THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 1984 by The American Society of Biological Chemists, Inc.
Vol. 259, No. 12, Issue of June 25, pp. 7947–7954, 1984
Printed in U.S.A.

The Complete Nucleotide Sequence of a β-Globin-like Structure, βh2, from the [*Hbb*]^d Mouse BALB/c*

(Received for publication, August 11, 1983)

Sandra J. Phillips‡, Stephen C. Hardies, Carolyn L. Jahn§, Marshall H. Edgell, and Clyde A. Hutchison III

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

We have determined the complete nucleotide sequence of βh2, a pseudogene in the mouse β-globin gene complex. The structure of βh2 is analogous to that of a normal β-globin gene, and its nucleotide sequence shares 72% homology with the coding regions of a reference mouse adult β-globin gene. A frame shift occurs in the first coding region for which a compensatory splicing scheme can be devised. The reading frame is not otherwise disrupted. All of the recognized transcription, translation, and splicing signals in βh2 are intact, with the exception of the "CCAAT box," which has been altered to GTAAC. We compared the predicted amino acid sequence of βh2 with other β-globin sequences. Evidence for a period of divergence without selection in the history of βh2 was found in a set of codons that are usually highly conserved in productive β-globin genes. An evolutionary tree constructed from nucleotide sequence suggests that βh2 originated from the adult genes at least 60 million years ago. After some period as a productive gene, βh2 was inactivated and has subsequently diverged without selection. Hybridization experiments demonstrated that βh2 and the surrounding region occur without major alteration in other rodent species. The sequence (AGCCA-4*n*-GTGT) occurs 5′ of the CCAAT box in βh2 and in many productive globin genes.

Four β-globin proteins are found in the blood of mice carrying the [*Hbb*]^d haplotype: two during embryogenesis and two during adulthood (1). When the mouse β-globin gene cluster was cloned and characterized, seven regions containing globin homology were discovered (2, 3). We report here the complete nucleotide sequence of one of the unanticipated globin-like regions named βh2. We show that βh2 is a pseudogene: an evolutionary remnant of a once active β-globin gene.

The pseudogene concept was put forward (4, 5) to explain the frequent occurrence of sequences that are homologous to known genes but which appear to be defective in some way. It was proposed that extra copies of genes are sometimes generated by gene duplication or other processes. Since conservative selection would not act on unnecessary sequences, inactivation through mutation of the control sites in these extra genes is eventually inevitable. Pseudogenes can be recognized because during divergence as neutral sequences they accumulate mutations at many sites that are conserved in functional genes. β-globin pseudogenes have been found in human (6), rabbit (7), and goat (8); and α-globin pseudogenes in mouse (9, 10) and human (11). They exhibit some or all of the properties expected from the model for their formation; that is, no correspondence to a known gene product, no transcription detected *in vivo* or *in vitro*, an accelerated rate of divergence, and sequence defects that would preclude expression of a typical globin protein.

The organization of the β-globin complex locus in the BALB/c mouse is *y*-βh0-βh1-βh2-βh3-β1^{dmajor}-β2^{dminor} (2, 3).[1] The gene, *y*, codes for the late embryonic β-globin, y (2, 12). β1^{dmajor} and β2^{dminor} are the two adult genes (3). The four remaining genes were named βh for "β homologous" because partial sequence determinations showed homology to, but not identity with the published amino acid sequence of known mouse β-globins (2). βh1 is now believed to code for the early embryonic β-globin z, in spite of its disagreement with the published amino acid sequence.[2] βh0 is now also known to be transcribed *in vitro*,[3] in induced murine erythroleukemia cells (13), and in early mouse embryos,[2] and may code for a minor embryonic β-globin. βh3 is a pseudogene with multiple sequence defects and apparently is a recombinant between two very different ancestral genes.[4] The fact that βh0 and βh1 are functional genes underscores the caution necessary in classifying genes as pseudogenes when the biology, structure, or history of the gene in question is not fully understood.

The nucleotide sequences of pseudogenes are valuable for a number of purposes. They have been used in an attempt to estimate the neutral mutation rate (14), which is important for a number of issues in molecular evolution. The history of the pseudogenes in a given cluster must be understood before the evolution of the cluster itself and its relation to the analogous loci in other species can be fully understood. Also, pseudogene sequences increase the DNA sequence data base thereby improving the accuracy of analytical methods for evolutionary studies, such as tree making algorithms.

This paper evaluates the evidence that βh2 is a pseudogene.

* 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, The Jackson Laboratory, Bar Harbor, ME 04609.

§ Present address, Department of Biological Sciences, University of Illinois at Chicago, P. O. Box 4348, Chicago, IL 60680.

[1] Nomenclature is according to that proposed by the Committee on Standardized Genetic Nomenclature for Mice (13).

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

[3] B. A. Brown, S. C. Hardies, B. Timmons, A. Hill, M. H. Edgell, and C. A. Hutchison III, manuscript in preparation.

[4] C. A. Hutchison III, S. C. Hardies, R. W. Padgett, S. G. Weaver, and M. H. Edgell, manuscript in preparation.

This is an Open Access article under the CC BY license.

An initial examination suggested a paradox. The sequence of *βh2* does not correspond to that of any known protein and is very divergent from the other mouse β-globins. We found that *βh2* shared only 63% homology by amino acid residues and 72% homology by nucleotides to its closest functional counterpart. However, the control signals thought to be important for transcription, translation, and splicing as well as the overall reading frame were not proportionately disrupted. Analysis of codons at positions where highly conserved residues normally appear in β-globins was used to clarify the nature of *βh2*. A model for the history of *βh2* is proposed to account for all of the features of its sequence.

EXPERIMENTAL PROCEDURES

Materials—Restriction endonucleases were purchased from New England Biolabs (Beverly, MA) or Bethesda Research Laboratories (Rockville, MD). *Escherichia coli* DNA polymerase I, large fragment (Klenow enzyme) and *E. coli* DNA polymerase I were purchased from Boehringer Mannheim. Calf intestinal phosphatase was from Worthington. Polynucleotide kinase was from Bethesda Research laboratories. [γ-^{32}P]ATP (specific activity 3000 Ci/mmol) and [α-^{32}P] deoxynucleotide triphosphates (specific activity >300 Ci/mmol) were purchased from New England Nuclear. Deoxynucleotide triphosphates were purchased from P-L Biochemicals.

