Vol. 75, No. 12, pp. 6187–6191, December 1978
Genetics

Comparison of cloned mouse α- and β-globin genes: Conservation of intervening sequence locations and extragenic homology

(recombinant DNA/evolution/mRNA/sequence analysis)

A. LEDER, H. I. MILLER, D. H. HAMER, J. G. SEIDMAN, B. NORMAN, M. SULLIVAN, AND P. LEDER

Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20014

Communicated by David R. Davies, October 10, 1978

ABSTRACT **We have cloned and characterized a 9.7-kilobase *Eco*RI fragment of mouse DNA that contains an α-globin gene. The gene is encoded in at least three discontinuous segments of DNA interrupted by two small intervening sequences that can be visualized as R-loop structures in the electron microscope. The size of the gene and its small intervening sequences fits well with the known size of the α-globin mRNA precursor, suggesting that these intervening sequences, like those of β-globin, are transcribed. Partial sequence analysis indicates that the larger intervening sequence interrupts the α-globin gene at a site exactly corresponding to that interrupted by the larger intervening sequences in both the β-globin major and minor genes. This observation suggests that these sequences were present when the α- and β-globin genes diverged in early vertebrate evolution, more than 500 million years ago. Furthermore, though α and β^{maj} genes are encoded on different chromosomes, when their sequences are compared directly by visualization of heteroduplex structures, only one 150- to 200-base-pair segment of homology is recognized. These homologous sequences are located on the 3′-flanking segments of both genes, about 1.5 kilobases from each.**

Globin production in developing erythroid cells involves the coordinate expression of at least two sets of genes located on different chromosomes: those corresponding to the α- and β-globin polypeptides (see review, ref. 1). By using cloned segments of BALB/c mouse genomic DNA we have shown that the two tightly linked adult β-globin genes, β^{maj} and β^{min}, are each interrupted by two intervening sequences of DNA (2–5) and that they share homology with one another in the regions immediately surrounding their coding and intervening sequences (5). Intervening sequences have also been found in rabbit globin genes (6) and in genes corresponding to *Drosophila* 28S ribosomal RNA (7–9), adenovirus (10–12), simian virus 40 (13), mouse immunoglobulin (14), yeast tRNA (15, 16), and chicken ovalbumin (17–20) and thus appear to be a common feature of eukaryotic gene organization. Because both mouse β-globin genes are coordinately expressed and direct the synthesis of mRNA precursors that include their intervening sequences (21–23), it is possible that their shared sequences play some critical role in their joint regulation and in the processing of their 15S mRNA transcripts.

Unlike the adult mouse β-globin genes, the mouse α-globin genes are expressed during both fetal and adult life (24). In addition, the mature form of the α-globin mRNA appears to be synthesized via a small 850-base precursor, approximately 250 bases larger than the cytoplasmic α-globin mRNA (25, 26). It is likely that certain aspects of the structure and organization of one α-globin gene will allow us to account for its joint expression with other α-globin genes located on the same chromosome (in *cis*) and with β-globin genes located on a different chromosome (in *trans*).

Here we describe the cloning and characterization of one of the *Eco*RI fragments of BALB/c mouse DNA that includes an α-globin gene. Like the β-globin genes, this α-globin gene is interrupted by at least two intervening sequences of DNA. Unlike the β-globin genes, however, both α intervening sequences are small, likely accounting for the smaller size of the α-globin mRNA precursor. Despite these differences, the one α intervening sequence whose location we establish occurs at a position that aligns precisely with those of the larger intervening sequences in the β-globin major and minor genes. We have also located a closely linked, flanking segment of DNA that is homologous between the α and β^{maj} genes. This shared sequence is especially interesting in view of the *trans* regulation of the α- and β-globin genes.

MATERIALS AND METHODS

Fragment Isolation and Hybrid Phage Formation. *Eco*RI restriction endonuclease fragments of BALB/c mouse DNA were prepared and isolated by RPC-5 chromatography and preparative agarose gel electrophoresis as described (27, 28). The purified fragments were cloned under P3-EK2 conditions in the EK2 vector λgt*WES*·λB (29) as described (27), and detected by using the technique of Benton and Davis (30). All *in situ* hybridizations were carried out by using ^{32}P-probes labeled by nick translation to a specific activity of about 40–80 cpm/pg and annealed to nitrocellulose blots according to Southern (31) as modified (5). Phage and plasmids were grown and DNA was prepared as described (2, 5, 27). α-Globin and β-globin cDNA clones (pCR1·βM9 and pCR1·αM10) in plasmids (32) were used as probes.

