THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 1984 by The American Society of Biological Chemists, Inc.

Vol. 259, No. 20, Issue of October 25, pp. 12881-12889, 1984
Printed in U.S.A.

The Mouse Globin Pseudogene *βh3* Is Descended from a Premammalian δ-Globin Gene*

(Received for publication, June 7, 1984)

Clyde A. Hutchison, III, Stephen C. Hardies, Richard W. Padgett, Steven Weaver‡, and Marshall Hall Edgell

From the Department of Microbiology and Immunology, Curriculum in Genetics, Program in Molecular Biology and Biotechnology, The University of North Carolina, Chapel Hill, North Carolina 27514

The *βh3* pseudogene of the BALB/c mouse contains sequence defects which prevent transcription and translation to produce a β-globin. Comparison with other globin gene sequences indicates that *βh3* arose by recombination between an adult β-globin gene and some significantly diverged globin sequence. Analysis of noncoding sequences shows that the 3′ end of mouse *βh3* and the human δ-globin gene are both descended from an ancestral gene, which we call proto-δ. The origin of proto-δ must predate the mammalian radiation. A member of the L1 family of interspersed repetitive elements is inserted into the 3′ untranslated δ-homologous sequence in *βh3* from BALB/c. *βh3* is a widespread feature of the rodent β-globin complex, which has been fixed in the genome for 35 million years. Independent inactivation events produced pseudogenes located between the adult and nonadult β-globin genes in the rodent, primate, rabbit, and goat lineages. One model to explain the abundance and evolutionary persistence of pseudogenes postulates that the mammalian genome simply has no efficient mechanism for deleting nonessential sequences. Consequently, the genomes of higher eukaryotes have been growing, by the accumulation of duplications, with doubling times of 200 ± 100 million years.

There are now many examples of DNA sequences which have obvious homology to some expressed gene, but which contain defects preventing the production of a related functional gene product. If the defective sequence is allelic to an essential gene, it often represents the determinant of a heritable disease. In surprisingly many cases, however, a defective gene is present in the population at very high frequency, has no known functional alleles, and does not produce any disease. Such evolutionarily fixed defective genes are commonly referred to as pseudogenes.

Pseudogenes have been discovered which are related to a wide variety of eukaryotic genes. These include the genes for 5 S RNA (Jacq *et al.*, 1977), small nuclear RNAs (Denison *et al.*, 1981), both α- and β-globin (for review see Little, 1982), interferon (Goeddel *et al.*, 1981), actin (Firtel *et al.*, 1979), and β-tubulin (Lee *et al.*, 1983), to cite a few examples. It seems reasonable to postulate that every vertebrate gene has related pseudogenes. Since gene inactivation is generally expected to produce disease, the widespread occurrence of pseudogenes is an intriguing phenomenon which requires explanation.

Although pseudogenes have no known function, the study of their structures has yielded insights into the mechanisms of genome evolution in eukaryotes. Analysis of pseudogene mutation rates has provided a base-line for determining selective pressures on mutations in expressed genes (Kimura, 1980; Li *et al.*, 1981; Miyata and Yasunaga, 1981). An α-globin pseudogene in the mouse lacks introns and appears to have been transported by a retrovirus (Lueders *et al.*, 1982). In this paper, we conclude that the structure of the mouse globin pseudogene *βh3* provides evidence for a "proto-δ" globin gene in the ancestor to all mammals.

A complex containing seven β-like globin genes from the BALB/c mouse has been cloned and mapped (Jahn *et al.*, 1980; Leder *et al.*, 1980). A map of 65 kb[1] spanning this locus is shown in Fig. 1. Three of the genes were easily identified with known globin products; gene y codes for the major late embryonic β-like chain, and $\beta 1^{dmaj}$ and $\beta 2^{dmin}$ code for the two adult β chains. Four other "β homologous genes" (*βh0*, *βh1*, *βh2*, and *βh3*) map between y and the adult genes (Fig 1). Recent studies on the *βh* genes in this laboratory, including complete DNA sequences of these genes, as well as transcriptional and translational studies, have clarified the role of these structures. *βh0* and *βh1* are transcriptionally active, both *in vitro*[2] and in mouse embryos (Brown *et al.*, 1982; Farace *et al.*, 1984). *βh1* specifies a major early embryonic transcript, which codes for the β-like globin designated "z". *βh0* may code for a less abundant, and therefore previously unobserved, early embryonic β-like globin. A combination of sequence and transcriptional data indicate that *βh2* is a pseudogene, although the defects in this structure are less obvious than in many other globin pseudogenes (Phillips *et al.*, 1984).

In contrast, the structure of *βh3* is clearly aberrant, and partial DNA sequence demonstrated that it cannot be translated to produce a normal β-globin (Jahn *et al.*, 1980). We report here the complete nucleotide sequence of *βh3*. We also present a method for tracking the evolutionary history of genes, which has led us to a model for the origin of this curious recombinant structure. This model has unexpected consequences for the evolutionary history of adult globin genes in all mammals. We also present speculative theories to explain the abundance and evolutionary persistence of pseudogenes.

* This research was supported by Public Health Service Grants AI08998 and GM21313 from the National Institutes of Health. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Present address, Department of Biological Sciences, University of Illinois at Chicago Circle, Chicago, IL 60680.

[1] The abbreviations used are: kb, kilobase; CB, coding block; IVS, intervening sequence.

[2] B. A. Brown, personal communication.

EXPERIMENTAL PROCEDURES

Clones and Subclones—*βh3* is contained in the BALB/c genomic clones CE17 (Jahn *et al.*, 1980), and CA4 (Edgell *et al.*, 1981). *Eco*RI fragment A from CE17 has been transferred to λ Charon 16A to produce the subclone CE17.6, and into the single-stranded phage vector M13mp2 in both orientations to produce MA1 and MA2 (these subclones are described by Jahn *et al.*, 1980).

DNA Sequencing—Chemical sequencing was performed as described by Maxam and Gilbert (1980). Sequencing by the chain terminator method of Sanger, Nicklen, and Coulson (1977), using the thin-gel technique of Sanger and Coulson (1978), was carried out with modifications described by Kan *et al.* (1979).

Sequence Transfer—The sequence of *βh3* will be made available in computer-readable form on request. The following information refers to the sequence appearing in Fig. 2. It is provided to assist verification of copied or transmitted sequence. The sequence has 2165 bases consisting of 696 A, 467 C, 432 G, and 570 T. The dinucleotide frequencies are: AA (247), AC (131), AG (166), AT (152), CA (176), CC (114), CG (9), CT (167), GA (152), GC (82), GG (110), GT (88), TA (121), TC (140), TG (146), and TT (163). These sum to 2164.

Computer Analysis—The analyses were performed on Z-80 microprocessor-based computers operating under CP/M (Digital Research). The dotplots were printed on a Diablo HyTerm Model 1620. The algorithms employed allow searches to be performed rapidly, with execution times independent of window size and stringency (White *et al.*, 1984). Programs will be made available upon request.

Mice—BALB/c mice are strain BALB/CJ obtained from Jackson Laboratories (Bar Harbor, ME). C57BL mice are C57BL/10ScSn and were obtained from G. Haughton (University of North Carolina). AU/SsJ mice were obtained from Jackson Laboratories. *Peromyscus maniculatus*, inbred at the *β*-globin locus, were obtained from Lee R. G. Snyder (University of California, Riverside). All other mice were from outbred stocks and therefore only one animal was used in each DNA preparation. *Mus caroli* were a gift of Verne Chapman (Roswell Park Memorial Institute). DNA preparations from the other wild *Mus* species were gifts of Charles Langley (National Institute of Environmental Health Sciences) who obtained the animals from Michael Potter (National Institutes of Health).

DNA Preparations—High-molecular-weight mouse DNA was prepared by a modification of the method of Blin and Stafford (1976). DNA was made from single individuals, either from the liver or from the whole animal, by a procedure to be described elsewhere.[3] Rat thymus DNA (*Rattus norvegicus*) was obtained from S. Harris (National Institute of Environmental Health Sciences).

Hybridization—DNA fragments were nick translated as described by Maniatis *et al.* (1975) using α-^{32}P-labeled nucleoside triphosphates with specific activities of about 850 Ci/mmol. Digests of genomic DNA were transferred to nitrocellulose and hybridized in 50% formamide at 42 °C by a modification of the procedure of Wahl, Stern, and Stark (1979) as described by Jahn *et al.* (1980). After washing, the filters were exposed to XR-1 film (Kodak) with a Dupont Cronex Lightning Plus intensifying screen for 1–4 days.