Isolation of Plasmid Clones from the Mouse β-Globin Cluster—The *Bam*HI fragments from CE19 (2) were subcloned into the *Bam*HI site of pBR322 (Fig. 1). The plus (+) orientation is defined to be when the globin gene is in the same polarity as the vector tetracycline gene. The names, contents, and insert orientation of the plasmids from left to right on the CE19 map are: pHE107 (3′ *βh0*, +), pHE109 (3′ *βh0*, −), pHE116 (*βh0*, *βh1* intergene region, −), pHE110 (5′ *βh1*, +), pHE111 (5′ *βh1*, −), pHE112 (3′ *βh1*, 5′ *βh2*, −), pHE113 (3′ *βh1*, 5′ *βh2*, +), and pHE115 (3′ *βh2*, −).

DNA Preparation—Plasmid DNA from the subclones described above was prepared by chloramphenicol amplification, sodium dodecyl sulfate lysis, and cesium chloride equilibrium centrifugation (15). Restriction digest conditions were those described by the enzyme suppliers. Fragments for sequencing were prepared by horizontal gel electrophoresis in 1% agarose or vertical gel electrophoresis in 5% polyacrylamide (1:29 cross-linked) in 0.04 M Tris acetate (pH 7.4), 0.02 M sodium acetate, 2 mM EDTA (15). Fragments were isolated from gel slices by horizontal electroelution in dialysis tubing coated with bovine serum albumin (15). Electroeluted DNA was further treated by phenol extraction, ether extraction, and ethanol precipitation.

Mouse DNA for genomic probings was prepared by the procedure in Ref. 16 modified for use on whole animals.

DNA Sequencing—End labeling of restriction fragments using *E. coli* DNA polymerase I, large fragment (Klenow enzyme) on recessed 3′ ends, or polynucleotide kinase on 5′ ends, and subsequent chemistry were performed according to Maxam and Gilbert (17). Reaction products were analyzed on 8 or 20% polyacrylamide/urea gels which were 0.3 mm thick (18). An average of 200 nucleotides was read per set of reaction products.

Genomic Probing—Fragments were nick translated with [α-^{32}P] nucleotide triphosphates to a specific activity of 1–3 × 10^8 cpm/μg using the method of Maniatis *et al.* (19). Genomic DNA (4 μg) was digested to completion with *Eco*RI, and subjected to gel electrophoresis in 1% agarose. The genomic digests were transferred to nitrocellulose filters by the method of Southern (20). Hybridization to the probes was carried out in 10% dextran sulfate and 50% formamide (21) with modifications described by Jahn *et al.* (2).

DNA Sequence Analysis—Some of the globin sequences were acquired in computer readable form from Dayhoff's nucleic acid sequence data base (22). Several computer programs used for sequence analysis were written in this lab for a Z-80 based microcomputer with a CP/M operating system. DNA sequence files were created and edited using a character oriented sequence editor (SED) written in Pascal. Fig. 2 was generated by a sequence comparison and display program (FIGMAKER) written in BASIC to run a DIABLO 1620 printer. Restriction site searches were done using the program ALLSITES (23).

Sequence comparisons were made using graphic dot matrix generating programs DIAGSRCH and DAIGPLOT (24). *βh2* and *β2^{dminor}* were aligned (Fig. 2) in regions (coding blocks, IVS[5] 1, and from the CCAAT box to the AUG) where the dot matrix displayed a string of local homologies of greater strength than the random background. The sequences were then edited to bring the local homologies into a single alignment. Because IVS 2 showed no homology between these genes, an arbitrarily positioned gap was added to *β2^{dminor}* to equalize the lengths. The extreme 5′ end was shifted to juxtapose the AGCCA box and the 3′ end was shifted to juxtapose the AATAAA box.

Evolutionary trees were made by the method of maximum parsimony (25) using an IBM 370 computer. A branch swapping program (MPN) was obtained from M. Goodman (Wayne State University) (26) and a second program (ALLPOS) that evaluates all possible trees for a small number of sequences was obtained from W. Fitch (University of Wisconsin). Branch lengths were calculated according to Fitch (27). Reasonable alternatives to the most parsimonious tree were found with ALLPOS and then further studied in the context of the full data set using MPN. Subsets of greater than 100 base pairs from the *βh2* sequence showed similar results to the full sequence.

A site frequency correction for positions that change amino acids (28) was used in the analysis of contact class residues. The number of changes to this class between human β and mouse *β2^{dminor}* was taken as the rate expected in productive genes.

RESULTS

Sequencing Strategy—*βh2* was sequenced by the chemical degradation method of Maxam and Gilbert (17). The original source of *βh2* DNA was CE19, a recombinant Charon 4A phage wtih a 17-kilobase mouse DNA insert spanning *βh0*, *βh1*, and *βh2* (2). Plasmid subclones of *Bam*HI fragments from CE19 were constructed (Fig. 1) to assist with restriction mapping of the *βh2* region and to act as a ready source of DNA restriction fragments for sequencing. The sequencing strategy (Fig. 1) covered 2 kilobase of DNA with sequence