Sequence Determination. The nucleotide sequence around the *Bam*HI site (Fig. 2) was determined by first isolating a *Sac* I-cleaved fragment containing the globin sequence, cleaving with *Bam*HI, enzymatically labeling the *Bam*HI ends with ^{32}P, isolating the appropriate fragments by polyacrylamide gel electrophoresis after cleavage with *Mbo* II and *Hae* III, and determining the fragment sequences as described by Maxam and Gilbert (33).

Electron Microscopy. The 9.7-kilobase (kb) α-1 *Eco*RI fragment was isolated from 1% agarose and annealed to globin mRNA and prepared for R-loop analysis as described (4). Heteroduplexes were formed between λgt*WES*·Mα1 (in both orientations) and λgt*WES*·βG2 (β^{maj}) and λgt*WES*·βG3 (β^{min}) by alkalai treatment of the purified phage in CsCl and subsequent renaturation in the presence of 50% formamide/0.1 M Tris·HCl, pH 7.5/10 mM EDTA for 1 hr at 25°C at a final concentration of 5 μg/ml. The DNA was diluted 5-fold in 70% formamide and spread onto a hypophase of H_2O.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: kb, kilobase; bp, base pair.

RESULTS AND DISCUSSION

Identification and Cloning of α-Globin Genes. Genetic evidence has suggested the presence of *at least* two adult α-globin genes in the BALB/c mouse (see review, ref. 32). We are able to identify and partially purify the *Eco*RI restriction endonuclease fragments of mouse embryonic DNA that encode elements of these genes by use of RPC-5 chromatography (26) and subsequent preparative agarose gel electrophoresis (27). Three fragments were detected by hybridization to an α-globin cDNA plasmid (Fig. 1). The largest of these (9.7 kb) was selected for cloning in the λgt*WES*·λB system (29). The isolated fragments (which had been purified approximately 500-fold) were ligated into the phage DNA and transfected to yield approximately 10,000 hybrid phage. These, in turn, were screened by using the filter blot technique of Benton and Davis (30) and a ^{32}P-labeled probe containing a cloned α-globin cDNA sequence (31). Two independent and ultimately indistinguishable hybrids containing α-globin sequences were obtained and designated Mα1. Neither cross-hybridized to a β-globin cDNA plasmid (32).

The α-Globin Gene Is Interrupted by At Least Two Intervening Sequences. The location and structure of the α-globin gene within the cloned 9.7-kb fragment (λgt*WES*·Mα1) was established by hybridization as noted above, by restriction endonuclease analysis (not shown), by visualization of the R-loop structure formed between the α-globin gene and globin mRNA (Fig. 2), and by partial sequence determination as noted below. The R-loop structure revealed two unequal R-loops separated by a small segment of double-stranded DNA for which there is no single-stranded DNA displacement loop. In addition, a smaller, double-stranded knot-like structure occurs within the larger R-loop. Both structures (indicated by the arrows in Fig. 2) occur at the same position in each of 31 molecules analyzed and both can be interpreted to represent DNA sequences in the α-globin gene that are missing from the mature α-globin mRNA—i.e., the α-globin gene appears encoded in at least three discontinuous segments of DNA separated by at least two intervening sequences. Interestingly, the relative position of each intervening sequence within the R-loop structure is very similar to that seen in the β-globin genes (see below), although the larger α intervening sequence is considerably smaller than that seen in two β-globin genes (2, 5, 34). Measurement of these structures suggests that the larger intervening sequence is approximately 150 ± 41 (SD) base pairs (bps) in length (this is likely an underestimate) whereas the smaller, which cannot be measured by using this technique, appears to be approximately the same size as the smaller intervening sequence visualized in the β-globin^{maj} gene (2). The small β^{maj} intervening sequence is 116 bps as measured by direct sequence determination (34). Taken together the R-loop and intervening sequences measure approximately 850 ± 170 bps, a length approximately equal to that of the α-globin mRNA precursor (25, 26).

In view of the fact that the 15S β-globin mRNA precursor contains transcripts of both its intervening sequences, it is likely that the α-globin precursor contains intervening sequences as well. The relatively smaller size of the α-transcript globin mRNA precursor would then be accounted for by the relatively smaller size of α-globin intervening sequences. Indeed, the difference in size between the second intervening sequence of β-globin and α-globin (≈650 bps and ≈150 bps, respectively), ≈500 bps, would account for the approximate difference in length between the two mRNA precursors.

