Polymorphism and Gene Conversion in Mouse α-Globin Haplotypes

Mark A. Erhart,[1] Kenneth Piller and Steven Weaver

Laboratory for Cell, Molecular and Developmental Biology, Department of Biological Sciences, University of Illinois, Chicago, Illinois 60680

Manuscript received August 4, 1986
Accepted November 28, 1986

ABSTRACT

We have cloned and characterized three distinct α-globin haplotypes obtained from inbred strains of the mouse, *Mus domesticus.* We report here the complete nucleotide sequence of the six α-globin genes that the haplotypes contain. Our analysis of these genes and those from one other previously described haplotype indicates that recurrent gene conversion events have played a major role in their history. The pattern of nucleotide substitutions suggests that conversions have occurred both within and between haplotypes. Limited segments of coding and noncoding DNA have been involved in these gene conversion events. In two of the haplotypes, the nonallelic genes of each maintain DNA sequence identity over discrete intervals and encode the same α-globin polypeptide. On the other hand, the coding regions of some genes have accumulated replacement changes that result in distinct α-globins. In one instance, these changes appear to reflect positive selection of advantageous mutations.

THE locus encoding α-hemoglobin production in members of the genus *Mus* features a large number of allelic alternatives. To date, 14 haplotypes have been identified in both wild populations and inbred strains of mice by means of immobilized gradient isoelectric focusing (WHITNEY *et al.* 1979, 1985). This sensitive technique revealed the conservative variations exhibited by the allelic complexes. These involve the number of distinct α-globin chains synthesized by a haplotype (one or two), the amino acid sequences of the globin chains produced (five different α-globin chains are known), and the relative proportions of α-chains synthesized. Tables 1 and 2 summarize this information for each of the eight haplotypes described in inbred strains. This biochemical and genetic evidence is intriguing in that, while a wide variety of haplotypes are present in mouse populations, their distinguishing features (*i.e.*, neutral amino acid changes at four positions, and levels of α-chain expression) are quite subtle and may be due to the presence of a select few alterations at the DNA level.

The existence of a variety of α-globin haplotypes offers an opportunity to examine the molecular events that led to this diversity in the mouse α-globin gene complex. While it has been shown that tandemly duplicated genes are responsible for the presence of two distinct α-chains in BALB/c mice (NISHIOKA and LEDER 1979; LEDER *et al.* 1981; A. LEDER, personal communication), the number of functional genes in other haplotypes is unclear. Since genomic blotting data indicate that the remaining haplotypes each possess two linked α-globin sequences (WHITNEY *et al.* 1981), the nature of the "single" haplotypes (*Hba*^{a}, *Hba*^{e}, *Hba*^{f}) can be accounted for by either of two mechanisms: gene silencing, where one gene is no longer expressed or makes a nonfunctional product, or gene conversion, whereby both genes are evolving concommitantly, each encoding the same α-globin chain. These alternatives can only be distinguished by examining these haplotypes at the DNA level via sequencing of genomic clones. Likewise, the basis for the variable amounts of globin chain expression in the two-chain haplotypes may be revealed by DNA sequencing.

From an evolutionary perspective, analysis of the contemporary variants of a gene cluster such as the mouse α-globins is an excellent way of uncovering the processes that affect the structure of a gene family. For example, recent studies by both HILL *et al.* (1985) and POWERS and SMITHIES (1986) have examined a number of alternative haplotypes in human embryonic α-globin and fetal β-globin genes, respectively. These analyses support the notion that both inter- and intrachromosomal gene conversion events are very frequent in globin gene families. In contrast to these human globin families, in which significant variation at the DNA level has resulted in little or no variation in expression or polypeptide structure, the mouse α-globin haplotypes show significant variability in both the level of gene expression and in the structure of the gene products. To ascertain whether or not selective forces played a role in the fixation of those changes in DNA structure that helped shape this variation requires an extensive DNA sequence comparison of these haplotypes.

[1] Present address: The Jackson Laboratory, Bar Harbor, Maine 04609.

The sequence data presented in this article have been submitted to the EMBL/GenBank Data Libraries under the accession number Y00306.

TABLE 1

Mouse α-globin haplotypes

Haplotype	Representative strain	α-Chains synthesized
Hba^{a}	C57BL/10	1
Hba^{b}	BALB/c	2,3 (50:50)
Hba^{c}	SWR/J	4,1 (70:30)
Hba^{d}	SM/J	2,1 (60:40)
Hba^{e}	NB (extinct)	4
Hba^{f}	CE/J	5
Hba^{g}	A/J	1,5 (50:50)
Hba^{h}	P/J	4,5 (?)

Haplotypes and α-chains listed are only those found in inbred strains of *M. domesticus.* Data are taken from WHITNEY *et al.* (1985).

TABLE 2

Amino acid polymorphisms

	Codon:			
α-Chain	25	62	68	78
1	Gly	Val	Asn	Gly
2	Gly	Val	Ser	Gly
3	Gly	Val	Thr	Gly
4	Val	Ile	Ser	Gly
5	Gly	Val	Asn	Ala

α-Chains listed are only those found in inbred strains of *M. domesticus.* Data are taken from HILSE and POPP (1968) and POPP *et al.* (1982).

In order to answer some of the questions raised concerning the polymorphism in the mouse α-globin genes, we have begun a systematic analysis of these haplotypes from DNA libraries of inbred strains of *Mus domesticus.* We report here the cloning of three haplotypes—*Hba*^{a} (from C57BL/10), *Hba*^{c} (from SWR/J), and *Hba*^{f} (from CE/J)—and the complete DNA sequence of the six α-globin genes and flanking regions that these haplotypes contain. Our analysis of these sequence data, together with that from a fourth haplotype (*Hba*^{b}, from BALB/c; A. LEDER, personal communication) indicates that gene conversion has operated often and over discrete segments of the α-globin genes and their flanking regions. The degree to which this process has taken place differs among the four haplotypes, with some alleles at either of the two loci (*a1* and *a2*) being slightly more divergent. In one instance, this divergence may be due to the positive selection of advantageous mutations in the coding region. Surprisingly, most of the polymorphic sites present among these genes are not random, but are clustered in three regions: the second coding block, the 5′ flanking region, and the 3′ flanking region. In contrast, the noncoding portions of the transcribed region are virtually substitution-free.

MATERIALS AND METHODS

Materials: All restriction enzymes were purchased from Boehringer Mannheim, as were *Escherichia coli* DNA polymerase I holoenzyme, Klenow fragment, and DNA sequencing primers. Dideoxy nucleoside triphosphates were obtained from P-L Biochemicals. ^{32}P-Labeled nucleoside triphosphates were obtained from New England Nuclear. Nitrocellulose filters were purchased from Schleicher & Schuell.