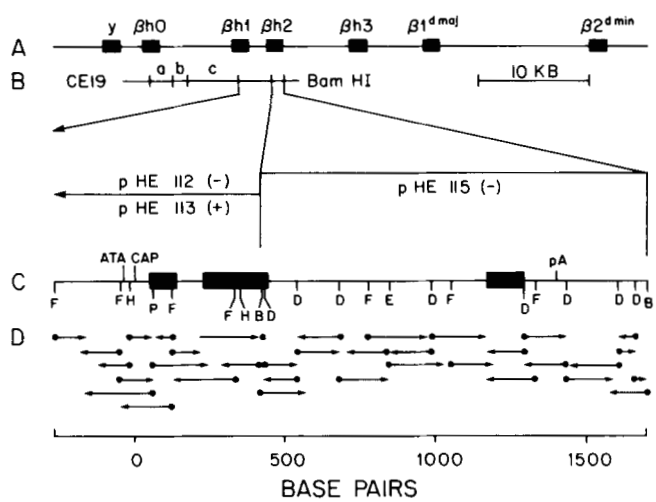

FIG. 1. **Strategy for sequencing *βh2*.** *A* is the mouse β-globin complex locus, [*Hbb*]^d. *B* shows the position within the cluster of the plasmid clones that were sequenced. The region that was found in the Charon phage clone CE19 (2) is displayed with its *Bam*HI sites marked. Subcloning of all of the *Bam*HI fragments from CE19 into pBR322 is described under "Experimental Procedures." Fragments marked *a*, *b*, and *c* are cloned in pHE107 and 109, pHE116, and pHE110 and 111, respectively. *C* displays the region whose sequence is reported in Fig. 2. The scale is in base pairs with the putative CAP site taken as coordinate 1. Regions that are analogous to coding blocks 1, 2, and 3 of other β-globins are marked with *boxes*. Sequences that are homologous to the ATA box, capping site, and poly(A) addition site are indicated by *ATA*, *CAP*, and *pA*, respectively. Restriction sites that were used in the sequence analysis are indicated as follows: *B*, *Bam*HI; *D*, *Dde*I; *E*, *Eco*RI; *F*, *Hin*fI; *H*, *Hae*III; and *P*, *Pst*I. *D* shows the extent and direction of individual sequence determinations. In each case the *dot* shows the position of the radioactive label.

[5] The abbreviation used is: IVS, intervening sequence.

determined from both strands for about 75% of the region. All but the most extreme 60 base pairs of the 5′ flanking DNA were sequenced several times from different sites to assure accuracy.

The sequence will be made available in computer readable form on request. The following is presented to aid verification of copied or transmitted sequence. The length is 1989. The base composition is 606 A, 398 C, 426 G, and 559 T. Dinucleotide frequencies are in Table I.

βh2 Coding Blocks—The sequence of *βh2* is structurally analogous to that of a typical β-globin gene (Fig. 2). Three regions of homology to β-globin coding blocks are separated by two sequences that are analogous in length and position to β-globin intervening sequences. The GT and AG dinucleotides usually found at splice sites (29) occur in *βh2* at positions analogous to the splice sites in other globin genes. There are start and stop codons at the appropriate positions and no internal in frame terminators. Although *βh2* is not a productive gene, for convenience we will refer to its "coding blocks" and "intervening sequences" by analogy to productive β-globin genes.

There are changes in the length of the *βh2* coding blocks 1 and 2 relative to mouse $\beta 2^{dminor}$ at positions 124 and 289, respectively (Fig. 2). The latter is a loss of three consecutive base pairs such that the reading frame is not disrupted. The one base deletion at position 124 shifts the frame of the last 6 codons of coding block 1. If splicing occurs after position 140 as in $\beta 2^{dminor}$, then the frame shift would be propagated into coding block 2 and lead to early termination of translation. However, another splice site satisfying the GT-AG rule (29) exists at position 144, which if used would allow resumption of the normal reading frame of coding block 2. Although this alternate splice site at position 144 does not conform well to the consensus splice sequence (30, 31), an analogous IVS 2 splice site is used in a rabbit β1 globin variant lacking a normal IVS 2 splice site (32). So the hypothesis that the alternate splice site could act to rescue *βh2* from the effects of the frame shift at 124 cannot be excluded. Therefore, neither of the length changes in *βh2* coding blocks can be taken as definitive evidence that the gene is defective.

Single base changes between *βh2* and $\beta 2^{dminor}$ are roughly evenly distributed. Homology to coding blocks 1, 2, and 3, is 69% (66/96), 74% (163/220), and 70% (88/126), respectively. In contrast to *βh3*, the neighboring mouse pseudogene,[4] *βh2* has not been recently altered by recombination or gene conversion.

The greatest local variation of homology occurs in coding block 2, where there is a region of 60% homology (bases 226–298) adjacent to a region of 81% homology (base 299 to the splice site). In the low homology region, there are two patches of high homology at 257–264 and 268–277. Each contains a phenylalanine codon (positions 260 and 269) that is conserved in all functional globins ranging from human α and β to root nodule leghemoglobin (33). This part of coding block 2 demonstrates the characteristic divergence pattern of *βh2*. That is, in spite of extensive divergence, the highly conserved features of functional β-globins are well preserved.

TABLE I
Frequencies of dinucleotides in βh2

First nucleotide	Second nucleotide			
	A	C	G	T
A	186	110	153	157
C	147	93	9	148
G	149	81	111	85
T	124	114	152	169

Intervening Sequences—The alignment of IVS 1 (Fig. 2) was made with the help of a dot matrix program (24). Five gaps covering a total of 43 positions were introduced. The remaining positions have 67% (57/85) homology. So, with respect to base substitutions, IVS 1 in *βh2* has diverged to a similar extent as the surrounding coding blocks. Such a relationship between IVS 1 and coding sequence divergence is also observed in other globin genes (34).

Except around the splice sites, IVS 2 of *βh2* and $\beta 2^{dminor}$ were so poorly related that no alignment was attempted. The IVS 2 in *βh2* does share homology with some globin genes (human δ, mouse *βh3*) but not with others (human β, rabbit β1, mouse $\beta 2^{dminor}$, and mouse $\beta 1^{dmajor}$).

3′ Sequences—In the 3′ untranslated region *βh2* had no significant homology to $\beta 2^{dminor}$ except for the sequence AATAAA-23*n*-TCA that is proposed to direct polyadenylation (35) (Fig. 2).

5′ Sequences—*βh2* has homology to $\beta 2^{dminor}$ 5′ to the initiation codon. The sequence CTTPyTG which is thought to be involved in capping (36) appears at position 6. Capping in $\beta 2^{dminor}$ occurs at the adenine homologous to position −1 in *βh2* (36). A mutation changes that A to G in *βh2* leaving position +1 as the nearest adenine for capping. An ATA box (37) appears at −31 in a region sharing homology to the ATA box of $\beta 2^{dminor}$. In addition to the ATA box, the sequence CCAAPy located 30–40 bases further upstream is thought to be involved in promoting transcription (34) and is often referred to as the CCAAT box. This sequence is changed to GTAAC in *βh2* and appears at −75. The two changed bases in this pentanucleotide are the most obvious damage to the control sequences of *βh2*. The homology between *βh2* and $\beta 2^{dminor}$ as well as among β-globin genes in general (38) is not confined to the three small sites described above. So there may be more subtle changes in the 5′ flanking region of *βh2* that restrict its potential for expression.

