Globin Genes: Evolution

Ross C Hardison, *Pennsylvania State University, University Park, Pennsylvania, USA*

Hemoglobins carry oxygen from the lungs to other tissues and are encoded by a family of globin genes that are differentially expressed during development. Defects in expression of these globin genes lead to inherited anemias called thalassemias. Other globin genes encode proteins involved in oxygen storage and other functions.

Advanced article

Article contents

- Hemoglobins and Related Proteins in Humans
- Evolution of Globin Genes in Humans and Other Species
- Insights into Gene Regulation from Evolutionary Comparisons
- Impact of Globin Genes on Genetics

doi: 10.1038/npg.els.0005134

Hemoglobins and Related Proteins in Humans

Hemoglobins were originally characterized as the abundant proteins in red blood cells of jawed vertebrates that bind and release oxygen reversibly. The major hemoglobin in adult humans, hemoglobin A (HbA), is a heterotetramer composed of two α-globin and two β-globin polypeptides with their associated heme groups. This multisubunit protein can bind oxygen cooperatively in the lungs (up to one oxygen per heme molecule) and then deliver the oxygen to other tissues. The ability to bind oxygen reversibly is critical to the physiological function of erythrocytes in picking up and delivering oxygen. During this process, neither the oxygen nor the heme groups to which it is bound change chemically (covalently), and the iron in the heme group stays in the reduced state.

Jawed vertebrates make different hemoglobins at progressive stages of development. All species examined have at least one embryonic-specific hemoglobin in primitive red blood cells, which are replaced by fetal and/or adult-specific hemoglobin in definitive red blood cells. The several developmentally regulated hemoglobins are heterotetramers, again with two subunits from the family of proteins related to α-globin and two subunits from the family of proteins related to β globin (**Figure 1**). All of these hemoglobins used for oxygen transport between tissues are produced abundantly but in a highly tissue-specific manner, exclusively in erythroid cells (red blood cells). (***See*** Gene Duplication: Evolution; Gene Families: Formation and Evolution.)

A related monomeric protein, called myoglobin, is an oxygen storage protein found most abundantly in different tissues, in particular skeletal and heart muscle. It may also play a role in delivery of oxygen to the respiring mitochondria within the muscle cells.

Additional globins have been discovered on analysis of the immense wealth of information in the sequence of the human genome and the segments within it that are expressed. Several different groups (e.g. Burmester *et al*., 2002; Gillemans *et al*., 2003; Trent and Hargrove, 2002) have discovered a globin-related protein called cytoglobin (also called histoglobin and stellate-cell activation-associated protein), encoded by the *CYGB* (*cytoglobin*) gene (**Figure 1**). In striking contrast to the specific expression pattern of hemoglobin and myoglobin genes, cytoglobin is expressed in all tissues examined. The most distantly related globin found in the human genome to date is called neuroglobin; its messenger ribonucleic acid (mRNA) is most abundant in brain tissue, but is observed in other tissues. The physiological role of the protein products of these newly discovered globin genes had not been defined by 2002; however, they show that the superfamily of globins in mammals is large and diverse.

In considering possible roles for the cytoglobin and neuroglobin, it is helpful to keep in mind the diversity of functions associated with hemoglobins in nonvertebrate species, including unicellular eukaryotes and prokaryotes. Proteins related to hemoglobin can have enzymatic functions, catalyzing oxidation–reduction reactions more characteristic of other heme-containing enzymes such as cytochromes. In these cases, substrates are oxidized, for example, nitric oxide is oxidized to nitrate, and the iron atom in the heme changes its oxidation state during catalysis. In contrast, the oxygen-binding and -unloading function of the erythroid hemoglobins and muscle myoglobins is not an enzymatic reaction, and the oxidation state of the iron atom in the heme is not changed. Indeed, one can view the erythroid hemoglobins and muscle myoglobins as having evolved by suppression of the ancient and virtually ubiquitous enzymatic activity of hemoglobins. In this light, it is intriguing that the heme bound to cytoglobin has a different coordination chemistry than the erythroid hemoglobins and muscle myoglobins, suggestive of a distinctive biochemical function (Trent and Hargrove, 2002).