Construction and screening of genomic libraries: Weanling male mice of the strains C57BL/10, CE/J, and SWR/J were obtained from the Jackson Laboratory, Bar Harbor, Maine. Genomic DNA was extracted from the entire carcass, after removal of skin and intestines, by the method of BLIN and STAFFORD (1976). Purified DNA was partially digested with *Mbo*I and size-selected on sucrose gradients by conventional techniques (MANIATIS, FRITSCH and SAMBROOK 1982). Fragments 16–20 kb in length were ligated into the *Bam*HI site of the λ-derived vector L47.1 (KARN *et al.* 1980) and packaged *in vitro* (HOHN and MURRAY 1977).

A total of 5 × 10^{5} plaque-forming units (pfu) per library were screened by plating the phage on ten 150-mm Petri plates. The plaques were transferred in duplicate to nitrocellulose filters (BENTON and DAVIS 1977) and probed with a nick-translated α-globin cDNA probe (1 × 10^{8} cpm/μg) derived from the plasmid pCR1αM10 (ROUGEON and MACH 1977). Hybridizations were carried out at 42° in the presence of 50% formamide, 6 × SSC, 10% dextran sulfate, 1 × Denhardts as described by WAHL, STERN and STARK (1979). After hybridization, the filters were washed twice at room temperature in 2 × SSC, 0.1% SDS and then five times at 55° in 0.1 × SSC, 0.1% SDS.

Characterization and sequencing of α-globin clones: Putative α-globin clones obtained from each of the three libraries (C57BL/10, SWR/J, and CE/J) were mapped using the restriction enzymes *Eco*RI, *Bam*HI, and *Hin*dIII (Figure 1). The α-globin genes were localized by hybridization of certain restriction fragments to the cDNA probe, and by alignment with the mapped BALB/c chromosome. Each of the genes was subcloned into M13mp18 or M13mp19 by taking advantage of conserved restriction sites within and flanking the *a1* and *a2* alleles. The *a1* alleles each lie within a conserved 3.1-kb *Sac*I fragment (data not shown), and the *a2* alleles are each present on a 3.4-kb *Hin*dIII fragment. By using the two *Bam*HI sites which are conserved in all mouse α-globin genes, each gene was subdivided into three regions and subcloned as three fragments.

Sequencing of the α-globin genes was carried out by the dideoxy chain termination method as described by SANGER *et al.* (1980). The strategy employed in sequencing the *a1* and *a2* genes is depicted in Figure 2. Using the subclones described above, we were able to sequence each of the six genes over a region extending from 255 bp upstream of the mRNA capping site through the termination codon. In order to sequence the 3′ nontranslated and flanking regions, it was necessary to make internal deletions of the 3′-containing subclones. We were then able to sequence over 800 bp from each of the 3′ subclones.

Because of the high level of sequence similarity among all of the genes, we obtained double-stranded sequence information for each region from only one gene—termed the reference sequence. Sequencing ladders of all the other genes were run on a gel next to the reactions of the corresponding strand from the reference sequence. Whenever a sequence differed, however, that particular region was sequenced on both strands.

Genetic nomenclature: The nomenclature used to describe the α-globin genes discussed in this paper follows the recommendations put forth at the Mouse Globin Nomenclature meeting (Jackson Laboratory, Bar Harbor, Maine, May

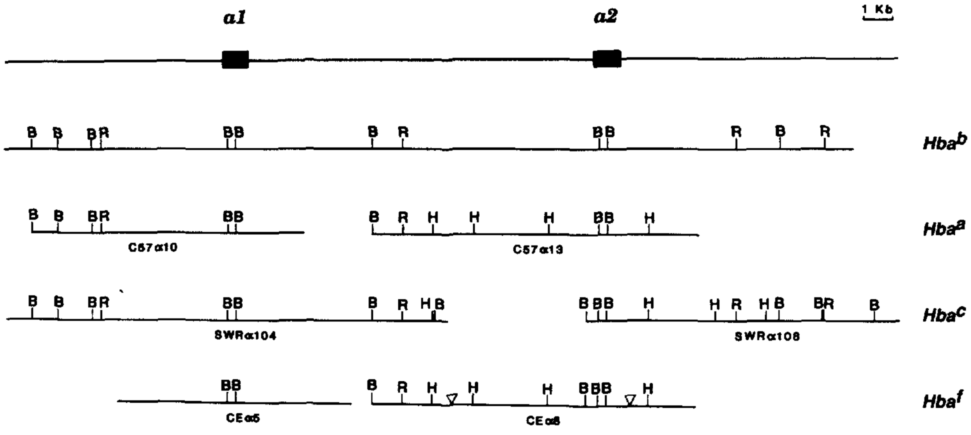


FIGURE 1. Restriction maps of cloned segments of the *Hba* haplotypes. The location of the *a1* and *a2* genes are shown as filled boxes on the line drawing at top. The data for Hba^b are taken from LEDER *et al.* (1981). The *Bam*HI (B), *Eco*RI (R) and *Hin*dIII (H) sites are indicated on the cloned segments from each of the three haplotypes described here. Clone designations are listed underneath each segment. The two inverted triangles above CEα8 denote insertions relative to the other haplotypes (see text). The direction of transcription is from left to right and the scale is shown at top right.

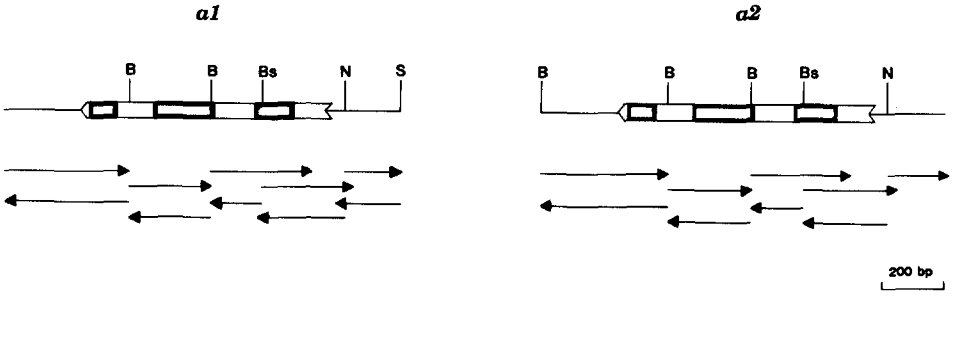


FIGURE 2. Sequencing strategy for the *a1* and *a2* genes. The *a1* (left) and *a2* (right) genes are shown divided into coding (thick-sided boxes), intervening (regular boxes), nontranslated (irregular areas), and flanking (horizontal lines) regions. Restriction sites used to generate M13 clones are as indicated: B (*Bam*HI), Bs (*Bst*EII), N (*Nco*I), and S (*Sst*I). Arrows below each gene detail the extent of individual sequencing runs over each segment. The direction of transcription is from left to right and the scale is shown at bottom right.

