stages is given at the bottom.

Figure 3.2. Structure and expression pathway of globin genes.

(**A**) General structure of globin genes. The coding sequences of all globin genes in humans and other animals are separated by two introns (white boxes) into three exons. The first exon has a short 5' untranslated region (gray box) followed by a coding region (black box). All of the central exon codes for protein, while the third exon begins with coding sequences and ends with a 3' untranslated region. The relative sizes of the portions of the genes are indicated by the sizes of the boxes, and codon numbers are given above the boxes. The consensus sequence for critical sequences used in splicing are shown under the second intron of the beta-globin gene, and similar sequences are present in all introns. The vertical arrows show the splice site junctions within the consensus sequences where cleavage occurs during the process of joining the exons.

(**B**) The pathway for expression of globin genes. The RNA transcript is shown with short boxes corresponding to the untranslated regions (gray), coding regions (black) and introns (white) as in A, with processing and splicing steps occurring in the nucleus to form the mature mRNA. The mRNA is translated in the cytoplasm to generate a globin polypeptide to which the heme (gray disk) will bind. The diagram of the folded globin structure was provided by Dr. John Blamire at the Brooklyn College of the City University of New York.

Figure 3.3. Detailed maps of the human globin gene complexes, including genomic features and representative deletions.

(**A**) Detailed map of the β-like globin gene complex and surrounding olfactory receptor genes. The globin genes are named both by the encoded globin polypeptide and the official gene name. Pseudogenes are shown on a line below the genes. The known *cis*-regulatory modules are separated into distal elements such as the locus control region (shown as five DNase hypersensitive sites or HSs), promoters and enhancers close to the 3' ends of *HBG1* and *HBB*. The next two tracks show two features derived from multiple alignments of the human genomic sequence with sequences from six other placental mammals (chimp, rhesus macaque, mouse, rat, dog, and cow). The regulatory potential measures the similarity of patterns in the alignments to those that are distinctive for known regulatory regions versus neutral DNA (57). The conservation score estimates the likelihood that an alignment is in the most constrained portion of the genome, likely reflecting purifying selection (phastCons, 56). Positions of deletions that cause beta-thalassemia or Hereditary Persistence of Fetal Hemoglobin (HPFH) are shown in the lower portion.

(**B**) Detailed map of the α-like globin gene complex and surrounding genes. The conventions and tracks are similar to those in panel A. Positions of the distal erythroid HSs are from Hughes et al. (26). The deletions are grouped by those with deletion of a single alpha-globin gene (alpha-thalassemia-2), deletion of both alpha-globin genes (alpha-thalassemia-1), and a representative deletion (Ti~) that removes the distal enhancer (HS-40) but no structural genes. Coordinates of the deletions were provided by Dr. Jim Hughes.

These figures were generated starting with output from the UCSC Genome Browser (121), using the following tracks in addition to ones already mentioned: UCSC Known Genes (122), ORegAnno for *cis*-regulatory modules (123), and Locus Variants for the deletions (14). For panel A, the Genome Browser output was rotated 180^{o} so that the 5'-3' transcriptional orientation is left to right (note that the genome coordinates are decreasing from left to right). Both figures were edited for clarity. Information on

deletions and other variants is available both on the Locus Variants track as well as HbVar (13).

Figure 3.4. Motifs and binding sites in *cis*-regulatory modules of globin genes.

(**A**) Motifs in the basal promoter, based on those defined in the review by Maston et al. (58). Numbers along the top are relative to the transcription start site as +1, and ATG denotes the translation start site. The top consensus sequence is from Maston et al. (58). Corresponding positions in the globin genes are given for each motif, followed by the consensus derived for the globin genes. Symbols for ambiguous nucleotides are S = C or G, W = A or T, R = A or G, Y = C or T, D = A or G or T, H = A or C or T, V = A or C or G, and N = A or C or G or T.

(**B**) Motifs in the regulatory regions immediately upstream of the basal promoters. Motifs are indicated by sequence (CCAAT, CACC, and GATA), the name of the element (βDRE, αIRE, γPE, OCT) or the protein name followed by bs for "binding site" (BP2bs, NF1bs, BB1bs). Boxes for each motifs found in several upstream regions are shaded. The boxes were placed in the correct order but spacing is not indicated. The thick line for the *HBA* upstream regions (both *HBA1* and *HBA2*) denotes that it is a CpG island.

(**C**) Motifs in the proximal enhancers.

(**D**) Motifs in distal positive regulators, including three hypersensitive sites of the β-globin LCR and HS-40 for the α-globin gene cluster.