During our examination of the region 5′ to the CCAAT box we discovered another site that may represent a conserved sequence. It has been observed (34) that there are sequences 5′ to the CCAAT site that are homologous among the adult β-globin genes of several species, but not between adult and fetal or embryonic globins. *βh2* does not share significant homology with $\beta 2^{dminor}$ immediately 5′ to the CCAAT box. However, sequence alignment of *βh2* with $\beta 2^{dminor}$ and with other globin genes revealed a region of homology at position −149 (Fig. 2). The consensus sequence at this position is AGCCA-4*n*-GTGT (Fig. 3). This is part of the homologous region among adult globins also reported by Hardison (38). We also find that adult globin genes have a good match to this sequence at a relatively consistent position. In the adult mouse α-globin gene, *α1*, there is a good match to this sequence, although its position is different, and overlaps the CCAAT box (Fig. 3). Human (39) and goat (40) adult α-globin genes similarly have a sequence like AGCCA-4*n*-GTGT overlapping with their CCAAT boxes. We were less successful in finding homology to this sequence in non-adult genes. However, goat γ is identical to the goat adult gene in this region including the match to AGCCA-4*n*-GTGT and a possible match was found in human ε (Fig. 3).

Evolution of βh2—We have examined the evolutionary relationship of *βh2* to other β-globin genes (Fig. 4). Because of the relatively great divergence of *βh2*, there is a larger than usual degree of error associated with placing it on an evolutionary tree. For example, using the method of maximum parsimony to place *βh2* onto a tree with the other globins, we found 3 positions where it could join at the cost of only a few

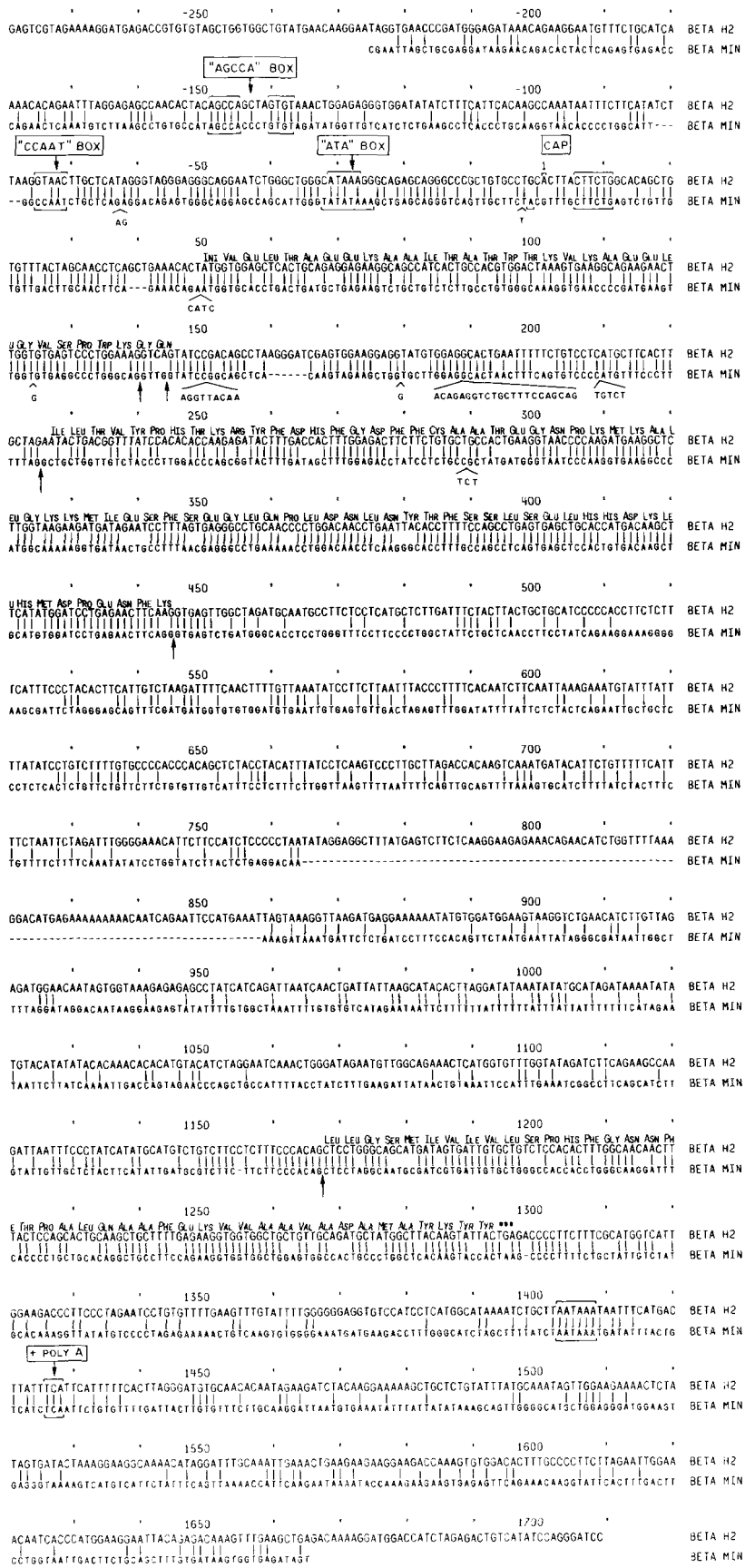

FIG. 2. **Nucleotide sequence of *βh2* compared with $\beta 2^{dminor}$.** *Dashes* represent gaps introduced into $\beta 2^{dminor}$ (46) for alignment. Splice sites are indicated by *arrows* under the $\beta 2^{dminor}$ sequence. Position 140 is the correct splice site for $\beta 2^{dminor}$. The alternate position shown at 144 would compensate for the frame shift in the first coding block of *βh2*. See "Experimental Procedures" for the alignment procedure. The gap in the second intervening sequence was placed arbitrarily. The numbering system includes a nucleotide numbered zero.

