Rabbits and Humans Indicates That the Gene Cluster 5'-ε-γ-δ-β-3' Predates the Mammalian Radiation

Ross C. Hardison
Department of Biochemistry, Microbiology, Molecular and Cell Biology,
Pennsylvania State University

The members of the rabbit and human β-like globin gene families have been compared both by a computer-generated dot matrix graphical analysis of each entire gene and by calculating divergences in the coding regions. The rabbit-human gene pairs β4-ε, β3-γ, ψβ2-δ, and β1-β were identified as orthologous on the basis of sequence similarities found in flanking and intervening sequences as well as by quantitative divergence calculations. The orthologous genes are in the same order on the chromosome in each species, which suggests that an ancestral family with the arrangement 5'-ε-γ-δ-β-3' preceded the mammalian radiation. Descendants of ancestral ε have diverged more slowly than other β-like genes and are expressed only in embryonic life. Descendants of ancestral γ and β diverged at a higher rate and are expressed at wider range of developmental times. Descendants of δ have undergone nonreciprocal recombination at a high frequency and are often pseudogenes. Paralogous comparisons among the rabbit β-like globin genes show that the β4-β3 and ψβ2-β1 pairs are most similar and that β4 and β3 are more closely related to β1 than to ψβ2. This fits with a branching pattern where the primordial β split into ancestral ε/γ and δ/β genes, which later split into ε and γ or δ and β, respectively. Rabbit genes β4 and β1 acquired similar 3' untranslated regions after the ε/γ split but prior to the mammalian radiation, presumably via a gene conversion event. The 5' end of β2 apparently converted with β1 after the radiation, and afterward it became a pseudogene.

Introduction

Organisms that use hemoglobin as an oxygen transport protein generally synthesize different hemoglobins at different stages of development. The globin polypeptides that comprise these different hemoglobins are encoded by families of related genes whose expression is developmentally regulated. In jawed vertebrates, the hemoglobin is composed of two α-like and two β-like globin polypeptides, each with a bound heme. The α-like and β-like globin gene expression is coordinately regulated so that equal amounts of both types are synthesized, but the individual members of the α-like and β-like gene families are expressed at different developmental stages. The α-like and β-like gene families are closely linked in the amphibian Xenopus (Jeffreys et al. 1980; Patient et al. 1980) but are located on different chromosomes in chickens (Hughes et al. 1979) and mammals (Russell and McFarland 1974; Deisseroth et al. 1977, 1978). Thus, in avians and mammals, the developmentally regulated members of the α- or β-globin gene families are closely linked, and the coordinately expressed members of each family (e.g., adult α- and β-globin) are on separate chromosomes.

1. Key words: β-like globin genes, ancestral gene families, divergence rates, rabbits, humans.

Address for correspondence and reprints: Ross C. Hardison, Biochemistry Program, 206 Althouse Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802.
The individual globin genes within these families share a common structure. Animal globin genes are split by two introns (intervening sequences) into three exons (message-coding blocks). The introns are located at equivalent positions in members of both α- and β-like globin gene families (Leder et al. 1978). The more distantly related gray seal myoglobin also retains two introns, albeit unusually large ones, in equivalent positions (Blanchetot et al. 1983). The soybean leghemoglobin has two introns in equivalent positions plus a third intron which separates the homologue of the animal central exon into two blocks that encode predicted protein domains (Gö 1981; Hyldig-Nielsen et al. 1982; Brisson and Verma 1982). Thus leghemoglobin could possibly be the descendant of a more primitive globin gene from which the central intron was lost in the parent to myoglobin and globin genes. Conservation of overall structure is a distinguishing characteristic of all types of globin genes.

The ready availability of hemoglobin proteins and, more recently, of cloned globin genes has facilitated determination of their sequence and subsequent use of these data in evolutionary comparisons (e.g., Dayhoff 1972; Czelusniak et al. 1982). These comparisons, usually based on single genes or polypeptides from each of two or more species, have been used to construct phylogenetic trees and to identify conserved sequences that may be important in the regulated expression of these genes. Comparisons have been published of β-like globin genes in humans (Efstratiadis et al. 1980), Old World monkeys (Martin et al. 1983), lemurs (Jeffreys et al. 1982), goats (Haynes et al. 1980; Cleary et al. 1981; Shapiro et al. 1983), mice (Jahn et al. 1980), and chickens (Roninson and Ingram 1982; Dodgson et al. 1983), and α-like globin genes in humans (Proudfoot et al. 1982), goats (Schon et al. 1982), and chickens (Engel et al. 1983). In no case, however, have all the members of a globin gene family in one species been compared with the members of that family in another species. Such an analysis of entire gene families should yield optimal information. For example, comparisons between individual members of related gene families in two different species will be of two types (Wilson et al. 1977): orthologous comparisons between genes which are derived from the same gene in the last common ancestral species, and paralogous comparisons between genes that arose by gene duplication either before or after divergence from the last common ancestor. By comparing the globin genes as members of linked families, one can assign orthologous and paralogous relationships and thereby obtain more complete information on the evolutionary history of the gene family. Comparisons between families may also reveal sequences which may be critical in the developmental regulation of the linked genes.

