18. A. J. Jeffreys and R. A. Flavell, *ibid.* **12**, 1097 (1977).
19. A. Leder, H. I. Miller, D. H. Hamer, J. G. Seidman, B. Norman, M. Sullivan, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 6187 (1978).
20. S. M. Tilghman, D. C. Tiemeier, J. G. Seidman, B. M. Peterlin, M. Sullivan, J. Maizel, P. Leder, *ibid.*, p. 725.
21. J. Abelson, *Annu. Rev. Biochem.* **48**, 1035 (1979).
22. S. A. Liebhaber, M. Goosens, R. Poon, Y. W. Kan, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
23. Y. Nishioka and P. Leder, *Cell* **18**, 875 (1979).
24. S. M. Tilghman, P. J. Curtis, D. C. Tiemeier, P. Leder, C. Weissman, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 1309 (1978); A. J. Kinniburgh, J. E. Mertz, J. Ross, *Cell* **14**, 681 (1978); A. J. Kinniburgh and J. Ross, *ibid.* **17**, 915 (1979).
25. R. A. Flavell, R. Bernards, G. C. Grosveld, H. A. M. Hooijmakers-VanDommelen, J. M. Kooter, E. De Boer, in *Eukaryotic Gene Regulation*, R. Axel, T. Maniatis, C. F. Fox, Eds. (Academic Press, New York, 1980), pp. 335–354.
26. L. E. Maquat, A. J. Kinniburgh, L. R. Beach, G. R. Honig, J. Lazerson, W. B. Ershler, J. Ross, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 4287 (1980).
27. J. A. Kantor, P. H. Turner, A. W. Nienhuis, *Cell* **21**, 149 (1980).
28. R. F. Weaver and C. Weissman, *Nucleic Acids Res.* **7**, 1175 (1979).
29. R. M. Lawn, A. Efstratiadis, C. O'Connell, T. Maniatis, *Cell* **21**, 647 (1980); R. Spritz, J. DeRiel, B. Forget, S. Weissman, *ibid.*, p. 639; F. E. Baralle, C. Shoulders, N. J. Proudfoot, *ibid.*, p. 621.
30. J. Slightom, A. Blechl, O. Smithies, *ibid.*, p. 627.
31. C. Benoist, K. O'Hare, R. Breathnach, P. Chambon, *Nucleic Acids Res.* **8**, 127 (1980).
32. M. Goldberg, thesis, Stanford University (1979).
33. D. A. Konkel, J. V. Maizel, P. Leder, *Cell* **18**, 865 (1979).
34. A. van Ooyen, J. van Den Berg, N. Mantei, C. Weissman, *Science* **206**, 337 (1979).
35. P. J. Farabaugh and J. H. Miller, *J. Mol. Biol.* **126**, 847 (1978).
36. E. A. Zimmer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2158 (1980).
37. M. O. Dayhoff, *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, D.C., 1972).
38. L. Hood, J. H. Campbell, S. C. R. Elgin, *Annu. Rev. Genet.* **9**, 305 (1975).
39. S. Orkin, J. Old, H. Lazarus, C. Altay, A. Gurgey, D. Weatherall, D. Nathan, *Cell* **17**, 33 (1979); S. Embury, R. Lebo, A. Dozy, Y. W. Kan, *J. Clin. Invest.* **63**, 1307 (1979).
40. M. Goossens, A. Dozy, S. Embury, Z. Zachariades, M. Hadjiminas, G. Stamatoyannopoulos, Y. W. Kan, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 518 (1980); D. R. Higgs, J. M. Old, L. Pressley, J. B. Clegg, D. J. Weatherall, *Nature (London)* **284**, 632 (1980).
41. Y. Nishioka, A. Leder, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2806 (1980); E. F Vanin, G. I. Goldberg, P. W. Tucker, O. Smithies, *Nature (London)*, in press.
42. R. Breathnach, C. Benoist, K. O'Hare, F. Gannon, P. Chambon, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 4853 (1978); M. Lerner, J. Boyle, S. Mount, S. Wolin, J. Steitz, *Nature (London)* **283**, 220 (1980).
43. C.-K. J. Shen, unpublished observations.
44. W. G. Jelinek *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 1398 (1980); F. E. Baralle and N. J. Proudfoot, unpublished observations; O. Smithies, unpublished observations.
45. C. M. Houck, F. P. Rinehart, C. W. Schmid, *J. Mol. Biol.* **132**, 289 (1979).
46. C. Duncan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 5095 (1979).
47. E. F. Fritsch, C. K. J. Shen, R. M. Lawn, T. Maniatis, *Cold Spring Harbor Symp. Quant. Biol.*, in press.
48. W. G. Jelinek and L. Leinwood, *Cell* **15**, 205 (1978).
49. H. D. Robertson, E. Dickson, W. Jelinek, *J. Mol. Biol.* **115**, 571 (1977).
50. C.-K. J. Shen and T. Maniatis, *Cell* **19**, 379 (1980).
51. A. Pellicer *et al.*, *Science* **209**, 1414 (1980); N. Mantee, W. Boll, C. Weissmann, *Nature (London)* **281**, 40 (1979); M. Wigler *et al.*, *Cell* **16**, 777 (1979); B. Wold, M. Wigler, E. Lacy, T. Maniatis, S. Silverstein, B. Axel, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 5684 (1979).
52. D. H. Hamer and P. Leder, *Nature (London)* **281**, 35 (1979); H. Mulligan and P. Berg, *Science* **209**, 1422 (1980).
53. P. A. Weil, D. S. Luse, J. Segall, R. G. Roeder, *Cell* **18**, 469 (1979).
54. J. L. Manley, A. Fire, A. Cano, P. A. Sharp, M. L. Gefter, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 3855 (1980).
55. C. Parker, personal communication.
56. B. Waslyk, C. Kedinger, J. Corden, O. Brison, P. Chambon, *Nature (London)* **285**, 367 (1980).
57. D. S. Luse and R. G. Roeder, *Cell* **20**, 691 (1980).
58. F. E. Baralle, *ibid.* **12**, 1085 (1977).
59. J. M. Old, N. J. Proudfoot, W. G. Wood, J. I. Longley, J. B. Clegg, D. J. Weatherall, *ibid.* **14**, 289 (1978).
60. L. Burns and A. Bank, unpublished observations.
61. M. H. M. Shander, unpublished observations.
62. E. B. Ziff and R. M. Evans, *Cell* **15**, 1436 (1978).
63. J. L. Manley *et al.*, in preparation.
64. S. L. Hu and J. L. Manley, in preparation.
65. J. Corden *et al.*, *Science* **209**, 1406 (1980).
66. C. C. Baker, J. Herisse, G. Courtois, F. Galibert, E. B. Ziff, *Cell* **18**, 569 (1979).
67. R. Grosschedl and M. L. Birnsteil, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 1432 (1980).
68. M. Wickens, S. Woo, B. O'Malley, J. Gurdon, *Nature (London)* **285**, 628 (1980).
69. V. Parker, N. J. Proudfoot, M. H. M. Shander, T. Maniatis, unpublished observations.
70. R. Bernards, J. M. Kooter, R. A. Flavell, *Gene* **6**, 265 (1979); L. H. T. van der Ploeg, A. Konings, M. Oort, D. Roos, L. Bernini, R. Flavell, *Nature (London)* **283**, 637 (1980).
71. H. Weintraub and M. Groudine, *Science* **193**, 848 (1976).
72. J. Stalder, A. Larsen, J. D. Engel, M. Dolan, M. Groudine, H. Weintraub, *Cell* **20**, 451 (1980).
73. S. L. Berger and C. S. Birkenmeier, *Biochemistry* **18**, 5143 (1979).
74. J. M. Bailey and N. Davidson, *Anal. Biochem.* **70**, 75 (1976).
75. J. G. Sutcliffe, *Cold Spring Harbor Symp. Quant. Biol.* **43**, 77 (1979).
76. R. Laskey, in *Methods Enzymol.*, **65**, 363 (1980).
77. H. Klenow and K. Overgaard-Hansen, *FEBS Lett.* **6**, 25 (1970).
78. N. J. Proudfoot, *Cell* **10**, 559 (1977).
79. A. Maxam and W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560 (1977).
80. T. R. Rutherford, J. B. Clegg, D. J. Weatherall, *Nature (London)* **280**, 164 (1979).
81. G. G. Brownlee, *Determination of Sequences in RNA: Laboratory Techniques in Biochemistry and Molecular Biology*, T. S. Work and E. Work, Eds. (North-Holland/American Elsevier, Amsterdam, 1972).
82. The authors are grateful to C. O'Connell, who was involved in the preliminary stages of this work, and to G. Attardi for providing HeLa cells. Supported by NIH grants.