The range of physiological functions of hemoglobins, even in humans, remains to be fully explored.

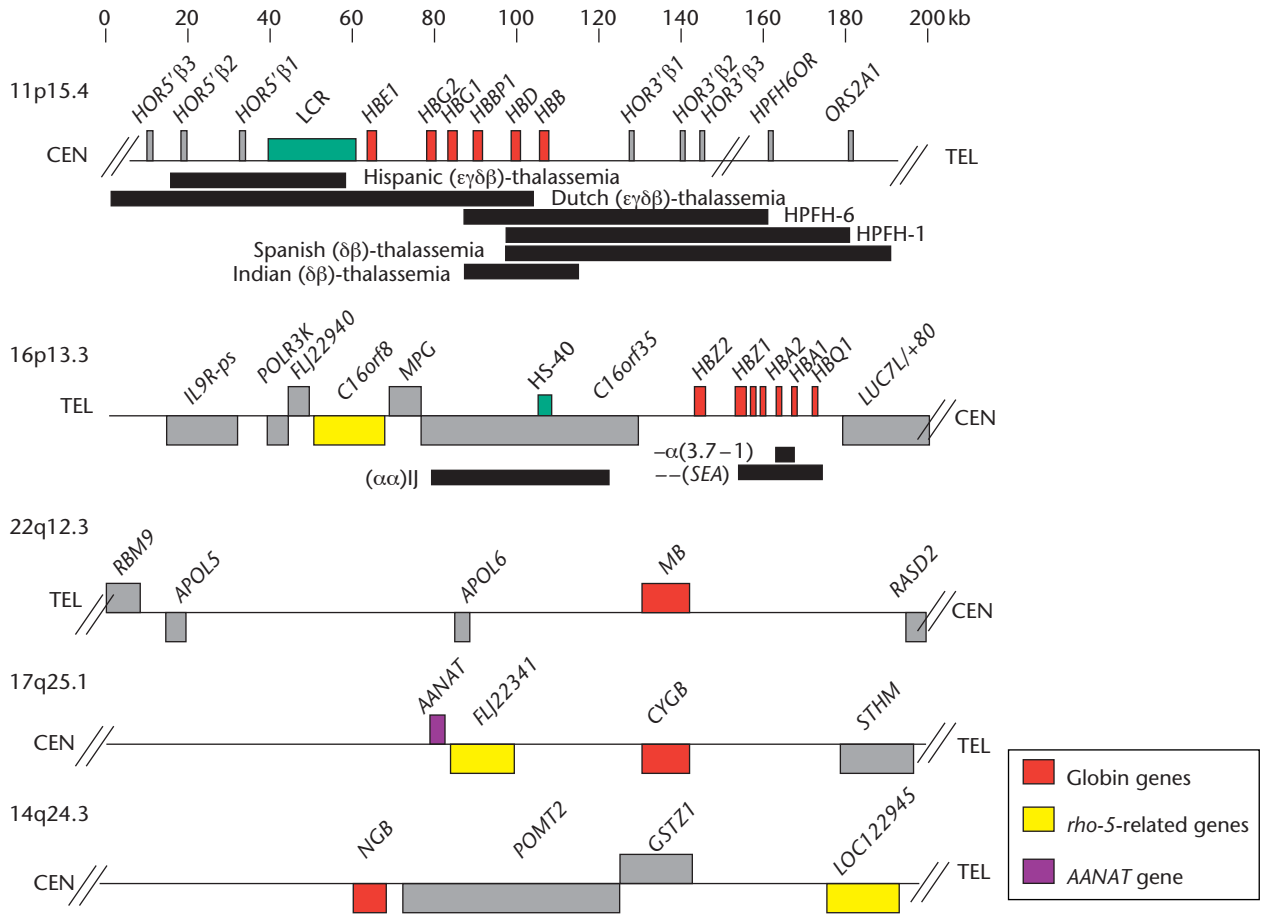

Figure 1 Maps of genes and gene complexes coding for globins in humans. An approximately 200-kb region surrounding the globin gene or gene complex is shown, with genes transcribed from left to right shown as boxes above the lines, and those transcribed in the opposite direction shown as boxes below the lines. Globin genes are hatched; genes related to the *Drosophila* rhomboid-5 (*rho-5*) gene (*C16orf8*) (*chromosome 16 open reading frame 8*) and *FLJ22341* (*hypothetical protein FLJ22341*) are white; and the *AANAT* (*arylalkylamine N-acetyltransferase*) gene is cross-hatched. These three comprise an ancient syntenic group. Chromosomal locations are given to the left of each map. CEN and TEL: centromeric and telomeric ends of each arm of the chromosomes. LCR: the distal locus control region for the globin gene complex (top map); HS-40: the major control region for the α-globin gene complex (second map). Regions deleted in selected β- and α-thalassemia mutations are shown as the bars beneath the top and second maps respectively. HPFH-1, HPFH-6: hereditary persistence of fetal hemoglobin; ps: pseudogene.

Proteins such as neuroglobin and cytoglobin were discovered recently and functional tests are underway. The role of myoglobin in oxygen transport to mitochondria within muscle cells remains controversial, whereas its role in oxygen storage is well established. The role of erythroid hemoglobins in carrying oxygen between tissues is unquestioned, and it is likely that the different forms of such hemoglobins are produced at different stages of development to enhance transport of oxygen from the mother to the developing embryo