The nucleotide sequence determination of each member of the rabbit (Oryctolagus cuniculus) β-like globin gene family has now been completed. This family is composed of four genes arranged in the order 5'-β4-β3-ψβ2-β1-3' (Lacy et al. 1979). Genes β4 and β3 are expressed only in embryonic erythrocytes; gene β1 is expressed in fetal and adult red cells (Hardison et al. 1979; Rohrbaugh and Hardison 1983); and ψβ2 is a pseudogene incapable of producing a functional globin (Lacy and Maniatis 1980). The sequences of two alleles of β1 (Hardison et al. 1979; van Ooyen et al. 1979) and one allele each of ψβ2 (Lacy and Maniatis 1980; Hardison and Margot 1984), β3 (Hardison 1981), and β4 (Hardison 1983) have been published. In this paper, each of the rabbit sequences is compared with the other and with the functional human globin genes: adult β (Lawn et al. 1980), minor adult δ (Spritz et al. 1980), fetal γ (Slightom et al. 1980; Shen et al. 1981), and embryonic e (Baralle et al. 1980). Each individual rabbit β-like globin gene is more related to one of the human globin genes than to any other rabbit gene. Extensive similarity in the introns was found to be a
good indicator of orthologous relationships, and the following rabbit-human gene pairs were identified as orthologous: $\beta_4$-$\epsilon$, $\beta_3$-$\gamma$, $\psi\beta_2$-$\delta$, and $\beta_1$-$\beta$. The orthologous genes are arranged in the same order along the chromosome, thus indicating that an ancestral globin gene family, $5'$-$\epsilon$-$\gamma$-$\delta$-$3'$, existed prior to the mammalian radiation. The descendants of each member of the ancestral family have accumulated mutations at different rates. Some ramifications of these conclusions for mammalian $\beta$-like globin gene evolution are discussed.

**Material and Methods**

**Computer-generated Dot Matrix Comparisons**

The nucleotide sequences of the rabbit and human $\beta$-like globin genes were compared graphically by a dot matrix analysis (Konkel et al. 1979), using the program DOT developed by Dr. Roy Britten. This program is designed for use with the Apple II microcomputer with an IDS 460 printer. One sequence file is arrayed along the horizontal axis and the other is arrayed along the vertical axis. Each nucleotide in one sequence is compared against each nucleotide in the other to test for matches that form strings of identical sequences. The parameters of minimum string length to form a match and amount of divergence allowed in each matched string can be set for any desired value. All the comparisons shown in this paper search for matched strings of five nucleotides allowing a 10% divergence (equivalent to no divergence for this short a string). In the present format, the comparisons are limited to about 600 nucleotides per file. The program has the unusual feature of showing both direct matches (negative slope) and matches with the complementary strand (positive slope).

**Divergence Calculations for Polypeptide-coding Regions**

The polypeptide-coding regions of the rabbit and human $\beta$-like globin genes were also compared to search for replacement site and silent site substitutions. The values for percent divergence have been corrected for multiple substitutions at a single base (Holmquist 1972; Kimura and Ohta 1972), using the methodology of Perler et al. (1980). We divided the percent divergence by 100 to present the data as nucleotide substitutions per site. Although some of the assumptions involved in these calculations are probably oversimplified (e.g., that transitions and transversions are equally probable), the resulting values can be directly compared with those in many recent reports (e.g., Efstradiadis et al. 1980; Cleary et al. 1981; Lemischka and Sharp 1982; Proudfoot et al. 1982).

The calculations were done on an Apple II microcomputer using programs written by Dr. Forrest Fuller and provided by Dr. Argiris Efstratiadis. Since the methodology does not analyze deletions, all comparisons involving rabbit $\psi\beta_2$ ignored the three codons (out of 146 total) containing deletions. This results in small underestimates for divergence from $\psi\beta_2$.

**Results**

**Pairwise Comparisons between Rabbit and Human $\beta$-like Globin Genes**

The chromosomal arrangement of rabbit and human $\beta$-like globin genes is shown in figure 1A. The multiple genes in each family have the same transcriptional orientation, and in these two families they are arranged $5' \rightarrow 3'$ in the order of their time of expression. The human gene family is larger; it has more genes and contains
Fig. 1.—Maps of the rabbit and human β-like globin gene families. Part A shows the chromosomal positions and the times of expression (or pseudogene status) for each gene in the rabbit and human gene clusters. Each gene is divided into message-coding exons (black boxes) by noncoding introns. Part B shows this canonical globin gene structure in more detail, with untranslated regions of exons denoted by crosshatching and polypeptide-coding regions shown as black boxes. The gene segments included in each sequence subfile are designated by the line under the canonical gene. Abbreviations for these gene segments are 5'F = 5' flanking, E1 = exon 1, I1 + E2 = intron 1 + exon 2, I2a = 5' half of intron 2, I2b = 3' half of intron 2, E3 = exon 3, and 3'F = 3' flanking region.