21–24, 1984; *Mouse News Letter,* February 1985, pp. 23–26). It is comprised of three components: the genetic locus designation (here *Hba*), a gene descriptor (*a1* or *a2*), and a superscript denoting the particular haplotype to which the gene belongs (here, either *a*, *b*, *c* or *f*). For example, the *a1* allele in the Hba^a haplotype is referred to as $Hba\text{-}a1^a$. For the sake of brevity, we will omit the locus designation (*e.g.*, $Hba\text{-}a1^a$ will be referred to as $a1^a$).

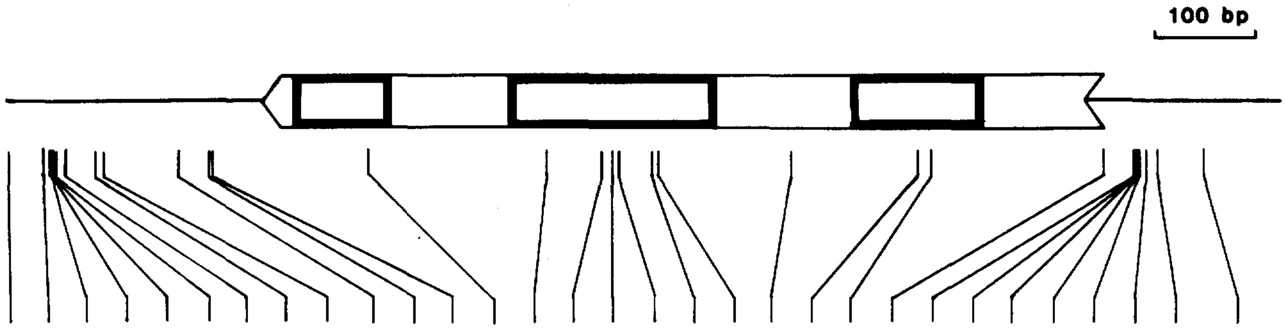
RESULTS AND DISCUSSION

Gross structure of α-globin haplotypes: Restriction mapping of genomic clones containing the α-globin gene sequences shows that each of the three *Hba* haplotypes examined here has a gross structure similar to that present in BALB/c (Hba^b) mice. As depicted in Figure 1, all four haplotypes consist of two genes that are apparently separated by a distance of 13 kb, assuming congruence with the physical map of BALB/c. There are relatively few polymorphisms with respect to the distribution of *Eco*RI, *Bam*HI, and *Hin*dIII sites along these chromosomes. Our mapping data of the CEα8 genomic clone reveals the presence of two small (150–200 bp) insertions, one on either side of the $a2^f$ gene. These are designated by inverted triangles in Figure 1. No additional length differences are apparent from our mapping data.

We determined the DNA sequence of the *a1* and *a2* genes from each of the Hba^a, Hba^c and Hba^f haplotypes for a length of about 1250 bp per gene. For the *a1* alleles, the sequenced region extends from a position about 250 bp upstream of the mRNA capping site to an *Sst*I site which lies 246 bp downstream of the poly(A) site. Similarly, the 5′ boundary of the sequence information obtained for the *a2* alleles lies near position −250. The 3′ boundary, however, extends about 346 bp downstream from the poly(A) site. The extent of the DNA sequence shown in Figure 3 and considered in our analysis lies between positions −256 and +1007. The sequence of the coding regions from each of the six genes allows the assignment of α-globin proteins to genes (see Figure 3). The coding regions from both genes of the Hba^a haplotype differ only at one silent position (codon 126); both genes code for the α_1-globin chain. Similarly, both genes in Hba^f mice are identical in the coding segment, each encoding alanine at position 78, characteristic of the α_5-globin polypeptide. Thus, each of these haplotypes possess two genes that specify identical proteins. In addition, the two genes from each haplotype are potentially expressible since there are no mutations in the known transcriptional or translational control signals which would interfere with proper globin chain synthesis. In the Hba^c haplotype, the *a1* gene codes for the α_4-globin chain and the *a2* gene encodes the α_1-chain. Finally, the complete DNA sequences of both genes of the Hba^b haplotype are known (NI-

```
         -250┐     -240┐     -230┐     -220┐     -210┐      -200┐     -190┐     -180┐     -170┐     -160┐
[a1^a] ggatactaacttcttcccaaactgccatcactggagacgtagtaaggggtaag-aagtgtgtccgggcaactgataaggattccctgcacccaggggaag
          c                                  a                 at   ca            t                   t     g
          └─[a2^c,a2^f]       [a1^b,a2^c,a2^f]─┘      [a1^b]─┘ |    ||            └─[a1^f]     [a1^b]─┘     |
                                                               └────┴┴─[a2^c,a2^f]                [a2^c,a2^f]─┘

          -150┐     -140┐     -130┐     -120┐     -110┐     -100┐      -90┐      -80┐      -70┐      -60┐
[a1^a] cacaacccagccccagaatctcaggggccctaacaagttttactgggtagagcaagcacaaaccagccaatgag-aactgctccaagggcgtgtccaccc
                                                                                   t
                                                                                   └─[a1^b]

                                                                                                 1
             -50┐      -40┐      -30┐      -20┐      -10┐       +1┐     +10┐      +20┐      +30┐ IniValLeu
[a1^a] tgcct-ggaggacacgcccttggagggcatataagtgctacttgctgcaggtccaagacACTTCTGATTCTGACAGACTCAGGAAGAAACCATGGTGCTC
            t -
 [a1^f,a2^f]─┘ └─[a2^b]

           5              10             15             20             25             30 ┌─ IVS 1
       SerGlyGluAspLysSerAsnIleLysAlaAlaTrpGlyLysIleGlyGlyHisGlyAlaGluTyrGlyAlaGluAlaLeuGluAr          +140┐
[a1^a] TCTGGGGAAGACAAAAGCAACATCAAGGCTGCCTGGGGGAAGATTGGTGGCCATGGTGCTGAATATGGAGCTGAAGCCCTGGAAAGgtgagaacaggacc
                                                                      T
                                                                     Val
                                                                      └─[a1^c]

          +150┐     +160┐     +170┐     +180┐     +190┐     +200┐     +210┐     +220┐     +230┐     +240┐
[a1^a] ttgatctgtaaggatcacaggatccaatatggacctggcactcgctcagtgggcagcttctaactatgcttttctgtgacctcaacttctcttctctcct

       IVS 1 ─┐ 32       35             40             45             50             55             60
               gMetPheAlaSerPheProThrThrLysThrTyrPheProHisPheAspValSerHisGlySerAlaGlnValLysGlyHisGlyLysLys
[a1^a] tctcccagGATGTTTGCTAGCTTCCCCACCACCAAGACCTACTTCCCTCACTTTGATGTAAGCCACGGCTCTGCCCAGGTCAAGGGTCACGGCAAGAAG
                                                   T
                                                  Phe
                                                   └─[a1^b]

                65             70             75             80             85             90
       ValAlaAspAlaLeuAlaAsnAlaAlaGlyHisLeuAspAspLeuProGlyAlaLeuSerAlaLeuSerAspLeuHisAlaHisLysLeuArgValAsp
[a1^a] GTCGCCGATGCTCTGGCCAATGCTGCAGGCCACCTCGATGACCTGCCCGGTGCCCTGTCTGCTCTGAGCGACCTGCATGCCCACAAGCTGCGTGTGGAT
       A           G     G──┐                              C     T
       Ile         Ala   Ser┴──[a1^b,a1^c]                 Ala   Leu
[a1^c]─┘             |   C──┐                    [a1^f,a2^f]─┘   └─[a1^b,a1^c]
        [a1^b,a1^c]──┘   Thr┴──[a2^b]

       95      99 ┌─ IVS 2
       ProValAsnPheLys +460┐     +470┐     +480┐     +490┐     +500┐     +510┐     +520┐     +530┐     +540┐
[a1^a] CCCGTCAACTTCAAGgtatgcgctgggacctggcaggcggcatctgggacccctaggaagggcttgggggtcctcgtgcccaaggcagggaacatagtggtc
                                                                                                t
                                                                                                └─[a1^c]

                                                    IVS 2 ─┐100            105            110            115
          +550┐     +560┐     +570┐     +580┐                LeuLeuSerHisCysLeuLeuValThrLeuAlaSerHisHisProAlaAspPhe
[a1^a] ccaggaaggggagcagaggcatcagggtgtccactttgtctccgcagCTCCTGAGCCACTGCCTGCTGGTGACCTTGGCTAGCCACCACCCTGCCGATTTC

           120            125            130            135            140
       ThrProAlaValHisAlaSerLeuAspLysPheLeuAlaSerValSerThrValLeuThrSerLysTyrArgTer          +730┐     +740┐
[a1^a] ACCCCCGCGGTGCATGCCTCTCTGGACAAATTCCTTGCCTCTGTGAGCACCGTGCTGACCTCCAAGTACCGTTAAGCTGCCTTCTGCGGGGCTTGCCTT
                  A              T
                 Val            Asp
           [a1^b]─┘              └─[a2^a,a2^b]

          +750┐     +760┐     +770┐     +780┐     +790┐     +800┐     +810┐     +820┐     +830┐     +840┐
[a1^a] CTGGCCATGCCCTTCTTCTCTCCCTTGCACCTGTACCTCTTGGTCTTTGAATAAAGCCTGAGTAGGAAGAAGCCTGCatgcctggttctctgcgtctgca
                                                                                                    a
                                                                                                    └─[a1^f]

          +850┐     +860┐     +870┐     +880┐     +890┐      +900┐     +910┐     +920┐     +930┐     +940┐
[a1^a] aaggtgtcatgtttagtgtggggatgcctctagctcatttagccatggggc-agtaaagacaaggttcagagcaaaaagcataattggatgcctacacac
                                   c  tgt-              -          c
              [a2^a,a2^b,a2^f]─┘  ||└─[a1^b]       └─[a1^b]      └─[a1^f]
                        [a2^b]──┘|└──[a2^a,a2^f]
                        [a1^b]───┘

             +950┐     +960┐     +970┐     +980┐     +990┐    +1000┐
[a1^a] acac--atatgtcttctgagtctgggcagcagtccctcccaagccctccactgacagccatgtgtcttc
       ----ac
       ├┴┴┘└┴─[a1^b]
       └─[a2^b]
```