and fetus. However, this does not rule out other possible physiological functions of these proteins that are extraordinarily abundant in erythroid cells. Some work has implicated hemoglobin in transport or regulation of nitric oxide in humans, thereby modulating the activity of this critical molecule in the vasoconstriction response.

Evolution of Globin Genes in Humans and Other Species

The large family of proteins related to hemoglobins are encoded at five different chromosomal locations in humans. The erythroid hemoglobins are encoded by multigene complexes on separate chromosomes in birds and mammals. In humans, the α-like globin

gene complex is at chromosomal location 16p13.3, whereas the β-like globin gene complex is at 11p15.4 (**Figure 1**). These gene families can be grouped together in some fish and amphibians. The members of these tightly linked gene complexes are developmentally regulated, whereas the expression of α-like and β-like globin genes between multigene complexes is coordinated, so that equal amounts of the two types of globin are produced in all erythroid cells. Myoglobin is encoded by the *MB* (*myoglobin*) gene at chromosomal position 22q12.3, cytoglobin is encoded by *CYGB* at chromosomal position 17q25.1 and neuroglobin is encoded by *NGB* (*neuroglobin*) at chromosomal position 14q24.3 (**Figure 1**). These locations are listed in genome browsers such as the UCSC Genome Browser (see Web Links section; Kent *et al*., 2002). *MB*, *CYGB* and *NGB* contain single globin genes, in contrast to the multigene complexes that have evolved to encode the erythroid hemoglobins. (***See*** Gene Families: Multigene Families and Superfamilies; Genome Databases; Homologous, Orthologous, and Paralogous Genes; Transcriptional Regulation: Coordination.)