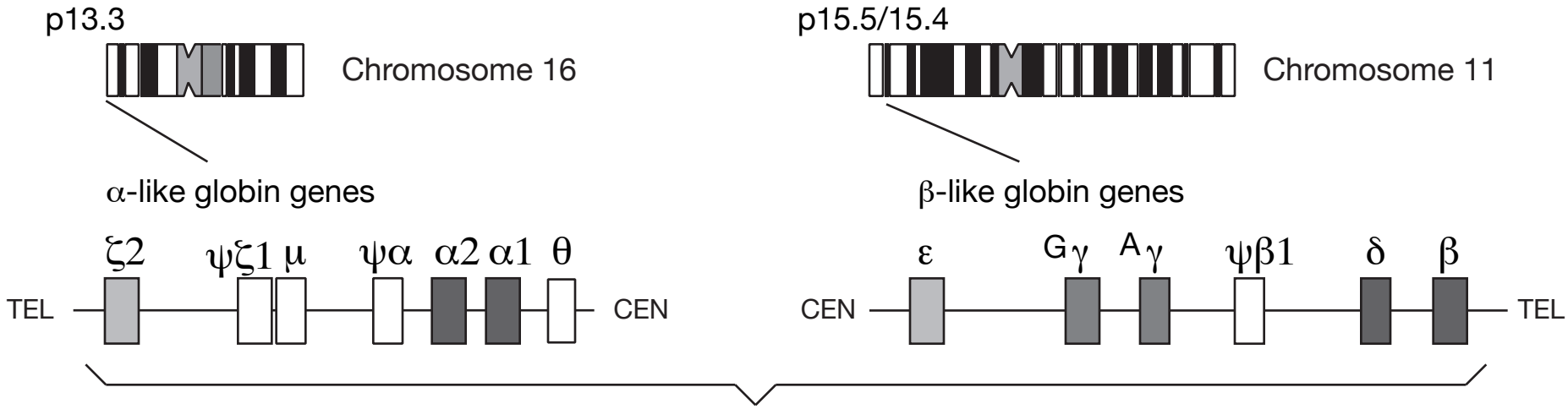
Figure 3.5. Conservation and mutations in globin gene promoters.

(**A**) Basal promoter and (**B**) Upstream promoter for *HBB*. In each panel, the sequence of an 80 bp segment is shown, along with positions of mutations associated with β-thalassemia, conservation scores, and alignments with many mammals, chicken and frog (*X. tropicalis*). The display is from the UCSC Genome Browser in genome coordinates (top line), and the direction of transcription is from right to left (opposite that used in previous figures). The start site of transcription is denoted by the vertical line leading to a leftward arrow. Boxes are drawn around motifs, which are labeled by name and proteins that bind to them (bottom line in each panel).

Figure 3.6. Wide range of conservation in globin gene enhancers.

(**A**) Proximal enhancer for *HBG1*, showing the sequence of part of the 3' enhancer, alignments with sequences of other anthropoid primates, the encompassing repetitive element, and binding motifs.

(**B**) Distal enhancer for the α-globin gene cluster, HS-40. The panel shows an 80 bp segment of the enhancer, along with the Ti~ α-thalassemia deletion that removes this DNA and more, the conservation track and alignments with several eutherian mammals and the marsupial opossum. Binding sites are boxed, and labeled by name and proteins binding to them.



Embryonic	Fetal	Adult
Hb Gower 1: $\zeta_2\varepsilon_2$	Hb F: $\alpha_2\gamma_2$	Hb A: $\alpha_2\beta_2$
Hb Gower 2: $\alpha_2\varepsilon_2$		Hb A_2: $\alpha_2\delta_2$
Hb Portland: $\zeta_2\gamma_2$		