21 July 1980

Mouse Globin System: A Functional and Evolutionary Analysis

Philip Leder, J. Norman Hansen, David Konkel, Aya Leder, Yutaka Nishioka, Carol Talkington

The globin genes have a special place among the systems first examined by means of recombinant technology (*1*). This is so because more than 50 years of intense study had created an array of questions that could now be dealt with at the molecular genetic level and also because nature had conveniently arranged for the development of the red blood cell to occur in such a way as to make globin messenger RNA (mRNA) abundantly available. The combination of interesting genetic phenomena and the availability of probes (globin mRNA) made the globins an early and promising target for gene cloning. Proudfoot *et al.* (*2*) have reviewed the progress made in understanding human globin genes. We describe what has been learned from studying the mouse.

The Mouse Globin Gene System

Globin gene expression in the mouse begins during intrauterine development with the appearance of a primitive population of nucleated red blood cells in the embryonic yolk sac. In contrast to adult mouse erythrocytes, these cells produce three embryonic globins, one α-like (X) and two β-like (Y and Z), whose expression is limited to this special cell population and this specific period of development (*3*). Adult α-globin also appears in yolk-sac red cells, but continues to be produced in nonnucleated adult erythrocytes, where it is accompanied by the appearance of adult β-globin. The two adult globins are produced in relatively equivalent amounts from the third and final week of gestation throughout the lifetime of the organism (*4*).

Genetic and structural studies had indicated that the BALB/c mouse ex-

The authors are investigators at the Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20205.

0036-8075/80/0919-1336$01.50/0

Summary. Structural and functional analysis of the mouse α-globin and β-globin genes reveals that the globin genes are encoded in discontinuous bits of coding information and that each gene locus is much more complex than was originally supposed. Each seems to consist of an array of several authentic genes as well as several apparently inactive pseudogenes. Comparison of the sequences of some of these genes to one another indicates that chromosomal DNA is a dynamic structure. Flanking and intervening sequences change in two ways: quickly, by duplication and extensive insertions and deletions, and slowly, by point mutation. Active coding sequences are usually limited to the slower mode of evolution. In addition to identifying fast and slow modes of evolution, it has also been possible to test the function of several signals that surround these genes and to identify those that appear to play a role in gene expression.

pressed at least two α and two β genes and that the α and β genes were located on separate chromosomes (*4*). Furthermore, while some form of coordinating regulation controlled the stoichiometric appearance of α-globin and β-globin, another form of regulation controlled the relative expression of the two β-globin genes (major and minor), adjusting their normal 4:1 level of expression to a more nearly equivalent ratio immediately postpartum or in the presence of profound anemia (*5*). Even though genetic and biochemical studies had prepared us to expect reasonably complex loci and elaborate regulatory systems, nothing really suggested the true complexity of a globin gene locus in terms of its expressed genes and pseudogenes, and nothing even faintly suggested that these genes would turn out to be encoded in discontinuous bits of coding DNA.

Structure and Assembly of Coding Information

The initial picture of the globin gene as an interrupted structure (Fig. 1) (*1, 6–8*) immediately raised a host of questions, some of which were easily answered. The interrupted nature of the β major globin gene turned out to be a general feature of the other mouse globin genes (*7, 9–11*) (see Fig. 2). Moreover, the globin genes were divided by intervening sequences (noncoding regions) at the same relative positions, indicating that the ancestral globin gene was identically interrupted and that this organization had been preserved for over 500 million years of vertebrate evolution (*12–14*). Indeed, as the mouse globin gene family was cloned and the sequences of many of its members were determined, it became clear that neither extensive regions of primary structure nor length were necessarily conserved within these intervening sequences (*12–14*).

The step at which the discontinued bits of coding information are assembled was easily identified. It had long been known that β-globin mRNA was synthesized via a 1.5-kilobase (kb) precursor that was processed to the 0.6-kb mature globin mRNA (*15*). By annealing purified precursor mRNA to the cloned β major gene, an R-loop (RNA-DNA heteroduplex) structure could be visualized that indicated that the entire gene, including the intervening sequences, was transcribed into the precursor (*16*). The precursor, in turn, must be processed by splicing out the intervening sequences and covalently joining the coding segments to form the coherent, mature form of the mRNA (Fig. 3). Considerable evidence supporting this view has been obtained (*17*).