A Striking Homology between the Sequence Interrupted by the Larger α- and β-Globin Intervening Sequences. When the complete nucleotide sequence of the β-globin gene was determined (34), it was noted that both β-globin intervening sequences interrupt the β-globin coding sequence between codons corresponding to the tripeptide Arg-Leu-Leu which is repeated at positions 30–32 and 104–106 in the β-globin polypeptide chain. We have noted above that the α-globin intervening sequences appear to occur at roughly the same positions relative to the R-loop structure as do the β-globin intervening sequences. In order to establish the exact position of the larger α-globin R-loop and the orientation of the gene, we determined the nucleotide sequence of a region of the gene surrounding the *Bam*HI site located at codons corresponding to amino acid positions 94 and 95 (ref. 35; see arrow in Fig. 3). The determined sequence (Fig. 3) specifies codons corresponding to amino acids 75 through 99 of α-globin and then begins a missense sequence that marks the 5′ border of the larger intervening sequence. The α-globin gene is therefore interrupted between codons 99 and 100.

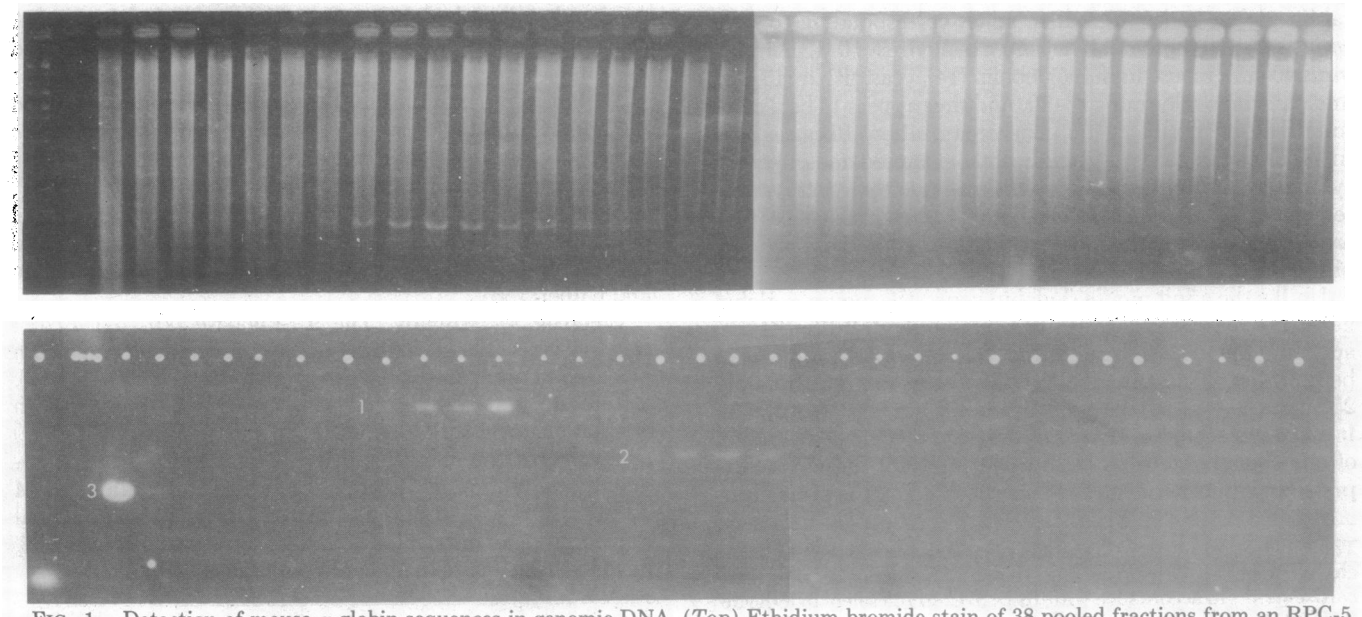
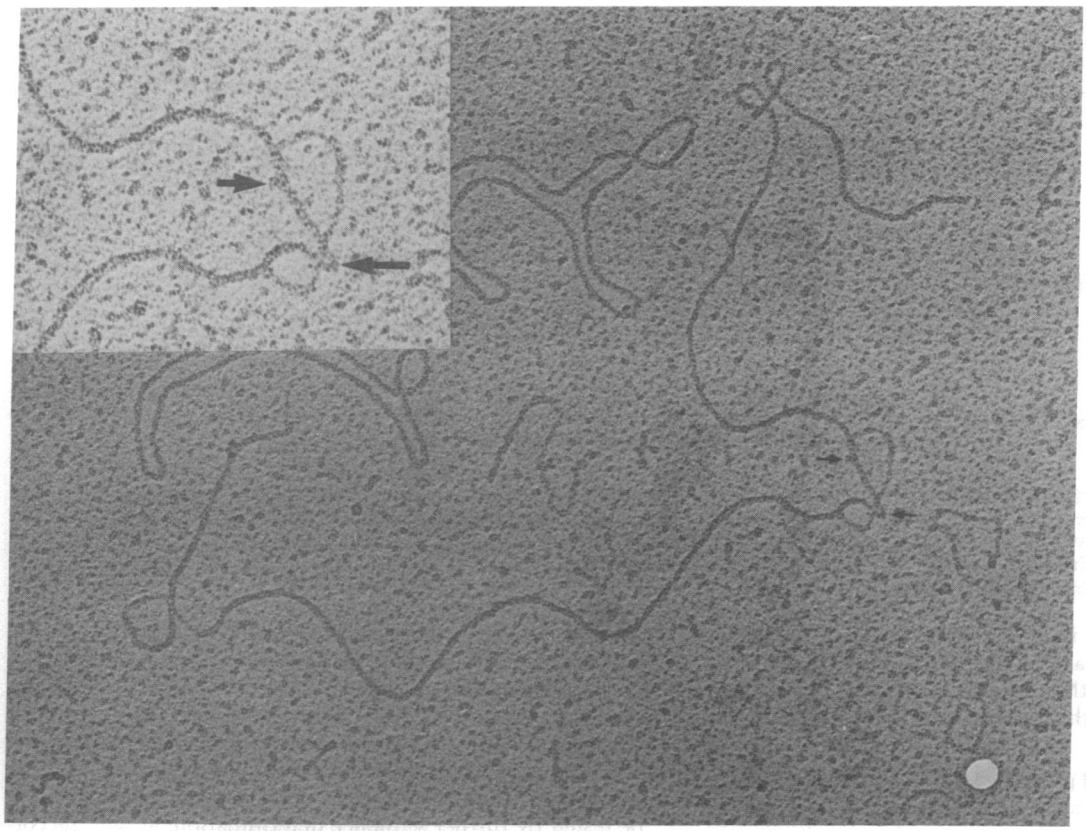