Fig. 1

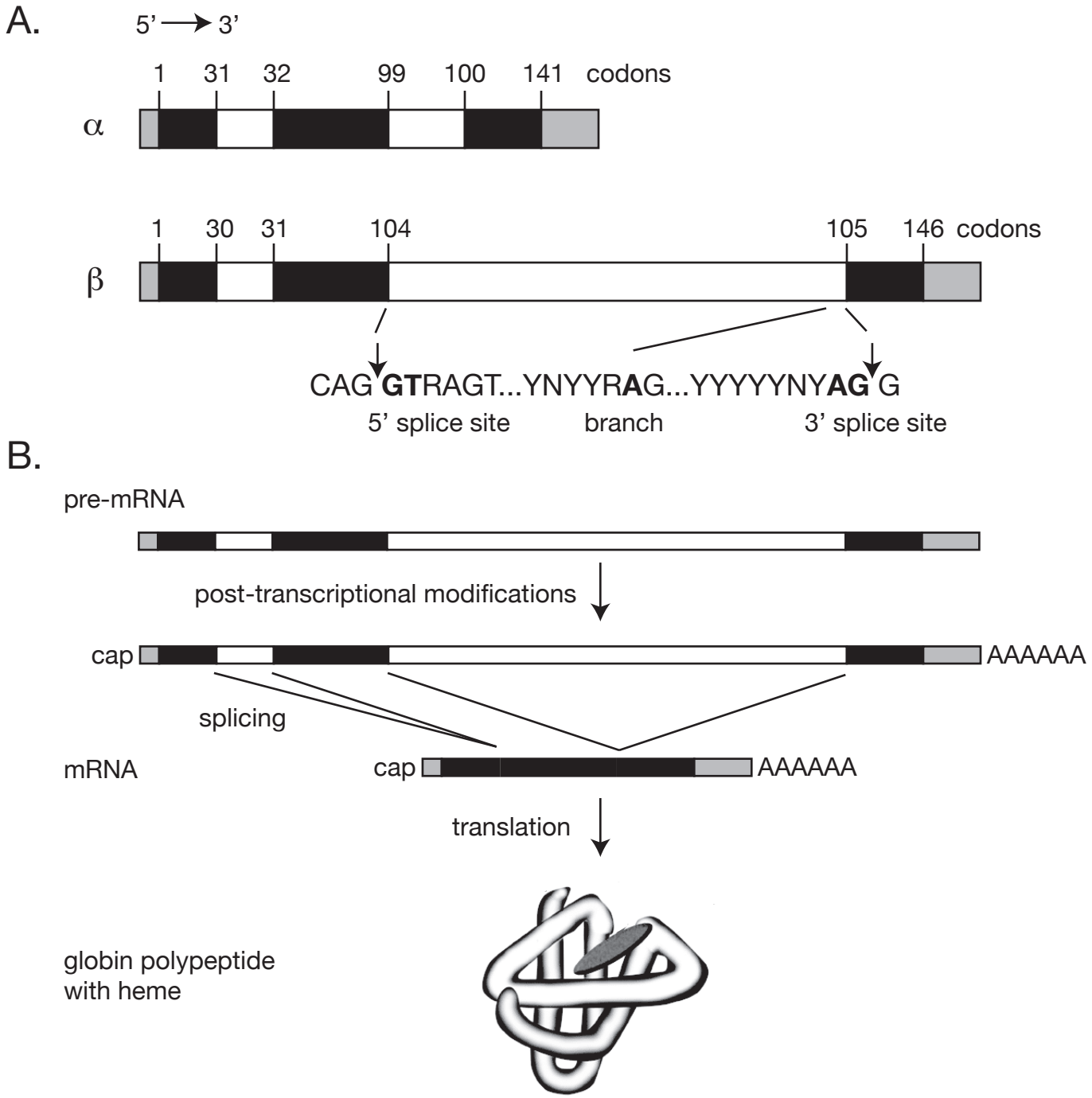

A.
5' → 3'
1 31 32 99 100 141 codons
α
1 30 31 104 105 146 codons
β
CAG GTRAGT...YNYYRAG...YYYYYNYAG G
5' splice site
branch
3' splice site
B.
pre-mRNA
post-transcriptional modifications
cap
AAAAAA
splicing
mRNA
cap
AAAAAA
translation
globin polypeptide
with heme