FIGURE 3.—Nucleotide sequence of the mouse α-globin gene and flanking regions. The DNA sequence of the *a1^a* gene is shown in its entirety, with only those nucleotides which differ in the other seven genes listed beneath the *a1^a* sequence. The numbering scheme is with respect to the mRNA capping site (+1) in flanking and intervening sequences; coding regions are numbered according to amino acid codons. Nucleotides which are part of the mRNA are listed in upper case; flanking and intervening regions are denoted by lower case letters. Polymorphic amino acids and nucleotides are indicated by boldface print.



	<---upstream										--->	<---codons									--->	<---downstream								--->
	252	218	204	203	200	199	189	166	159	84	54	52	25	43	62	65	68	78	80	IVS	121	126	834	866	869	870	871	872	882	893	935
$\underline{a1}^{a}$	A	G	-	A	T	G	A	C	A	-	-	G	G	C	G	T	A	G	C	G	G	C	C	T	C	T	C	T	A	-	$(AC)_5$
$\underline{a1}^{b}$	A	$\boxed{A}$	$\boxed{A}$	A	T	G	A	$\boxed{T}$	A	$\boxed{T}$	-	G	G	$\boxed{T}$	G	$\boxed{G}$	$\boxed{G}$	G	$\boxed{T}$	G	$\boxed{A}$	C	C	T	C	$\boxed{G}$	C	$\boxed{-}$	$\boxed{-}$	-	$\boxed{(AC)_6}$
$\underline{a1}^{c}$	A	G	-	A	T	G	A	C	A	-	-	G	$\boxed{T}$	C	$\boxed{A}$	$\boxed{G}$	$\boxed{G}$	G	$\boxed{T}$	$\boxed{T}$	G	C	C	T	C	T	C	T	A	-	$(AC)_5$
$\underline{a1}^{f}$	A	G	-	A	T	G	$\boxed{T}$	C	A	-	$\boxed{T}$	G	G	C	G	T	A	$\boxed{C}$	C	G	G	C	$\boxed{A}$	T	C	T	C	T	A	$\boxed{C}$	$(AC)_5$
$\underline{a2}^{a}$	A	G	-	A	T	G	A	C	A	-	-	G	G	C	G	T	A	G	C	G	G	$\boxed{T}$	C	$\boxed{C}$	C	T	$\boxed{T}$	T	A	-	$(AC)_5$
$\underline{a2}^{b}$	A	G	-	A	T	G	A	C	A	-	-	$\boxed{-}$	G	C	G	T	$\boxed{C}$	G	C	G	G	$\boxed{T}$	C	$\boxed{C}$	$\boxed{T}$	T	C	T	A	-	$\boxed{(AC)_3}$
$\underline{a2}^{c}$	$\boxed{C}$	$\boxed{A}$	-	$\boxed{T}$	$\boxed{C}$	$\boxed{A}$	A	C	$\boxed{G}$	-	-	G	G	C	G	T	A	G	C	G	G	C	C	T	C	T	C	T	A	-	$(AC)_5$
$\underline{a2}^{f}$	$\boxed{C}$	$\boxed{A}$	-	$\boxed{T}$	$\boxed{C}$	$\boxed{A}$	A	C	$\boxed{G}$	-	$\boxed{T}$	G	G	C	G	T	A	$\boxed{C}$	C	G	G	C	C	$\boxed{C}$	C	T	$\boxed{T}$	T	A	-	$(AC)_5$
													[R]	[S]	[R]	[S]	[R]	[R]	[S]		[S]	[S]									