FIG. 1. Detection of mouse α-globin sequences in genomic DNA. (*Top*) Ethidium bromide stain of 38 pooled fractions from an RPC-5 chromatograph of *Eco*RI DNA analyzed by electrophoresis in 1% agarose. The leftmost lane is an *Eco*RI digest of λcI857 DNA in which fragment sizes in kb, from top to bottom, are 21.3, 7.4, 5.8, 5.4, 4.7, and 3.3. (*Bottom*) The DNA was transferred to a Millipore filter and hybridized to pCR1·αM10 [^{32}P]DNA. A radioautogram of the filter is shown. The numbers refer to the three detected fragments.



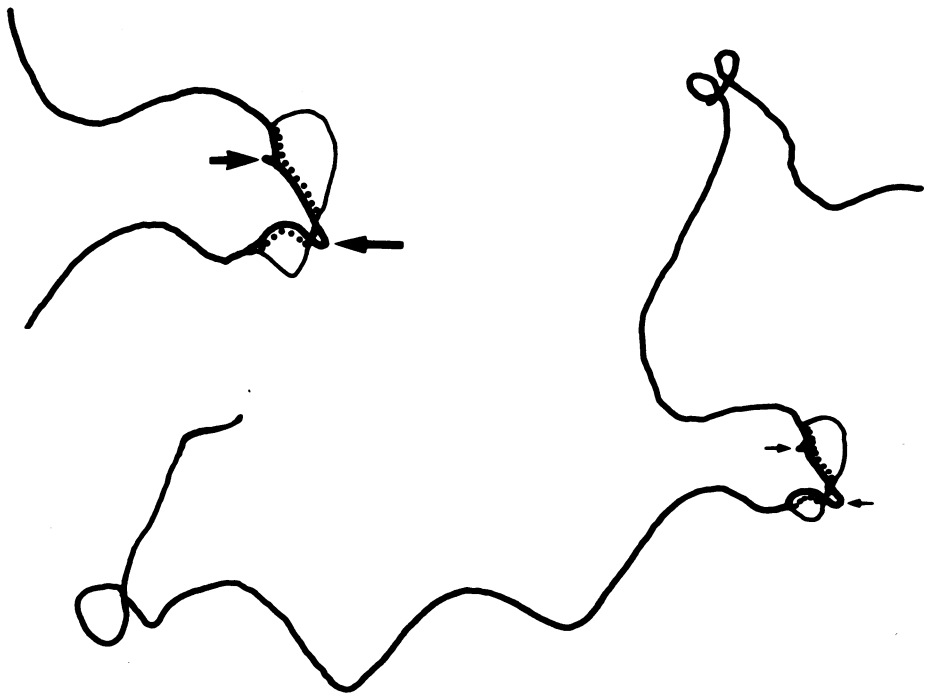


FIG. 2. Hybrids formed between the α-globin gene and globin mRNA. The figure illustrates R-loops formed between the 9.7-kb *Eco*RI fragment isolated from λgt *WES*·Mα1 and globin mRNA. The arrows indicate double-stranded DNA structures in the R-loop that do not anneal to mRNA (intervening sequences). The *Inset* is a magnification of the R-loop. The line drawing is a diagram in which the heavy line represents double-stranded DNA or DNA·RNA and the thin line, single-stranded DNA. The dotted line represents annealed mRNA. Thirty-one molecules were measured and the results of these measurements are summarized in Fig. 3 and the text.

Interestingly, when we align the α- and β-globin genes to maximize homology in their amino acid sequences and to match their functionally specific amino acid residues (36), amino acid positions 99 and 100 of α-globin correspond to amino acid positions 104 and 105 of β-globin. Thus, the larger intervening sequences interrupt the α and both β-globin genes at the same relative position in each gene. Furthermore, the nucleotide sequence immediately preceding the 5′ border of the larger intervening