The nonglobin genes found nearby on the same chromosome (close synteny) give important insights into the evolution of the five globin loci (Flint *et al*., 2001; Gillemans *et al*., 2003). Approximately 100 kb telomeric to the α-like globin gene complex in humans is the gene *C16orf8* (*chromosome 16 open reading frame 8*), which encodes a homolog of the *Drosophila melanogaster* rhomboid-5 protein, an integral membrane protein. Orthologs of this gene and the intervening genes (*MPG* (*N-methylpurine-DNA glycosylase*); and *C16orf35* (*chromosome 16 open reading frame 35*) (also known as *CGTHBA*)) are also found in the α-globin gene complex of puffer fish, showing that this syntenic group predates divergences early in the vertebrate lineage. The gene *AANAT* (*arylalkylamine N-acetyltransferase*) is adjacent to the rhomboid-5 homolog in puffer fish; this gene encodes arylalkylamine *N*-acetyltransferase, which carries out the penultimate step in melatonin biosynthesis. Interestingly, *AANAT* and another homolog of rhomboid-5 (*FLJ22341* (*hypothetical protein FLJ22341*)) are located approximately 30 kb away from the *CYGB* gene on chromosome 17. This syntenic group is paralogous to the *HBA* (hemoglobin, alpha) region on chromosome 16, indicating that this syntenic group is even older than the early divergences in jawed vertebrates (**Figure 2**). (***See*** Comparative Genomics.)

A model for the evolution of these loci containing globin genes that is consistent with the synteny relationships and extent of amino acid sequence (Burmester *et al*., 2002; Gillemans *et al*., 2003) homology is shown in **Figure 2**. The globin gene family is very old, but the repeated gene duplications to generate linked groups of multiple globin genes are characteristic of the erythroid globin gene complexes. The expansion and contraction of the number of genes has occurred independently for the α- and β-like globin gene complexes. This continues in contemporary time, as demonstrated by α- and β-thalassemias, which are inherited anemias resulting from a loss of one or more globin genes, and the fact that the human population is polymorphic for the number of α-globin genes.

The syntenic group, including *AANAT* (a homolog of *rho-5*) and a globin gene, probably predates the divergence of jawed and jawless vertebrates; if so it would be ancestral to all the globin genes except perhaps *NGB*. The absence of *AANAT* and a rhomboid-5 homolog from the region around *MB* can be explained by a chromosomal translocation close to *MB* (corresponding to a break in the conservation of synteny between human and mouse) and the apparent transposition of the β-like globin gene complex into an olfactory receptor gene cluster (Bulger *et al*., 2000). (***See*** Racism, Ethnicity, Biology and Society.)

Some structural features are conserved in all globin genes. All the active genes have at least three exons. Two introns are found in the erythroid globin genes and *MB*; although they differ dramatically in size, they are in homologous locations (**Figure 3**). Furthermore, even more distantly related globin genes, such as *CYGB* and *NGB* in vertebrates and plant hemoglobin genes, have introns in positions homologous to those in *HBB* (*globin, beta*), *HBA* and *MB*. This was one of the first observed examples of conservation of introns in eukaryotic genes and it was interpreted functionally, for example, that the presence of the introns facilitated the shuffling of exons during protein evolution. Allied with this hypothesis was the notion that exons encoded structural and/or functional domains of proteins.

Although several notable examples of the latter have been demonstrated, and indeed the presence of introns in such cases could reasonably make it easier for new protein structures to form by recombination, it is also apparent that this need not be the case for all introns. For instance, the human *CYGB* and *NGB* genes each have an additional intron, but these are not in homologous locations; in fact they are at different ends of the genes. Additional introns in other plants and animals can interrupt the globin gene exons in different locations; for example, the central exon observed in the plant leghemoglobin gene and some invertebrate hemoglobin genes. Intron positions in globin genes of unicellular eukaryotes are highly variable. As this issue has been examined over a wider phylogenetic distance and for other gene families, many examples of intron acquisition have been elucidated. Thus a role for introns in facilitating exon shuffling cannot apply uniformly to all introns, but it may be true in some cases. (***See*** Exons: Shuffling; Introns: Movements.)