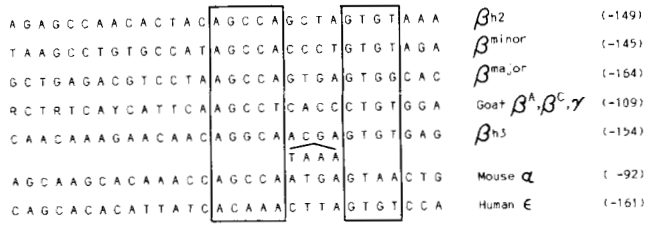

FIG. 3. **Conserved site among globin genes in the 5′ flanking region.** The coordinates are such that transcription initiation is at number 1 and there is a base number zero. References for the sequences together with the coordinate of the cited sequence in the original numbering system are: mouse $\beta 2^{dminor}$, 77 (46); mouse $\beta 1^{dmajor}$, 16 (46); goat, −109 (47); mouse *βh3*, 29;[4] mouse α, 280 (48); and human ε, 225 (49).

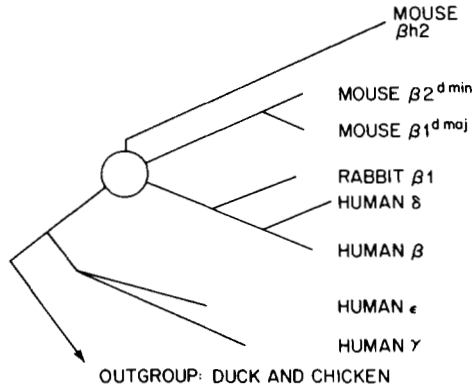

FIG. 4. **Evolutionary relationship between *βh2* and other β-globin genes.** The tree was established by the method of maximum parsimony (see "Experimental Procedures"). Three different branching arrangements for *βh2* yield trees of nearly equal total length. The circle intersects the three corresponding positions where *βh2* would join the rest of the tree. Moving the position of *βh2* disturbs the lengths of the three branches in the vicinity by about the diameter of the circle. The branch lengths are scaled according to total number of mutations. The length of rabbit *β1* to its first node is 23 mutations.

mutations over the most parsimonious solution (Fig. 4). These positions are clustered around the point where the mouse adult genes join those of the other mammals. The choice of the most parsimonious position depended on which other sequences were included in the analysis. We conclude that the true relationship of *βh2* to the other globins could be reflected by a joining anywhere within the area defined by these three positions.

We draw several conclusions about *βh2* from this evolutionary tree (Fig. 4). First, *βh2* is on the adult lineage. That is, *βh2* falls on the adult side of the division that occurred about 200 million years ago separating most present day adult β-globins from non-adult β-globins. Second, *βh2* is old. It intersects the common adult ancestral amino acid sequence at the time of the mammalian radiation (60 million years) or even before. Much of the observed divergence of *βh2* from the other adult genes would be expected of a gene this old, even if it were fully functional. There does appear to be extra length in the *βh2* branch, indicating accelerated divergence due to loss of function. However, the exact amount of extra divergence in *βh2* is difficult to determine due to its uncertain age.

We then asked how the point changes in the coding sequences of *βh2* were distributed with respect to the features that confer function on a β-globin. Among functional β-globins, the number of different residues found at each position within the protein follows a non-random profile (Fig. 5, *closed boxes*). Regions of low divergence are found where the β-chain makes contact with the α chain, heme, or diphosphoglycerate. Excluding the frame shifted region, there are 36 positions within the coding sequence of *βh2* where amino acid residues are specified that do not appear in the other β-globins examined (Fig. 5, *open boxes*). These positions are distributed smoothly over the sequence rather than showing the characteristic profile of the functional β-globins.

We compared *βh2* to a consensus β-globin sequence which was constructed from the most frequent residue found at each position in a set of nine functional β-globins. Differences from the consensus were scored for heme, αβ, or diphosphoglycerate contact residues, and non-contact residues (Table II). A number of adult and non-adult functional genes are shown as controls. They are considerably less divergent in their contact residues than in their non-contact residues presumably due to the effects of selection. In *βh2*, on the other hand, divergence in the contact class has risen to a level similar to that seen in the non-contact class.

We then asked if there was a conflict between the extract divergence observed in the contact residues of *βh2* and the relatively good preservation of its structure and control sequences. The following control sequences seem to be nearly totally conserved in functional β-globins: CCAAT box, ATA box, capping site, start codon, GT-AG at each intervening sequence, terminator codon, and poly(A) addition site. These comprise a set of 35 highly conserved bases from which *βh2* has three changes: CC to GT in the CCAAT box and the loss of the normal adenine for capping. This is consistent with a history in which about 10% of available sites are mutated during a time of evolution after loss of function. The extent of divergence in the contact residue class was converted to the number of base pairs changed per available site by the method of Miyata and Yasunaga (28). When this was done we also found about 10–15% of the available sites changed beyond what is expected for a functional β-globin. So, the amino acid residue analysis does not conflict with the analysis of control bases.

The effect on contact residues *versus* non-contact residues is amplified because the non-contact codons have tended to accumulate multiple mutations that would change the identity of the amino acid more often than have the contact codons. Neither the analysis of contact residues or control bases are greatly affected by the uncertainty in determining the true age of *βh2*, because the background of changes at these sites in functional genes is so low. We conclude that all of the data about *βh2* is consistent with a gene that arose sometime around the mammalian radiation and was under functional constraints during its early history. More recently, selective pressure was lost and sufficient time has since passed for an extra 10% of the bases in the gene to have been changed over what would be expected for a functional β-globin.

Detection of βh2 in Other Rodent Species—Based on the history of *βh2* described above, a related pseudogene should be widespread among rodents. We have examined the DNA from a variety of other species by hybridization to a probe (pHE112) that includes *βh2*, *βh1*, and the entire intergene region (Fig. 6*A*). The different species are ordered from left to right in Fig. 6 according to their genetic relatedness to the source of the probe. The probe is from BALB/cJ, an inbred strain of *Mus domesticus*, with the haplotype $[Hbb]^d$. C57BL/10J, another inbred *M. domesticus* strain, carries a different haplotype ($[Hbb]^s$). *Mus castaneus, Mus spretus,* and *Mus cervicolor* are other species in the same subgenus as *M. domesticus*. *Mus pahari* is in a different subgenus, while *Peromyscus maniculatus* is in a different family (41, 42).

