MOLECULAR CLONING OF THE HUMAN α-GLOBIN GENE FAMILY

Thesis by

Joyce Ellen Lauer

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1981

(Submitted July 24, 1980)

Quoted by Denise Levertov
in With Eyes at the Back of Our Heads

The true artist: draws out all from
his heart,
works with delight, makes things
with calm, with sagacity,
works like a true Toltec, composes
his objects, works dextrously,
invents,
arranges materials, adorns them,
makes them adjust

The carrion artist: works at
random, sneers at the people,
makes things opaque, brushes
across the surface of the face of
things,
works without care, defrauds
people, is a thief

Acknowledgements

I am grateful to Tom Maniatis for kindness, guidance, and conversations, and for teaching me about DNA and cloning. It has been a wonderful privilege to be a member of this lab.

Acknowledgements

I was supported by a training grant from the National Institutes of Health and received funds for the preparation of this thesis from the Jean Weigle Memorial Fund.

Abstract

During human development a series of α-like and β-like subunits of hemoglobin are produced. Five β-like polypeptides, embryonic (ε), fetal ($^{G}\gamma$, $^{A}\gamma$) and adult (δ, β), and two α-like polypeptides, embryonic (ζ) and adult (α), have been identified. A structural analysis of the gene family encoding the human α-like globins is described.

An improved gene isolation procedure was developed. This procedure makes it possible to isolate genes and their flanking sequences, and thus to directly determine the linkage relationship between genes, without requiring partial purification of genes. Large random genomic DNA fragments resulting from physical shear of DNA were joined to phage lambda vectors using synthetic DNA linkers. The resulting recombinant DNA molecules were packaged in vitro into viable phage to yield a collection of cloned overlapping DNA fragments which were then amplified to produce a permanent library of genomic DNA sequences. This library can be screened repeatedly using nucleic acid probes and an in situ plaque hybridization procedure in order to isolate many different genes.

Clones containing the duplicated human α-globin genes (α1 and α2) were isolated from a library of human DNA. Also present on these clones are an α-like pseudogene (ψα1) and an embryonic α-like gene (ζ1) which were identified by blot hybridization experiments and DNA sequence analysis. Genomic blotting using the ζ1 coding sequence as probe identified a second embryonic gene (ζ2). The ζ2 gene was isolated by cloning a DNA fragment which overlaps the clones containing the other α-like genes. All five genes are transcribed from the same DNA strand and are arranged in the order 5'-ζ2-ζ1-ψα1-α2-α1-3'.

Comparison of α1 and α2 by restriction mapping and heteroduplex analysis of DNA fragments containing α1 or α2 plus 5' flanking sequences demonstrated that each gene is located within an approximately 4 kb region of homology interrupted by two short regions of nonhomology. The association of these large blocks of homology with

genes which are thought to have duplicated long ago suggests the existence of a mechanism for sequence matching.

Two types of deletions invariably occur during propagation of clones containing α1 and α2. The breakpoints of these two types of deletions are located within the two blocks of α1-α2 homology. The positions and the precise lengths of these deletions indicate that deletion occurs by homologous but unequal crossing-over between corresponding regions of α1 and α2. The lengths and positions of these deletions are indistinguishable from those of the two types of deletions which are associated with α-thalassemia 2, suggesting that this common genetic disease results from homologous but unequal crossing-over between regions within and/or surrounding the adult α-globin genes.

Table of Contents

Introduction

The human globin gene family is a paradigm for studying differential gene activity during development and the molecular basis of inherited disorders in gene expression. Globin polypeptides and genes from a number of species have been characterized (Tilghman et al., 1977; Lawn et al., 1978; Dodgson et al., 1979; Lacy et al., 1979) and cloned globin genes from any of these species can be manipulated in various ways in order to study gene expression (Mulligan et al., 1979; Weil et al., 1979; Wigler et al., 1979; Manley et al., 1980). Unique to the human globin gene system, however, is the existence of a large number of well-characterized natural mutants in globin gene expression (for review, see Maniatis et al., 1980, which is presented as Chapter 3 of this thesis). Mutants analyzed thus far include examples of defects in transcription, RNA processing and translation, although in most cases the exact DNA sequence alteration responsible for the defect remains to be identified.