RESULTS AND ANALYSIS

Sequencing Strategy for the BALB/c βh3 Gene—The *βh3* structure has been cloned in two segments. The Charon 4A clone CE17 contains the 4.8-kb *Eco*RI fragment A. This fragment contains most of *βh3*, from the 5′ end of the gene to the RI site at a position homologous to codons 120–122, and a partial sequence has been reported previously (Jahn *et al.*, 1980). The remainder of *βh3* is included within CA4, a Charon 4A clone of adult liver DNA. An *Eco*RI fragment 0.9 kb in size contains the 3′ end of the gene. Although no clone has been isolated which overlaps the portions of *βh3* isolated in these two clones, genomic mapping experiments (Jahn *et al.*, 1980) led to an estimate of approximately 5 kb for the distance between *Eco*RI fragment A and the 7-kb *β1*^{dmaj} *Eco*RI fragment. The clone CA4 contains *Eco*RI fragments 0.9 and 4.5 kb in size 5′ to the 7-kb *Eco*RI fragment. It was therefore unlikely that any additional RI fragments had been missed. This interpretation is confirmed by the sequence itself, which is homologous to the adult *β*-globin coding sequence across the RI site in question.

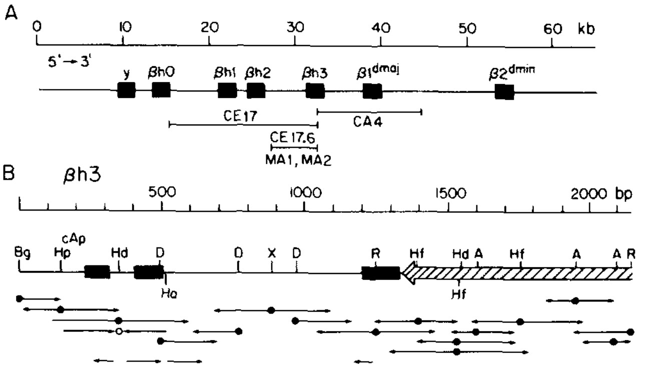

FIG. 1. **Restriction map and sequencing strategy for *βh3*.** *Panel A* shows the linkage arrangement of the BALB/c mouse *β*-globin gene complex. The *horizontal brackets* indicate the position of the phage subclones used in sequencing. The *scale* above the cluster is given in kilobases. *Panel B* displays the restriction map and sequence strategy used in sequencing *βh3*. Regions homologous to the capping site (*cAp*), and the coding regions (*solid boxes*) are indicated. The *cross-hatched arrow* indicates sequence homologous to the L1Md repetitive element. Abbreviations for restriction enzymes are: *Bg* (*Bgl*I), *Hp* (*Hpa*II), Hd (*Hin*dIII), *D* (*Dde*I), *Ha* (*Hae*III), *X* (*Xba*I), *R* (*Eco*RI), Hf (*Hin*fI), and A (*Ava*II). *Each arrow* (drawn in the 5′ to 3′ direction along the strand displayed in the sequencing gel) indicates the sequence determined from a particular restriction-site terminus. *Circles* indicate the position of the radiolabeled end in chemical sequencing experiments (*solid circles*, 5′ label; *open circles*, 3′ label). *Arrows without circles* indicate experiments using the chain-terminator method.

The λ and M13 phage clones and subclones used in sequencing *βh3* are shown in Fig. 1*A*. Both the enzymatic-chain-terminator method and the chemical method were used. Fig. 1*B* indicates the experiments from which the sequence shown in Fig. 2 was derived.

The Structure of βh3—Preliminary analysis, based on a partial sequence of *βh3*, demonstrated an aberrant structure (Jahn *et al.*, 1980), but did not fully define its relationship to the normal globin gene. We observed obvious homology to adult gene sequence in the 3′ portion of the gene (starting at codon 75), and only short patches of homology in the 5′ portion of *βh3*. We also noted a short, almost perfect, match near the mRNA cap site of the *β1*^{dmaj} sequence, and suggested that *βh3* might contain a remnant of the 5′ end of a globin gene. Further sequence analysis has shown this to be the case (Edgell *et al.*, 1981; Miyata and Hayashida, 1981). Here we report a more extensive analysis based on the completed nucleotide sequence of *βh3*.

We have found the homology matrix, or "dotplot" method of comparing sequences to be a powerful tool for investigating the evolutionary relationships between DNA sequences. In a dotplot, two perpendicular axes represent the sequences under comparison. Each dot within the matrix represents homology between the two sequences at positions corresponding to the coordinates of the dots. Thus, extended homologies are represented by diagonal lines of dots. This method, long used to compare amino acid and nucleotide sequences (Gibbs and McIntyre, 1970), has found renewed use with the proliferation of DNA sequence data (for example see Maizel and Lenk, 1981; Staden, 1982). The advantage of the particular programs used in the analysis presented here is that they are designed for very rapid execution on microcomputers (White *et al.*, 1984). The speed of the programs has allowed us to experiment with the parameters of the comparison, in order to achieve

[3] R. W. Padgett *et al.*, manuscript in preparation.

```
Bgl I    '         '         '         '         50        '         '         '         '        100
GGCCTCAGCCCCTCCAACAAAGAACAACAGGCAACTAAAGAGTGTGAGAGCTAGAGAAATAGTCTTCCCCAGGGAAGAGCACACCAAGCCTTCAGACTTG

         '         '         '         '        150  ATA Box'         '         '     cAp homology 200
AAAACATACATGCAAGTAACATTACATGGAAGTCAGGTAGTCCGGGATGGGCATTAAAGGTAGAGCAGATGCCTGCTGTGCCTGCATCTGCTTCTGACAC

         '         '                CB1-(-1f.s.)---250---------'---------'---------'---------'-------300
AGGTTGTTCACTAGCAGCAGTCCTACACCAGGGTGTACTGACTGACAAAGAGAAGTTGGCCATCAGTGGCCTATGGGACACAGTAGACATGGGAAACATT

---------'---------- IVS1    '         '        350  Hind III         '         '         '        400
GATGGCGAGTCACTGGAGAAGTTGTAATAGACTGCATAAGCAGGACACATGGAAGCTTGGTAGGTGAAGAGAGAGCAGAATGTCTGGCAGGAACTGATTT
                    **
         //CB2-----'---------'---------'-------450-------(+1f.s.)----'---------'---------'-------500
TCTCTACCCACCTGAAACACCTGGCCATCCTCAAGGGCACTTTTCTGGCATGAATGAAGCTGCACTGTGACAAGCTGCATGTAGATCCTGAGAACTTCA

-- IVS2  '         '         '         '        550        '         '         '         '        600
GGGTGAGTGGGCCAGATGCTCTGGCTTCTGCTTCTGCTCCTGCTCGTTTTATCACTTCTTTTTCCATTTGCCTTTTTTCCCCACAGTTTCCTCTACTGTT
  **
         '         '         '         '        650        '         '         '         '        700
CTGTATTTCTTCACTTAACACTCTTCTTTTGTTAACTTTCCTTCACAAGTCAACTTTCCCCCTTTTCCCCCCATAATCTTCTTTAACAAATCTTTTCCTC

         '         '         '         '        750        '         '         '         '        800
CTTCTTCCTCTCTCTCCCCCAATTCCCTTCCTTCATTCATCTTCAAATCCCCTTGGTAGGATCACCTCCTGAGTTATACATACTATCTCTCATCTACTAT

         '         '         '         '        850        '         '         '  Xba I  '        900
ATCTACATCTGGAAAACATCCTTCTTATTGCTACCATTGCTTGAGCTTTTGAGAGTCTCCTTAATGAAGAGGCACGTGTGGTTCTAGAGTCTTTGAATCA

         '         '         '         '        950        '         '         '         '       1000
TTCTTTAAATAATAAGAATTTCATGAATTTAGGCAGAGTAAGTGCAAAGATAAGGAAAGAAGAGTTAATGTCTAAGGATGGAAGACCTGAAATTCATATT

         '         '         '         '       1050        '         '         '         '       1100
AGAGATAGCATAACAGTAGATGGTAAAAAGGAGCCCACCATTGAACTAATCAATTAAGTAGTAAATATAAATATTATATAAAAAATACAAAGACACACAT

         '         '         '         '       1150        '         '         '         '       1200
TTTAGAAGAAATATCCTGGAAATTCATAGTGTTGGTGCAATTCCTTAAAAGACTAGCTTGATTTCTGATACCCAGGGGTAAATGTGTGTGCTCTTCTCCA

   CB3---'---------'---------'---------'------1250-Eco RI--'---------'---------'---------'------1300
CAGTTCTTGGGAAACATAATAGTGTTTGTGCTGTTCCTCCACTTTGGCAAGGAATTCACCCCAGCAGTGCAGGTTGCTTTTCAGGTGTTGGTGGCTTCTG
 **
---------'---------'---------Ter        '      1350L1Md repetitive sequence      '         '       1400
TAACCATGGCTCTAGCTCACAAGTACCACAGAGGCCTTGGACTATTTCCTCAAGATTACTGGAGCAAATGGCTATCTTGCCAAAAGCAATCTACAGATTC

         '         '         '         '       1450        '         '         '         '       1500
AATGCAATCCCCATTAAAATTCCAACTCAATTCTTCAACGAATTAGAAGGAGCAATTTGCAAATTCATCTGGAATAACAAAAAACCGAGGATAGCAAAAA

         '         '         '         '       1550        '         '         '         '       1600
CTCTTCTCAAGGATAAAAGAACCTCTGGTGGAATCACCATGCCTGACCTAAAGCTTTACTACAGAGCAATTGTGATAAAAACTGCATGGTACTGGTATAG

         '         '         '         '       1650        '         '         '         '       1700
AGACAGACAAGTGGACCAATGGAATAGAATTGAAGACCCAGAAATGAACCCACACACCTATGGTCACTTGATCTTCGACAAGGGAGCCAAAACCATCCAG

         '         '         '         '       1750        '         '         '         '       1800
TGGAAGAAAGACAGCATTTTCAACAATTGGTGCTGGCACAACTGGTTGTTATCATGTAGAAGAATGTGAATCGATCCATACTTATCTCCTTGTACTAAGG

         '         '         '         '       1850        '         '         '         '       1900
TCAAATCTAAGTGGATCAAGGAACTTCACATAAAACCAGAGACACTGAAACTTATAGAGGAGAAAGTGGGGAAAAGCCTTGAAGATATGGGCACAGGGGA

         '         '         '         '       1950        '         '         '         '       2000
AAAATTCCTGAACAGAACAGCAATGGCTTGTGCTGTAAGATCGAGAATTGACAAATGGGACCTAATGAAACTCCAAAGTTTCTGCAAGGCAAAAGACACT

         '         '         '         '       2050        '         '         '         '       2100
GTCTATAAGACAAAAAGACCACCAACAGACTGGGAAAGGATCTTTACCTATCCTAAATCAGATAGGGGACTAATATCCAACATATATAAAGAACTCAAGA

         '         '         '         '       2150       Eco RI    '         '         '       2200
AGGTGGACCTCAGAAAATCAAATAACCCCCTTAAAAAATGGGGCTCAGAACTGAACAAAGAATTC
```