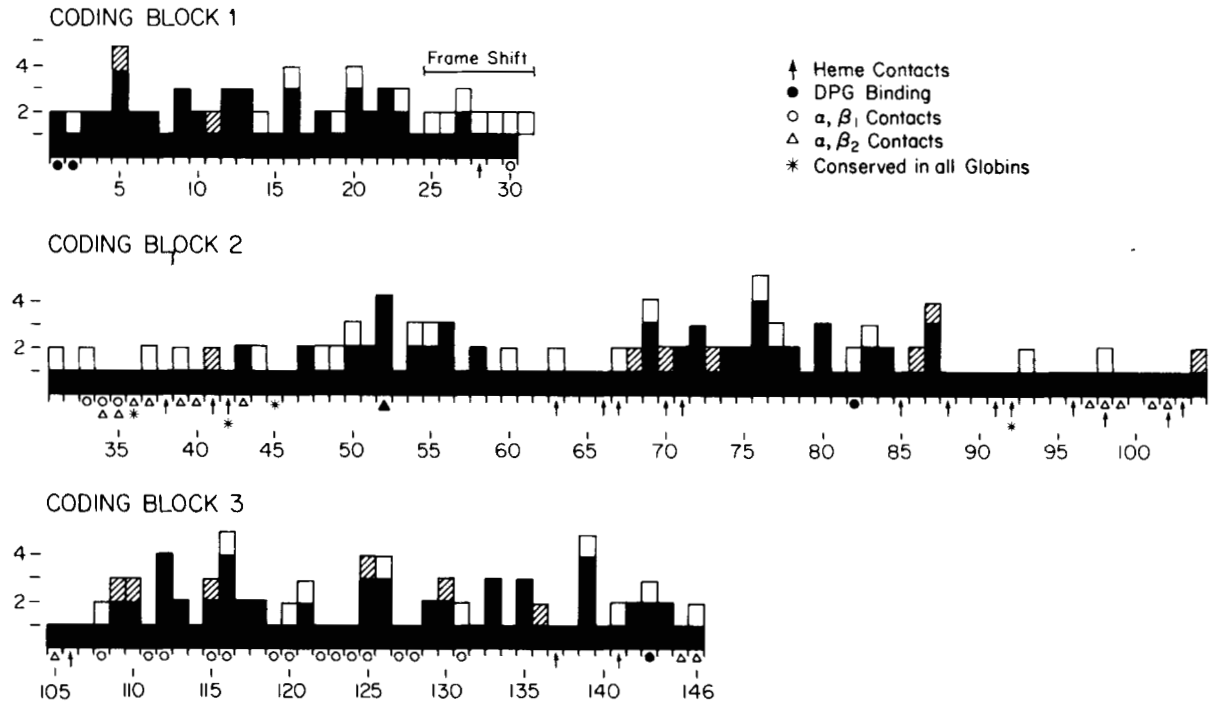

FIG. 5. **Conservation of amino acid sequence in *βh2 versus* other β-globins.** The number of different amino acids appearing at a given position among nine reference β-globins is presented as the height of a *solid* or *solid* and *hatched bar*. An additional *open box* appears where *βh2* codes for an amino acid residue that does not occur in the reference β-globins at that position. The upper unit on a *bar* is *hatched* if *βh2* codes for a residue that occurs in one of the other reference genes, but is not the most common (or tied for the most common) residue at that position. Residues involved in intermolecular contacts are displayed as tabulated by Eaton (44). *DPG* is diphosphoglycerate; α,β_1, and α,β_2 refer to contact with each of the two α subunits in the hemoglobin tetramer. The four positions marked by *asterisks* are conserved in all known functional globins (33). The numbering system is that of the reference genes which have no deletions or insertions relative to one another. The frame-shifted sequence at the end of coding block 1 of *βh2* has one extra amino acid before encountering the alternate splice site. *βh2* does not have a homologous amino acid for position 52 due to a three base deletion. The nine reference globin sequences were taken from the published nucleotide sequences. They were human $^{A}\gamma$ and $^{G}\gamma$ (50), mouse $\beta 1^{dmajor}$ (51) and $\beta 2^{dminor}$ (46), human δ (52) and β (53), human ε (49), and two alleles of rabbit β1 (54).

TABLE II

Divergence of residues by intermolecular contact class

The percentage of residues that differ from a consensus globin sequence is tabulated for each of three classes distinguished by type of intermolecular contact. There were 19, 34, and 88 residues in the heme contact, other contact, and non-contact classes, respectively. Positions where *βh2* is frame shifted or deleted were excluded from the analysis. The classification scheme, literature references for each sequence, and the definition of the consensus globin sequence are described in the legend to Fig. 5.

Residue type	Human			Mouse	
	ε	$^{G}\gamma$	δ	$\beta 2^{dmin}$	*βh2*
			%		
Heme contact	5	10	0	5	32
Other contact	3	12	9	6	41
Non-contact	26	22	13	21	35

The $[Hbb]^d$ and $[Hbb]^s$ haplotypes of *M. domesticus* share the same *Eco* RI fragments except for a small length difference in the intergene region. This is in agreement with previous mapping studies (2, 16). The next two most closely related species (*M. castaneus* and *M. spretus*) have a pattern identical to one or both of the *M. domesticus* haplotypes. Therefore, the region including *βh2* is not greatly altered among these three *Mus* lineages. The ancestor to these three lineages must have had an ancestral *βh2* gene in the same relative position to *βh1* as do its present day descendents.

The more distantly related species show successively weaker hybridization to fragments of different sizes (Fig. 6*A*). This loss of intensity is expected due to greater divergence. Fragment sizes may have changed by insertion/deletion or by mutation at the defining restriction sites. The presence or absence of *βh2*, specifically, in the more distantly related species is not determined because the hybridization could be due to *βh1* or intergene sequences.

Repeated Sequences 3′ of βh2—A second probe (pHE115) was used to analyze these same DNAs (Fig. 6*B*). It contains the 3′ end of *βh2* from the *Bam*HI site in coding block 2 (position 427, Fig. 2) to the *Bam*HI site 408 base pairs downstream from the termination codon at 1709. The hybridization data shows that a repetitive element is present within the region covered by this probe. Comparison of the *βh2* sequence to that of the large *Bam*HI repetitive element[6] (43) showed strong homology with sequences 3′ to the poly(A) addition site (Fig. 2, positions 1550–1709). Hybridization with this repetitive element also decreased in the more distantly related species (Fig. 6*B*) reflecting greater divergence.

DISCUSSION

We report here the complete nucleotide sequence of *βh2*, a gene-like structure found in the murine β-globin complex locus. The sequence has homology and structural similarity

[6] C. Voliva, personal communication.