Particularly intriguing is a class of mutants, of which numerous examples are known in both the α-like and β-like globin gene families, in which deletion of DNA in one region of a cluster of linked globin genes disrupts the regulation of a globin gene located many kilobases (kb) distant from the deletion. Whereas DNA sequence alterations which affect globin gene transcription or globin translation are perhaps analogous to well-understood prokaryotic examples, this class of deletion mutants suggests the existence of mechanisms of gene regulation unique to eukaryotes (for discussion and review, see Maniatis et al., 1980). As a first step towards studying the various aspects of eukaryotic gene regulation, we have isolated the human α-like and β-like globin gene clusters. This thesis describes our approach to gene isolation, and structural characterization of the α-like globin gene cluster.

The Ontogeny of Human Globin Gene Expression

The α-like and β-like subunits of hemoglobin (Hb) are encoded by a small group of genes which are expressed sequentially during development (Weatherall and Clegg, 1979). The earliest embryonic hemoglobin tetramer, Gower 1, consists of ε (β-like) and ζ (α-like) polypeptide chains. Beginning at approximately eight weeks of gestation, the embryonic chains are gradually replaced by the adult α-globin chain and two different fetal β-like chains, designated $^{G}\gamma$ and $^{A}\gamma$. The γ chains differ only in the presence of glycine or alanine at position 136, respectively. During the transition period between embryonic and fetal development, Hb Gower 2 ($\alpha_2\varepsilon_2$) and Hb Portland ($\zeta_2\gamma_2$) are detected. Hb F ($\alpha_2\gamma_2$) eventually becomes the predominant Hb tetramer throughout the remainder of fetal life. Beginning just prior to birth, the γ-globin chains are gradually replaced by the adult β- and δ-globin polypeptides. At six months after birth, 97-98% of the hemoglobin is Hb A ($\alpha_2\beta_2$), while Hb A_2 ($\alpha_2\delta_2$) accounts for approximately 2%.

α-Like Globin Gene Number

The existence of multiple human α-like globin genes was established on the basis of genetic studies, analysis of the structure and function of embryonic hemoglobins, and the isolation of the α-like globin gene cluster by molecular cloning. Genetic and structural studies of α-globin polypeptide chain variants provided the first evidence for duplication of the human α-globin genes (Lehmann and Carrell, 1968; Brimhall et al., 1970; Lie-Injo et al., 1974; for reviews see Weatherall and Clegg, 1976; Bunn et al., 1977; Weatherall and Clegg, 1979). The hypothesis of α-globin gene duplication was strongly supported by cDNA/DNA solution hybridization experiments (Kan et al., 1975). In most cases the two genes are expressed at approximately the same level, since in individuals heterozygous for an α-chain variant, the variant polypeptide usually represents approximately one-quarter of the α-globin protein (Lehmann and Carrell, 1968). Since only one α-globin amino acid sequence is found in

normal individuals (Dayhoff, 1972), the two α-globin genes must encode identical polypeptides.

Until relatively recently, it was thought that there was only one α-type globin polypeptide, which was produced throughout all stages of human development. This contrasted to the β-like globin gene family, where two switches in globin gene expression, from embryonic to fetal and from fetal to adult, had been identified. The misconception about the number of α-type globin polypeptides resulted from the hypothesis that the earliest embryonic hemoglobin, Gower 1, consisted of a tetramer of epsilon chains (ε_4) (Huehns et al., 1961, 1964). The embryonic ζ-globin polypeptide chain was first identified as a component of Hb Portland ($\zeta_2\gamma_2$) (Capp et al., 1967, 1970). Peptide comparisons suggested that ζ is an α-like chain (Kamuzora and Lehmann, 1975; Huehns and Farooqui, 1975). On the basis of the oxygen dissociation behavior of embryonic blood, it was proposed that the structure of Gower 1 is in fact $\zeta_2\varepsilon_2$ (Huehns and Farooqui, 1975). Peptide analysis of purified Gower 1 confirmed this hypothesis (Gale et al., 1979). Although only one ζ-globin polypeptide sequence has been identified (J. Clegg, personal communication), other evidence indicates the existence of two highly homologous ζ-globin genes (see below).

Genetic Disorders in α-Globin Gene Expression

Inherited disorders in α-globin gene expression (α-thalassemias) usually result from α-globin gene deletion. One or both α-globin genes on a chromosome can be deleted, and diploid combinations of the "single-gene" and "zero-gene" chromosomes cause a variety of syndromes whose severity reflects the number of α-globin genes affected (for review, see Maniatis et al., 1980). The apparent cause of the deletions which generate the "single-gene" or α-thalassemia 2 chromosome, and the association of "zero-gene" chromosomes with alterations in ζ-globin gene regulation, are described below.