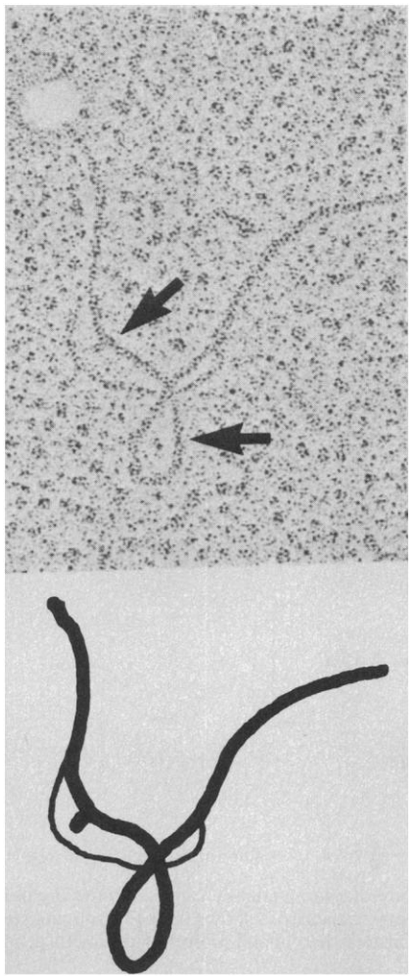


Fig. 1. Electron microscopic visualization of a hybrid formed between the mouse β major globin gene and its RNA. The arrows point to two intervening segments of DNA that are absent in the mRNA sequences. The molecule is represented diagrammatically below the figure, the thicker line is double-stranded DNA or RNA-DNA hybrid. The thin line represents the single-stranded DNA that has been displaced by the binding of globin mRNA to gene. The photograph is from Leder *et al.* (*6*).

The splicing reaction itself and the signals that specify it have not yet been fully characterized. In contrast to the enzyme involved in maturation of transfer RNA's (*18*), no soluble enzyme system has as yet been obtained that is capable of catalyzing the mRNA splicing reaction. The hierarchy of events that lead to splicing have, however, been roughly outlined. Capping (the addition of 7′-methyl guanosine to the 5′ end of the globin transcript) and 3′ polyadenylation seem to precede splicing, which is largely completed within the nucleus (*19*). A consensus sequence that might be involved in signaling the splicing reaction has been identified by comparison of intervening-sequence borders from many genes (*20*). The collected mouse globin sequences are shown in Table 1 and are displayed in relation to one such splicing rule (*21*). Most recently it has been suggested that sequences contained within ubiquitous species of low-molecular-weight nuclear RNA might serve to facilitate splicing because of complementarity to the intervening-sequence borders (*22*). This and other attractive ideas obviously await the development of suitable splicing systems for the application of critical experimental tests.

Function of Intervening Sequences

A question that has generated considerable attention is what roles, if any, intervening sequences might play in the evolution and in the regulated expression of genes. It has been suggested, for example, that intervening sequences might serve as sites for illegitimate recombination, thereby facilitating evolution by joining different functional coding domains (*23*). This idea fits well with the repeated domain structure of immunoglobulin heavy chain genes (*24*). On the other hand, we have suggested that evolutionary drift between intervening sequences in repeated genes (such as those of

Table 1. Comparison of the intervening-sequence (IVS) borders of mouse globin genes.

Gene	IVS	Sequence
β major	1	GCAG(GTTGG.....TTTAG)GCTG
β minor	1	GCAG(GTTGG.....TTTAG)GCTG
α_1	1	AAAG(GTGAG.....CCCAG)GATG
β major	2	CAGG(GTGAG.....CACAG)CTCC
β minor	2	CAGG(GTGAG.....CACAG)CTCC
α_1	2	CAAG(GTATG.....CGCAG)CTCC

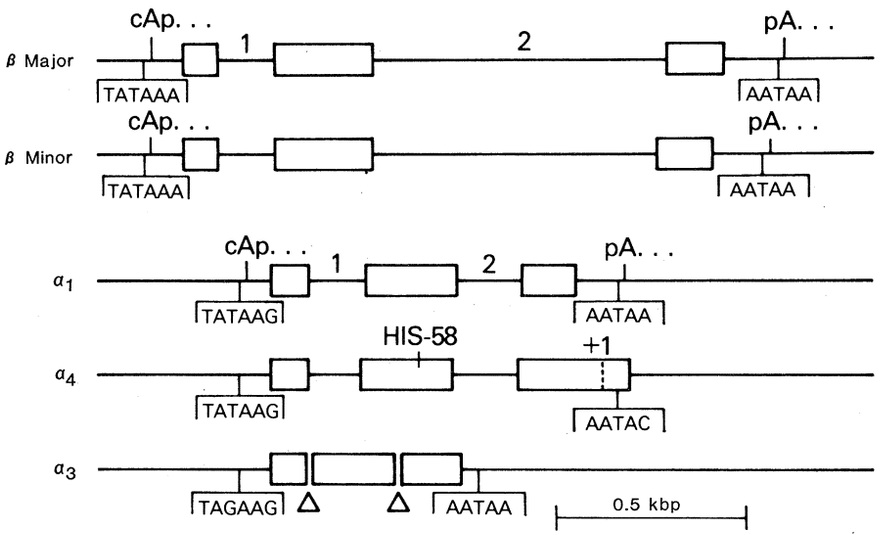


Fig. 2. Diagrammatic representation of the mouse β and α genes and pseudogenes. The boxes represent coding sequences; the thin lines, flanking or intervening sequences. The numbers within the intervening sequences of these structures identify the intervening sequences as the first or second (5′ to 3′); the letters *cAp . . .* identify the cap site; *pA . . .* identifies the poly(A) addition site. The conserved and possibly functionally significant 5′ and 3′ sequences are noted. The pseudogenes α_4 and α_3 are specially annotated. Position 58 is noted in α_4; it is normally a His, but in α_4 this position encodes Tyr, which would convert this gene into a pathologic methemoglobin sequence. The +1 noted at the termination site of α_4 indicates a +1 frame shift with respect to the normal termination codon that is indicated by a broken line. The next (read-through) terminator codon is indicated by the extended box. The fact that the intervening sequences are lost from the α_3 gene is indicated by the triangles below the sites from which they have disappeared. Note that the conserved promoter or Hogness box in α_3 contains a G in place of T in what is otherwise a highly conserved region.