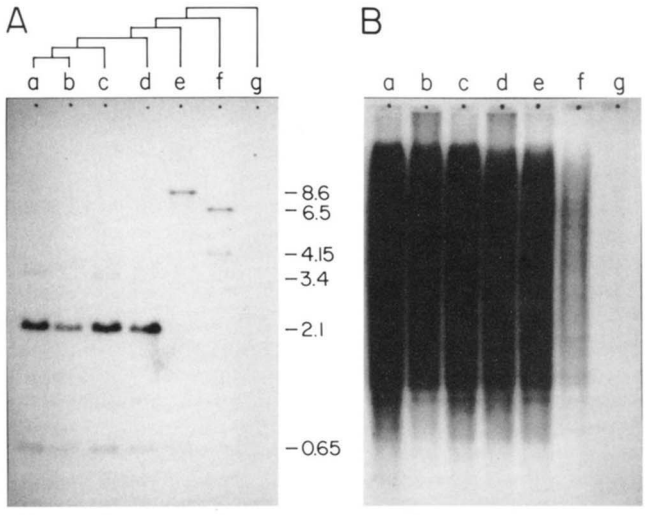

FIG. 6. **Presence of *βh2* and surrounding sequences in other rodents.** *A* is a Southern blot of *Eco*RI-digested genomic DNA from a variety of rodent species probed with the *Bam*HI fragment from pHE112 containing most of *βh2*, part of *βh1*, and the intergene region. Fragment sizes, in kilobases, were determined by comparison to fragments of known size in an adjacent lane. Mapping of the globin clusters of the mice in *lanes a* and *b* (2, 16) shows that the 2.1-kilobase band contains *βh2*, whereas the 3.4- and 0.65-kilobase bands contain *βh1*. The DNA came from the following rodents: *a*, *M. domesticus* BALB/cJ; *b*, *M. domesticus* C57BL/10J; *c*, *M. castaneus*; *d*, *M. spretus*; *e*, *M. cervicolor popaeus*; *f*, *M. pahari*; *g*, *P. maniculatus*. These species are of increasing phyletic and genetic distance from the source of the probe (BALB/c). (41, 42). *B* is a Southern blot of the same DNAs probed with the *Bam*HI fragment from pHE115. This probe contains the 3′ end of *βh2* and a repetitive element in the 3′ flanking region.

to functional β-globin genes. *βh2* was suspected to be an evolutionary remnant of a once active β-globin gene because of its extensive divergence and lack of correspondence to any known mouse β-globin protein. However, the general structural integrity and intactness of control signals stood in contrast to the overall high degree of divergence.

βh2 has a frame shift near the end of coding block 1 for which a compensatory alternate splicing scheme can be devised. Expression of *βh2* was not found *in vitro*,[3] or in induced murine erythroleukemia cells (13), or in mouse embryos.[2] This lack of transcription might be explained by the change of the usual promoter sequence CCAAT to GTAAC in *βh2*. Because other mouse β-globin genes were unexpectedly found to be productive (13), we sought more definitive evidence that *βh2* is a pseudogene.

The evolutionary relationship of *βh2* to the other β-globins suggests an explanation for the apparent paradox of high sequence divergence without major structural aberration. *βh2* last shared a common ancestral sequence with modern functional β-globins at about the time rodents and primates split. This long independent history provided adequate time for *βh2* to acquire much of its sequence divergence through evolution under selection. Another mouse pseudogene, *βh3*, has a sequence that shares only 55% homology to present day mouse adult genes.[4] This divergence occurred over the same time interval during which *βh2* was reduced to 72% homology. So *βh2* has not evolved at the maximum rate possible and therefore must have been constrained by selection over some of its history.

Analysis of the codons that are most highly conserved during β-globin evolution showed an excess of mutations in *βh2* that could be attributed to a period of divergence without selection. Structural studies of the hemoglobin molecule have identified the amino acids in β-globin that contact the heme cofactor or the α chain (44). The nucleotides that specify the identity of those amino acids change very slowly in functional β-globins. Therefore, the difference in divergence rates between an active and inactive gene at these codons is considerably larger than for other codons. Also, the error in estimating the background of changes acquired before inactivation is much smaller. The divergence within the codons for contact residues in *βh2* is indicative of loss of selection for β-globin function sometime in the latter part of its history.

As described under "Results," the overall divergence of *βh2*, the minimal damage to the control sequences, and the changes to the most conserved codons are all consistent with a long history of divergence under selection followed by a subsequent period of divergence without selection. This two-part history is consistent with the observation that *βh2* is intermediate between productive genes and most pseudogenes in the ratio of base substitutions that change amino acids to those that are silent (26). Formally, the evidence shows the loss of function as a β-globin and does not exclude models where *βh2* has been recruited to some atypical role. However, the simplest explanation is that *βh2* became a pseudogene after a long history as a working β-globin.

Pseudogenes contain information about the state of gene clusters in the past. *βh2* shows that, historically, active genes have been distributed in the β-globin cluster much differently than they are today. Expression must have occurred in the early mouse from the site now occupied by *βh2*. Unfortunately, it is hard to determine if this gene arose before or after the mammalian radiation. The former possibility would point to a previously unknown gene in the ancestor to mammals. In that case, the homology between δ and *βh2* in IVS 2 would suggest a separate ancestor for *βh2* and human δ from that for mouse $\beta 1^{dmajor}$, $\beta 2^{dminor}$, and human β. Similarly, analysis of the other mouse pseudogene, *βh3*, is being used to uncover relationships between the organization of ancestral and modern mammalian β-globin clusters (45).[4]

We originally discovered *βh2* in the process of sequencing another gene, *βh1*. Our experience with *βh2* suggests that other divergent structures may exist that have not been detected by hybridization techniques.

Finally, comparison of *βh2* to other β-globins revealed a conserved sequence, AGCCA-4*n*-GTGT, located approximately 200 nucleotides upstream from the initiation codons. This site is 5′ to the CCAAT box and overlaps with conserved sequences extensively discussed by Hardison (38). We also noted a similar sequence overlapping the CCAAT box of α-globins. Other control signals in *βh2* stand out as well conserved homologies on a background of generally poor homology. This property is also displayed by the AGCCA-4*n*-GTGT suggesting that it too was highly conserved during the productive period of *βh2*s history.