Gene Isolation by Molecular Cloning

Globin genes were chosen as a model system in which to study the regulation of gene expression in eukaryotes. The initial phase of the research described in this thesis was the development of a modified procedure which makes it possible to efficiently isolate genes and their flanking sequences and to determine the linkage relationship between genes. In this procedure, DNA is fragmented randomly and high molecular weight (16-20 kb) DNA fragments are isolated. These are inserted into a bacteriophage lambda vector using DNA linkers and the recombinant DNA molecules are packaged in vitro into viable phage to yield a collection of cloned overlapping DNA fragments. These recombinants are amplified in order to establish a permanent library of DNA segments representing the entire genome.

For such a library to be truly representative, it is desirable that the initial DNA fragmentation be as random as possible, and this is best accomplished by physically shearing the DNA. Sheared DNA molecules are likely to have single-stranded tails which must be removed using S1 nuclease before linkers can be attached by flush-end ligation. I constructed a library of Drosophila melanogaster DNA by this method. The construction and characterization of the Drosophila library are described in Chapter 1 (Maniatis et al., 1978).

Trimming DNA molecules using S1 nuclease is a relatively inefficient process, leading to low cloning efficiency. As a result, subsequent library construction in our lab employed combined partial digestion with several restriction enzymes that generate flush ends which can be efficiently ligated to DNA linkers. The human DNA library from which α-globin genes were isolated was constructed by the partial digestion method (Lawn et al., 1978; Maniatis et al., 1978).

There are several advantages to the library approach to gene isolation: 1) If numerous gene-specific probes or a mixed (e.g., cDNA) probe is available, a number of genes can be isolated following a single cloning step; 2) Because the genomic DNA

inserts are large (16-20 kb), one can isolate linked genes for which no probe is available by their presence in the same clone as a gene for which there is a probe. The isolation of ζ1 and of ψα1 illustrate this (see below); 3) By using the terminal fragment of a previously-isolated clone, one can rescreen a library in order to isolate an adjacent region of DNA. The isolation of ζ2 by screening with the terminal fragment of λHαG1 illustrates this. (Although problems with the original human library necessitated construction of a second, Bgl II library, this point is still valid.)

Structural Analysis of the Human α-Like Globin Gene Cluster

Structural analysis of the human α-like globin gene cluster, as derived from cloned DNA, is presented in Chapter 2 (Lauer et al., 1980, and Appendix). The results of this analysis may be summarized as follows.

1. The α-like genes are linked in the order 5'-ζ2-ζ1-ψα1-α2-α1-3' which parallels the 5' to 3' embryonic-fetal-adult order of genes within the β-like globin gene cluster (Fritsch et al., 1980; Figure 1).

2. ψα1 is a pseudogene for which no globin polypeptide can be identified; it probably resulted from gene duplication followed by divergence. Analysis of the complete nucleotide sequence of ψα1 (Proudfoot and Maniatis, 1980) reveals sequence alterations which would prevent the production of an α-globin polypeptide.

3. Fine-structure mapping indicated extremely close sequence homology, except for a small insertion, between the coding, intervening and immediately-adjacent flanking regions of α1 and α2. Furthermore, the repeated positions of certain more distant restriction enzyme sites, and the invariable occurrence of two types of deletions during propagation of clones containing α1 and α2, suggested the existence of additional regions of homology within the several kb 5' to each gene. To verify this, I collaborated with J. Shen to examine heteroduplexes between DNA fragments containing α1 or α2 plus 5' flanking sequences. Each gene is located within an approximately 4 kb region of homology interrupted by two short regions of non-

homology. The association of these large blocks of homology with genes which are thought to have duplicated long ago suggests the existence of a mechanism for sequence matching.

4. The breakpoints of the two types of deletions which occur in cloned DNA are located within the two blocks of α1-α2 homology. The positions and the precise lengths of these deletions indicate that deletion occurs by homologous but unequal crossing-over between corresponding regions of α1 and α2. The lengths and positions of these deletions are indistinguishable from those of the two types of deletions which are associated with α-thalassemia 2 (Embury et al., 1980), suggesting that this common genetic disease results from homologous but unequal crossing-over between regions within and/or surrounding the adult α-globin genes.