The comparison between rabbit β1 (Hardison et al. 1979; van Ooyen et al. 1979) and human β (Lawn et al. 1980) is shown in figure 2. Extensive similarity is seen in the 5' flanking region, exon 1 including the 5' untranslated region, intron 1, exon 2, and exon 3. The 3' untranslated region of rabbit β1 has suffered a deletion of 39 base pairs, indicated by the shift in the diagonal of matches in this region. The second introns in the two genes are also similar in nucleotide sequence, indicated by the broken diagonal in panels I2a and I2b. The similarity between introns is not as
FIG. 2.—Dot matrix comparison of rabbit β1 and human β. Each panel shows the computer output for the comparison of the designated segment of two genes. Boxes along the axes show the exon regions. untranslated segments are crosshatched. Rabbit gene β1 is arrayed along the horizontal axis and human β is along the vertical axis. Dots are placed at 10-nucleotide intervals along each axis. Sequence similarities emerge as a set of diagonals with descending slopes.
extensive as that between exons and is interrupted by regions of dissimilarity, but
the diagonal of matches is quite apparent. Intron 2 of rabbit β1 is 277 base pairs
shorter than that of human β, suggesting that a deletion of this size occurred in rabbit
β1. The diagonal in panel I2a ends at around nucleotide 210, which is the probable
position of the β1 intron 2 deletion. Since intron 2 of β1 is so much shorter than
that of other β-like genes, the two sequence subfiles covering intron 2 were overlapped
by 149 nucleotides. Therefore, the last 149 nucleotides that do not match in panel
I2a are the same as the first 149 nucleotides of panel I2b that do show partial similarity.
Thus the sequences of rabbit β1 and human β are similar throughout, from the 5'
flanking region to the 3' flanking region. Moschonas et al. (1982) have shown that
the similarity in the 5' flanking region extends beyond that shown in figure 2, to at
least 370 base pairs before the cap site.

The comparison between rabbit pseudogene ψβ2 (Lacy and Maniatis 1980; Har-
dison and Margot 1984) and human δ (Spritz et al. 1980) is shown in figure 3. These
two genes are also similar throughout, with the greatest similarity in the exons, including
the 5' untranslated region, and intron 1, with less similarity in intron 2 and the 3'
untranslated region. Little similarity is seen in the 5' flanking region; we have proposed
that this region was included in a gene conversion event between rabbit ψβ2 and β1
(Hardison and Margot 1984) and so would not be expected to match between ψβ2
and δ. The similarity in the 3' half of intron 2 (panel I2b) is apparent for the last
100 nucleotides but is lost in a region of simple sequence, (TA)n, in both genes, which
shows as a set of black rectangles in the matrix. The δ-globin gene is longer than ψβ2
in this region, which indicates that some of the (TA)n tracts were deleted in ψβ2 or
were duplicated in δ. Nucleotides 110 through 170 in subfile I2b of rabbit ψβ2 match
with nucleotides 40 through 100 of δ; the first 70 nucleotides in I2b of ψβ2 match
the last 70 nucleotides in I2a of δ (not shown). An alignment derived by inspection
of intron 2 sequences of ψβ2 and δ also shows 52 to 68% similarity throughout this
segment (Hardison and Margot 1984). These two genes, therefore, are similar in every
segment except the 5' flanking region.

Rabbit gene β3 (Hardison, 1981) is compared with human ^γ (Shen et al. 1981)
in figure 4. These two genes are similar through all the available sequence, which
extends from 282 nucleotides before the cap site to 203 nucleotides past the poly A
addition site. (The comparison with the first 80 nucleotides of β3 is not shown.) The
similarity in intron 2 is less extensive than in the 5' flanking region, the exons, or
intron 1, but a clear diagonal of short matches is seen in both panels I2a and I2b.
A deletion occurs about 260 base pairs before the 3' end of intron 2 in β3, as shown
by the offset of the diagonal in panel I2b. This corresponds to a segment containing
the repeating dinucleotide (GT)n in γ. This has been proposed as a hot spot for
recombination in human γ's because it marks one boundary of the conversion unit
between ^γ and Gγ (Slightom et al. 1980); perhaps it recombined in the lagomorph
lineage to delete most of itself.

The most extensive blocks of DNA that could be compared are rabbit β4 (Hardison
1983) and human ε (Baralle et al. 1980), where sequences from 552 base pairs before
the cap site to 301 base pairs past the poly A site were analyzed. Figure 5 shows that
all these segments matched between β4 and ε. The 5' flanking similarity falls off
around 400 base pairs before the cap site (nucleotide 140 in panel 5'F) in a region
containing the repeating dinucleotide (GT)n in β4; the (GT)n is not found in this
segment of human ε. Some interrupted matches precede the (GT)n segment, as in-
dicated by the patchy diagonal involving nucleotides 30–70 of β4 in this panel. The
Fig. 3.—Dot matrix comparison between rabbit \(\psi\beta2\) and human \(\delta\)
FIG. 4.—Dot matrix comparison between rabbit $\beta 3$ and human $\Delta \gamma$
Fig. 5.—Dot matrix comparison between rabbit β4 and human e
apparent deletion of a region containing (GT)$_n$ in the β3 intron 2 (fig. 4) and in the ε 5' flanking segment (fig. 5) in separate events supports the proposal that recombinations occur frequently in this repeating dinucleotide. The similarity in the 3' flanking region continues through the available sequence data, with a 40 nucleotide mismatch from positions 60 to 100 in panel 3'F.