REFERENCES

1. Russell, E. S., and McFarland, E. C. (1974) *Ann. N. Y. Acad. Sci.* **241,** 25–38
2. 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
3. Leder, P., Hansen, J. N., Konkel, D., Leder, A., Nishioka, Y., and Talkington, C. (1980) *Science (Wash. D. C.)* **209,** 1336–1342
4. Proudfoot, N. (1980) *Nature (Lond.)* **286,** 840–841
5. Jacq, C., Miller, J. R., and Brownlee, G. G. (1977) *Cell* **12,** 109–120
6. Fritsch, E. F., Lawn, R. M., and Maniatis, T. (1980) *Cell* **19,** 959–972
7. Lacy, E., and Maniatis, T. (1980) *Cell* **21,** 545–553
8. Cleary, M. L., Schon, E. A., and Lingrel, J. B. (1981) *Cell* **26,** 181–190

9. Nishioka, Y., Leder, A., and Leder, P. (1980) *Proc. Natl. Acad. Sci. U. S. A.* **77,** 2806–2809
10. Vanin, E. F., Goldberg, G. I., Tucker, P. W., and Smithies, O. (1980) *Nature (Lond.)* **286,** 222–226
11. Proudfoot, N. J., and Maniatis, T. (1980) *Cell* **21,** 537–544
12. Hansen, J. N., Konkel, D. A., and Leder, P. (1982) *J. Biol. Chem.* **257,** 1048–1052
13. 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
14. Miyata, T., and Hayashida, H. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78,** 5739–5743
15. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning—A Laboratory Manual,* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
16. Weaver, S., Comer, M. B., Jahn, C. L., Hutchison, C. A., III, and Edgell, M. H. (1981) *Cell* **24,** 403–411
17. Maxam, A. M., and Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560
18. Sanger, F., and Coulson, A. R. (1978) *FEBS Lett.* **87,** 107–110
19. Maniatis, T., Jeffrey, A., and Kleid, D. G. (1975) *Proc. Natl. Acad. Sci. U. S. A.* **72,** 1184–1188
20. Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517
21. Wahl, G. M., Stern, M., and Stark, G. R. (1979) *Proc. Natl. Acad. Sci. U. S. A.* **76,** 3683–3687
22. 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, Georgetown University, Washington D. C.
23. Lautenberger, J. A., White, T. C., Haigwood, N. L., Edgell, M. H., and Hutchison, C. A., III (1980) *Gene* **9,** 213–231
24. White, C. T., Hardies, S. C., Hutchison, C. A., III, and Edgell, M. H. (1984) *Nucleic Acids Res.* **12,** 751–766
25. Fitch, W. M. (1977) *Am. Nat.* **111,** 223–257
26. Czelusniak, J., Goodman, M., Hewett-Emmett, D., Weiss, M. L., Venta, P. J., and Tashian, R. E. (1982) *Nature (Lond.)* **298,** 297–300
27. Fitch, W. M. (1971) *Syst. Zool.* **20,** 406–416
28. Miyata, T., and Yasunaga, T. (1980) *J. Mol. Evol.* **16,** 23–36
29. Benoist, C., O'Hare, K., Breathnach, R., and Chambon, P. (1980) *Nucleic Acids Res.* **8,** 127–142
30. Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L., and Steitz, J. A. (1980) *Nature (Lond.)* **283,** 220–224
31. Mount, S. M. (1982) *Nucleic Acids Res.* **10,** 459–472
32. Wieringa, B., Meyer, F., Reiser, J., and Weissmann, C. (1983) *Nature (Lond.)* **301,** 38–43
33. Lesk, A. M., and Chothia, C. (1980) *J. Mol. Biol.* **136,** 225–270
34. Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C., and Proudfoot, N. J. (1980) *Cell* **21,** 653–668
35. Proudfoot, N. J., and Brownlee, G. G. (1976) *Nature (Lond.)* **263,** 211–214
36. Baralle, F. E., and Brownlee, G. G. (1978) *Nature (Lond.)* **274,** 84–87
37. Goldberg, M. (1979) Ph.D. dissertation, Stanford University
38. Hardison, R. C. (1983) *J. Biol. Chem.* **258,** 8739–8744
39. Liebhaber, S. A., Goossens, M. J., and Kan, Y. W. (1980) *Proc. Natl. Acad. Sci. U. S. A.* **77,** 7054–7058
40. Schon, E. A., Wernke, S. M., and Lingrel, J. B. (1982) *J. Biol. Chem.* **257,** 6825–6835
41. Marshall, J. T. (1981) in *The Mouse in Biomedical Research* (Foster, H. L., Small, J. D., and Fox, J. G., eds) Vol. 1, pp. 17–26, Academic Press, New York
42. Sage, R. D. (1981) in *The Mouse in Biomedical Research* (Foster, H. L., Small, J. D., and Fox, J. G., eds) Vol. 1, pp. 39–90, Academic Press, New York
43. Fanning, T. G. (1982) *Nucleic Acids Res.* **10,** 5003–5013
44. Eaton, W. A. (1980) *Nature (Lond.)* **284,** 183–185
45. Hardies, S. C., Edgell, M. H., and Hutchison, C. A., III (1984) *J. Biol. Chem.* **259,** 3748–3756
46. Konkel, D. A., Tilghman, S. M., and Leder, P. (1978) *Cell* **15,** 1125–1132
47. Schon, E. A., Cleary, M. L., Haynes, J. R., and Lingrel, J. B. (1981) *Cell* **27,** 359–369
48. Nishioka, Y., and Leder, P. (1979) *Cell* **18,** 875–882
49. Baralle, F. E., Shoulders, C. C., and Proudfoot, N. J. (1980) *Cell* **21,** 621–626
50. Slightom, J. L., Blechl, A. E., and Smithies, O. (1980) *Cell* **21,** 627–638
51. Konkel, D. A., Maizel, J. V., Jr., and Leder, P. (1979) *Cell* **18,** 865–873
52. Spritz, R. A., DeRiel, J. K., Forget, B. G., and Weissman, S. M. (1980) *Cell* **21,** 639–646
53. Lawn, R. M., Efstratiadis, A., O'Connell, C., and Maniatis, T. (1980) *Cell* **21,** 647–651
54. Hardison, R. C., Butler, E. T., III, Lacy, E., Maniatis, T., Rosenthal, N., and Efstratiadis, A. (1979) *Cell* **18,** 1285–1297