FIG. 2. **DNA nucleotide sequence of *βh3* from the BALB/c mouse.** *Underlined* or *overlined* nucleotides denote regions of possible significance (see text). *Asterisks* (**) indicate splicing sites which conform to the GT/AG rule (Breathnach, *et al.*, 1978). Base composition and dinucleotide frequency data for the sequence appears under "Experimental Procedures" for use in verifying copies of the sequence. A partial sequence of *βh3* has been determined for the C57BL/10 mouse. This includes regions homologous to nucleotides 307–466 and 1197–1257 of the sequence shown in this figure. The C57BL/10 sequence differs from that shown at the following positions: 324 (C), 329 (G), 331 (G), 423 (A or T), 440 (T), 441 (C), 447 (T or C), 1223 (C), and 1224 (A).

optimal results for various comparisons.

Fig. 3 shows comparisons of the *βh3* sequence with the sequence of the adult $\beta 1^{dmaj}$ gene at four different "stringencies." When matches of 8 out of 10 nucleotides are plotted, we detected only the same obvious homologies noted earlier. A somewhat higher stringency (10 out of 12 nucleotides) reduced the background in the plot but did not reveal any new homology. A dramatic new homology was revealed by increasing the search "window," and simultaneously reducing the degree of homology required. A clear homology of the *βh3* sequence to the first coding block of β^{dmaj} (CB1) is revealed at a stringency of 30 out of 50 nucleotides (Fig. 3*c*). This plot shows virtually no spurious background. If the degree of homology required in the comparison is increased slightly (to 35 out of 50), this CB1 homology disappears from the plot (Fig. 3*d*).

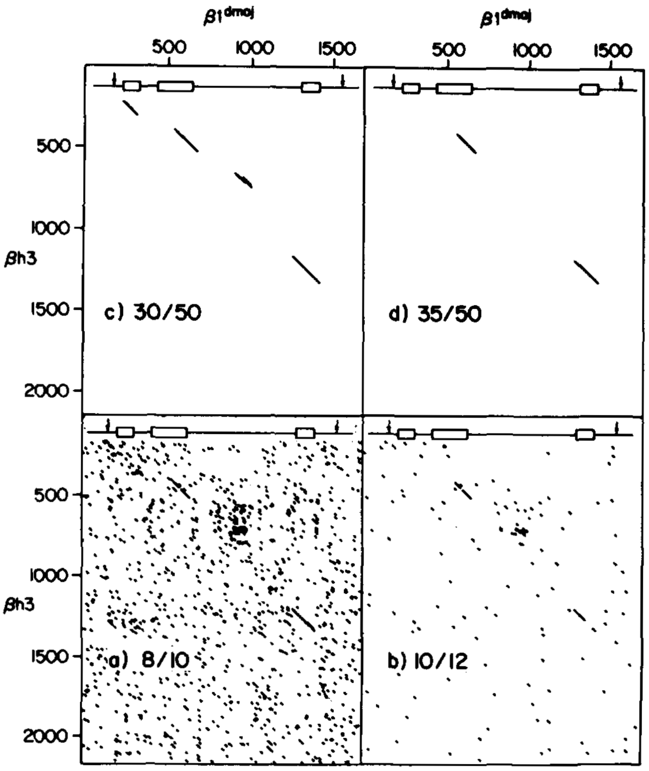

FIG. 3. **Dotplots comparing the *βh3* sequence with that of $\beta 1^{dmaj}$ (Konkel *et al.*, 1979).** Stringencies of comparison, expressed as the ratio of required matches to the length of the comparison window, appear in each panel. A scale diagram of the $\beta 1^{dmaj}$ gene is included in each panel for reference. *Arrows* indicate the positions of the 5′ and 3′ termini of mRNA, and *open boxes* indicate coding regions.

The dotplot of Fig. 3*c*, which shows homology between *βh3* and all three coding regions of $\beta 1^{dmaj}$ serves as a basis for aligning the two sequences. The displacement of the line indicating a CB2 match, from the diagonal on which the CB1 match appears, indicates a deletion of 154 nucleotides in the *βh3* structure.

In Fig. 4*a*, the portion of the *βh3* sequence flanking the deletion is shown aligned with the $\beta 1^{dmaj}$ sequence. Two different alignments of $\beta 1^{dmaj}$ which show the CB1 and CB2 homologies are displayed above and below the *βh3* sequence. The alignment which maximizes CB1 homology can be extended through most of the small intervening sequence (IVS1) by assuming a total of four bases deleted from the *βh3* sequence. The most dramatic defect in the *βh3* sequence is a deletion of 22 nucleotides of IVS1 plus 132 nucleotides of CB2. The transition from good to poor sequence homology to *βh3* is quite sharp at both deletion endpoints within the *βh3* structure. However, both alignments show the "patchy" homology beyond the region of strong homology, which we have previously noted for the CB2 alignment (Jahn *et al.* 1980). It is clear from the dotplots, as well as from inspection of the alignment, that the homology to the adult $\beta 1^{dmaj}$ sequence is much higher on the 3′ side of the deletion than on the 5′ side. A diagrammatic representation of the *βh3* structure compared to the $\beta 1^{dmaj}$ gene is shown in Fig. 4*b*. The degree of sequence homology within each segment of the gene is indicated.

In addition to the deletion described above, a number of other specific defects make it impossible for *βh3* to code for a protein resembling a β-globin. Both the "CCAAT" box and the "ATA" box sequences are altered in ways which probably inactivate the *βh3* promoter. There is no initiation codon at the beginning of CB1. As noted previously (Jahn *et al.*, 1980), a single base insertion at a position homologous to codon 90 puts termination codons in frame with the sequence homologous to codons 75–90. There is no termination codon at the end of the *βh3* sequence homologous to CB3. The first termination codon in frame with CB3 occurs at residue 185. However, this terminator occurs in a repetitive sequence located 3′ to *βh3*. There is no canonical polyadenylation addition sequence. These defects are listed in Table I, along with the characteristics of other β-globin pseudogenes.