The comparisons examined so far, between β1 and β, ψβ2 and δ, β3 and γ, and β4 and ε, confirm previous reports that these pairs of genes are related (Efstratiadis et al. 1980; Hardison 1981, 1983; Moschonas et al. 1982; Hardison and Margot 1984). In all cases it appears that the block of chromosomal DNA containing them, including 5' and 3' flanking and intervening sequences, retains sequence similarity. The similarity in intron 2 is particularly significant since it has not been observed in most globin gene comparisons. Efstratiadis et al. (1980) found no intron 2 homology in paralogous comparisons of human β-like globin genes (except in the recently duplicated γ genes). The observation of intron 2 similarity in the four comparisons above may indicate that these gene pairs are orthologous, i.e., descended from the same gene in the last common ancestor. This proposal was tested by comparing the 5' half of intron 2 (subfile I2a in fig. 1B) between all four rabbit and all four human genes. The matrix of 16 comparisons presented in figure 6 shows that obvious intron 2 similarity is seen only between β4 and ε, β3 and γ, ψβ2 and δ, and β1 and β. The other comparisons in figure 6 do not form lengthy diagonals; in fact, there is little difference between the direct homology searches (negative slopes) and the reverse complement searches (positive slopes). Thus, these four gene pairs are the most closely related between rabbits and humans and, therefore, are the orthologous pairs.

Comparisons among the Rabbit β-like Globin Genes

Paralogous comparisons between pairs of rabbit β-like globin genes reveal similarities within the exons but not in the introns (except intron 1 of β1 and ψβ2; see below). Only the comparisons between gene segments with similar sequences are shown in figure 7; comparisons between flanking and intervening sequences that lack similar sequences are not shown. (Complete gene comparisons are available upon request.)

The comparison between rabbit β1 and ψβ2 (fig. 7, first row of panels) shows that matching sequences are found in all three exons, and some patchy matches are seen in the 5' flanking sequence and intron 1. The similarity in the 5' flanking region was not observed in the comparison between rabbit ψβ2 and human δ (fig. 3), which fits the proposal that ψβ2 was originally a δ-like gene whose 5' end converted with rabbit β1 (Hardison and Margot 1984). The comparison between rabbit β3 and β4 (fig. 7, second row of panels) shows that the sequences are most similar in exon 2, with less, but detectable, similarity in the 5' flanking region and in exons 1 and 3. Earlier heteroduplex and hybridization analysis (Hardison et al. 1979) showed hybrid formation only in exon 2; apparently exons 1 and 3 are not sufficiently similar for DNA heteroduplex formation at the stringency used previously. Rabbit genes β3 and β1 (fig. 7, third row) are similar only in exons 2 and 3, whereas genes β4 and β1 (fig. 7, fourth row) are similar in all three exons. For the comparisons in figure 7, the exon 3 similarities are limited to the translated region except for the β4/β1 comparison, which shows a broken diagonal of matches in the 3' untranslated region. Comparison of genes β3 and β4 with ψβ2 (data not shown) shows matches only in exon 2 and the translated portion of exon 3, not in exon 1.

These data show that the closest matches among β-like globin gene pairs in
rabbits are β1 with ψβ2 and β3 with β4. Gene β4 is closer to β1 than is β3, and both β3 and β4 are approximately equally divergent from ψβ2. The segments of similarity seen between β4 and β1 in the 3' untranslated region are not seen between either β4 and β3 or β3 and β1. This indicates that ancestor of either β4 or β1 has acquired the 3' untranslated segment of the other. This event would have preceded the deletion suffered by β1 in this region (fig. 2).

Quantitative Comparison of Polypeptide-coding Regions

Nucleotide substitutions that accumulate in a gene are of two types: replacement (nonsilent) site substitutions that lead to an amino acid replacement in the encoded polypeptide, and silent site substitutions that do not alter the encoded amino acid sequence. The polypeptide-coding portions of the rabbit and human β-like globin
FIG. 7.—Dot matrix comparison between pairs of rabbit β-like globin genes. Only the panels showing the comparisons of the immediate 5′ flanking region and exon 1 (E1), intron 1 (I1) and exon 2 (E2), and exon 3 (E3) and the immediate 3′ flanking regions are shown. No similarities were observed in the intron 2 comparisons. Gene sequences arrayed along the horizontal and vertical axes are identified at the beginning of each row of panels.

genes were compared using the algorithm of Perler et al. (1980), and the frequency of replacement site and silent site substitutions are tabulated in table 1. For the comparison of a single rabbit gene with each human gene, the gene pairs with the least substitutions per silent site (upper right of the matrix) correspond to the orthologous pairs identified by dot matrix analysis: β4-ε, β3-γ, ψβ2-δ, and β1-β. This is also true of the substitutions per replacement site (lower left of the matrix) except for β3, which shows 0.13 substitutions per replacement site when compared with γ
Table 1
Divergence Comparisons of the Coding Portions of Rabbit and Human β-like Globin Genes:
Nucleotide Substitutions per Site