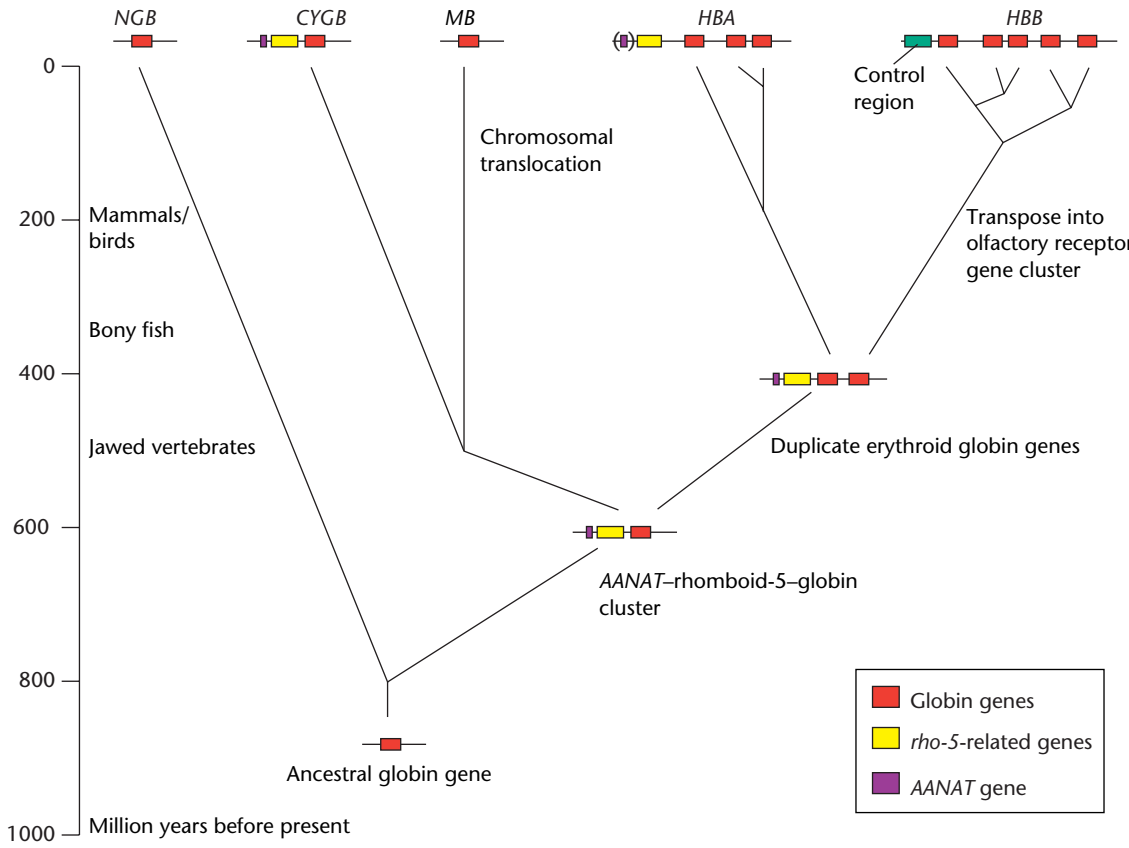

Figure 2 Model of the evolution of vertebrate globin genes. The deduced times of duplication and divergence are shown along the horizontal axis, and contemporary human globin genes or gene complexes are shown at the top. Major events in globin gene evolution are noted along the tree, and time of origin of some major vertebrate groups is indicated along the horizontal axis. Globin genes are hatched; genes related to the *Drosophila* rhomboid-5 gene are white; and the *AANAT* gene is cross-hatched. (Revised and redrawn from a figure in Gillemans *et al*. (2003).)