The Origin of βh3—We have been interested in identifying the ancestral genes from which *βh3* was formed by recombination. A major difficulty in this identification arises from the frequent occurrence of gene conversion within β-globin clusters (Slightom *et al.*, 1981; Weaver, *et al.*, 1981; and Martin *et al.*, 1983). In gene conversion, sequence from one gene

TABLE I

Evolutionary diagnostic sequences and sequence defects in β-globin pseudogenes

a.a., amino acid; f.s., frame shift; n, nucleotide; ?, unknown; homol., homology to.

Evolutionary diagnostic	Mouse *βh2*	Mouse *βh3*	Rabbit $\psi\beta 2$	Goat $\psi\beta^x$	Lemur $\psi\delta$	Human $\psi\beta 1$
Coding sequence	Adult	Adult	Adult	Adult	5′ nonadult, 3′ adult	γ
5′ diagnostic	δ	Unrecognizable	δ	δ	$\psi\beta 1$	ε/γ
IVS2	δ	δ	δ	β	δ	?
3′ untranslated	δ	δ (truncated)	δ	δ (truncated)	δ	?
Defect in:						
"CAT" Box	GTAAC	Unrecognizable	Deleted	TTATT	ACAAT	TCAAT
"ATA" Box	CATAAA	TTAAAA	CATAAA	CTCACA	AATAAA	AGTAAA
Cap site	CTTCTG	CTTCTG	TATTTG	CTTTTG	CTTCTA	CTTCTG
Initiator	ATG (OK)	AGG	ATG (OK)	ATG (OK)	GTG	GTA
CB1	1–30 or 31?	−1 f.s. at 3, 60% homol. adult	−1 f.s. at 20	+1 f.s. at 12	−1 f.s. at 10	Terminator
IVS1	Alternate splice?	22 n deleted at 3′ end	No AG splice	GT->GC splice	GT->GC splice	OK
CB2	Terminator at 31? or 31–104?	31–74 deleted +1 f.s. at 90	31–104	Missing His 64,93	+4, +1, −1 f.s.	31–104
CB3	105–146	Extended	−1 f.s.	7 extra a.a.	105–146	105–146
Terminator	TGA (OK)	AGA	TAA (OK)	TGG:6 a.a.:TAA	TGA (OK)	TGA (OK)
+poly(A)	AATAAA (OK)	Missing (L1Md insertion)	AATAAA (OK)	AATTAA	AATAAA (OK)	AATAAA (OK)

a

```
  '        '       50        '        '        '        '      100   CAT Box
GGCAC....AGCATGTCCAGG...GAGAAATATCGCTTCGTCCTCACCGAAGCCTGATTCCGTAGAGCCACACCCTGGTAAGGGCCAATCTGCTCACA..  5'Beta dmaj
::::     ::  :::    :   ::::::::        :: ::: ::  ::::               : : : :::  :: :    :    :  ::  : :     ::
GGCAACTAAAGAGTGTGAGAGCTAGAGAAATAG....TCTTCCCCAGGGAAG........AGCACA.CCAAGCCTTCAGACTTGAAAACATACATGCAAG    Beta h3
'      '        50        '        '        '        '      100       '

     '       '        '      150 ATA Box '      '        '        '      200      '
...CAGGATAGAGAGGGCAGG.AGCCAGGGCAGAGCATATAAGGTGAGGTAGGATCAGTTGCTCCTCACATTTGCTTCTGACATAGTTGTGTTGACTCAC  5'Beta dmaj
   ::  : :  :: : :::: :: : :::  : ::::  :::::   : ::   :       : :   ::: ::::::::::: :: : :::: :::  :
TAACATTACATGGAAGTCAGGTAGTCCGGGATGGGCATTAAAGGTAGAGCAGATGCCTGCTGTGCCTGCATCTGCTTCTGACACAGGT.TGTTCACTAGC    Beta h3
  '      '        '      150        '        '        '        '      200      '

     '       '     CB1-'-------250---------'---------'---------'---------'-------300---------'---
AACCCCAGAAACAGACATCATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGTCTCTTGCCTGTGGGGAAAGGTGAACTCCGATGAAGTTGGTGGTGAG  5'Beta dmaj
: :      :  :  ::: :: :::: :: :::::::    ::::::   ::  ::  : :::: ::::  :  ::  ::   :   :  ::: ::: :::
AGC.....AGTCCTACACCAGGGTGTAC.TGACTGACAAAGAGAAGTTGGCCATCAGTGGCCTATGGGACACAGTAGACATGGGAAACATTGATGGCGAG    Beta h3
     '       '        '      250        '        '        '        '      300

------'---- IVS1'      '       350        '        '        '        '      400      '
GCCCTGGGCAGGTTGGTATCCAGGTTACAAGGCAGCTCACAAGAAGAAGTTGGGTGCTTGGAGACAGAGGTCTGCTTTCCAGCAGACACTAACTTTCAGT  5'Beta dmaj
 : ::::  : ::::  ::   :     : ::: :::  :::: :  :::: : :::     :: ::: ::::          : ::  ::::  ::: : ::::  :
TCACTGGAGAAGTTGTAATAGACTGCATAAG.CAGGACACATG..GAAGCTTGGTAGGTGAAGAGAGAGCAGAA.TGTCTGGCAGGAACTGATTTTCTCT    Beta h3
'      '        '        '      350        '        '        '        '      400
 :  ::::::  :           :::: :  :  :       ::::   ::       : ::::::     :    ::     ::    :::::  :::  :
GCTTTGGAGACCTATCCT....CTGCCTCTGCTAT....CATGGGTAAT..GCCAAAGTGAAGGCCCATGGCAAGAA...GGTGATAACTGCCTTTAACG  3'Beta dmaj
(CB2)---'-----------500---------'------------'-----------'---------'----------550---------'-------

    * junction '        'CB2----450---------'---------'---------'---------'-------500
GTCCCCTGTCTATGTTTCCCTTTTAGGCTGCTGGTTGTCTACCCTTGGACCCAGCGGTACTTTGATAGCTTTGGAGACCTATCCTC                  5'Beta dmaj
  :: :      :    :  :  :      :    :: :    : :: :  ::: : ::: ::: :   :    :     :
ACCCACCTGAAACACCTGGCCATCCTCAAGGGCACTTTTTCTGGCATGAATGAAGCTGCACTGTGACAAGCTGCATGTAGATCCTGAGAACTTCAGGGTG    Beta h3
   '       '        '        '      450        '        '        '        '      500
:    :::::: ::: ::: :: :::::::::::: ::: :   :: : : ::: ::: ::::::::::::::::::::: ::::::::::::::::::::
AT.GGCCTGAATCACTTGGACAGCCTCAAGGGCACCTTTGCCAGCCTCAGTGA.GCTCCACTGTGACAAGCTGCATGTGGATCCTGAGAACTTCAGGGTG  3'Beta dmaj
-'---*-junction-------'-------600---------'---------'----------'---------'-------650---------'--- IVS2

         '        '        '        '      550        '        '        '        '
AGTG....GGCCAGATGCTCTGGCTTCTGCTTCTGCTCCTGCTCGTTTTATCA..CTTCTTTTTCCATTT........GCCTTTTTTCCCCACAGTTTC    Beta h3
:::     :: ::  : ::  :  : :: :  :::       :::  : :  :::  :::: : :     :             ::  :: :        :::: ::
AGTCTGATGGGCACCTCCTGGGTTTCCTTCCCCTG.....GCTATTCTGCTCAACCTTCCTATCAGAAAAAAAGGGGAAGCGATTCTAGGGAGCAGTCTC  3'Beta dmaj
   '       '        '      700        '        '        '        '      750      '
```