<table>
<thead>
<tr>
<th></th>
<th>Hε</th>
<th>H^γ</th>
<th>Hδ</th>
<th>Hβ</th>
<th>Rβ4</th>
<th>Rβ3</th>
<th>Rβ2</th>
<th>Rβ1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hε</td>
<td></td>
<td>0.61</td>
<td>0.62</td>
<td>0.61</td>
<td>0.58</td>
<td>1.13</td>
<td>0.67</td>
<td>0.60</td>
</tr>
<tr>
<td>H^γ</td>
<td>0.10</td>
<td>0.75</td>
<td>0.74</td>
<td>0.69</td>
<td>0.62</td>
<td>0.55</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>Hδ</td>
<td>0.17</td>
<td>0.19</td>
<td>0.32</td>
<td>0.93</td>
<td>1.08</td>
<td>0.43</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>Hβ</td>
<td>0.16</td>
<td>0.18</td>
<td>0.037</td>
<td>0.78</td>
<td>1.00</td>
<td>0.55</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>Rβ4</td>
<td>0.081</td>
<td>0.12</td>
<td>0.19</td>
<td>0.18</td>
<td>1.00</td>
<td>0.99</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>Rβ3</td>
<td>0.11</td>
<td>0.13</td>
<td>0.20</td>
<td>0.19</td>
<td>0.13</td>
<td>1.01</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Rβ2</td>
<td>0.28</td>
<td>0.27</td>
<td>0.20</td>
<td>0.20</td>
<td>0.28</td>
<td>0.28</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>Rβ1</td>
<td>0.18</td>
<td>0.18</td>
<td>0.073</td>
<td>0.056</td>
<td>0.20</td>
<td>0.19</td>
<td>0.17</td>
<td>0.17</td>
</tr>
</tbody>
</table>

NOTE.—Percent divergence for replacement (nonsilent) site (lower left) and silent site (upper right) comparisons were calculated by the Perler et al. (1980) algorithm and divided by 100 to present the data as nucleotide substitutions per site. H = human, R = rabbit.

but only 0.11 substitutions per site when compared with ε. This is a very small difference, and the substantial blocks of similarity in intron and flanking regions of β3 and γ shown in figure 4 argue for their being orthologous. The intron 2 similarity between ψβ2 and δ indicates that they are derived from a common ancestor, even though ψβ2 shows 0.20 substitutions per replacement site when compared with either δ or β.

The divergence data for the comparisons among rabbit β-like genes also confirm the conclusions based on dot matrix analysis. The replacement site divergences show that the most similar pairs of genes are β3 and β4 (0.13 substitutions per replacement site) and β1 and ψβ2 (0.17). Genes β3 and β4 are equally distant from ψβ2 (0.28 for both). Although β3 and β1 are slightly less divergent (0.19) than β4 and β1 (0.20) in replacement sites, β4 is closer to β1 than is β3 (.80 vs. 0.99) in silent sites.

Rabbit gene β3 has accumulated an unusually large number of silent site substitutions, with an average (±SD) of 1.04 ± 0.06 substitutions per silent site observed for paralogous comparisons with both rabbit and human β-like genes (table 1). Comparisons involving rabbit β4 also show a high value for silent site substitutions (average of 0.86 ± 0.13 for paralogous comparisons). The average value for substitutions per silent site for other comparisons in table 1 is 0.62 ± 0.08, omitting comparisons between orthologous gene pairs or those proposed to have been involved in recent gene conversions.

A History of β-like Globin Gene Evolution in Rabbits and Humans

The rate of replacement site substitution in globin genes is about 1 × 10^-9 substitutions per replacement site per year (Efstratiadis et al. 1980; Perler et al. 1980). One can divide the replacement site divergence values in table 1 by this rate to obtain a rough estimate of the time when the two genes last diverged, by a gene duplication or by a gene conversion or by other recombination. This gives divergence times that agree in some, but not all, respects with the trees presented by Dayhoff (1972) and Efstratiadis et al. (1980). As outlined in figure 8, the ancestor of mammalian β-like globin genes duplicated (or a gene pair last converted) 180–200 Myr ago to give the ancestral genes of the ε/γ and δ/β lineages (c.g., see the β4/β, β1/ε, and β4/β1 comparisons in table 1). Around 110 (β3/ε) to 130 (β4/β3) Myr ago, the ε/γ parent diverged to establish the ε and γ lineages.
FIG. 8.—A history of the β-like globin gene family in the ancestors to rabbits and humans. The primordial β-like globin gene (top line) diverged into the ancestors of the ε/γ and δ/β lineages about 190 Myr ago. These split into the ancestors of ε and γ or δ and β, respectively, before the mammalian radiation. The acquisition of 3' untranslated sequences from β by ε (or vice versa) is denoted by the blackened region of the crosshatched ε. γ is dotted, δ is striped, and β is black. The replacement site divergence (in substitutions per site) for orthologous genes in rabbits and humans is given between the branching arrows. The parentheses along the arrows to contemporary descendants of δ and β denote the proposed conversion events involving the 5' halves of rabbit ψβ2 and β1 and human δ and β. Abbreviations are R = rabbit and H = human. Approximate times in millions of years are given from top to bottom.