FIGURE 4.—Polymorphic nucleotide positions in the mouse α-globin genes and flanking regions. The region being compared is the same as that listed in Figure 3. The line drawing at top is a schematic representation of the α-globin gene: flanking regions (horizontal lines), nontranslated regions (irregular areas), coding blocks (thick-sided boxes) and intervening sequences (regular boxes). The position of each polymorphic nucleotide is indicated by a connecting line beneath the diagram, with the corresponding base at each position listed for every gene; nonconsensus nucleotides are boxed. The numbers under "upstream" denote nucleotide distances 5′ to the cap site (+1), those under "downstream" indicate nucleotide distances 3′ to the cap site, and codon positions are numbered under "codons." The notation [R] or [S] refers to replacement and silent changes, respectively. "IVS" denotes the lone substitution in the second intervening sequence. Dashes denote gaps in the sequence alignment.

SHIOKA and LEDER 1979; A. LEDER, personal communication); *al*^{b} encodes the α_2-globin, and *a2*^{b} encodes the α_3-chain.

Polymorphisms in the α-globin sequences are nonrandomly distributed: A composite depiction of the polymorphic positions is presented in Figure 4. The average similarity among the eight α-globin sequences is about 98% over the 1263 bp being compared. However, the nature and distribution of those substitutions that do exist is novel. The polymorphisms are mostly found in three regions: a 93-bp upstream segment between −252 and −159, a 110-bp region in the second coding block, and a 100-bp downstream segment between +834 and +935. Within the upstream and downstream segments there exists a smaller cluster of substitutions (from −204 to −199 and from +866 to +872). There is nothing obvious about the DNA sequence environment in which these clusters are embedded to explain the locally high density of substitutions. Both sets of changes lie outside the putative transcription unit (defined as extending from the CAAT box to the poly(A) addition site), so any effect that these mutations may have on gene expression is unclear. Even more striking than this clustering of polymorphisms is the nearly absolute dearth of nucleotide changes in the noncoding portions of the transcribed region. The lone substitution in any of these segments is a G to T transversion in IVS 2 of the *al*^{c} sequence. One would expect to see some nucleotide changes in these areas since most mutations here would presumably have no significant effect on α-globin expression. Moreover, other pairs of tandemly duplicated mammalian globin genes typically show abrupt divergence either within or just beyond their 3′ nontranslated regions (ERHART, SIMONS and WEAVER 1985; LIEBHABER and BEGLEY 1983; MICHELSON and ORKIN 1980; PROUDFOOT, GIL and MANIATIS 1982; SCHON, WERNKE and LINGREL 1982). Although the significance of this pattern of abrupt divergence is unclear, it could indicate that the 3′ boundaries of these globin gene duplication units tend to be situated very near to the 3′ end of the genes. The fact that the duplicated mouse α-globin genes do not conform in this respect to the general mammalian pattern suggests that the mouse α-globin duplication unit extends further downstream than does any previously studied mammalian globin gene duplication unit.

In the transcribed region, the uneven distribution of polymorphisms is evident in two regards. First, the number of differences within the coding blocks (nine) exceeds the number within the introns (one). Second, the numbers of synonymous (five) and replacement (four) substitutions are similar. This phenomenon is most pronounced in the middle of coding block two, where six variant positions (three synonymous, three nonsynonymous) are clustered within a 110-bp region. We refer to this type of DNA polymorphism as a "short-term" evolutionary pattern in that it is very different from the trend that is usually observed when more distantly related homologous genes are compared. For example, in a study involving distantly related genes (PERLER *et al.* 1980), silent changes were shown to exceed replacements by sevenfold. However, several analyses of closely related genes (*i.e.*, within a species or a genus) reveal a pattern of "short-term" DNA sequence evolution which is similar to that exhibited by the mouse α-globin genes. Alleles of the β-globin locus of both mice (ERHART, SIMONS and WEAVER 1985) and rabbits (EFSTRATIATIS, KAFATOS and MANIATIS 1977; HARDISON *et al.* 1979; VAN OOYEN *et al.* 1977), and the constant regions of the immunoglobulin κ-chain genes in several species of rats (FRANK *et al.* 1984) are all examples of orthologous genes in which coding regions are diverging at a higher rate than noncoding regions, with replacement changes accumulating as fast as, or faster than, silent changes.

Gene conversion has occurred within and between haplotypes: The relatedness of the α-globin genes is diagrammed in Figure 4 by boxing the rarer alternative at each of the 31 varying sites. About half of the polymorphic positions (14 of 31) have variants that are shared by at least two sequences. It is these shared variants that are potentially informative concerning the historical relationship between the α-globin haplotypes, as they mark regions of genes with recent common ancestry. Private variants, on the other hand, signal regions that have not interacted with other members of the group since the substitution occurred. Note that there are no variant nucleotides that are characteristic of either the *a1* or *a2* locus. This suggests that either (1) no mutations were fixed separately in either *a1* or *a2* immediately following their duplication, yet before the contemporary haplotypes appeared, or (2) any such changes that did occur have been erased by subsequent gene conversion events. There are only two positions at which more than two alternatives are present: the second position of codon 68 (either A, G or C), and the length of the repeating dinucleotide AC at position +935. With regard to the latter, it is puzzling that only the *Hba*^{b} genes have an aberrant number of repeats. The *a1*^{b} sequence has six and the *a2*^{b} gene has only three. This fluctuation in repeat number is probably due to slipped mispairing during DNA replication (MOORE 1983).

The distribution of the shared substitutions suggests that gene correction has occurred both within and between haplotypes. The patterns of similarity, and hence apparent historical relatedness, exhibited by these eight genes are different for different segments. In other words, shared substitutions are clustered. A diagram of this pattern is illustrated in Figure 5, in which clustered shared variants are represented by hatching and shading. The genes in the *Hba*^{f} haplotype have apparently experienced a gene conversion event recently. In this haplotype, a region of identity exists from position −158 to position +833. Within this region there are two positions at which both sequences share a nucleotide substitution found in no other mouse α-globin sequence: the insertion of a T residue at position −54, and a replacement change in codon 78, which results in an alanine at position 78, diagnostic of the α5-globin chain (POPP *et al.* 1982). Whitney *et al.* (1979) showed that hemoglobin tetramers with chain 5 have a more basic isoelectric point than tetramers containing chain 1. The neutral substitution of alanine for glycine apparently alters the net charge of the hemoglobin molecule. That these polymorphisms are found only in the two *Hba*^{f} genes is strong evidence that these markers were transmitted by gene conversion. There also may have been a recent conversion involving the genes in the *Hba*^{a} haplotype. Although there is no positive evidence (*i.e.*, shared variants) to support this notion, both sequences are identical over a region extending from the 5′ border of our sequence data to codon 126. However, this segment of identity conforms perfectly to the consensus; this makes it difficult to distinguish between gene conversion and simple lack of divergence as a means of explaining a portion of consensus DNA sequence that is shared by two or more genes. So, the extensive regions of consensus sequence in a number of the mouse α-globin genes may be a result of a lack of divergence since these genes last shared a common ancestor, or they may be a product of a "network" of gene conversions involving consensus donor sequences.