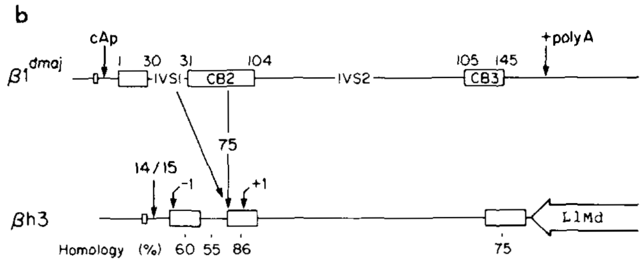

FIG. 4. **Structure of *βh3*.** *Panel A*, alignment of the *βh3* sequence with $\beta 1^{dmaj}$ across the recombination-deletion boundary. The *βh3* sequence is shown aligned in two different registers with $\beta 1^{dmaj}$. The *βh3* sequence is numbered as in Fig. 2, and the $\beta 1^{dmaj}$ sequence (Konkel *et al.*, 1979) is numbered from 1 to 1726. *Periods* (.) have been added to the sequences to provide gaps necessary for alignment. The MATCH program described by Dayhoff *et al.* (1981) with a gap penalty of 2, was used to align the 5′ $\beta 1^{dmaj}$ sequence and *βh3* in the region 5′ to the junction (indicated by *), and to align 3′ $\beta 1^{dmaj}$ and *βh3* in the region 3′ to the junction. The program only aligns two sequences. In the region where three sequences are shown, the third sequence was aligned by inspection. *Colons* (:) indicate sequence homology. The CAT Box and ATA Box sequences of $\beta 1^{dmaj}$ are underlined. Coding regions of $\beta 1^{dmaj}$ are *overlined* or *underlined.* Splice signals following the GT/AG rule are underlined, as is the initiation codon (ATG). Patches of homology between *βh3* and the poorly aligned sequence, in the region where three sequences are shown, are indicated by *underlining. Panel B*, diagram of the *βh3* structure compared to the expressed *β*-globin gene, $\beta 1^{dmaj}$. *Numbers* on the $\beta 1^{dmaj}$ map refer to codons in the mature globin. Coding regions (*CB*) and introns (*IVS*) are labeled. The capping site (*cAp*) and the poly(A) addition site are indicated by *arrows*. In the *βh3* map, *open boxes* indicate the sequence homologous to coding blocks, while the *lines* connecting them are homologous to the intron sequence. The *arrows* labeled *−1* and *+1* indicate the positions of a single base deletion and insertion. The *arrow* labeled *14/15* indicates a sequence homologous to the $\beta 1^{dmaj}$ capping site found in the *βh3* sequence (see text). *Two long arrows* originating within IVS1 and at codon 75 at $\beta 1^{dmaj}$ indicate the portion of the gene deleted from *βh3*. The *large open arrow* in the *βh3* diagram represents the L1Md repetitive element.

within a cluster replaces sequence in a second gene, thereby obscuring the evolutionary history of the second gene. In the analysis of primate *β*-globins, it has been shown that noncoding sequences are converted less frequently than coding sequences, and are therefore better indicators of the time of the *δ*-*β* duplication (Martin *et al.*, 1983; Jeffreys *et al.*, 1982). We have applied the same reasoning to trace the ancestry of the *βh3* structure. We have used the dotplot method to compare

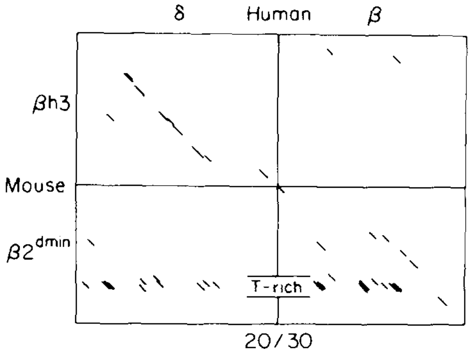

FIG. 5. **Dotplots of the mouse *βh3* and $\beta 2^{dmin}$ IVS2 sequences compared to those of human δ and β globins.** The stringency of comparison was 20 matches required, with a window size of 30 nucleotides. A portion of the $\beta 2^{dmin}$ sequence rich in T gives rise to a band of spurious matches for statistical reasons, and is indicated by the label "T-rich" in the diagram.

```
---------- Delta - like ---><---------- L1Md- like ----------------

L1Md    ATGCTCATGGATTGGCAGGATCAACATT....GTAAAAATGGCTATCTTGCCAAAAGCAA   (65)
        |    | | |          |||| |||    |   ||||||||||||||||||||||||
Beta h3 AGAGGCCTTGGACTATTTCCTCAAGATTACTGGAGCAAATGGCTATCTTGCCAAAAGCAA (1389)
         |||  | |||||| |||||| |  |  |   ||  |      |  | |
Delta   TGAGATCCTGGACTGTTTCCTGATAACCATAAGAAGACCCTATTTCCCTAGATTCTATTT (1690)
           |  | |||||| |||||| 14/18 (78%)
Beta h3 ---GGCCTTGGACTATTTCCT
           |  |     ||   |     5/18 (28%)
Beta    ---GCTCGCTTTCTTGCTGTC
            | | ||||| ||||  |  12/18 (67%)
Bldmaj  ---ACCCCCTTTCCTGCTCTT
        Ter
```

FIG. 6. **Alignments of the 3′ untranslated regions of *βh3*, human δ, $\beta 2^{dmin}$, and human β.** The first three nucleotides of the human δ-globin sequence shown are the termination codon (TGA). *Numbers* to the *right* of the L1Md, *βh3*, and δ sequences are the coordinate of the last nucleotide shown. L1Md is the sequence of clone MIFC37 from Brown and Piechaczyk (1983). *βh3* is numbered as in Fig. 2. The human δ sequence of Spritz *et al.* (1980) is numbered from 1 to 1976.

the noncoding sequences of *βh3* with those of human δ and β.

The region of *βh3* homologous to the large intron, IVS2, is compared to the corresponding human sequences in Fig. 5. IVS2 from an expressed mouse adult gene, $\beta 2^{dmin}$, is also compared with the human sequences, using the same stringency of comparison (20 out of 30). The comparison of Fig. 5 clearly shows that the IVS2 of *βh3* is more closely related to human δ than to β. Similarly, $\beta 2^{dmin}$ is more β-like than δ-like. (In the case of $\beta 2^{dmin}$, a region high in T introduces a background band of spurious matches to both δ and β, as indicated in Fig. 5.)

The 3′ untranslated sequence of *βh3* provides a second diagnostic region. In Fig. 6, the 3′ sequences are compared for mouse genes *βh3* and $\beta 1^{dmaj}$, and for human δ and β. Only a very short region of 18 nucleotides is available for comparison, because a large repetitive element is located immediately 3′ to *βh3* (see below). Within this region, the *βh3* sequence matches δ much better than β (78% compared to 28%). Similarly, $\beta 1^{dmaj}$ is clearly β-like rather than δ-like (67% compared to 33%).

Taken together, these results clearly indicate that the 3′ end of the *βh3* structure is descended from a δ-like ancestor. This implies that the duplication leading to δ and β predates the speciation of mouse and man.

We have been unable to identify any donor gene for the 5′ portion of *βh3* by using dotplots. Comparisons with other globin gene sequences by this method do not distinguish whether this region is adult or nonadult in origin, presumably because this sequence is too divergent from other globin sequences to permit tracing its ancestry.

A Member of the L1 Family of Interspersed Repetitive Sequences Is Immediately 3′ to βh3—The 0.9-kb *Eco*RI fragment which contains the 3′ end of the *βh3* structure hybridizes with nick-translated mouse genomic DNA under conditions which only display repetitive sequences.[4] The sequence flanking the 3′ end of *βh3* can be aligned with a published sequence from the mouse repetitive family MIF-1 (Brown and Piechaczyk, 1983). Homology to this repetitive sequence picks up immediately following the disappearance of homology to the 3′ untranslated sequence of the δ gene, as shown in Fig. 6. This MIF-1 sequence is a subset of the long interspersed repetitive element which we have called L1Md (LINES One of *Mus domesticus*; see Voliva *et al.*, 1983; Martin *et al.*, 1984).

Evolutionary Fixation of βh3 in the Rodent Lineage—Although it is seriously defective, the *βh3* structure appears to be fixed at a frequency of 100% in the mouse (*M. domesticus*) genome. The BALB/c sequence reported in this paper carries the $[Hbb]^d$ haplotype at the β-globin locus. We have also determined a partial sequence of *βh3* from C57BL, which carries the $[Hbb]^s$ haplotype. The Charon 4A clone BA-11 was isolated from a library of C57BL/10 liver DNA. This clone contains a 4.8-kb *Eco*RI fragment homologous to the *Eco*RI A fragment of BALB/c.[5] Homology between these two mouse strains is great enough that most restriction sites mapped in BALB/c *βh3* also occur in C57BL. Preliminary sequence data has been obtained for portions of *βh3* from C57BL. This sequence which spans the deletion within the *βh3* structure (see below) differs from the BALB/c sequence by a few single base substitutions, as indicated in the legend to Fig. 2. This sequence therefore demonstrates that *βh3* in the $[Hbb]^s$ haplotype has the same recombinant structure as described for $[Hbb]^d$.