5. Using a fragment of the embryonic ζ1 gene to probe genomic blots, I identified a second embryonic gene, ζ2, located nearly 12 kb 5' to ζ1. Hybridization-melting experiments indicated very close sequence homology between these two genes, which recalls the high degree of α1-α2 homology.

6. ζ1 had been identified by the correspondence between its DNA sequence and the sequence of ζ polypeptide. Using my genomic blotting map of ζ2, this second gene was shown to be functional by blotting DNA from an infant with the α-thalassemia syndrome, hydrops fetalis (Pressley et al., 1980). In this individual, there is a homozygous deletion of α1, α2, ψα1 and ζ1, but ζ2 remains and ζ polypeptide is produced in large amount. Ordinarily ζ-globin is not detectable after 10 weeks gestation. The persistence of ζ-globin production in cases of hydrops fetalis appears to be the α-locus analogue of persistent expression of γ-globin caused by the β-locus disease HPFH. In this syndrome, a DNA rearrangement many kb 3' to the γ genes removes a control which would ordinarily switch off these genes during the normal developmental program (for review, see Maniatis et al., 1980).

References

Brimhall, B., Hollan, S., Jones, R., Koler, R., Stocklen, Z. and Szelenyi, J. (1970). Multiple alpha-chain loci for human hemoglobin. Clin. Res. **18**, 184.

Bunn, H. F., Forget, B. G. and Ranney, H. M. (1977). Human Hemoglobins. Philadelphia: W. B. Saunders Co.

Capp, G., Rigas, D. and Jones, R. (1967). Hemoglobin Portland 1: a new human hemoglobin unique in structure. Science **157**, 65-66.

Capp, G., Rigas, D. and Jones, R. (1970). Evidence for a new hemoglobin chain (ζ-chain). Nature **228**, 278-280.

Dayhoff, M. O. (1972). Atlas of Protein Sequence and Structure. Washington, D.C.: National Biomedical Research Foundation.

Dodgson, J. B., Strommer, J. and Engel, J. D. (1979). Isolation of the chicken β-globin gene and a linked embryonic β-like globin gene from a chicken DNA recombinant library. Cell **17**, 879-887.

Embury, S. H., Miller, J. A., Cozy, A. M., Kan, Y. W., Chan, V. and Todd, D. (1980). Two different molecular organizations account for the single α-globin gene of the α-thalassemia 2 genotype. Manuscript submitted.

Fritsch, E. F., Lawn, R. M. and Maniatis, T. (1980). Molecular cloning and characterization of the human β-like globin gene cluster. Cell **19**, 959-972.

Gale, R., Clegg, J. and Huehns, E. (1979). Human embryonic hemoglobins Gower 1 and Gower 2. Nature **280**, 162-164.

Huehns, E., Dance, N., Beaven, G., Keil, J., Hecht, F. and Motulsky, A. (1964) Human embryonic hemoglobins. Nature **201**, 1095-1097.

Huehns, E., Flynn, F., Butler, E. and Beaven, G. (1961). Two new hemoglobin variants in a very young human embryo. Nature **189**, 496-497.

Huehns, E. and Farooqui, A. (1975). Oxygen dissociation properties of human embryonic red cells. Nature **254**, 335-337.

Kamuzora, H. and Lehmann, H. (1975). Human embryonic hemoglobins including a comparison by homology of the human ζ and α chains. Nature **256**, 511-513.

Kan, Y. W., Dozy, A., Varmus, H., Taylor, J., Holland, J., Lie-Injo, L., Ganesan, J. and Todd, D. (1975). Deletion of α-globin genes in hemoglobin H disease demonstrates multiple α-globin structural loci. Nature **255**: 255-256.

Lacy, E., Hardison, R. C., Quon, D. and Maniatis, T. (1979). The linkage arrangement of four rabbit β-like globin genes. Cell **18**, 1273-1283.

Lauer, J., Shen, C.-K. J. and Maniatis, T. (1980). The chromosomal arrangement of human α-like globin genes: sequence homology and α-globin gene deletions. Cell **20**, 119-130.

Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. and Maniatis, T. (1978). The isolation and characterization of linked δ- and β-globin genes from a cloned library of human DNA. Cell **15**, 1157-1174.

Lehmann, H. and Carrell, R. (1968). Differences between α- and β-chain mutants of human hemoglobin and between α- and β-thalassemia. Possible duplication of the α-chain gene. Br. Med. J. **4**, 748-750.