Fig. 2

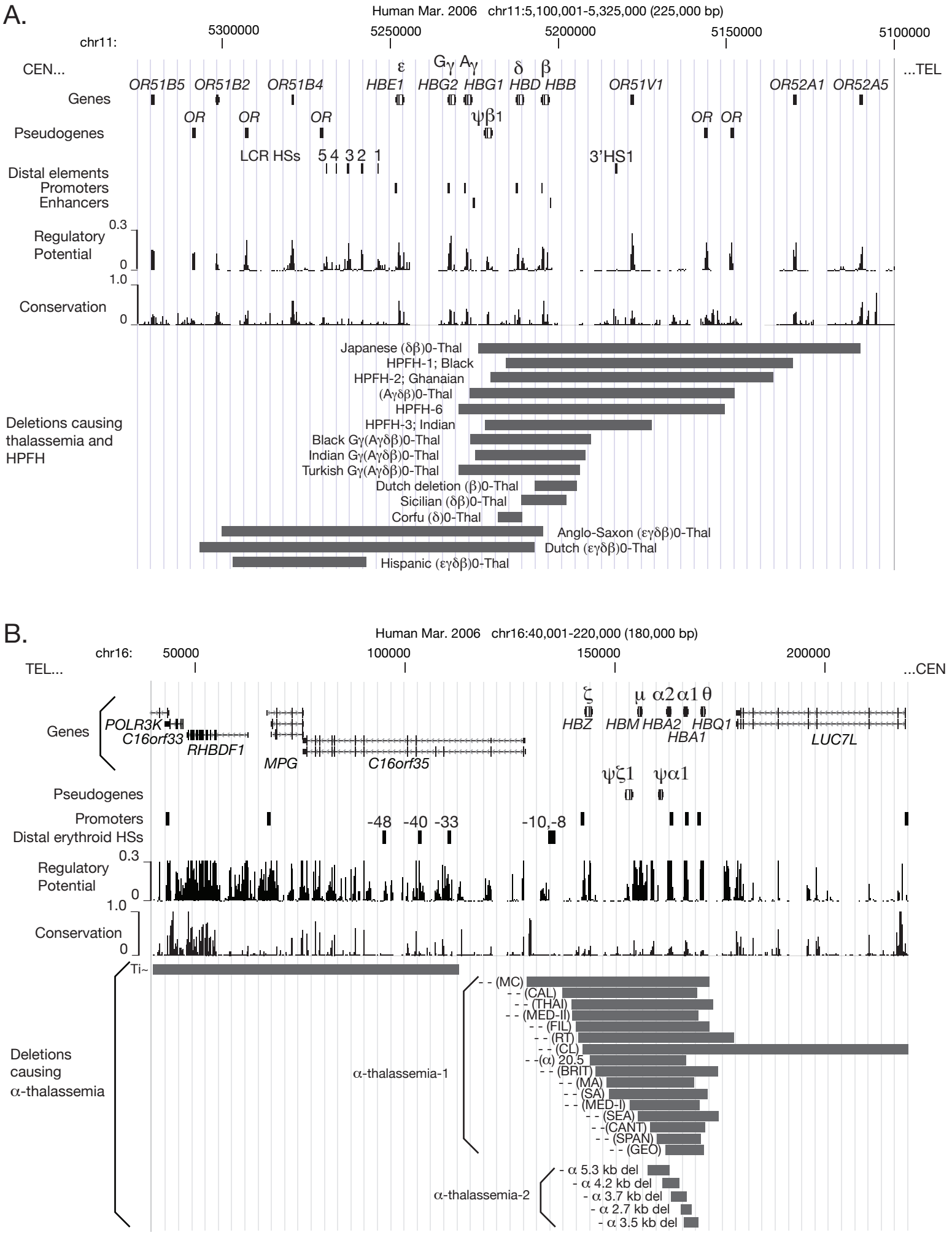
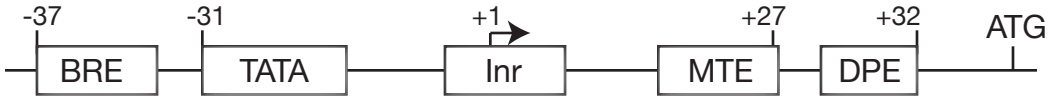

A.
Human Mar. 2006 chr11:5,100,001-5,325,000 (225,000 bp)
chr11:
5300000
5250000
5200000
5150000
5100000
CEN...
...TEL
ε Gγ Aγ δ β
OR51B5 OR51B2 OR51B4 HBE1 HBG2 HBG1 HBD HBB OR51V1 OR52A1 OR52A5
Genes
OR OR OR ψβ1 OR OR
Pseudogenes
LCR HSs 5 4 3 2 1
3'HS1
Distal elements
Promoters
Enhancers
Regulatory Potential
0.3
0
Conservation
1.0
0
Deletions causing thalassemia and HPFH
Japanese (δβ)0-Thal
HPFH-1; Black
HPFH-2; Ghanaian
(Aγδβ)0-Thal
HPFH-6
HPFH-3; Indian
Black Gγ(Aγδβ)0-Thal
Indian Gγ(Aγδβ)0-Thal
Turkish Gγ(Aγδβ)0-Thal
Dutch deletion (β)0-Thal
Sicilian (δβ)0-Thal
Corfu (δ)0-Thal
Anglo-Saxon (εγδβ)0-Thal
Dutch (εγδβ)0-Thal
Hispanic (εγδβ)0-Thal
B.
Human Mar. 2006 chr16:40,001-220,000 (180,000 bp)
chr16:
50000
100000
150000
200000
TEL...
...CEN
ζ μ α2 α1 θ
Genes
POLR3K C16orf33 RHBDF1 MPG C16orf35 HBZ HBM HBA2 HBQ1 HBA1 LUC7L
ψζ1 ψα1
Pseudogenes
Promoters
-48 -40 -33 -10,-8
Distal erythroid HSs
Regulatory Potential
0.3
0
Conservation
1.0
0
Deletions causing α-thalassemia
Ti~
- - (MC)
- - (CAL)
- - (THAI)
- - (MED-II)
- - (FIL)
- - (RT)
- - (CL)
- -(α) 20.5
- - (BRIT)
- - (MA)
- - (SA)
- - (MED-I)
- - (SEA)
- -(CANT)
- - (SPAN)
- - (GEO)
α-thalassemia-1
- α 5.3 kb del
- α 4.2 kb del
- α 3.7 kb del
- α 2.7 kb del
- α 3.5 kb del
α-thalassemia-2