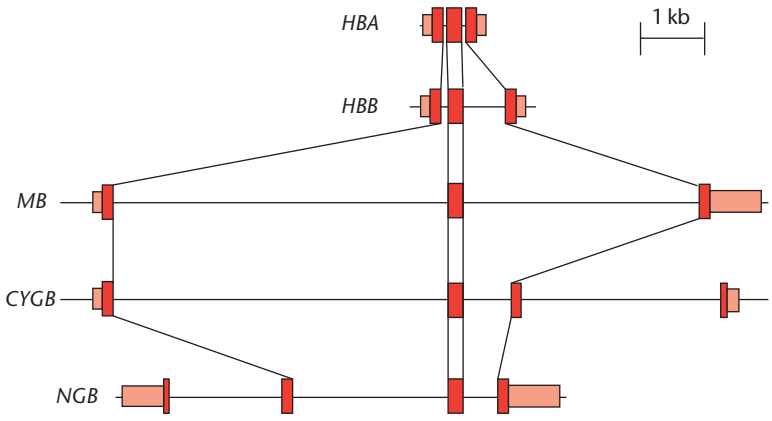

Figure 3 Similarities and differences in globin gene structure. The protein-coding exons of the indicated genes are drawn as red rectangles, and the untranslated regions of the exons are drawn as orange rectangles. All genes are oriented with their direction of transcription from left to right. Lines are drawn to connect homologous splice junctions.

Insights into Gene Regulation from Evolutionary Comparisons

The model of the evolution of globin genes (**Figure 2**) reveals several points at which the pattern of expression changed, presumably by acquisition of new regulatory elements. Although *CYGB* is expressed in all cells, *MB* was recruited to more muscle-specific expression, and the *HBA* and *HBB* genes and their relatives were recruited to high-level, erythroid-specific expression. All these genes share a common ancestor, but the regulatory elements have changed considerably in each lineage.

This suggests that a search for conserved regulatory elements should be restricted to one lineage outlined in **Figure 2**. Indeed, although almost no similarity is observed in alignments of promoter regions between *HBA* and *HBB*, examination of the promoters of orthologous *HBB* genes in eutherian mammals shows obvious regions of high conservation. Indeed, comparisons between noncoding genomic DNA sequences of human and mouse frequently reveal conservation in the promoter and enhancer regions (Waterston *et al*., 2002); this is not just a property of the globin gene complexes. Noncoding regions whose level of conservation probably reflects selection are strong candidates for regulatory elements (Gumucio *et al*., 1996). Methods for analyzing the likelihood that an aligned segment reflects selection are being developed and can be accessed at the UCSC Genome Browser. (***See*** Phylogenetic Footprinting; Similarity Search; Transcriptional Regulation: Evolution.)

Although the promoters and some enhancers of globin genes are close to the 5′ ends of the genes, other regulatory elements can be far away. Well-characterized, distal regulatory elements for mammalian globin genes include the β-globin locus control region (LCR), covering approximately 17 kb, located 60 kb centromeric to the *HBB* gene and lying between the β-like globin genes and the olfactory receptor genes (**Figure 1**). Another is the major regulatory element (MRE or HS-40) of the α-globin gene complex, a much smaller (approximately 300 bp) regulatory element located in an intron of the *C16orf35*, which is telomeric to all the α-like globin genes (**Figure 1**). These distal regulatory elements were discovered by the confluence of several lines of investigation, including mapping DNase-hypersensitive sites in chromatin, the effects of naturally occurring thalassemic mutations (**Figure 1**), and sequence conservation in mammals that is substantially higher than that seen for most other intergenic regions. The deletion of the β-globin LCR in the Hispanic and Dutch thalassemias and the deletion of the α-globin HS-40 in (αα)IJ α-thalassemia caused a substantial decrease in the level of expression of the still-intact *HBB* gene and *HBA* genes respectively (**Figure 1**). Inclusion of these distal regulatory elements in globin transgenes, including large constructs with the entire gene clusters, conferred the ability for these transgenes to be expressed at a high level in erythroid cells at most sites of integration (Grosveld *et al*., 1987; Higgs *et al*., 1990). Hence these distal elements are strong enhancers, and they have been implicated in reducing the effects of the site of integration on expression. The thalassemia phenotypes of the deletions, combined with a very large number of studies in transgenic animals and transfected cells, have shown that these are powerful regulators of globin gene expression. Some of these studies were directed by the pattern of sequence conservation, and several novel components of the LCR were discovered by this approach. Indeed, the success of these applications of the evolutionary studies to finding new aspects of regulation in globin genes is being recapitulated in other systems and has helped drive the approach genome-wide (Waterston *et al*., 2002). (***See*** Chromatin Structure and Domains; Thalassemias.)