sequence (IVS2, Fig. 4) is closely preserved between α- and β-globin genes (85% identity), whereas the more distal portion of the coding sequence and the intervening sequence have diverged considerably (49% and 29% identity, respectively) between the two genes.

Possible Evolutionary Significance: Predictive Value of the Conserved Interruption. The fact that both mouse and rabbit β-globin genes are interrupted by intervening sequences that occur at the same relative positions (A. Efstratiadis, T. Maniatis, and E. Lacy, personal communication; ref. 37) suggests that these intervening sequences existed from the time that rabbit and mouse began to diverge as separate species approximately 70 million years ago (36). Furthermore, the occurrence of intervening sequences at the same site in the α-globin and β-globin genes suggests that the evolutionary antiquity of these sequences extends back approximately 500 million years to a point in early vertebrate evolution at which the globin gene is thought to have undergone its first duplication

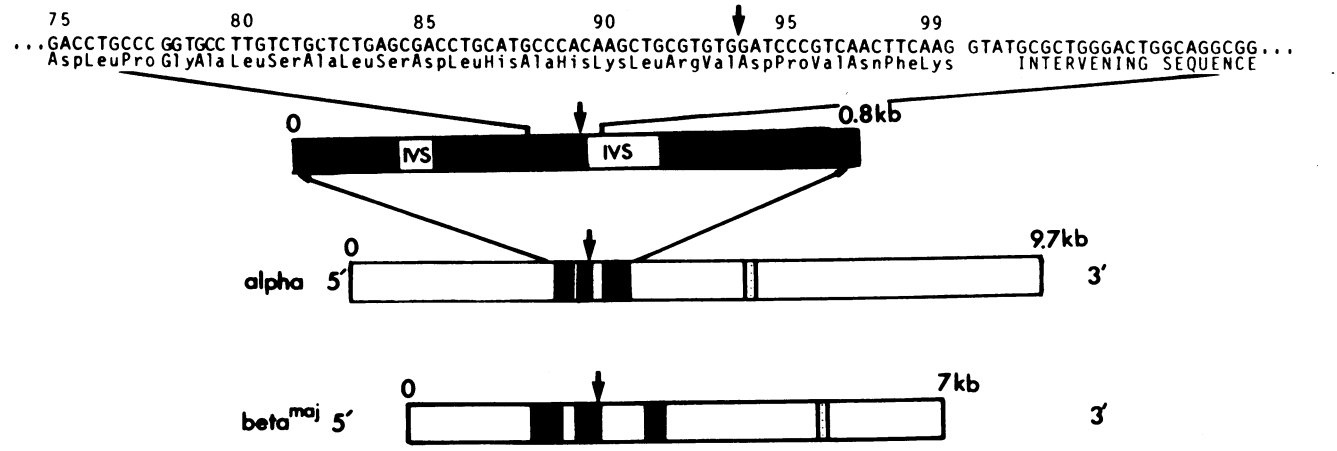


FIG. 3. Diagrammatic representation of the map of the cloned *Eco*RI fragment containing the α gene compared to the cloned *Eco*RI fragment containing the β^{maj} gene. The α and β genes are identified in the figure. The filled portions represent sequences present in mature globin mRNA. The open regions (IVS) represent intervening sequences. The dotted region is the shared homologous segment present on both α and β^{maj} genes. The arrow indicates the position of a *Bam*HI site at codon 94 in α-globin and 99 in β-globin. The sequence surrounding this region is indicated. The scale and 5′ → 3′ orientation are also indicated.