In order to investigate the existence of *βh3* in other mouse strains, and related rodents, we have probed genomic DNA digests with *βh3* specific probes. Jahn *et al.* (1980) showed that unique genomic probes can be derived from *βh3*, in spite of demonstrable homology to adult coding sequences. We have prepared probes which are specific for the 5′ and the 3′ regions of the *βh3* structure. These have been used to probe *Eco*RI digests of genomic DNA from various rodents. If *βh3* exists in an animal, then we expect a single *Eco*RI fragment to hybridize with both probes. Fig. 7 shows the probes used in these experiments, along with some representative data.

In a mouse of the haplotype $[Hbb]^p$ (AU/SsJ), an *Eco*RI fragment indistinguishable in size from *Eco*RI fragment A (4.8 kb) hybridized with both probes (data not shown). The same result was obtained with individuals of several other species of the genus *Mus*, *molossinus*, *cookii*, and *hispanicus* (Fig. 7). A fragment 5.9 kb in size was observed with both probes in *M. caroli* (Fig. 7), consistent with the existence of *βh3* in this species. In *Mus pahari*, two fragments, 7.2 and 5.5 kb, were observed with the 3′ probe, but only the smaller of these was seen when the 5′ probe was used (Fig. 7). This result would be obtained if *pahari* carries a *βh3* structure containing an *Eco*RI site within the region of the 3′ probe. Alternatively, the two bands could represent alleles at the *βh3* locus, since the animal was not inbred. The most distantly related genomes we have probed with *βh3* are the rat, *R. norvegicus*, and the deermouse, *P. maniculatus*. Both of these genomes show specific hybridization to BALB/c *βh3* although the bands observed are much fainter than with *Mus* DNAs. In the rat, a single band 4.0 kb was seen using both probes

[4] C. F. Voliva, S. L. Martin, C. A. Hutchison, III, and M. H. Edgell, manuscript in preparation.

[5] S. Weaver, manuscript in preparation.

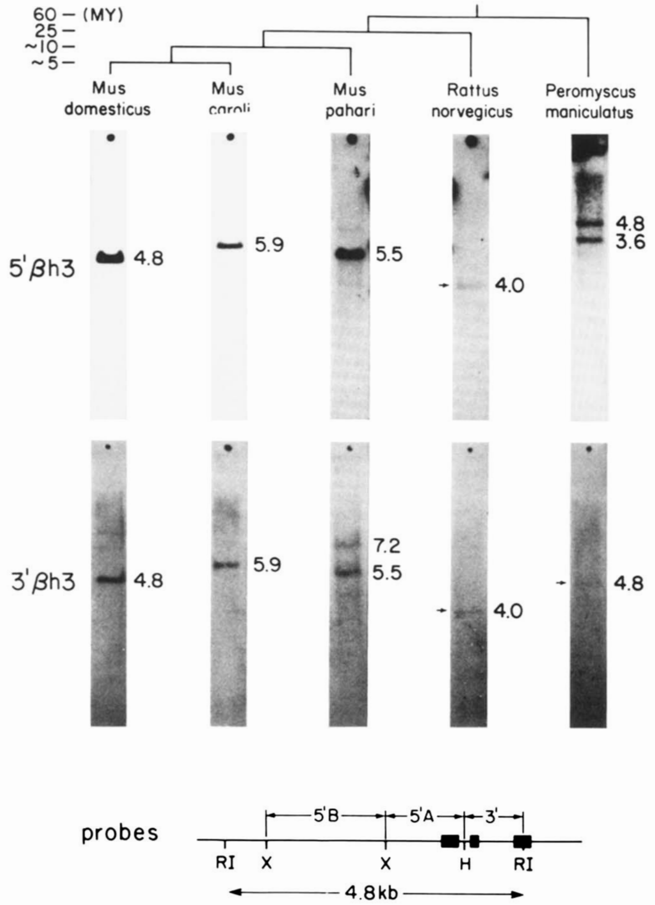

FIG. 7. **Various rodent genomic DNAs probed with 5′ and 3′ sequences of *βh3*.** Total genomic DNA from the indicated sources was digested with *Eco*RI and electrophoresed in 1.0% agarose, then transferred to nitrocellulose. The *M. domesticus* DNA used in this experiment is from strain AU/SsJ. The *diagram* at the *bottom* of the figure indicates the source of the *βh3* specific hybridization probes used in these experiments. Restriction site abbreviations are: *RI* (*Eco*RI), *X* (*Xba*I), and *H* (*Hin*dIII). The *5′B* probe was used in the upper *P. maniculatus* lane. All other *upper lanes* were visualized with probe "*5′A*". The probed labeled *3′* was used for all *lanes* in the *bottom row*. DNA fragments were sized by comparison with marker fragments (not shown). The 5′-probed *P. maniculatus* lane was electrophoresed for a significantly shorter time than all of the other experiments shown.

(Fig. 7). This strongly suggests the existence of the recombinant *βh3* structure in the rat. In one *Peromyscus* individual, the 3′ probe showed a single faint band indistinguishable from *Eco*RI fragment A (4.8 kb), while a probe from the 5′ flanking region of *βh3* (see Fig. 7) showed two bands, 4.8 and 3.6 kb. This result would be obtained if *Peromyscus* carries a *βh3* containing an *Eco*RI site in its 5′ region. Since this animal was inbred at the *β*-globin locus, it is unlikely that these two bands represent two alleles of *βh3* in *Peromyscus*. Another possible explanation for these results is that *βh3*-related sequences in *Peromyscus* are contained in more than one gene within the *β*-globin complex. These results indicate that *βh3*-like sequences are a universal feature of the rodent *β*-globin gene complex. However, our results do not prove that the observed hybridization in all rodents is due to a recombinant structure identical to *βh3* of BALB/c.

DISCUSSION

βh3 Is a Recombinant Structure—The peculiar structure of the *βh3* pseudogene immediately suggests models for its origin. The most significant defect in the *βh3* sequence is a deletion of 154 nucleotides, compared to the functional adult gene $\beta 1^{dmaj}$. This deletion includes 22 nucleotides of the small intron (IVS1) and more than half (132 nucleotides) of the second coding region (CB2). The sequence 3′ to the deletion is highly homologous to $\beta 1^{dmaj}$ (86% match for the remaining 100 nucleotides of CB2); however, the 5′ end of *βh3* is considerably more divergent (60% match for CB1, and 55% match for the IVS1 alignment shown in Fig. 4*a*.). This difference in homology to adult gene sequence suggests that the two parts of *βh3* have different origins. It seems reasonable to conclude that the structure arose by recombination between an adult gene and a gene which was already divergent prior to the fusion event. We cannot distinguish whether the recombination and deletion resulted from two separate events, or from a single unequal crossover between two globin genes.

It is curious that the portion of CB2 deleted in *βh3* corresponds closely to a "structural unit" (Go, 1981), and that the 3′ endpoint of the deletion is located very near the position of an intron unique to leghemoglobin (Østergaard Jensen *et al.*, 1981). The deletion endpoint in *βh3* is within codon 74, Go identified a structural unit boundary at residue 66–71, and the leghemoglobin intron is located between codons 68 and 69. All of these phenomena may possibly be related to a hotspot for recombination at this location within globin genes.

The 3′ Portion of βh3 Is Descended from a "Proto-δ" Gene—We have presented evidence that the adult-like 3′ portion of *βh3* is more closely related to the human *δ*-globin gene than to other adult *β*-globin genes. In this analysis, we have made use of intron (IVS2) and 3′ untranslated sequence. These sequences are less likely than coding sequences to have undergone recent gene conversion events, and are therefore uniquely useful in tracking the evolutionary history of genes. Our results indicate that the 3′ portion of *βh3* and the human *δ*-globin gene are descended from a common ancestor.

Pseudogenes which are *δ*-like have also been seen in the lemur (Jeffreys *et al.*, 1982) and in the rabbit (Lacy and Maniatis, 1980). Our analysis of the mouse *β*-globin locus indicates that *βh2* is also a *δ*-like pseudogene (Hardies *et al.*, 1984).

To account for these findings, we propose that there were at least two adult genes in the ancestor to all present day mammals. We call these "proto-*δ*" and "proto-*β*". We refer to these genes as adult because all of their known modern descendents are adult, with the exception of the goat fetal globin gene (Schon *et al.*, 1981). The analysis of coding sequences has not previously revealed two ancestral genes because gene conversion has caused them to evolve in concert.

The proto-*δ* and proto-*β* genes must have existed for a significant period prior to the mammalian radiation. This is necessary in order for the diagnostic sequences to have diverged significantly by the time of the mouse-human split. A significant divergence prior to this split could explain why the large number of changes since this event have not obscured the lineages of the genes. We believe that the diagnostic untranslated sequences are not evolving at a rate dramatically higher than that found for coding sequences. This has been found for sequences flanking primate adult *β*-globin genes (Martin *et al.*, 1981). A similar result comes from the comparison of homology between mouse *βh3* and human delta within IVS2 (60%, not counting insertions and deletions), and within coding sequence (80% for the 3′ portion of CB2 plus CB3).