Impact of Globin Genes on Genetics

Studies of globin genes have had a substantial impact on genetics, just as studies of the hemoglobin protein has greatly influenced biochemistry. Disorders of the hemoglobins are the most common inherited diseases in humans. Many variants of hemoglobin have been studied structurally and functionally, such that mutations in almost every amino acid position in α-globin and β-globin have been found in humans (and some positions have been mutated multiple times). A useful database of the hemoglobin variants is at the Globin Gene Server (see Web Links section). Although most of these variants cause no obvious disease, some do. The most common is sickle cell hemoglobin, HbS, which is caused by a replacement of a glutamate at position 6 of β-globin with a valine. This was the first mutation ever defined at the molecular level, in classic work completed by Dr Vernon Ingram in the 1950s. A consequence of this amino acid replacement is that the HbS will form large polymers and precipitate when deoxygenated. This leads to a change in morphology of the erythrocyte so that it is sickle- rather than disc-shaped, and it becomes more rigid. This in turn makes it difficult for the sickle cells to traverse capillaries, and the consequent vaso-occlusion is the cause of much of the pathophysiology of this too-common disease. Other diseases result from mutations in the globin polypeptides that, for example, render the hemoglobin unstable, increase its oxygen affinity excessively or cause the heme iron to be oxidized. (***See*** Globin

genes: Polymeric Variants and Mutations; Human Variation Databases; Sickle Cell Disease as a Multifactorial Condition.)

Inherited anemias can be caused by inadequate production of α globin or β globin, referred to as α-thalassemia and β-thalassemia respectively. Mutant alleles of *HBB* with a thalassemia phenotype have been discovered that affect almost every step in the pathway of gene expression, including transcription, splicing and translation. Analysis of such naturally occurring mutations was instrumental in uncovering some of the important steps in gene expression. Thalassemia can also result from deletion of globin genes (Spanish and Indian (δβ)-thalassemia; $-\alpha(3.7–1)$; and $--(SEA)$ α-thalassemia in **Figure 1**), or from deletion of major enhancer or LCR sequences, as previously discussed. Some deletions of *HBB* are not so detrimental, because the expression of *HBG1* (*globin, gamma A*) and *HBG2* (*globin, gamma G*), which is normally restricted to fetal life, continues after birth in a condition known as hereditary persistence of fetal hemoglobin (HPFH-6 and HPFH-1 in **Figure 1**).

Although disorders of hemoglobin such as sickle cell disease and thalassemias are common in some human populations, more effective therapies are needed. Studies of the detailed mechanisms of globin gene regulation are pursued with the hope that this information will allow development of effective gene therapy strategies, or that they will lead to the design of pharmaceuticals that modulate the regulation in a therapeutic manner. For instance, compounds that caused fetal hemoglobin production to continue after birth could be effective in counteracting the effects of mutant or insufficient adult hemoglobin. (***See*** Carrier Screening for Inherited Hemoglobin Disorders in Cyprus and the United Kingdom; Hemoglobin Disorders: Gene Therapy.)

See also

Globin Genes: Polymorphic Variants and Mutations
Major Histocompatibility Complex (MHC) Genes: Evolution
Thalassemias

References

Bulger M, Bender MA, von Doorninck JH, *et al*. (2000) Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse β-globin gene clusters. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 14 560–14 565.

Burmester T, Ebner B, Weich B and Hankeln T (2002) Cytoglobin: a novel globin type ubiquitously expressed in vertebrate tissues. *Molecular Biology and Evolution* **19**: 416–421.