At some time prior to the mammalian radiation, which was about 85 Myr ago (Romero-Herrera et al. 1973), the δ/β parent also diverged into separate δ and β lineages. This estimate for the δ/β divergence is substantially earlier than has been reported previously (Efstratiadis et al. 1980); it is based on the observation of δ-like genes in both rabbits (ψβ2) and humans (δ), as well as in mice (βh2 and βh3; Edgell et al. 1983; Hardies et al. 1984). The previous estimate of 40 Myr for the δ/β divergence is actually the time of a proposed gene conversion event involving the 5' halves of the genes (Jeffreys et al. 1982; Martin et al. 1983). The silent site divergence between δ and β is 1.12 in exon 3 (Efstratiadis et al. 1980), which was not involved in the conversion. Using the globin divergence rate of 8 × 10⁻⁹ substitutions per silent site per year, one can calculate that δ and β diverged about 140 Myr ago. This is a rough estimate, since the accumulation of silent site mutations is linear only over about 80 Myr (Efstratiadis et al. 1980). Also, human δ may have been a pseudogene for part of its history (Martin et al. 1983), which will tend to exaggerate the calculated time since its divergence. This rough calculation suggests that the s/p split may have occurred at about the same time as the e/γ split (fig. 8).

The sequence similarity between rabbit β4 and β1 in the 3' untranslated region indicates that a conversion event occurred between the ancestral ε and β after the ε/γ split. The 3' untranslated regions of human ε and β are also similar, which indicates that the proposed conversion occurred before the lagomorph/primate divergence and argues against the similarity arising by recent, parallel mutations. The β-globin sequence comparisons in Efstratiadis et al. (1980) show that the 3' untranslated region is not highly conserved in evolution.

The branching pattern shown in figure 8 would establish the ancestral four-gene family, 5'-ε-γ-δ-β-3', prior to the mammalian radiation. After the split between the lagomorph and primate lineages, the orthologous genes began to diverge, but they have not diverged at the same rate. Rabbit β1 has diverged from human β by only 0.056 substitutions per replacement site, whereas β4 has diverged from ε by 0.081,
and β3 has diverged from γ by 0.13, which is 1.6 times the value for β4/ε (fig. 8). Rabbit ψβ2 has diverged from δ by 0.20 substitutions per replacement site, but this more rapid evolution could result from the loss of purifying selection because of its pseudogene status. These results show that even the active β-like globin genes have not fixed mutations at the same constant rate, at least since the lagomorph-primate split.

Intrachromosomal gene conversions apparently have occurred between rabbit ψβ2 and β1 and between human δ and β well after the mammalian radiation (fig. 8). Martin et al. (1983) estimate the time of conversion at around 40 Myr, prior to the split to hominoids and Old World monkeys. Lacy and Maniatis (1980) estimate the divergence (conversion) between ψβ2 and β1 at about 55 Myr ago. The human γ globin region duplicated around 34 Myr ago, and the two genes corrected against each other (converted) about 1 Myr ago (Shen et al. 1981).

Discussion

The principal conclusions from this analysis are that (1) the rabbit-human gene pairs β4-ε, β3-γ, ψβ2-δ, and β1-β are orthologous; (2) an ancestral family in the arrangement 5'-ε-γ-δ-β-3' existed prior to the mammalian radiation; (3) this ancestral family was derived from distinct ε/γ and δ/β lineages which formed from the primordial β about 190 Myr ago and which split into ε and γ, or δ and β, respectively, about 130 Myr ago; and (4) the pairs of orthologous genes have diverged at different rates since the split between the lagomorph and primate lineages. The assignment of orthologous relationships was based primarily on similarities in introns, especially intron 2. Pairwise comparisons of paralogous globin genes failed to reveal similarity in intron 2, so finding such similarity is a good indication of an orthologous pair. A current estimate of the rate of nucleotide substitutions at neutral sites is about 5 to 7 \times 10^{-9} per site per year (Li et al. 1981). At this rate, it would take 120–160 Myr for two segments of DNA to diverge to 50% difference (excluding effects of insertions, deletions, and recombinations), so intron similarity is a good indicator of orthology between genes in species that diverged less than about 120 Myr ago. It is not a necessary criterion for orthology, since some other region, perhaps in the flanking segments, may reveal a close similarity. It is also not a sufficient criterion, since two paralogous genes can acquire intron homology either by a recent duplication or by a gene conversion event. The test for orthology applied in the present analysis is whether the two genes occupy a segment of genomic DNA, including flanking regions, that shows sequence similarity throughout. This block of DNA in each species is presumably derived from the same segment of genomic DNA in the last common ancestor of the two species.

Although the introns and flanking regions show detectable similarity in the orthologous comparisons, the exons are always more similar. This is consistent with a greater selective constraint placed on these regions. From the paralogous comparisons among the rabbit genes, it appears that exon 2 is the most highly conserved, followed by exon 3 and then exon 1. Exon 2 encodes the heme-binding domains of globins and is highly conserved in all globin gene comparisons.