In addition to the instance of a recent gene conversion event within *Hba*^{f}, the relatedness of portions of certain sequences from different haplotypes suggests that gene conversion has taken place between nonsister chromatids. The resultant shared substitutions are schematically represented by similar patterns of shading in Figure 5. The most conspicuous instance is the 5′ flanking regions of *a2*^{c} and *a2*^{f}. These genes share six polymorphisms between −252 and −159 found in no other sequence. This observation can be best explained by assuming that a gene conversion involving this region took place between the predecessors of *a2* and *a2*^{f} in a heterozygote. Also, the segment from codon 126 to position +871 features three positions where variants are shared between at least two of the *a2* sequences. The silent substitution in codon 126 is present in both *a2*^{a} and *a2*^{b}. These two alleles, along with *a2*^{f}, also share a C residue at position +866. Finally, both *a2*^{a} and *a2*^{f} have a T residue in common at position +871. Another example of a cluster of polymorphisms shared by two genes from different haplotypes is a 40 bp segment of coding

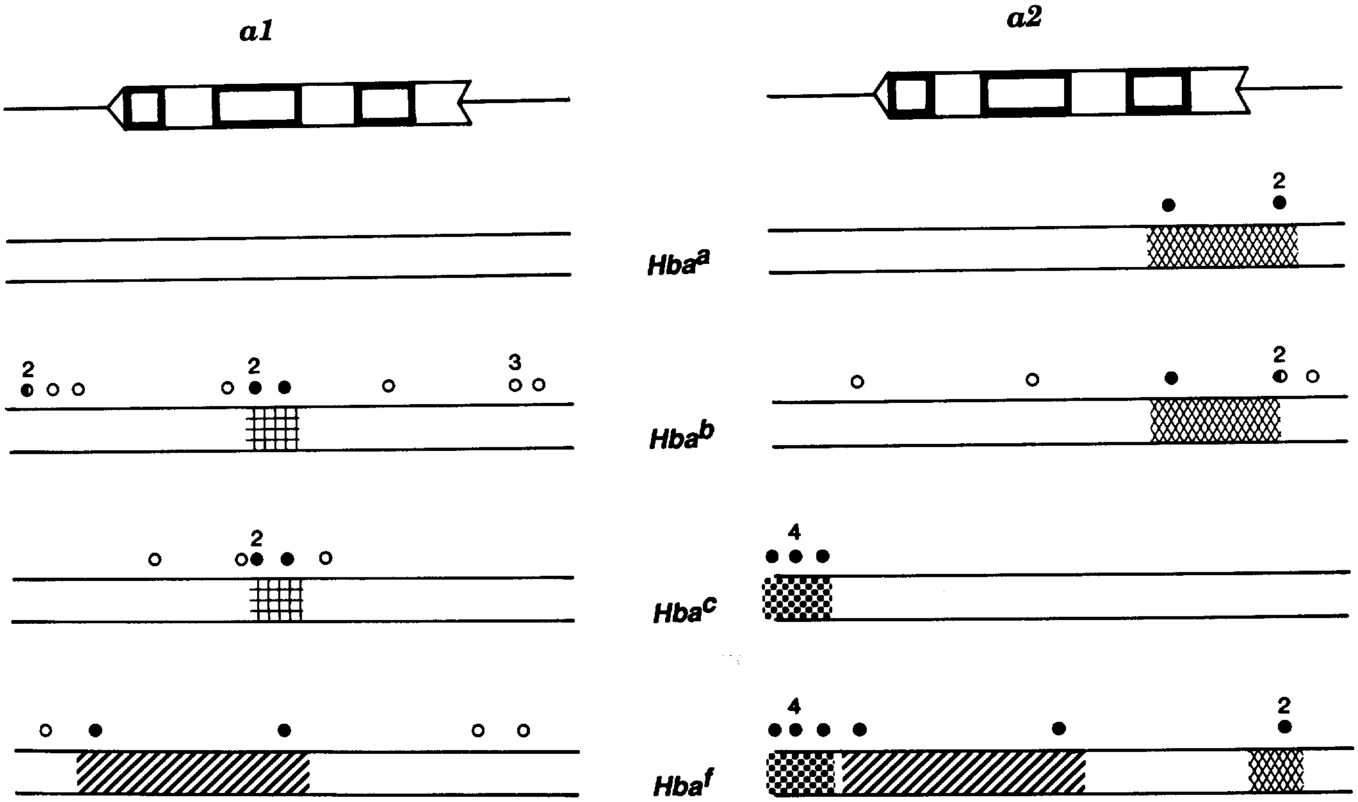


FIGURE 5.—Schematic representation of the relationship among the eight mouse α-globin sequences. At top is a line drawing depicting the structure of the *a1* and *a2* genes and flanking sequences. Beneath both genes are regions that are shaded in a manner that denotes the relatedness of the various sequences. The open area (for example, all of $a1^a$) designates consensus sequence. Nucleotide variants in each sequence are indicated by circles above the variable positions: The numbers above certain circles designate a cluster of nucleotide substitutions. Filled circles denote shared variants, and open circles signify private variants. The various shaded areas indicate segments of DNA that are alike due to gene conversions (see text).

block two in which $a1^b$ and $a1^c$ share three substitutions, one replacement (codon 68) and two synonymous (codons 65 and 80).

It is obvious from the data presented here that both intrachromosomal (within a haplotype) and interchromosomal (between haplotypes) gene conversion has taken place during the evolutionary history of the *Hba* complex in mice. Previously, the best evidence for intrachromosomal conversion was provided by LIEBHABER *et al.* (1984) in their analysis of a mutant α-globin haplotype in humans, while interchromosomal conversion events have been well illustrated in the human ζ-globin genes (HILL *et al.* 1985). Clearly, our results provide support for the notion that both types of conversion can help to shape the variability of a single gene family.