The 5′ Portion of βh3 Is Descended from a Different Ancestral Gene—The 5′ part of the *βh3* structure is the most divergent *β*-globin sequence known. Parsimony analysis of *β*-globin sequences shows that CB1 of *βh3* is not recently derived from any other known globin gene (results not shown). Dot-plot comparisons with both adult and nonadult genes have

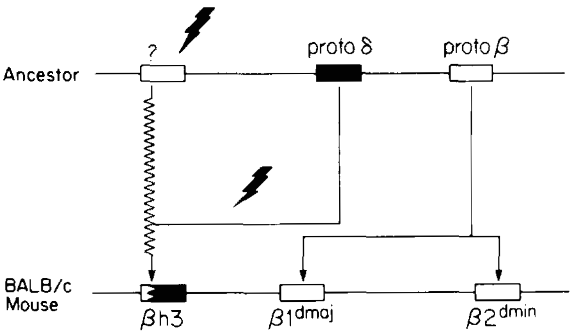

FIG. 8. **Diagram showing the origin of βh3 from the "proto-δ" gene and an unknown ancestral gene, indicated by ?.** *Arrows* indicate lines of descent. *Jagged arrows* indicate evolution as a pseudogene. *Lightning bolts* indicate gene-inactivating events.

also failed to identify any donor for the 5′ βh3 sequences. The divergent nature of the 5′ sequence can be explained by assuming that it is derived from a pseudogene which had been evolving without selection for a long time prior to the formation of βh3. A model for the evolution of the adult β-globin genes has been developed (Hardies *et al.*, 1984) in which the 5′ βh3 sequences are descended from a proto-β gene. This assignment is based mainly on criteria other than sequence homology, such as conserved intergenic distances.

A diagrammatic representation of our model for the formation of βh3 is shown in Fig. 8. An implication of the model is that some species may exist today carrying the two parental genes which formed βh3, in unfused form. We are looking for these genes in rodents distantly related to the laboratory mouse.

A Repetitive Element Is Immediately 3′ to βh3—Strong homology to L1Md family (see below) repetitive DNA sequence begins 18 nucleotides beyond the position of the CB3 terminator of βh3. This sequence is part of a repetitive element with a maximum length of about 7 kb[6] (Fanning, 1983; Voliva *et al.*, 1983). This element in the mouse has been called the *Bam*HI family (Fanning, 1983), and recently has been shown to be homologous to the human *Kpn*I family (Singer *et al.*, 1983; Burton *et al.*, 1983). We have used the name L1 (LINES One) as a general name for this family, and L1Md (LINES One of *M. domesticus*) for the murine form of the repeat (Voliva *et al.*, 1983; Martin *et al.*, 1984), based upon the terminology of Singer (1982). Voliva *et al.* (1983) have shown that most members of the L1Md family are truncated at variable distances from a conserved, A-rich 3′ endpoint. The particular example adjacent to βh3, termed L1Md-4, is one of seven occurrences in the BALB/c β-globin complex, and is truncated with a length of 2.1 kb. Recently, we have observed that the L1Md family sequence contains a long open-reading frame (Martin *et al.*, 1984). Following the TGA termination codon at residue 1767 of the βh3 sequence, an open-reading frame of 978 nucleotides is found, extending beyond the sequence presented here into a 4.5-kb *Eco*RI fragment (Martin *et al.*, 1984). Comparison with homologous sequences from other species indicates that this sequence is evolving as a protein-coding sequence. It should be noted that the first terminator 5′ to, and in frame with, the TGA at 1767, is located at position 1179 of the βh3 sequence, within IVS2 of βh3. We believe that a consensus sequence for the *Bam*HI family would not contain a terminator at position 1767, but would contain a single open-reading frame greater than 1.6 kb in length.[6] The implications of these open-reading frames are being pursued.

The Abundance and Evolutionary Persistence of Pseudogenes—β-globin pseudogenes similar to βh3 are so common that they require an explanation. Our genomic probing experiments indicate that the recombinant βh3 structure was genetically fixed in the common ancestor to the genus *Mus*. These experiments also suggest the presence of βh3 in rat and the deermouse (*Peromyscus*). All other β-globin gene clusters mapped in detail also have pseudogenes located between the adult and nonadult genes. Table I lists the defects found in these structures, along with what is known of their ancestry. Five of the six pseudogenes listed, from mouse, rabbit, goat, lemur, and human, are at least partially derived from the ancestral proto-δ gene. Only the human lineage, in which the δ gene remains active, has acquired a pseudogene (ψβ1) by inactivation of a nonadult gene. In the lemur, ψβ1 has apparently fused with the δ gene to form the ψδ pseudogene (Jagadeeswaran *et al.*, 1982). Details of the tabulated defects differ, indicating that the δ gene was independently inactivated in the mouse, rabbit, goat, and some primates. It would also appear that βh2 and βh3 of the mouse are the results of independent inactivations (see Hardies *et al.*, 1984).

Does βh3 Have a Function?—One explanation for the persistence of pseudogenes is that they may have some function in globin gene regulation or expression. However, there is no evidence to support this idea. We have failed to detect transcripts from the region of the βh3 sequence homologous to functional globin gene promoters, under conditions where expressed globin genes are transcribed in vitro.[2] We do not know if the repetitive sequence 3′ to βh3 can be transcribed. It is possible, of course, that a pseudogene might have a function which does not require transcription. For example, it might contain a binding site for a regulatory protein.

Another class of models does not require that the pseudogene play an active role in gene expression, but allows for its persistence to be selected. It is clear that gene clusters containing adult and nonadult genes arise by a series of duplications. Initially, the cluster would contain a group of active, closely spaced genes. Perhaps tissue specific domains tend to be larger than the average duplicated region. Genes in the middle of the cluster may not have room to be independently regulated. They therefore tend to be inactivated, but are retained to provide a spacer between the adult and nonadult region.

Finally, we propose an explanation for the abundance of pseudogenes which is independent of any functional role for these structures. There seems to be an element of truth to the following argument, whether or not pseudogenes have some, as yet undiscovered, function.

Genome Expansion and the Persistence of Pseudogenes—Perhaps pseudogenes are an incidental consequence of the continuous expansion of the eukaryotic genome. This model assumes that duplications become evolutionarily fixed, in higher eukaryotes, at a much higher rate than deletions. This would be expected, for example, if most deletions have deleterious effects, but most duplications do not. Once a functional gene has duplicated, one copy could be inactivated by mutation without resulting in strong negative selection. The inactive gene may then become an evolutionarily fixed pseudogene by genetic drift. The pseudogene remains because there simply is no efficient mechanism for deleting it.

A consequence of an excess of duplications over deletions is an increase in genome size with time. Consider the history of the mouse β-globin complex. The whole region has effectively duplicated three times during its 500-million-year history, since the α-β split. The primitive β gene had increased

[6] F. H. Burton, D. D. Loeb, C. F. Voliva, S. L. Martin, M. H. Edgell, and C. A. Hutchison, III, manuscript in preparation.

to a total of eight genes, one of which was lost during the fusion which formed *βh3*. The number of *β*-globin genes varies somewhat from one mammal to another. From the numbers of genes for mouse (7), rabbit (4), goat (9), and man (6), we obtain an average of 6.5 *β*-globin genes per mammalian cluster. This corresponds to 2.7 net doublings in the *β*-globin region. If the behavior of the *β*-globin complex is representative of the genome as a whole, then the mammalian genome has been doubling in size about every 185 million years (2.7 doublings in 500 million years).

This is consistent with an argument based on quite different considerations. It has been estimated that the earliest eukaryotes arose about 1.5 × 10^9 years (1.5 billion years) ago (Margulis, L., 1970). It has been proposed that such organisms were ancestors to modern eukaryotes, and that they were formed by symbiosis of prokaryotes. We will assume that the primitive eukaryote had a genome complexity similar to *Escherichia coli* (4 × 10^3 kb). (Presumably, eukaryotes evolved from prokaryotes, whether or not the theory of endosymbiosis is correct.) The present mammalian genome is approximately 1000 times larger. Assuming a constant rate of growth from the primitive eukaryote to the mammal, we arrive at 10 doublings in 1.5 billion years (2^{10} is approximately 1000), or a doubling time of about 150 million years. Nei (1969) used similar reasoning with somewhat different assumptions to arrive at an estimate of 300 million years for the doubling time for the mammalian genome.