(36). The preservation of intervening sequences at this site in globin genes over such a long period of evolutionary time supports the view that these sequences serve some critical function or functions in the expression of these genes.

Table 1. Homologies and predicted homologies between spliced regions of mouse α- and β-globin genes

Globin	IVS	Sequence				
α	2	...C	AAG	(CTX	CTX	$\frac{U}{A}$...)*
			Lys	Leu	Leu	
			99	100		
β	2	...C	AGG	CTC	CTG	G...
			Arg	Leu	Leu	
			104	105		
			31	32		
α	1	(...$\frac{A}{G}$	AGG	ATG	TT$\frac{T}{C}$	G...)*
			Arg	Met	Phe	
β	1		30	31		
			Arg	Leu	Leu	
		...C	AGG	CTG	CTG	G...

( )*, Predicted; IVS, intervening sequence.

The occurrence of the larger intervening sequences at analogous sites in both α- and β-globin genes suggests that we might be able to predict, by comparison to the β-globin gene, the site in the α-globin gene that is interrupted by the smaller intervening sequence. The smaller intervening sequence interrupts the β-globin gene after an Arg codon at position 30. The analogous position in the α-globin gene is an Arg codon at position 31, followed by Met and Phe codons. Because this site fits well with the measured distance between the orientation of the two α-globin intervening sequences, one might expect the smaller α-globin intervening sequence to interrupt the gene at this point. The comparative α- and β-globin sequences and the predicted sites are summarized in Table 1 and their comparative maps are indicated in Fig. 3. Obviously this prediction will be tested by further sequence determination, as will the corollary prediction that all vertebrate globins will be interrupted by two intervening sequences that occur in analogous positions in α- and β-globin genes.

A Small Common Sequence Occurring Close to α- and β-Globin^{maj} Genes. We have already referred to the fact that α- and β-globin genes, although residing on different chromosomes, are coordinately expressed. Understanding the molecular basis of the *trans* regulation of these genes constitutes one of the most interesting problems presented by the globin system. A possible mechanism governing the *trans* regulation of these genes might depend upon the recognition of primary sequences common to both. Substantial stretches of homology (>25 bps) of potential interest between these genes can be detected by *in situ* hybridization or the visualization of heteroduplex molecules formed between cloned DNA fragments containing the α- and β-globin genes. By using these techniques, we have searched for homology between the cloned fragment containing the α-globin gene and the 7-kb or 14-kb fragment that contains, respectively, the β^{maj} or β^{min} gene. When α and β^{maj} segments were compared both by *in situ* hybridization (not shown) and heteroduplex analysis, a 165 ± 39-bp (SD) segment of homology could be identified (Fig. 5) approximately 1.5–2 kb from the 3′ end of both coding sequences (see map, Fig. 3). No extensive homology was found between the fragments that contained the α and the β^{min} genes. The significance of this homology between α and β^{maj} segments, if any, remains to be determined. It may constitute one of the moderately (or highly) reiterated

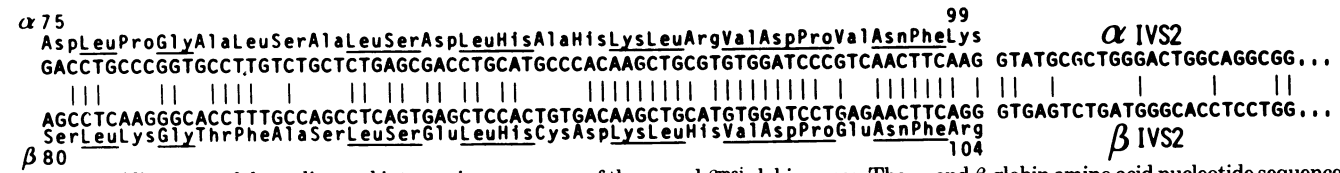


FIG. 4. Alignment of the coding and intervening sequences of the α- and β^{maj} globin genes. The α- and β-globin amino acid nucleotide sequences are identified in the figure. The α sequences are as indicated in Fig. 3. The β^{maj} sequence was determined by Konkel *et al.* (34). The amino acid sequences were determined by Popp (38) and Popp and Bailiff (39) for α and β^{maj}, respectively. Perpendicular lines indicate identical nucleotides. Underscored amino acids represent identical amino acids. IVS2 refers to the second (numbering 5′ to 3′), larger intervening sequence of each indicated gene.