If an ancestral gene family, 5'-ε-γ-δ-β-3', existed prior to the mammalian radiation, then all mammalian globin gene families should retain vestiges of this arrangement, unless some elements have been lost through nonreciprocal recombination (e.g., conversion or unequal crossing-over). The summary of β-like globin gene families in figure 9 illustrates this point. The human and rabbit families can easily be seen as
derived from the proposed ancestral family. The lemur family also shows a close relationship to the proposed ancestral family. It contains \( \epsilon, \gamma, \psi \beta, \) and \( \beta \) genes; the \( \psi \beta \) is a hybrid between either \( \epsilon \) or \( \gamma \) and \( \delta \) (Jeffreys et al. 1982). The mouse family (haplotype Hbb\( ^{d} \); Jahn et al. [1980]) can also be described in terms of the four-gene ancestor. The 5'-most gene, embryonic \( \epsilon^{e} \), is closely related to human \( \epsilon \) (Jahn et al. 1980; Hansen et al. 1982), embryonic \( \beta^{0} \) and \( \beta^{1} \) are related to human \( \gamma \) (Czelusniak et al. 1982; Hill et al. 1984), pseudogene \( \beta^{2} \) is related to \( \delta \) (Hardies et al. 1984), \( \beta^{3} \) is a fusion gene between a \( \beta \)- and a \( \delta \)-like gene (Edgell et al. 1983; Hardies et al. 1984), and \( \beta \) major and \( \beta \) minor are closely related to human and rabbit \( \beta \) (Konkel et al. 1979; van Ooyen et al. 1979; Efstratiadis et al. 1980). Although van Ooyen et al. (1979) interpreted the \(-50\%\) match in intron 2 between rabbit \( \beta^{1} \) and mouse \( \beta^{1} \) major as being essentially random, this comparison actually shows the vestiges of homologous sequences that are diverging rapidly. Also, Moschonas et al. (1982) found sequence similarity between human \( \beta \) and mouse \( \beta \) major and \( \beta \) minor in the 5' flanking regions. The goat \( \beta \)-like globin gene family is derived from a duplication, and possibly a triplication, of a four-gene family (Townes et al. 1984). In each four-gene unit, the 5' gene (\( \epsilon^{I} \) in the first set) is orthologous to human \( \epsilon \) (Shapiro et al. 1983), and the 3' gene (\( \beta^{C} \)) is most related to human \( \beta \) (Schon et al. 1981). The ancestry of the 3' \( \epsilon \) gene (\( \epsilon^{II} \)) and the \( \psi \beta \) genes are not as clear. Goat gene \( \epsilon^{II} \) is about equally similar (73%-79% similarity) to both \( \epsilon \)-like and \( \gamma \)-like genes from other species, but it is slightly more related to the \( \gamma \)-like genes in the 5' flanking regions (Shapiro et al. 1983). Thus it could be a derivative of \( \gamma \) or could result from a separate duplication from the ancestral \( \epsilon \); in either event, it has diverged substantially from either proposed parent. The \( \psi \beta \) genes are estimated to have diverged from the \( \beta \) genes about 42 Myr ago (Cleary et al. 1981); this could have been either a \( \beta \) gene
duplication or a conversion between a β gene and a δ-like gene so that almost all the δ sequences were lost. A segment of the 5' flanking region is similar between goat ψδ and human δ, which supports this conversion model (Hardies et al. 1984). On the whole, it appears that the β-like globin genes from eutherian mammals examined so far can be described as descendants of the proposed ancestral family.

Although the descendants of the ancestral gene family have diverged at different rates, the genes at an equivalent position in the families from different species have evolved at a similar rate and in a similar mode. The descendants of ancestral e (fig. 9) have evolved most slowly in both coding and intervening sequences. The coding region comparisons presented in table 2 show that these genes have a significantly lower replacement site divergence among the set than do the descendants of β (average divergence of 0.086 vs. 0.123, respectively) or the descendants of γ (average divergence of 0.155). The slower accumulation of replacement site substitutions in e-derived genes has also been noted by Shapiro et al. (1983). The similarities between the introns

Table 2
Divergence Comparisons of Mammalian β-like Globin Genes at Equivalent Chromosome Positions