Note that the *a1* genes of the Hba^b and Hba^c haplotypes are significantly different from one another and from any other gene in the set. The $a1^b$ gene is clearly the most divergent of the eight sequences considered here, harboring eight private substitutions. In the $a1^c$ gene, the distribution of substitutions is curious in that the coding portion of this gene is the most divergent of any mouse α-globin gene, but elsewhere, $a1^c$ is identical to the consensus. The fact that $a1^c$ has six substitutions over the middle third of its sequence and absolutely no changes to either side of this region suggests that either (1) mutations have been fixed very rapidly in this middle region for some reason, or (2) the areas outside of the middle region have been corrected against a "consensus gene" (*e.g.*, $a1^a$) thereby erasing any changes that would have accumulated there.

The resultant "mosaic" nature of the $a1^c$ gene may be due to the action of selective forces. This gene is unique in that it encodes an α-globin with unusual amino acid residues at two highly invariant positions: 25 and 62. In almost all vertebrate globin chains examined, the amino acids at these positions are glycine and valine, respectively (DICKERSON and GEIS 1983). The glycine residue is located at an intrachain contact point, while the valine residue forms part of the hydrophobic pocket in which the heme moiety resides. The fact that both of these conserved positions have been substituted in the mouse α4 protein (valine at position 25, isoleucine at position 62), and that precisely the same pair of substitutions occurs in another mammalian globin [the α-chain of the gray kangaroo (GILMAN 1974)], suggests that these two changes are compensatory and that only their combination can be tolerated as well as, if not better than, the usual amino acid residues.

Evolutionary history of the mouse α-globin genes: To determine when the gene duplication which gave rise to the putative ancestral two-gene haplotype took place, one needs to analyze the nucleotide sequence divergence between the *a1* and *a2* loci. Because both loci apparently have been subjected to gene conversion over the entire length of DNA sequence reported here, the precise age of this gene duplication event cannot be determined based on this data. Although it has been hypothesized that there was one duplication event which led to the ancestor of the two adult gene arrangement present in both mammals and birds (ZIMMER *et al.* 1980), a recent analysis by HARDISON and GELINAS (1986) has shown that the relationships among the various mammalian adult α-globin genes are not obviously orthologous. Thus, there may have been independent α-globin gene duplications in the lineages leading to contemporary mammals. The presence of a primate-specific *Alu* element in the human α-globin duplication unit (HESS *et al.* 1983) is strong evidence that the duplication which led to the human genes occurred subsequent to the divergence of rodents and primates. In addition to the absence of *Alu*-like elements upstream of the mouse *a1* and *a2* genes (M. A. ERHART and S. WEAVER, unpublished results), there are two other features that distinguish the human and mouse α-globin gene duplication units. First, the distance separating the mouse genes is over three times that which separates the human genes. Of all the mammalian species from which the α-globin locus has been characterized, mice have the only intergenic distance exceeding 5 kb (HARDISON and GELINAS 1986). Second, while the 3′ boundary of the duplication unit in humans lies very near to the 3′ nontranslated region (HESS *et al.* 1983; LAUER, SHEN and MANIATIS 1980; MICHELSON and ORKIN 1983), the 3′ boundary in mice extends for at least another 800 bp beyond the 3′ nontranslated region (M. A. ERHART and S. WEAVER, unpublished results). At the gross level, it appears that other mammalian adult α-globin clusters are more similar to the human arrangement than to that of the mouse. Whether the α-genes in other mammals descended from the same duplication event as did the human genes will require further analysis. However, the evidence is compelling that an independent duplication or additional chromosomal rearrangement has occurred in the rodent lineage. Interestingly, a comparison of mammalian β-globin gene clusters revealed that mice have experienced a duplication of the adult β-globin gene, whereas other mammals possess only one adult β gene (HARDIES, EDGELL and HUTCHISON 1984). In contrast to the mouse β-globin gene duplication, which has been estimated at 65–80 MYA (ERHART, SIMONS and WEAVER 1985; HARDIES, EDGELL and HUTCHISON 1984), the mouse α-globin genes appear to have been diverging for a much shorter period, based on the DNA sequence divergence in the flanking regions of the nonallelic gene pairs. There are two explanations for this. One is that the α-globin gene duplication actually occurred much more recently in the rodent lineage than the β-globin gene duplication. An alternative explanation is that gene correction events are more frequent in *Mus* α-globin gene evolution, and/or encompass more extensive stretches of flanking DNA. Additional sequencing of the α-globin flanking regions is now in progress in order to identify the boundaries of the *Mus* α-globin duplication unit, if they have been preserved. The time of the duplication event can then be estimated directly based on the level of DNA sequence divergence.

We are in the process of cloning and sequencing the genes from the three remaining *Hba* haplotypes found in inbred mice (the strain that carried the eighth haplotype, *Hba^e*, apparently is extinct). The nature of these haplotypes will provide us with both a broader and more complete data base from which to make evolutionary inferences. In addition, we are extending our analysis to other species of the genus *Mus*, including some that have been separated from *M. domesticus* for 15 million years. This will not only allow us to continue our examination of, for example, gene conversion and positive selective forces in *Mus* α-globin gene evolution, it will enable us to establish more rigorously the relatedness of the many α-globin haplotypes within the *Mus* genus.

We thank Dr. AYA LEDER for kindly providing the unpublished sequence of the *a2^b* gene, BERNADETTE HOLDENER-KENNY for preparing mouse genomic DNA, SUSAN SHYMAN for critical reading of the manuscript, and Dr. BRIAN NICHOLS for many stimulating discussions.

This work was supported by a grant from the National Institutes of Health (GM 30140). M. A. E. is supported by a Laboratory for Cell, Molecular and Developmental Biology fellowship.

LITERATURE CITED

BENTON, W. D. and R. W. DAVIS, 1977 Screening λgt recombinant clones by hybridization to single plaques in situ. Science **196:** 180–182.

BLIN, N. and D. W. STAFFORD, 1976 Isolation of high molecular-weight DNA. Nucleic Acids Res. **3:** 2303–2308.

DICKERSON, R. E. and I. GEIS, 1983 *Hemoglobin: Structure, Function, Evolution, and Pathology.* Benjamin/Cummings, Menlo Park, California.

EFSTRATIADIS, A., F. C. KAFATOS and T. MANIATIS, 1977 The primary structure of rabbit β-globin mRNA as determined from cloned DNA. Cell **10:** 571–585.

ERHART, M. A., K. S. SIMONS and S. WEAVER, 1985 Evolution of the mouse β-globin genes: a recent gene conversion in the Hbb^s haplotype. Mol. Biol. Evol. **2:** 304–320.

FRANK, M. B., R. P. BESTA, P. R. BAVERSTOCK and G. A. GUTMAN, 1984 Kappa chain constant region gene sequences in genus *Rattus*: coding regions are diverging more rapidly than non-coding regions. Mol. Biol. Evol. **1:** 489–501.

GILMAN, J. G., 1974 Rodent hemoglobin structure: a comparison of several species of mice. Ann. N. Y. Acad. Sci. **241:** 416–433.