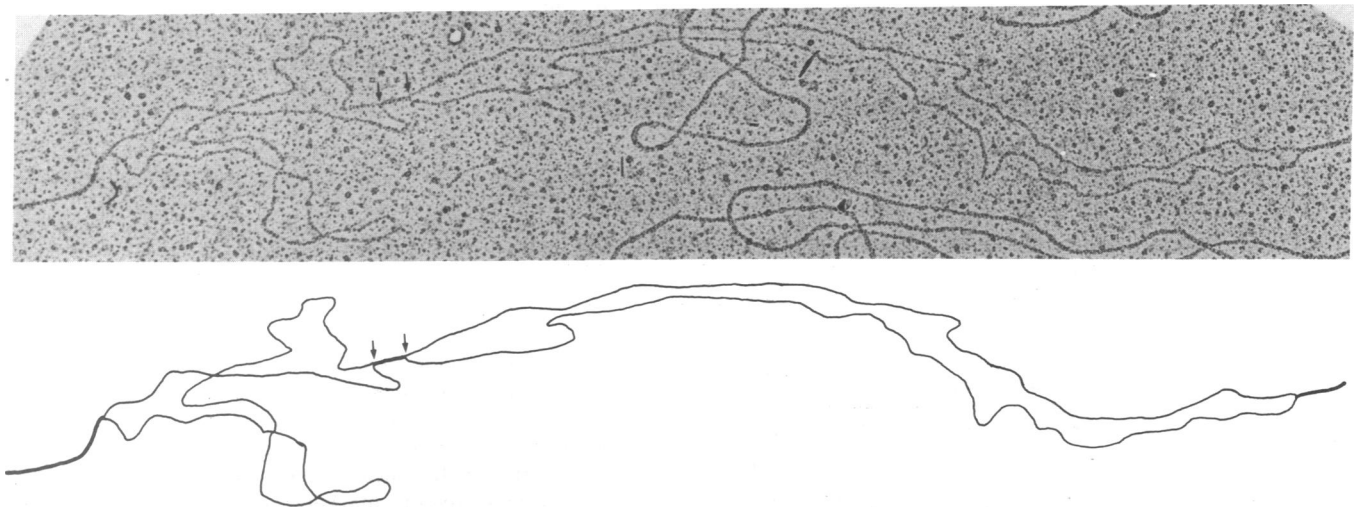


FIG. 5. Heteroduplex structure formed between λ phage hybrids containing mouse α- and β-globin genomic segments. Heteroduplex structures were formed between λgt*WES*·Mα1 and λgt*WES*·βG2 (α and β^{maj}) as indicated in *Materials and Methods*. Twenty-five heteroduplex molecules were visualized and measured. The only region of double-stranded DNA (homology) is indicated by the arrows in the electron micrograph and the line drawing. Homologous structures were seen only when fragments were aligned in the same 5′ → 3′ orientation with respect to their coding sequences. Twenty-one molecules were measured. The length of the homologous sequence is 165 ± 39 (±SD) bp and is located relative to the coding sequences of both genes as indicated in Fig. 3.

sequences identified in the studies of Davidson and Britten (40) and may be present in many segments of the genome. On the other hand, such a sequence might be restricted to a smaller set of genes, say those involved in the process of erythrodifferentiation. Both alternatives are interesting and both can be tested. The proximity of such shared sequences to genes of known function make them an especially useful substrate for elucidating the role that reiterated sequences might play in gene expression.

Note Added in Proof. We have recently located the site of the small intervening sequence in the α-globin gene by direct sequence determination (Y. Nishioka and P. Leder, unpublished results). It is as predicted in Table 1, strengthening the notion that all adult vertebrate globin genes, α and β, will be interrupted by two intervening sequences at analogous positions.