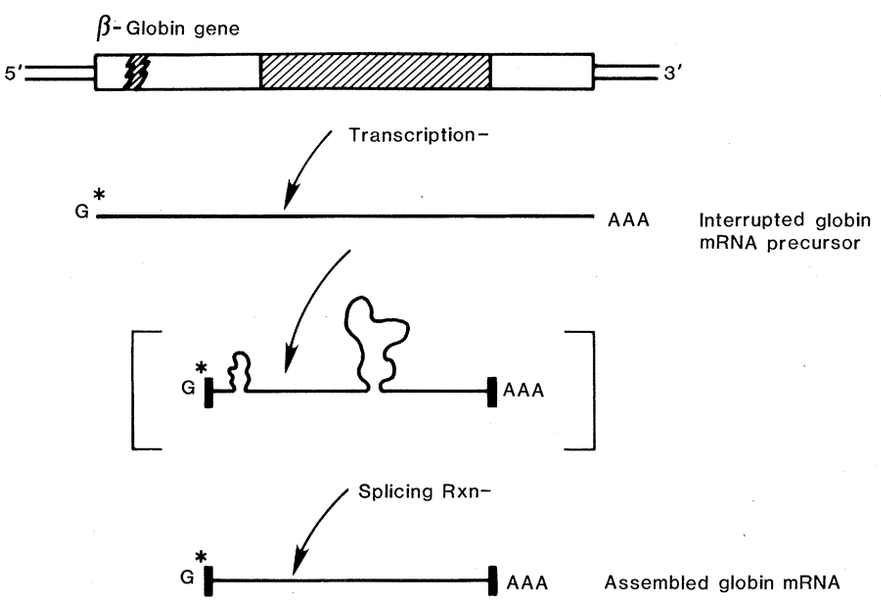


Fig. 3. Pathway of mRNA maturation. The top figure diagrammatically represents the β-globin gene with its two intervening sequences. The primary transcript is shown as is the splicing step that results in the mature form of mRNA; G* indicates the capped 5′ end of the mRNA and AAA indicates the 3′-poly(A) tail.

mouse β-globin) might serve to reduce sequence homology between them and thereby stabilize those genes by reducing the likelihood of unequal crossing-over and subsequent gene loss or amplification (*10*). Tandem genes that conserve considerable intervening and flanking sequence homology, such as the mouse (*25*) and human α-globin genes (*26*) and families of immunoglobulin κ variable region genes (*27*), might be less stable, undergoing more frequent gene loss or replication because of long stretches of shared homology between tandem sequences. Indeed, the α genes and immunoglobulin variable genes appear to behave in just this way (*28*).

Apart from the role intervening sequences might play in evolution, their role in gene expression has been tested by making use of hybrids formed between mouse globin genes and the simian virus SV40 (*29*). Hybrid viruses that contained intervening sequences derived from SV40 or the globin gene (or both) directed the synthesis of stable RNA transcripts in suitably infected African green monkey kidney cells. In the absence of functional intervening sequences, the inserted globin sequence was transcribed, but no stable transcript appeared in the cytoplasm of infected cells (Table 2). Studies with other derivatives of SV40 showed similar results (*30*). Evidently the presence of a functional splice signal is required for the expression of these genes, at least in this system.

Although this result suggests that intervening sequences play some essential role in the expression of globin genes, we cannot confidently generalize from this observation: clearly, histone genes are uninterrupted and readily expressed (*31*). Recently one of the early adenovirus primary transcripts has been found to be colinear with its mRNA (*32*). It may be that there are two (or more) classes of genes that depend upon two (or more) different mechanisms to move their transcripts from nucleus to cytoplasm in a stable way.

Unexpected Complexity of Globin Gene Loci

As we have indicated above, genetic and biochemical experiments had led us to expect four adult and three embryonic globin genes. This simple phenotype, however, is derived from two unexpectedly complex loci, encoding many ostensibly silent copies of globin-like genes. Randomly generated libraries of fragments of mouse DNA cloned in phage λ

enabled us to identify nine hybrid phages which contained overlapping sequences that could be rearranged into the physical map of the β locus (shown in Fig. 4). Six discrete coding regions are shown. Edgell and Hutchison and their colleagues (*33*), who have created a similar map of this locus, have evidence to suggest that the B4 coding region (Fig. 4) is actually two discrete globin-like sequences, one of which is only a fragment of a coding sequence. Presumably two of the genes encode the embryonic globins, X and Y, whereas others have no known phenotypic counterpart.

The α locus is equally complex, though not yet organized into a complete physical map. We have cloned six discrete α-globin-like genes and determined the sequences of four of them. Two correspond to adult α-globins [one (α_1) is shown diagrammatically in Fig. 2] and two are pseudogenes; that is, they contain sequences quite homologous to the α coding sequence, but with enough base substitutions to create either missense or nonsense mutations (or both). One of these pseudogenes [α_4 in Fig. 2 (*34*)] contains point mutations that would substitute a Tyr (tyrosine) for a His (histidine) in amino acid position 58. It also has undergone changes that shift the normal termination codon out of phase so as to continue the globin reading frame for an additional 40 amino acids (α_4 in Fig. 2). If α_4 were expressed in the BALB/c mouse, it would direct the synthesis of pathologic hemoglobin analogous to both methemoglobin Boston and to hemoglobin Constant Spring. We have found no abnormal hemoglobins in these mice and therefore assume that this gene is not expressed in adult red cells. The second pseudogene [α_3 in Fig. 2 (*11, 35*)] has a most surprising structure in that it entirely lacks intervening sequences. Since we have been unable to find significant amounts of mRNA that correspond to this gene in either embryonic or adult reticulocytes, we assume that this gene is also inactive in red cells.

Below, we consider the mechanisms that may have given rise to these genes and what they tell us about the mechanisms that are available for chromosomal evolution. It is already evident that the globin gene locus consists not only of its phenotypically active genes, but also of an array of ostensibly inactive genes that probably arose as the result of ancient duplications. These genes, released from the selective pressures that conserve globin coding sequences, began to drift and break up in a way that now reflects very fundamental mechanisms that act on all DNA sequences.

Evolution at Two Speeds

The two mouse β-globin genes, major and minor, provide a useful evolutionary model by which to gauge the progress and reconstruct the mechanisms of evolutionary drift among related genes. The minimal amino acid differences between these genes (9 of 146 amino acids) suggest that they arose by a duplication event and have been apart for 30 million to 50 million years. Both genes have been cloned, and their sequences have been completely determined (*13*). The degree of conserved homology between the genes can be visualized by forming a heteroduplex molecule between cloned segments encoding each gene (Fig. 5) (*10*). Both genes are embedded in nonhomologous segments of DNA, but have conserved homology—in addition to their coding sequences—in a few hundred bases bordering their structural genes and in the border regions of their intervening sequences. Such conserved homology suggests a functional role for these flanking sequences, a possibility to which we shall return below. The region of nonhomology within the gene is especially interesting; it corresponds to a large segment of the second intervening sequence.