<table>
<thead>
<tr>
<th>ANCESTRAL GENE</th>
<th>Replacement (Silent)</th>
<th>Average ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hε</td>
<td>Geδ</td>
</tr>
<tr>
<td>ε:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R/δ4</td>
<td>0.08 (.58)</td>
<td>0.09 (.74)</td>
</tr>
<tr>
<td>Hε</td>
<td>0.06 (.59)</td>
<td>0.09 (.60)</td>
</tr>
<tr>
<td>Geδ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R/δ3</td>
<td>0.13 (.62)</td>
<td>0.17 (.65)</td>
</tr>
<tr>
<td>Hγ</td>
<td>0.15 (.72)</td>
<td>0.15 (.69)</td>
</tr>
<tr>
<td>Geδ</td>
<td>0.15 (.72)</td>
<td>0.15 (.69)</td>
</tr>
<tr>
<td>δ:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R/δ2</td>
<td>0.20 (.43)</td>
<td>0.33 (.66)</td>
</tr>
<tr>
<td>Hδ</td>
<td>0.21 (.48)</td>
<td></td>
</tr>
<tr>
<td>β:</td>
<td>Hβ</td>
<td>GeδC</td>
</tr>
<tr>
<td>R/δ1</td>
<td>0.06 (.42)</td>
<td>0.13 (.44)</td>
</tr>
<tr>
<td>Hβ</td>
<td>0.12 (.51)</td>
<td>0.13 (.49)</td>
</tr>
<tr>
<td>GeδC</td>
<td>0.13 (.49)</td>
<td>0.18 (.56)</td>
</tr>
</tbody>
</table>

Note.—Genes have been grouped as descendants of ancestral ε, γ, δ, or β (see fig. 9), and the nucleotide substitutions per replacement (or silent) site calculated by the method of Perler et al. (1980) for each pairwise comparison. H = human, R = rabbit, G = goat, M = mouse.

* Using Student's t-test, the average replacement site divergence for genes at these positions is significantly different:
  - ε vs. γ, P < 0.005;
  - ε vs. β, P < 0.05;
  - γ vs. β, P < 0.05.
of rabbit $\beta_4$ and $\varepsilon$ (fig. 5) are the most extensive of all the pairwise comparisons done; extensive intron similarity is also seen when goat $\varepsilon^1$ is compared with human $\varepsilon$ (Shapiro et al. 1983). The descendants of $\varepsilon$ are always expressed in embryonic red cells.

The descendants of $\gamma$ usually occupy the next position in the gene family. They fix nonsilent mutations more rapidly than the $\varepsilon$-derived genes (1.8-fold higher; table 2) or the $\beta$-derived genes (1.3-fold higher). Comparison between $\gamma$-derived genes usually show a higher silent site divergence than do descendants of $\beta$ and $\varepsilon$ (table 2). The $\gamma$-derived gene has been duplicated in mice and triplicated in humans (fig. 9). In different species it can be expressed in either embryonic (rabbit, mouse, goat) or fetal (human) life.

The descendants of the ancestral $\delta$ occupy the third position in the gene family and have undergone a nonreciprocal recombination event in every mammalian order examined (reviewed in Hardison and Margot [1984]). They have been duplicated in the mouse (fig. 9). The large divergence for these genes shown in table 2 at least partly results from the fact that they are usually pseudogenes, except for human $\delta$, which is a minor adult gene.

The descendants of ancestral $\beta$ occupy the 3'-most position in the family. They show a replacement site divergence (average of 0.123; table 2) greater than the $\varepsilon$ descendants (0.086 average) but less than the $\gamma$ descendants (0.155 average). Note that the replacement site divergence for the rabbit $\beta_1$-human $\beta$ comparison is unusually low. The silent site divergences among the $\beta$-derived genes are usually lower than those among the $\gamma$-derived genes (average of 0.51 versus 0.98, respectively). The intervening sequences are also not as well conserved in the descendants of $\beta$ as they are in $\varepsilon$-derived genes (compare figs. 2 and 5). The $\beta$-derived genes are almost always expressed in adult (postnatal) erythroid cells, and some can also be expressed in fetal erythrocytes (the mouse $\beta$'s and rabbit $\beta_1$). The goat $\gamma$ is expressed exclusively in fetal life.

In summary, the most highly conserved genes, the descendants of $\varepsilon$, also are the most limited in their span of developmental expression. Genes descended from ancestral $\gamma$ and $\beta$ show more plasticity in structure, which may allow them to function at different developmental times in different species. In other words, the higher rate of mutation fixation at these chromosome positions may provide the diversity of protein structures needed to accommodate the patterns of developmental expression in different species, e.g., for the rabbit $\beta_3$ globin to function in embryonic red cells and human $\gamma$-globin to function in fetal red cells.

The different rates of evolution for globin genes at each position in the family further complicate the calibration of molecular clocks for globin genes. The commonly used clocks (e.g., Wilson et al. 1977; Efstratiadis et al. 1980) assume that all globin genes are accumulating mutations at equal rates. However, it is apparent that correction factors for either position in the family or perhaps stage of expression may be required in order to derive times of divergence from paralogous sequence comparisons. Thus, the branching patterns derived for mammalian globin genes are probably in the correct order (fig. 8), but the dates assigned to the branch points may need to be revised as we gain more detailed knowledge of the rates of evolution.

Acknowledgments

I thank S. Hardies and S. Shapiro for communicating results prior to publication and for helpful discussions, R. Britten, A. Efstratiadis, and F. Fuller for sequence
analysis programs, R. Deering for use of his Apple II microcomputer, and T. Peters for typing the manuscript. Early sequence comparisons used the program SEQ in the SUMEX AIM/GENE\textsc{et} project at Stanford University. This work was supported by grants from the National Institutes of Health.

LITERATURE CITED


MASATOSHI NEI, reviewing editor

Received September 27, 1983; revision received January 20, 1984.