Lie-Injo, L., Ganesan, J., Clegg, J. and Weatherall, D. (1974). Homozygous state for Hb Constant Spring (slow moving Hb X components). Blood **43**, 251-259.

Maniatis, T., Fritsch, E. F., Lauer, J. and Lawn, R. M. (1980). The molecular genetics of human hemoglobins. Ann. Rev. Genetics, in press.

Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. and Efstratiadis, A. (1978). The isolation of structural genes from libraries of eucaryotic DNA. Cell **15**, 687-701.

Manley, J. L., Fire, A., Cano, A., Sharp, P. A. and Gefter, M. L. (1980). DNA-dependent transcription of adenovirus genes in a soluble whole-cell extract. Proc. Natl. Acad. Sci. USA, in press.

Mulligan, R. C., Howard, B. H. and Berg, P. (1979). Synthesis of rabbit β-globin in cultured monkey kidney cells following infection with a SV40 β-globin recombinant clone. Nature **277**, 108-114.

Pressley, L., Higgs, D., Clegg, J. and Weaterhall, D. (1980). Gene deletions in α-thalassemia prove that the 5' ζ locus is functional. Proc. Natl. Acad. Sci. USA, in press.

Proudfoot, N. J. and Maniatis, T. (1980). The structure of a human α-globin pseudogene and its relationship to α-globin gene duplication. Cell, submitted.

Tilghman, S. M., Tiemeier, D. C., Polsky, F., Edgell, M. H., Seidman, J. G., Leder, A., Enquist, L. W., Norman, B. and Leder, P. (1977). Cloning specific segments of the mammalian genome: bacteriophage λ containing mouse globin and surrounding sequences. Proc. Natl. Acad. Sci. USA **74**, 4406-4410.

Weatherall, D. and Clegg, J. (1976). Molecular genetics of human hemoglobin. Ann. Rev. Genetics **10**, 157-178.

Weatherall, D. J. and Clegg, J. B. (1979). Recent developments in the molecular genetics of human hemoglobin. Cell **16**, 467-479.

Weil, P. A., Luse, D. S., Segall, J. and Roeder, R. G. (1979). Selective and accurate initiation of transcription at the Ad2 major late promotor in a soluble system dependent on purified RNA polymerase II and DNA. Cell **18**, 469-484.

Wigler, M., Sweet, R., Sim, G. K., Wold, B., Pellicer, A., Lacy, E., Maniatis, T., Silverstein, S. and Axel, R. (1979). Transformation of mammalian cells with genes from procaryotes and eucaryotes. Cell **16**, 777-785.

Figure 1. Linkage Arrangement of Human β-Like and α-Like Globin Genes.

The positions of the embryonic (ε), fetal ($^{G}\gamma$, $^{A}\gamma$) and adult (δ, β) β-like globin genes and of the two β-like pseudogenes (ψβ1, ψβ2) are shown on the upper line. The positions of the embryonic (ζ2, ζ1) and adult (α2, α1) α-like globin genes and of the α-like pseudogene (ψα1) are shown on the lower line. For each gene, the black and white boxes represent the coding (exon) and noncoding (intron) sequences, respectively.