The nucleotide sequence data confirm the impression gained from the heteroduplex analyses and allow us to discern two different modes of evolution (*13*). The sequence homology and divergence is summarized diagrammatically in Fig. 5. The coding sequences differ mainly by point mutations and have changed very little. Their divergence has been slow, largely one base at a time. Deletions and insertions that would grossly affect protein structure or cause out-of-phase missense reading do not appear. It is from the second intervening sequence, however, that we gain our greatest insight into the mechanisms that affect chromosomal DNA. Here the two sequences have diverged greatly, propelled principally by insertions and deletions that alter very large segments of DNA. Occasionally a deletion can be traced back to small repeated sequences in the nondeleted gene, suggesting that the segment was lost by homologous recombination within highly mutable sites (hot spots) (*36*). Thus, strongly selected sequences change slowly, if at all, and then by point mutations. Unselected flanking and intervening sequences change quickly, altering large pieces of DNA at a single stroke.

From what we have learned by comparing the sequences of the α genes and pseudogenes (*11, 14, 37*), it is clear that coding sequences are not immune from deletions, but rather that selection prevents them from appearing in essential genes. Evidently large segments of the second intervening sequences of the β major and minor genes are not subject to the same selective pressures that operate on its coding sequences to tightly conserve their homology. Interestingly, the first, and smaller, intervening sequence contains no in-phase termination codons and therefore could be read through entirely if it were not spliced out during RNA maturation. This sequence is well conserved between β major and β minor genes.

Pseudogenes

Pseudogenes may be a misnomer when it is used to refer to sequences that bear an obvious evolutionary relationship to an active gene but that do not seem to share its activity. As it turns out, two well-characterized examples that occur at the mouse α locus have been found. Each has undergone sufficient

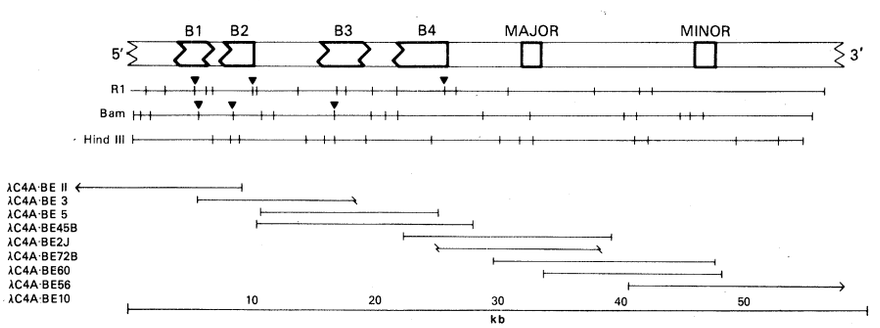


Fig. 4. Physical map of the BALB/c mouse β-like globin gene locus. The map was derived from restriction endonuclease digestion of the cloned segments of DNA shown in the figure. The extent and identifying number of each cloned fragment is shown, as are some restriction sites. Triangles represent sites that occur within a coding sequence which have been confirmed by sequence analysis. Each box represents an area containing coding sequences. Jagged lines represent borders of coding sequences not yet precisely determined; straight lines represent precisely determined borders. The physical map is approximately 60 kbp in length.

evolutionary drift so as to either encode a pathologic globin (α_4) or a coding sequence no longer translatable into a globin-like polypeptide chain (α_3). As we shall see, these pseudogenes both raise and answer questions about evolutionary mechanisms. They also conveniently provide a ready array of naturally occurring mutant gene segments that can be compared to analogous segments in active genes.

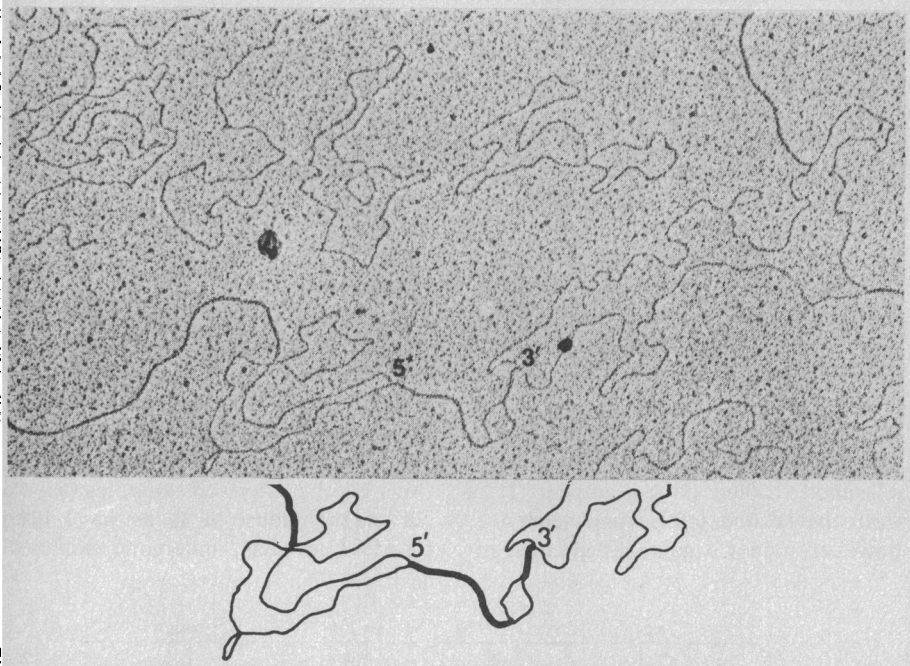
The array of pseudogenes suggests that a genetic locus is in a dynamic state, amplifying gene sequences and releasing extra copies from selection so as to undergo further mutational drift than would otherwise be permitted. These extra copies might, for example, provide the spare sequences from which new modifications or functions are fashioned within functional signals for transcription, processing, and regulation. In short, they might serve as a bank for the eventual reclamation of new genes. Were this a general feature of evolution, one might expect to encounter such complex gene-pseudogene loci frequently. In addition, one might also expect to find active genes that map close to a given locus, bearing a recognizable evolutionary relationship to its genes, but now serving a completely different function in the organism. A possible example of this might be the interesting evolutionary relationship that Koussis *et al.* (*38*) have discovered between serum albumin and α fetal protein. These two genes apparently differ in function but appear to have evolved from a common ancestor.