Flint J, Tufarelli C, Peden J, *et al*. (2001) Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Human Molecular Genetics* **10**: 371–382.

Gillemans N, McMorrow T, Tewari R, *et al*. (2003) A functional and comparative analysis of globin loci in puffer fish and man. *Blood* (in press).

Grosveld F, van Assendelft GB, Greaves D and Kollias G (1987) Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell* **51**: 975–985.

Gumucio D, Shelton D, Zhu W, *et al*. (1996) Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the β-like globin genes. *Molecular Phylogenetics and Evolution* **5**: 18–32.

Higgs D, Wood W, Jarman A, *et al*. (1990) A major positive regulatory region located far upstream of the human α-globin gene locus. *Genes and Development* **4**: 1588–1601.

Kent WJ, Sugnet CW, Furey TS, *et al*. (2002) The human genome browser at UCSC. *Genome Research* **12**: 996–1006.

Trent III JT and Hargrove MS (2002) A ubiquitously expressed human hexacoordinate hemoglobin. *Journal of Biological Chemistry* **277**: 19 538–19 545.

Waterston RH, Lindblad-Toh K, Birney E, *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

Further Reading

Dickerson RE and Geis I (1983) *Hemoglobin: Structure, Function, Evolution and Pathology*. Menlo Park, CA: Benjamin/Cummings Publishing Company, Incorporated.

Forget BG, Higgs DR, Steinberg M and Nagel RL (2001) *Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management*. Cambridge, UK: Cambridge University Press.

Forrester WC, Epner E, Driscoll MC, *et al*. (1990) 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 and Development* **4**: 1637–1649.

Garry DJ, Ordway GA, Lorenz JN, *et al*. (1998) Mice without myoglobin. *Nature* **395**: 905–908.

Gilbert W (1978) Why genes in pieces? *Nature* **271**: 501.

Godecke A, Flogel U, Zanger K, *et al*. (1999) Disruption of myoglobin in mice induces multiple compensatory mechanisms. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 10 495–10 500.

Hardison R (1998) Hemoglobins from bacteria to man: evolution of different patterns of gene expression. *Journal of Experimental Biology* **201**: 1099–1117.

Hardison R and Miller W (1993) Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. *Molecular Biology and Evolution* **10**: 73–102.

Hardison R, Slightom JL, Gumucio DL, *et al*. (1997) Locus control regions of mammalian β-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights. *Gene* **205**: 73–94.

Hardison RC, Chui DHK, Giardine B, *et al*. (2001) HbVar: a relational database of human hemoglobin variants and thalassemia mutations at the Globin Gene Server. *Human Mutation* **19**: 225–233.

Ingram VM (1956) A specific chemical difference between globins of normal human and sickle-cell anemia hemoglobin. *Nature* **178**: 792.

Web Links

Globin Gene Server. The function of DNA sequences, especially those involved in production of hemoglobin
http://globin.cse.psu.edu

UCSC Genome Bioinformatics. Genome browser
http://genome.ucsc.edu

AANAT (*arylalkylamine N-acetyltransferase*); LocusID: 15. Locus Link: http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l = 15.
CYGB (*cytoglobin*); LocusID: 114757. Locus Link: http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l = 114757.
HBB (*hemoglobin, beta*); LocusID: 3043. Locus Link: http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l = 3043.
MB (*myoglobin*); LocusID: 4151. Locus Link: http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l = 4151.
NGB (*neuroglobin*); LocusID: 58157. Locus Link: http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l = 58157.

AANAT (*arylalkylamine N-acetyltransferase*); MIM number: 600950. OMIM: http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispmim?600950.
HBB (*hemoglobin, beta*); MIM number: 141900. OMIM: http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispmim?141900.
MB (*myoglobin*); MIM number: 160000. OMIM: http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispmim?160000.
NGB (*neuroglobin*); MIM number: 605304. OMIM: http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispmim?605304.