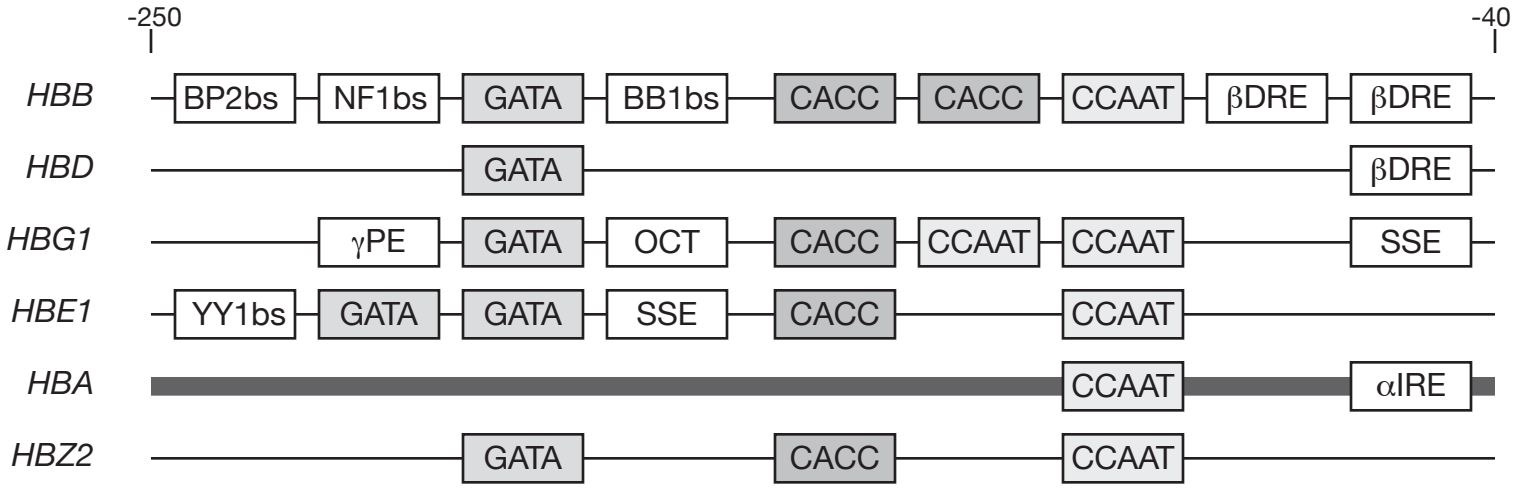
Fig. 3

A. Basal promoter

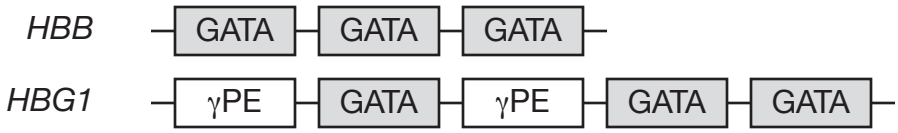


Consensus	SSRCGCC	TATAWAAR	YY**A**NWYY	SSAACGS	RGWYV
HBB	GGGCTGG	CATAAAAG	TTACATT	ACAACTG	AGCAA
HBD	GGACACA	CATAAAAG	TTACACT	ATAACAG	AGCAA
HBG1	CGGCGGC	AATAAAAG	ACACACT	GGAACGT	AGGTT
HBE1	CTTCGGC	AATAAAAG	GCACATA	GCAATCA	AGCAA
HBA	CCGCGCC	CATAAACC	GCACTCT	CAGACTC	AGAGA
HBZ2	CAGCTCC	TATATAAG	CTACCAA	CCAACTC	AGTGC
Consensus for globin genes	SSRCDSC	HATAAAAG	NYACAYW	VNAACNN	AGNTW