HARDIES, S. C., M. H. EDGELL and C. A. HUTCHISON III, 1984 Evolution of the mammalian β-globin gene cluster. J. Biol. Chem. **259:** 3748–3756.

HARDISON, R. C. and R. E. GELINAS, 1986 Assignment of orthologous relationships among mammalian α-globin genes by examining flanking regions reveals a rapid rate of evolution. Mol. Biol. Evol. **3:** 243–261.

HARDISON, R. C., E. T. BUTLER III, E. LACY, T. MANIATIS, N. ROSENTHAL and A. EFSTRATIADIS, 1979 The structure and transcription of four linked rabbit β-like globin genes. Cell **18:** 1285–1297.

HESS, J. F., M. FOX, C. SCHMID and C.-K. J. SHEN, 1983 Molecular evolution of the human adult α-globin-like gene region: insertion and deletion of Alu family repeats and non-Alu DNA sequences. Proc. Natl. Acad. Sci. USA **80:** 5970–5974.

HILL, A. V. S., R. D. NICHOLLS, S. L. THEIN and D. R. HIGGS, 1985 Recombination within the human embryonic ζ-globin locus: a common ζ-ζ chromosome produced by gene conversion of the $\psi\zeta$ gene. Cell **42:** 809–819.

HILSE, K. and R. A. POPP, 1968 Gene duplication as the basis for amino acid ambiguity in the alpha-chain polypeptides of mouse hemoglobins. Proc. Natl. Acad. Sci. USA **61:** 930–936.

HOHN, B. and K. MURRAY, 1977 Packaging recombinant DNA molecules into bacteriophage particles in vitro. Proc. Natl. Acad. Sci USA **74:** 3259–3264.

KARN, J., S. BRENNER, L. BARNETT and G. CESARENI, 1980 Novel bacteriophage λ cloning vector. Proc. Natl. Acad. Sci. USA **77:** 5172–5176.

LAUER, J., C.-K. J. SHEN and T. MANIATIS, 1980 The chromosomal arrangement of human α-like globin genes: sequence homology and α-globin gene deletions. Cell **20:** 119–130.

LEDER, A., D. SWAN, F. RUDDLE, P. D'EUSTACHIO and P. LEDER, 1981 Dispersion of α-like globin genes of the mouse to three different chromosomes. Nature **293:** 196–200.

LIEBHABER, S. A. and K. A. BEGLEY, 1983 Structural and evolutionary analysis of the two chimpanzee α-globin mRNAs. Nucleic Acids Res. **11:** 8915–8928.

LIEBHABER, S. A., E. F. RAPPAPORT, F. E. CASH, S. K. BALLAS, E. SCHWARTZ and S. SURREY, 1984 Hemoglobin I mutation encoded at both α-globin loci on the same chromosome: concerted evolution in the human genome. Science **226:** 1449–1451.

MANIATIS, T., E. F. FRITSCH and J. SAMBROOK (Editors), 1982 *Molecular Cloning. A Laboratory Manual.* Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

MICHELSON, A. M. and S. H. ORKIN, 1980 The 3′ untranslated regions of the duplicated human α-globin genes are unexpectedly divergent. Cell **22:** 371–377.

MICHELSON, A. M. and S. H. ORKIN, 1983 Boundaries of gene conversion within the duplicated human α-globin genes. J. Biol. Chem. **258:** 15245–15254.

MOORE, G. P., 1983 Slipped-mispairing and the evolution of introns. Trends Biochem. Sci. **8:** 411–414.

NISHIOKA, Y. and P. LEDER, 1979 The complete sequence of a chromosomal mouse α-globin gene reveals elements conserved throughout vertebrate evolution. Cell **18:** 875–882.

PERLER, F., A. EFSTRATIADIS, P. LOMEDICO, W. GILBERT, R. KOLODNER and J. DODGSON, 1980 The evolution of genes: the chicken preproinsulin gene. Cell **20:** 555–566.

POPP, R. A., E. G. BAILIFF, L. C. SKOW and J. B. WHITNEY, 1982 The primary structure of genetic variants of mouse hemoglobin. Biochem. Genet. **20:** 199–208.

POWERS, P. A. and O. SMITHIES, 1986 Short gene conversions in the human fetal globin gene region: a by-product of chromosome pairing during meiosis? Genetics **112:** 343–358.

PROUDFOOT, N. J., A. GIL and T. MANIATIS, 1982 The structure of the human zeta-globin gene and a closely linked, nearly identical pseudogene. Cell **31:** 553–563.

ROUGEON, F. and B. MACH, 1977 Cloning and amplification of α and β mouse globin gene sequences synthesised in vitro. Gene **1:** 229–239.

SANGER, F., A. R. COULSON, B. G. BARELL, A. J. H. SMITH and B. A. ROE, 1980 Cloning in single stranded bacteriophage as an aid to rapid sequencing. J. Mol. Biol. **143:** 161–178.

SCHON, E. A., S. M. WERNKE and J. B. LINGREL, 1982 Gene conversion of two functional goat α-globin genes preserves only minimal flanking sequences. J. Biol. Chem. **257:** 6825–6835.

VAN OOYEN, A., J. VAN DEN BERG, N. MANTEI and C. WEISSMAN, 1979 Comparison of total sequence of a cloned rabbit β-globin gene and its flanking regions with a homologous mouse sequence. Science **206:** 337–344.

WAHL, G. M., M. STERN and G. R. STARK, 1979 Efficient transfer of large DNA fragments from agarose gels to diazobenzloxymethyl-paper and rapid hybridization by using dextran sulfate. Proc. Natl. Acad. Sci. USA **76:** 3683–3688.

WHITNEY, J. B. III, G. T. COPLAND, L. C. SKOW and E. S. RUSSELL, 1979 Resolution of products of the duplicated hemoglobin α-chain loci by isoelectric focusing. Proc. Natl. Acad. Sci. USA **76:** 867–871.

WHITNEY, J. B. III, J. MARTINELL, R. A. POPP, L. B. RUSSELL and W. F. ANDERSON, 1981 Deletions in the α-globin gene complex in α-thalassemic mice. Proc. Natl. Acad. Sci. USA **78:** 7644–7647.

WHITNEY, J. B. III, R. R. COBB, R. A. POPP and T. W. O'ROURKE, 1985 Detection of neutral amino acid substitutions in proteins. Proc. Natl. Acad. Sci. USA **82:** 7646–7650.

ZIMMER, E. A., S. L. MARTIN, S. M. BEVERLY, Y. W. KAN and A. C. WILSON, 1980 Rapid duplication and loss of genes coding for the α chains of hemoglobin. Proc. Natl. Acad. Sci. USA **77:** 2158–2162.

Communicating editor: R. E. GANSCHOW