The fact that this growth rate is in agreement with that based on the history of the *β*-globin complex means that genome growth in the mammal could be entirely accounted for by local tandem duplications. It is not necessary to invoke duplication of the whole genome, or of chromosomes, to explain the present size of the mammalian genome. Of course there is considerable variation in the size of present day eukaryotic genomes, ranging from 1.6 × 10^5 kb for *Drosophila*, to 2.4 × 10^8 kb in *Lilium longiflorum* (reviewed by Holliday, 1970). This variation can be explained by doubling times which are within a factor of 2 of that calculated above for the mammalian lineage. We estimate doubling times of 280 million years for the line leading to *Drosophila*, and 95 million years for *Lilium*, assuming that genome expansion was constant over the last 1.5 billion years. (This constant rate assumption seems unlikely to be strictly accurate.)

We conclude that higher eukaryotes have adopted an evolutionary strategy in which tandem duplications exceed deletions, resulting in net genome growth with doubling times of 200 ± 100 million years. Inactivated duplicate genes therefore persist in the genome as pseudogenes.

REFERENCES

Blin, N., and Stafford, D. W. (1976) *Nucleic Acids Res.* **3,** 2303–2308

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P. (1978) *Proc. Natl. Acad. Sci. U. S. A.* **75,** 4853–4857

Brown, B. A., Padgett, R. W., Hardies, S. C., Hutchison, C. A., III, and Edgell, M. H. (1982) *Proc. Natl. Acad. Sci. U. S. A.* **79,** 2753–2757

Brown, S. D. M., and Piechaczyk, M. (1983) *J. Mol. Biol.* **165,** 249–256

Burton, F. H., Voliva, C. F., Edgell, M. H., and Hutchison, C. A., III (1983) *DNA (NY)* **2,** 82 (abstr.)

Dayhoff, M. O., Schwartz, R. M., Chen, H. R., Hunt, L. T., Barker, W. C., and Orcutt, B. C. (1981) *Nucleic Acid Sequence Database*, Vol. 1, National Biomedical Research Foundation, Washington, D. C.

Denison, R. A., Van Arsdell, S. W., Bernstein, L. B., and Weiner, A. M. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78,** 810–814

Edgell, M. H., Weaver, S., Jahn, C. L., Padgett, R. W., Phillips, S. J., Voliva, C. F., Comer, M. B., Hardies, S. C., Haigwood, N. L., Langley, C. H., Racine, R. R., and Hutchison, C. A., III (1981) in *Organization and Expression of Globin Genes* (Stamatoyannopoulos, G., and Nienhuis, A., eds) pp. 69–88, Alan R. Liss, Inc., New York

Fanning, T. G. (1983) *Nucleic Acids Res.* **11,** 5073–5091

Farace, M. G., Brown, B. A., Raschella, G., Alexander, J., Gambari, R., Fantoni, A., Hardies, S. C., Hutchison, C. A., III, and Edgell, M. H. (1984) *J. Biol. Chem.* **259,** 7123–7128

Firtel, R. A., Timm, R., Kimmel, A. R., and McKeown, M. (1979) *Proc. Natl. Acad. Sci. U. S. A.* **76,** 6206–6210

Gibbs, A. J., and McIntyre, G. A. (1970) *Eur. J. Biochem.* **16,** 1–11

Gō, M. (1981) *Nature (Lond.)* **291,** 90–92

Goeddel, D. V., Leung, D. W., Dull, T. J., Gross, M., Lawn, R. M., McCandliss, R., Seeburg, P. H., Ullrich, A., Yelverton, E., and Gray, P. W. (1981) *Nature (Lond.)* **290,** 20–26

Hardies, S. C., Edgell, M. H., and Hutchison, C. A., III (1984) *J. Biol. Chem.* **259,** 3748–3756

Holliday, R. (1970) *Symp. Soc. Gen. Microbiol.* **20,** 359–380

Jacq, C., Miller, J. R., and Brownlee, G. G. (1977) *Cell* **12,** 109–120

Jagadeeswaran, P., Pan, J., Forget, B. G., and Weissman, S. M. (1982) *Cold Spring Harbor Symp. Quant. Biol.* **47,** 1079–1086

Jahn, C. L., Hutchison, C. A., III, Phillips, S. J., Weaver, S., Haigwood, N. L., Voliva, C. F., and Edgell, M. H. (1980) *Cell* **21,** 159–168

Jeffreys, A. J., Barrie, P. A., Harris, S., Fawcett, D. H., Nugent, Z. J., and Boyd, A. C. (1982) *J. Mol. Biol.* **156,** 487–503

Kan, N. C., Lautenberger, J. A., Edgell, M. H., and Hutchison, C. A., III (1979) *J. Mol. Biol.* **130,** 191–209

Kimura, M. (1980) *J. Mol. Evol.* **16,** 111–120

Konkel, D. A., Maizel, J. V., Jr., and Leder, P. (1979) *Cell* **18,** 865–873

Lacy, E., and Maniatis, T. (1980) *Cell* **21,** 545–553

Leder, P., Hansen, J. N., Konkel, D., Leder, A., Nishioka, Y., and Talkington, C. (1980) *Science (Wash. D. C.)* **209,** 1336–1342

Lee, M. G.-S., Lewis, S. A., Wilde, C. D., and Cowan, N. J. (1983) *Cell* **33,** 477–487

Li, W-H., Gojobori, T., and Nei, M. (1981) *Nature (Lond.)* **292,** 237–239

Little, P. F. R. (1982) *Cell* **28,** 683–684

Lueders, K., Leder, A., Leder, P., and Kuff, E. (1982) *Nature (Lond.)* **295,** 426–428

Maizel, J. V., Jr., and Lenk, R. P. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78,** 7665–7669

Maniatis, T., Jeffrey, A., and Kleid, D. G. (1975) *Proc. Natl. Acad. Sci. U. S. A.* **72,** 1184–1188

Margulis, L. (1970) Origin of Eukaryotic Cells, p. 66, Yale University Press, New Haven

Martin, S. L., Zimmer, E. A., Davidson, W. S., Wilson, A. C., and Kan, Y. W. (1981) *Cell* **25,** 737–741

Martin, S. L., Vincent, K. A., and Wilson, A. C. (1983) *J. Mol. Biol.* **164,** 513–528

Martin, S. L., Voliva, C. F., Burton, F. H., Edgell, M. H., and Hutchison, C. A., III (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81,** 2308–2312

Maxam, A. M., and Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560

Miyata, T., and Hayashida, H. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78,** 5739–5743

Miyata, T., and Yasunaga, T. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78,** 450–453

Nei, M. (1969). *Nature (Lond.)* **221,** 40–42

Østergaard Jensen, E., Paludan, K., Hyldig-Nielsen, J. J., Jorgensen, P., and Marcker, K. A. (1981) *Nature (Lond.)* **291,** 677–679

Phillips, S. J., Hardies, S. C., Jahn, C. L., Edgell, M. H., and Hutchison, C. A., III. (1984) *J. Biol. Chem.* **259,** 7947–7954

Sanger, F., and Coulson, A. R. (1978) *FEBS Lett.* **87,** 107–110

Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467

Schon, E. A., Cleary, M. L., Haynes, J. R., and Lingrel, J. B. (1981) *Cell* **27,** 359–369

Singer, M. F. (1982) *Cell* **28,** 433–434

Singer, M. F., Thayer, R. E., Grimaldi, G., Lerman, M. I., and Fanning, T. G. (1983) *Nucleic Acids Res.* **11,** 5739–5745

Slightom, J. L., Blechl, A. E., and Smithies, O. (1981) *Cell* **21,** 627–638

Spritz, R. A., DeRiel, J. K., Forget, B. G., and Weissman, S. M. (1980) *Cell* **21,** 639–646

Staden, R. (1982) *Nucleic Acids Res.* **10,** 2951–2961

Voliva, C. F., Jahn, C. L., Comer, M. B., Hutchison, C. A., III, and Edgell, M. H. (1983) *Nucleic Acids Res.* **11,** 8847–8859

Wahl, G. M., Stern, M., and Stark, G. R. (1979) *Proc. Natl. Acad. Sci. U. S. A.* **76,** 3683–3687

Weaver, S., Comer, M. B., Jahn, C. L., Hutchison, C. A., III, and Edgell, M. H. (1981) *Cell* **24,** 403–411

White, C. T., Hardies, S. C., Hutchison, C. A., III, and Edgell, M. H. (1984) *Nucleic Acids Res.* **12,** 751–766