Table 2. Effect of intervening sequences on the ability to form stable transcripts in globin gene–SV40 hybrid virus-infected cells [adapted from (*29*)]

Virus	Source of IVS		
	SV40	Globin	Stable RNA
1	+	+	+
2	+	−	+
3	−	+	+
4	−	−	−

In addition to suggesting the possibility of an evolutionary collection of spare genes, these pseudogenes, unlike the active genes from which they are derived, reveal an unexpected range of mutational mechanisms that operate on apparently unselected sequences of DNA. The most striking example of such a process is provided by the structure of the α_3 gene (Fig. 2) (*11*, *35*). In contrast to all globin genes thus far examined, this pseudogene lacks intervening sequences. Moreover, intervening sequences are missing in accordance with the GT-AG (A, adenine; G, guanine; T, thymine) rule of RNA splicing (*21*); that is, the gene sequence is homologous to that of the mature α-globin mRNA. A further comparison of the α_3 sequence to that of the adult α gene revealed a close (>80 percent) homology, which indicated that this gene had diverged from the α gene lineage after the α-β divergence. Since α and β genes are interrupted by intervening sequences at homologous positions, their common ancestor must have carried these two intervening sequences and, therefore, rather than representing an uninterrupted primitive globin gene sequence, it is far more likely that α_3 has lost both its intervening sequences.

Given the conserved nature of α-globin and β-globin gene organization, loss of an intervening sequence must be a fairly infrequent event. From studies of globin gene sequences cloned in SV40, it would seem that such a loss would result in the inactivation of a globin gene (*29*). From the way the coding segments of α_3 have diverged (unselected mode involving insertions and deletions) and because we are unable to find α_3 mRNA in embryonic or adult reticulocytes, we tentatively conclude that α_3 is an inactive gene. This is in contrast to the apparently normal activity exhibited by the rat in-



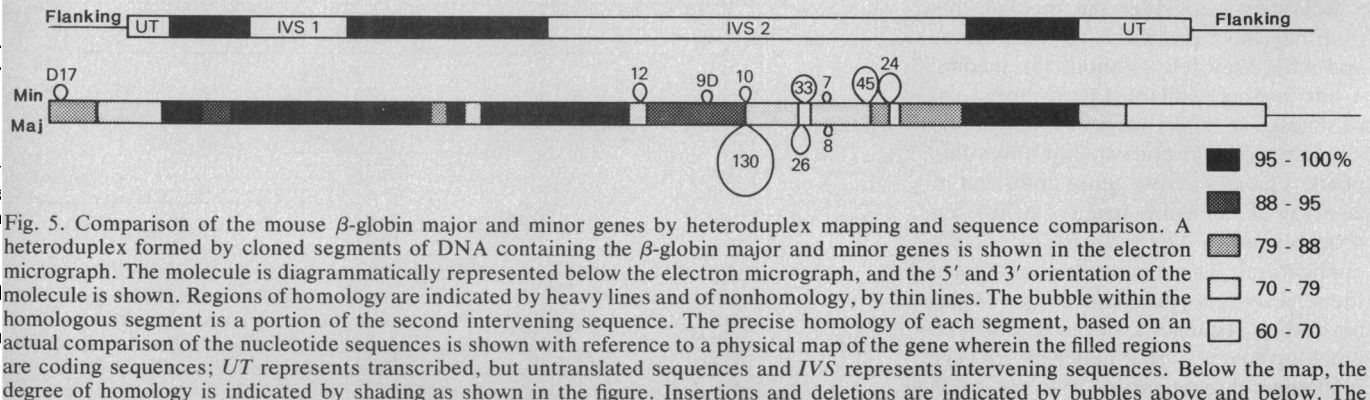


Fig. 5. Comparison of the mouse β-globin major and minor genes by heteroduplex mapping and sequence comparison. A heteroduplex formed by cloned segments of DNA containing the β-globin major and minor genes is shown in the electron micrograph. The molecule is diagrammatically represented below the electron micrograph, and the 5′ and 3′ orientation of the molecule is shown. Regions of homology are indicated by heavy lines and of nonhomology, by thin lines. The bubble within the homologous segment is a portion of the second intervening sequence. The precise homology of each segment, based on an actual comparison of the nucleotide sequences is shown with reference to a physical map of the gene wherein the filled regions are coding sequences; *UT* represents transcribed, but untranslated sequences and *IVS* represents intervening sequences. Below the map, the degree of homology is indicated by shading as shown in the figure. Insertions and deletions are indicated by bubbles above and below. The numbers indicate the number of bases involved. Data are from (*10*) and (*13*).

sulin I gene, which has lost only one of its two intervening sequences (*39*). These two examples suggest that while intervening-sequence loss is an infrequent event, it is not altogether rare.

One can imagine a number of possible mechanisms by which intervening sequences could be lost, and several of these have been dealt with (*11, 35*). A critical clue resides in the fact that the intervening sequences have been lost in accordance with the GT-AG splicing rule; this suggests that the RNA splicing mechanism has at least indirectly mediated the loss of sequences from chromosomal DNA. Even with this assumption, a number of mechanisms are possible. For example, chromosomal integration of a complementary DNA (cDNA) copy of the globin mRNA would produce such a structure. The fact that sequence homology of the α_3 and α_1 genes extends beyond the 5′ capping site of the genes (*11, 35*) suggests that, if such an integration event occurred, it did not occur by illegitimate recombination at some nonglobin locus, but required recombination involving a double crossover with a normal globin gene. The many steps required by this process, including the presence of cDNA in a germ-line cell, make this mechanism seem unlikely. It is also possible that the RNA splicing enzymes can operate on DNA; even though there is nothing to rule out this possibility, and it is in some respects quite attractive, we might expect far more variation in the occurrence of intervening sequences in globin genes than seems to be the case.

In the absence of definitive evidence, we favor another mechanism that also invokes the mediation of normal splice mechanisms, but indirectly, through a gene conversion event involving a heteroduplex structure formed between globin mRNA and a replicating globin gene. The model is shown diagrammatically in Fig. 6. Imagine a growing fork at a DNA replication site and a globin mRNA hybridized to one of the single-stranded regions encoding the interrupted globin gene. The putative structure could contain one or two looped-out, single-stranded intervening sequences; two are indicated in the figure. A nicking activity could readily attack either or both single-stranded regions containing the intervening sequences. The nicked strands would then be removed by an exonuclease, and the covalent structure would be restored by a closing enzyme. If the RNA strand were nicked—and this should be a much rarer event—the intervening sequence would be filled in by repair enzymes and no change in the gene structure would occur. In this way, either or both intervening sequences could be removed. We imagine that both loops were resolved by deletion in the α_3 gene, whereas only one appears to have been resolved by deletion in the rat insulin I gene.