B. Upstream regulation



C. Proximal enhancers 3' to genes



D. Distal positive regulators 5' to genes

HBB locus control region

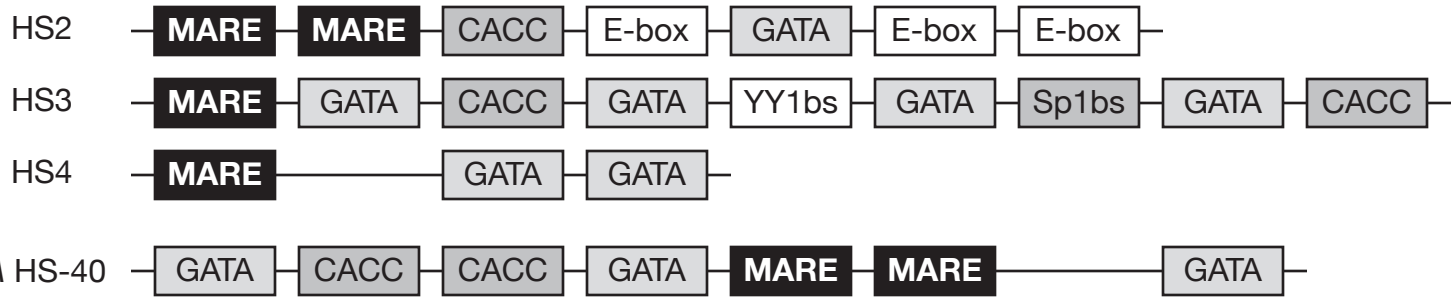
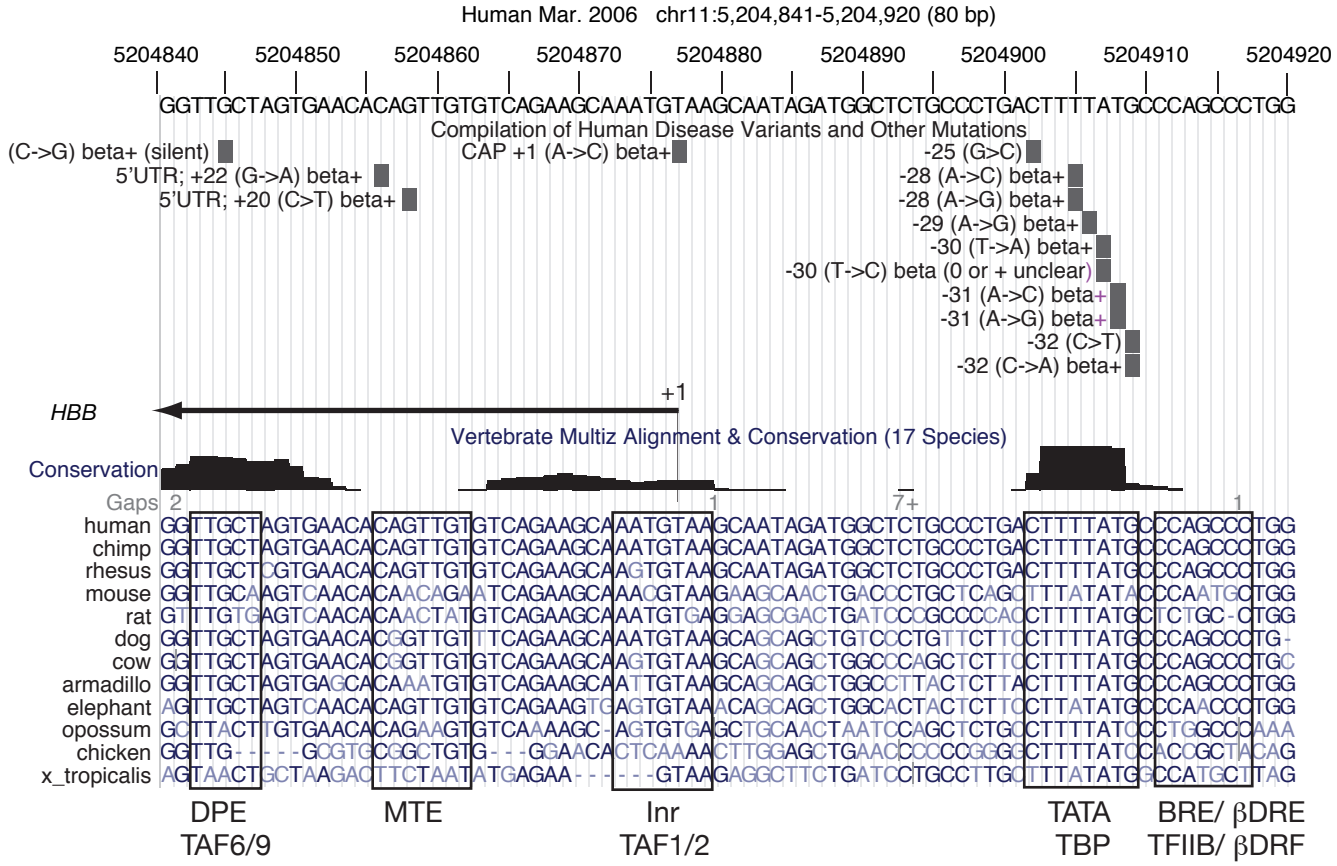