1. Russell, E. S. & McFarland, E. C. (1974) *Ann. N. Y. Acad. Sci.* **241,** 24–38.
2. Tilghman, S. M., Tiemeier, D. C., Polsky, F., Edgell, M. H., Seidman, J. G., Leder, A., Enquist, L. W., Norman, B. & Leder, P. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 4406–4410.
3. Leder, P., Tilghman, S. M., Tiemeier, D. C., Polsky, F. I., Seidman, J. G., Edgell, M. H., Enquist, L. W., Leder, A. & Norman, B. (1977) *Cold Spring Harbor Symp. Quant. Biol.* **42,** 915–920.
4. Tilghman, S. M., Tiemeier, D. C., Seidman, J. G., Peterlin, B. M., Sullivan, M., Maizel, J. V. & Leder, P. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 725–729.
5. Tiemeier, D. C., Tilghman, S. M., Polsky, F. I., Seidman, J. G., Leder, A., Edgell, M. H. & Leder, P. (1978) *Cell* **14,** 237–245.
6. Jeffreys, A. J. & Flavell, R. A. (1977) *Cell* **12,** 1077–1108.
7. Glover, D. M. & Hogness, D. S. (1977) *Cell* **10,** 167–176.
8. White, R. & Hogness, D. S. (1977) *Cell* **10,** 177–192.
9. Wellauer, P. K. & Dawid, I. B. (1977) *Cell* **10,** 193–212.
10. Berget, S. M., Moore, C. & Sharp, P. A. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 3171–3175.
11. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. (1977) *Cell* **12,** 1–8.
12. Kitchingman, G. R., Lai, S.-P. & Westphal, H. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 4392–4395.
13. Aloni, Y., Dhar, R., Laub, O., Horowitz, M. & Khoury, G. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 3686–3690.
14. Brack, C. & Tonegawa, S. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5652–5656.
15. Goodman, H. M., Olson, M. V. & Hall, B. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5453–5457.
16. Valenzuela, P., Venegas, A., Weinberg, F., Bishop, R. & Rutter, W. J. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 190–194.
17. Doel, M. T., Hoghton, M., Cook, E. A. & Carey, N. H. (1977) *Nucleic Acids Res.* **4,** 3701–3713.
18. Breathnach, R., Mandel, J. L. & Chambon, P. (1977) *Nature (London)* **270,** 314–319.
19. Lai, E., Woo, S.-L., Dugaiczyk, A., Catterall, J. & O'Malley, V. W. (1978) *Proc. Natl. Acad. Sci. USA* **74,** 2203–2209.
20. Weinstock, R., Sweet, R., Weise, M., Cedar, H. & Axel, R. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 1299–1303.
21. Tilghman, S. M., Curtis, P. J., Tiemeier, D. C., Leder, P. & Weissmann, C. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 1309–1313.
22. Kinniburgh, A. J., Mertz, J. E. & Ross, J. (1978) *Cell* **14,** 681–693.
23. Smith, K. & Lingrel, J. B. (1978) *Nucleic Acids Res.,* **5,** 3295–3302.
24. Fantohni, A., Bank, A. & Marks, P. A. (1967) *Science* **157,** 1327–1330.
25. Curtis, P. J., Mantei, N., van den Berg, J. & Weissmann, C. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 3184–3188.
26. Ross, J. (1976) *J. Mol. Biol.* **106,** 403–420.
27. Tiemeier, D. C., Tilghman, S. M. & Leder, P. (1977) *Gene* **2,** 173–191.
28. Polsky, F., Edgell, M., Seidman, J. & Leder, P. (1978) *Anal. Biochem.* **87,** 397–410.
29. Tiemeier, D., Enquist, L. & Leder, P. (1976) *Nature (London)* **263,** 526–527.
30. Benton, W. D. & Davis, R. W. (1977) *Science* **196,** 180–182.
31. Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517.
32. Rougeon, F. & Mach, B. (1977) *Gene* **1,** 229–239.
33. Maxam, A. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
34. Konkel, D. A., Tilghmann, S. M. & Leder, P. (1978) *Cell* **15,** 1125–1132.
35. Gilman, J. G. (1974) *Ann. N. Y. Acad. Sci.* **241,** 416–433.
36. Dayhoff, M. O. (1972) *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, DC).
37. van den Berg, J., van Ooyen, A., Mantei, N., Schamböck, A., Grosveld, G., Flavell, R. A. & Weissmann, C. (1978) *Nature (London)* **276,** 37–44.
38. Popp, R. A. (1967) *J. Mol. Biol.* **27,** 9–16.
39. Popp, R. A. & Bailiff, E. G. (1973) *Biochim. Biophys. Acta* **303,** 61–67.
40. Davidson, E. H. & Britten, R. J. (1973) *Q. Rev. Biol.* **48,** 565–613.