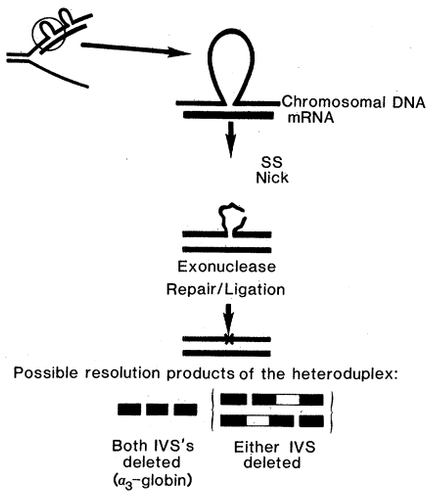


Fig. 6. Diagram indicating the possible modes of resolution of a hypothetical recombinant heteroduplex formed between a single-stranded region of genomic DNA to which globin mRNA is hybridized. A mRNA forms a heteroduplex structure (for example) at a replication fork. One or both intervening sequences (*IVS*'s) loop out as a single-stranded structure that is nicked, degraded, repaired, and covalently joined in a sequence corresponding to the mature mRNA, thereby deleting the intervening sequences in accordance with the GT-AG splice rule.

The requisite globin mRNA and the enzymes required by this model are, in fact, likely to be present in an appropriate germ cell. Globin mRNA sequences have been detected in *Xenopus* oocytes (*40*) which is consistent with the notion that the RNA populations of primitive cell types are extremely complex (*41*). We assume the presence of the requisite DNA repair system, extrapolating from the experimental demonstration of gene conversion in SV40-infected African green monkey kidney cells (*42*).

Functional Assessment of Surrounding and Interrupting Sequences

A correlation of the structural information we have gathered about the mouse globin genes allows us to identify specific segments that are likely to be of functional significance. Above we noted how conservation of sequences at the globin intervening-sequence borders and their comparison to similarly placed sequences in other genes allowed the derivation of a consensus sequence that might serve as all or part of a splicing signal (Table 1). Similarly, comparisons of sequences that lie to the 5′ and 3′ sides of globin genes allow us to identify sequences that might be concerned with transcriptional initiation and termination or polyadenylate [poly(A)] addition (Fig. 2). Notice that the sequence TATAAA (or TATAAG) occurs approximately 30 bases to the 5′side of the cap site of each active globin gene. This sequence had previously been identified on the 5′ side of several histone genes and has been found similarly located in a variety of viral and chromosomal sequences. It was presumed to be involved in transcriptional initiation (*43*). The 3′ nucleotide AATAAA had also previously been found preceding the poly(A) addition site in many mRNA's and it was therefore suggested that it served as a signal for this process (*44*). Such correlations are useful in that they rely upon selection to conserve regions of functional significance. However, while these correlations provide important clues, they cannot prove a function or display the broad spectrum of relative activities that can be encoded into such signals. This kind of information comes from direct functional tests.

Of the two basic approaches to functional testing, in vivo and in vitro systems, the latter is obviously easier to manipulate, while the former offers at least the hope of being able to arrange for gene expression in its authentic cellular context. Perhaps the most successful in vivo system for manipulating the function of mammalian genes comes from forming hybrid SV40 viruses (*45, 46*). With this system, it was shown that the RNA splicing reaction can occur across species and organ barriers, that as few as 18 nucleotides on the 5′ side of the β-globin intervening-sequence border suffice for RNA splicing, and that intervening sequences seem to be essential for expression of certain genes (*29, 45, 47*). In addition it could be shown that the α-globin gene promoter resided on the α-globin gene-containing fragment used to construct the hybrid (this DNA included 500 bp of 5′ flanking sequence) (*48*).

The gene mapping of the promoter activity, however, is much more easily determined by use of one of the cell-free transcriptional systems that have recently been developed and that accurately initiate transcription at RNA polymerase II promoters (*49*). With the use of a suitably modified system, truncated segments of the α_1-globin gene were used to show that the α_1 promoter activity re-

sided on a 148-bp segment that included the TATAAG box referred to above and extended only seven nucleotides to the 3′ side of α_1 cap site (the putative initiation point of transcription) (*50*). In addition, the cell-free system was used to assay analogous regions derived from the two mutant genes or pseudogenes. Both were found to be inactive. While a number of base substitutions occur between the authentic α sequence and its pseudogenes, very few occur in positions that are otherwise conserved between the α and β genes in this region. Perhaps most interesting is the single-base substitution that occurs in the α_3 pseudogene, converting the position exactly 30 nucleotides from the cap site from T to G (TAGAAG). Obviously, this approach, coupled with the ability to use other pseudogenes and to form hybrids between active and inactive promoters offers a means of completely defining the structure of a promoter site. The subtler definitions, those seeking to answer critical questions about relative and temporal expression of globin (and other) genes, depend on the development of more complex systems or more precise means of introducing genes (or mutations) into their authentic chromosomal context in living cells.

Recombinant DNA technology has become such an essential and commonplace tool of genetic investigation that its success will eventually cease to astonish us. Future surprises will have to come from understanding solutions that nature devised for genetic problems. In this regard, there is a special irony in the uninterrupted structure of the α_3 gene that we discussed above. Had it been discovered in 1977, we would have been left with the comfortable feeling that evolution created no surprises when dealing with fundamental structures like genes. Just 3 years later, such a finding creates exactly the opposite impression.