Fig. 4

A. Basal promoter for *HBB*



B. Upstream promoter for *HBB*

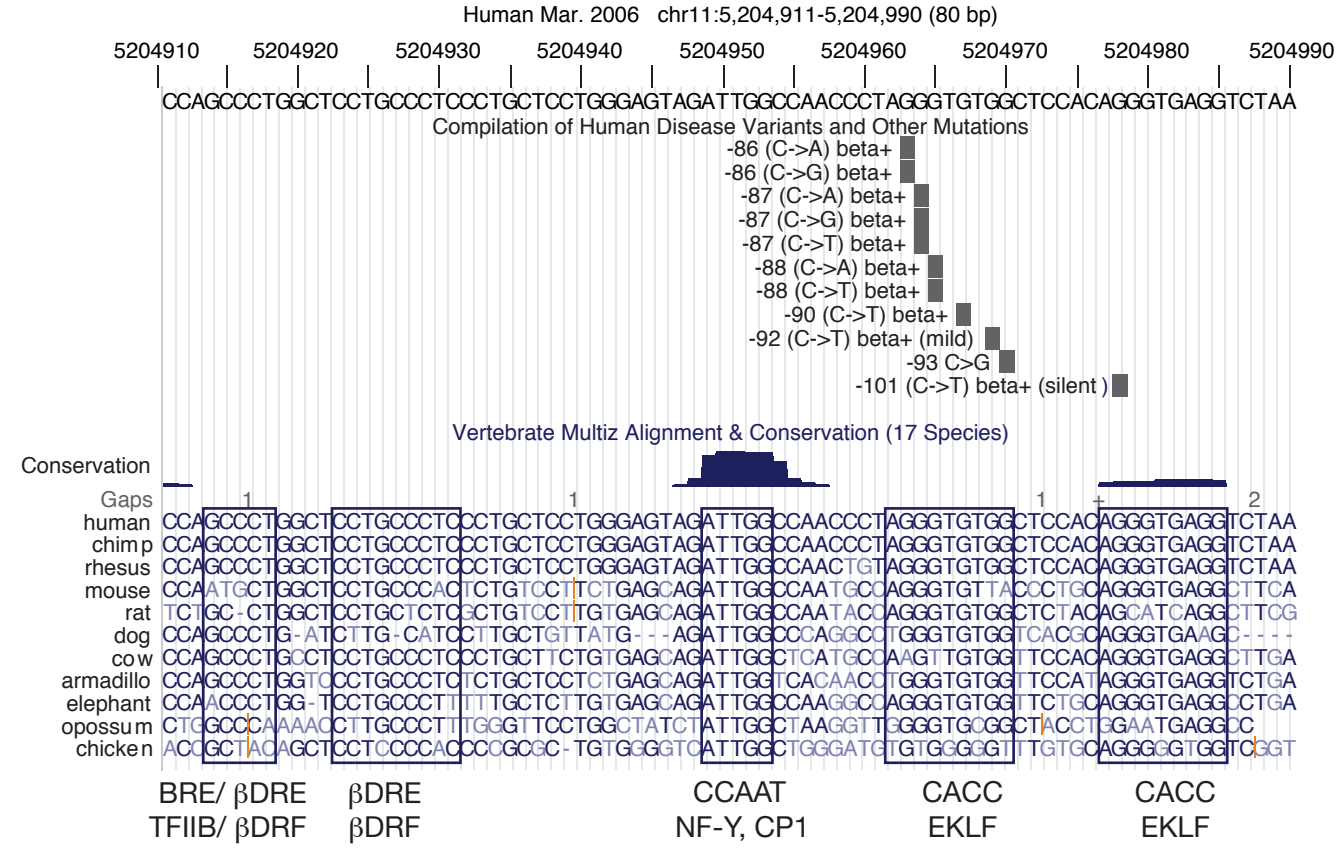
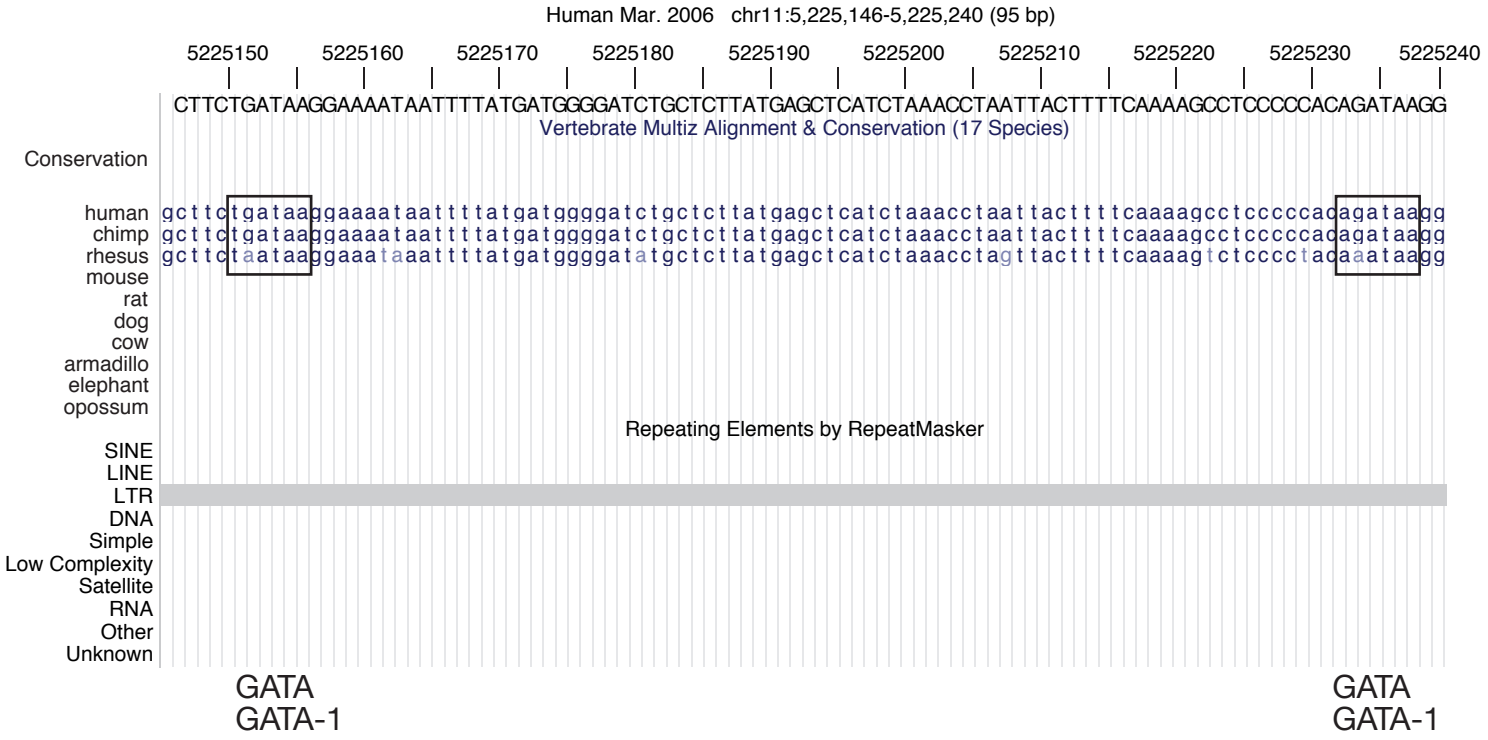


Fig. 5

A. 3’ enhancer for *HBG1*



B. Distal enhancer for *HBA*

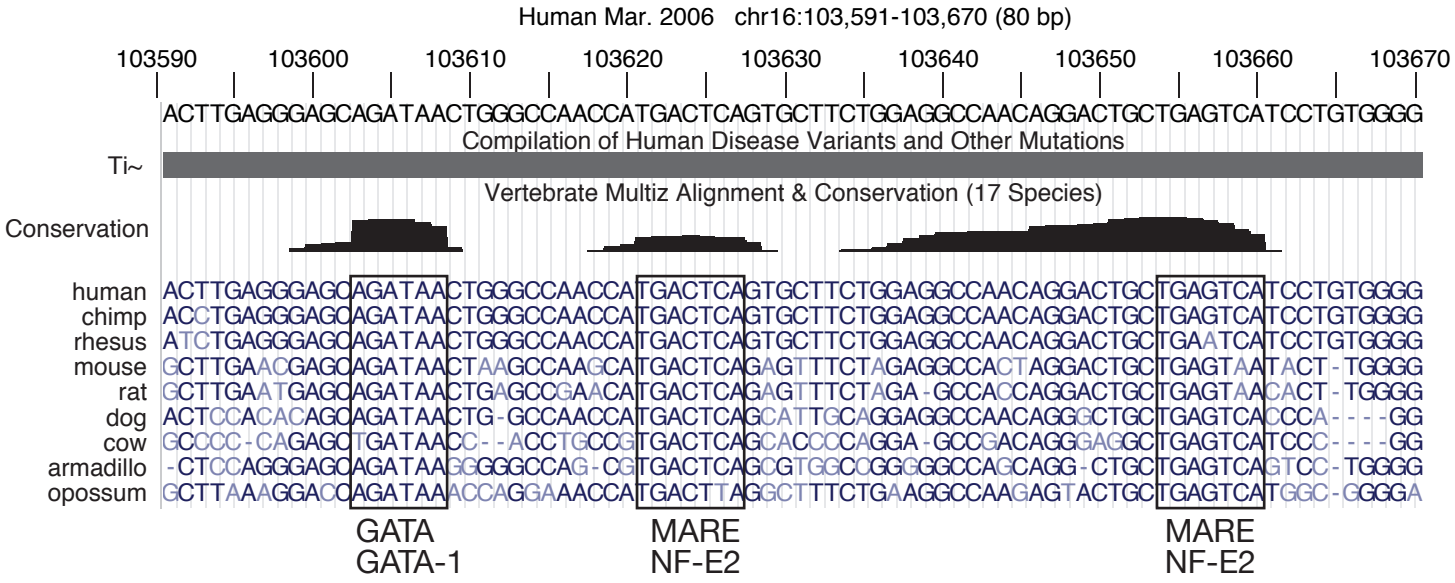


Fig. 6