References and Notes

1. S. M. Tilghman, D. C. Tiemeier, F. Polsky, M. H. Edgell, J. G. Seidman, A. Leder, L. W. Enquist, B. Norman, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 4406 (1977).
2. N. J. Proudfoot, M. H. M. Shander, J. M. Manley, M. L. Gefter, T. Maniatis, *Science* **209**, 1329 (1980).
3. A. Fantoni, A. Bank, P. A. Marks, *ibid.* **157**, 1327 (1967).
4. E. S. Russell and E. L. McFarland, *Ann. N.Y. Acad. Sci.* **241**, 25 (1974).
5. J. B. Whitney, *Cell* **12**, 863 (1977).
6. P. Leder, S. M. Tilghman, D. C. Tiemeier, F. I. Polsky, J. G. Seidman, M. H. Edgell, L. W. Enquist, A. Leder, B. Norman, *Cold Spring Harbor Symp. Quant. Biol.* **42**, 915 (1977).
7. S. M. Tilghman, D. C. Tiemeier, J. G. Seidman, B. M. Peterlin, M. Sullivan, J. V. Maizel, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 725 (1978).
8. A. J. Jeffreys and R. A. Flavell, *Cell* **12**, 1097 (1977).
9. A. Leder, H. I. Miller, D. H. Hamer, J. G. Seidman, B. Norman, M. Sullivan, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 6187 (1978).
10. D. C. Tiemeier, S. M. Tilghman, F. I. Polsky, J. G. Seidman, A. Leder, M. H. Edgell, P. Leder, *Cell* **14**, 237 (1978).
11. Y. Nishioka, A. Leder, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2806 (1980).
12. D. Konkel, S. M. Tilghman, P. Leder, *Cell* **15**, 1125 (1978); A. van Ooyen, J. van den Berg, N. Mantei, C. Weissmann, *Science* **206**, 337 (1979).
13. D. Konkel, J. V. Maizel, P. Leder, *Cell* **18**, 865 (1979).
14. Y. Nishioka and P. Leder, *ibid.*, p. 875.
15. P. J. Curtis and C. Weissmann, *J. Mol. Biol.* **106**, 1061 (1976); J. Ross, *ibid.*, p. 403; S.-P. Kwan, T. G. Wood, J. Lingrell, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 178 (1977).
16. S. M. Tilghman, P. J. Curtis, D. C. Tiemeier, P. Leder, C. Weissmann, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 1309 (1978).
17. A. J. Kinniburgh, J. E. Mertz, J. Ross, *Cell* **14**, 681 (1978).
18. R. C. Ogden, J. S. Beckman, J. Abelson, H. S. Kang, *ibid.* **17**, 399 (1979).
19. P. J. Curtis, N. Mantei, C. Weissmann, *Cold Spring Harbor Symp. Quant. Biol.* **42**, 971 (1977).
20. I. Seif, G. Khoury, R. Dhar, *Nucleic Acids Res.* **6**, 3387 (1979).
21. R. Breathnach, C. Benoist, K. O'Hare, F. Gannon, P. Chambon, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 4853 (1978).
22. M. R. Lerner, J. A. Breyle, S. M. Mount, S. L. Wolin, J. A. Steitz, *Nature (London)* **283**, 220 (1980); M. B. Mathews, *ibid.* **285**, 575 (1980).
23. W. Gilbert, *ibid.* **271**, 501 (1978).
24. P. W. Early, M. M. Davis, D. B. Kaback, N. Davidson, L. Hood, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 857 (1979); H. Sakano, J. H. Rogers, K. Huppi, C. Brack, A. Traunecker, R. Maki, R. Wall, S. Tonegawa, *Nature (London)* **277**, 627 (1979).
25. A. Leder, personal communication.
26. J. Lauer, C.-K. J. Shen, T. Maniatis, *Cell* **20**, 119 (1980).
27. J. G. Seidman, A. Leder, M. Nau, B. Norman, P. Leder, *Science* **202**, 11 (1978).
28. P. Leder, J. G. Seidman, E. Max, Y. Nishioka, A. Leder, B. Norman, M. Nau, *Miami Winter Symp.* **16**, 133 (1977); E. A. Zimmer, S. L. Martin, S. M. Beverley, Y. W. Kan, A. C. Wilson, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2158 (1980).
29. D. H. Hamer and P. Leder, *Cell* **18**, 1299 (1979).
30. C.-J. Lai and G. Khoury, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 71 (1979).
31. I. Sures, J. Lowry, L. H. Kedes, *Cell* **15**, 1033 (1978); W. Schaffner, G. Kunz, H. Daetwyler, J. Telford, H. O. Smith, M. L. Birnstiel, *ibid.* **14**, 655 (1978).
32. P. Alestrom, G. Akusjarvi, M. Perricaudet, M. B. Mathews, D. F. Klessig, U. Pettersson, *ibid.* **19**, 671 (1980).
33. M. H. Edgell and C. Hutchison, personal communication.
34. Y. Nishioka and P. Leder, personal communication.
35. E. E. Vann, G. I. Goldberg, P. W. Tucker, O. Smithies, *Nature (London)* **286**, 222 (1980).
36. P. J. Farabaugh and J. H. Miller, *J. Mol. Biol.* **126**, 847 (1978).
37. Y. Nishioka, A. Leder, P. Leder, in preparation.
38. D. Koussis, F. Eifferman, P. van de Rijn, M. B. Gorin, R. S. Ingram, S. M. Tilghman, in preparation.
39. P. Lomedico, N. Rosenthal, A. Efstratiadis, W. Gilbert, R. Kolodner, R. Tizard, *Cell* **18**, 545 (1979); B. Cordell, G. Bell, E. Tischer, F. M. DeNoto, U. Ullrich, R. Pietet, W. J. Rutter, H. M. Goodman, *ibid.*, p. 533.
40. S. M. Perlman, P. J. Ford, M. M. Rosbach, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 3835 (1977).
41. G. G. Galau, W. H. Klein, M. M. Davis, B. J. Wold, R. J. Britten, E. H. Davidson, *Cell* 7, 487 (1976).
42. C.-J. Lai and D. Nathans, *Virology* **66**, 70 (1975).
43. M. Goldberg and D. Hogness, personal communication.
44. N. J. Proudfoot and G. G. Brownlee, *Nature (London)* **263**, 211 (1976).
45. D. H. Hamer and P. Leder, *ibid.* **281**, 35 (1979).
46. R. Mulligan, E. Howard, P. Berg, *ibid.* **277**, 108 (1979).
47. D. H. Hamer, K. D. Smith, S. H. Boyer, P. Leder, *Cell* **17**, 725 (1979); D. H. Hamer and P. Leder, *ibid.*, p. 737.
48. D. H. Hamer, M. Kaehler, P. Leder, *ibid.*, in press.
49. P. A. Weil, D. S. Luse, J. Segall, R. G. Roeder, *ibid.* **18**, 469 (1979); J. Manley, M. Gefter, P. Sharp, personal communication.
50. C. Talkington, Y. Nishioka, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.*, in press.

51. This work was carried out in its initial and subsequent stages by several valuable, inventive colleagues. In particular we thank S. Tilghman and D. Tiemeier, who began these studies, and D. Hamer, who developed the SV40 system and then, together with M. Kaehler, used it as a vehicle for exploring the activity of globin genes. We also thank O. Smithies and E. Vanin, who made sequence data (they also cloned and sequenced an α_3-like gene) available to us prior to publication, M. Edgell, who shared his BALB/c β-globin gene mapping data with us, thereby saving us at least one mapping error, and T. Broderick for the expertise, patience, and good humor she displayed in the preparation of this manuscript.

11 August 1980