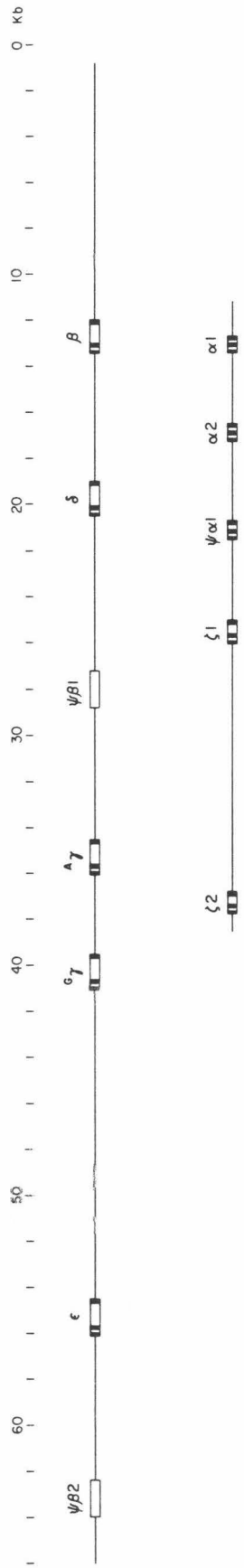


Figure 1

Chapter 1

The isolation of structural genes from libraries of eukaryotic DNA

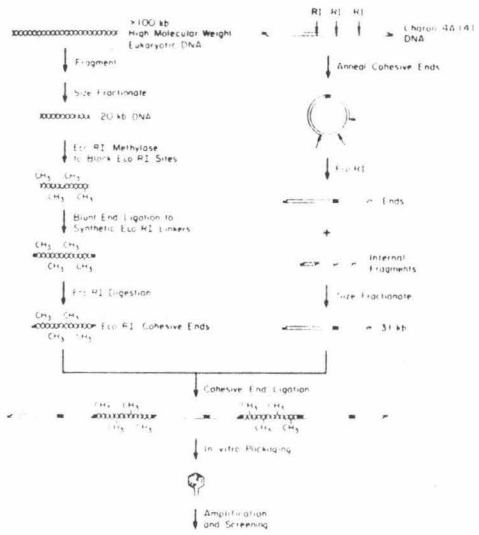


Figure 1. Schematic Diagram Illustrating the Strategy Used to Construct Libraries of Random Eucaryotic DNA Fragments

al., 1977). For this purpose, high molecular weight eucaryotic DNA was fragmented and fractionated by size to obtain molecules of approximately 20 kb. This reduces the number of plaques required for screening an entire genome and minimizes the possibility of small eucaryotic DNA fragments joining to each other and subsequently to the vector. Although a nonlimit Eco RI digestion is the most direct method of obtaining large DNA fragments with Eco RI cohesive ends (Glover et al., 1975), we were concerned that a nonrandom distribution of Eco RI sites within and adjacent to particular structural genes might result in their selective loss from the population of size-fractionated DNA. We therefore used two methods to obtain random eucaryotic DNA fragments, both of which generate molecules with blunt ends that can be joined to Eco RI linkers and subsequently inserted into the cloning vector.

In one method, DNA was fragmented by shearing and the ends were trimmed using S1 nuclease. Unfortunately, this method of preparing blunt-ended DNA fragments is rather inefficient (see also Scheller et al., 1977b; Seeburg et al., 1977). An alternative, and in fact more efficient, approach is to perform a nonlimit restriction endonuclease digestion with two enzymes that cleave frequently and generate blunt-ended molecules – that is, Hae III (GGCC) and Alu I (AGCT) (Roberts, 1976). The combined activities of these two enzymes under conditions of nonlimit digestion should generate a collection of large fragments approaching the random sequence representation of sheared DNA more closely than the products of a nonlimit Eco RI digestion; the reason for this is that a recognition site for each of the two enzymes on DNA should occur once every 256 (4^4) nucleotides, whereas Eco RI recognition sites should be present an average of once every 4096 (4^6) nucleotides (uncorrected for base composition). The greater the number of possible cleavage sites, the larger the number of possible ways of generating a 20 kb fragment by nonlimit digestion of a higher molecular weight molecule, and thus the more random the collection of resulting fragments.

The results of an electrophoretic analysis of nonlimit Hae III and Alu I digests of rabbit DNA are shown in Figure 2. Three reactions with different enzyme to DNA ratios were performed separately for each enzyme, and the digestion products containing a substantial fraction of 20 kb fragments were pooled and fractionated on a 10–40% sucrose gradient.

Modification of Eco RI Sites in Eucaryotic DNA

Since the synthetic Eco RI linkers attached to eucaryotic DNA must be cleaved by Eco RI to generate cohesive ends, the Eco RI sites within the DNA fragments must first be rendered resistant to cleavage. This was accomplished by reacting the eucaryotic DNA with the Eco RI modification-methylase in the presence of S-adenosyl-L-methionine (Greene et al., 1975). Small aliquots of the methylation reaction mixture were taken before and after the addition of methylase and mixed with λ DNA to monitor the extent of methylation; following electrophoresis, the appearance of discrete λ DNA bands reveals incomplete methylation. The results of a typical assay are shown in Figure 3. The mixture of λ and eucaryotic DNAs taken before the addition of methylase is digested to completion, while the methylated DNA is totally resistant to Eco RI cleavage.

Joining of Synthetic DNA Linkers to Eucaryotic DNA

Duplex DNA linker molecules bearing restriction endonuclease recognition sites have been chemically synthesized (Bahl et al., 1977; Scheller et al., 1977a) and used to insert a number of different DNA molecules into plasmids (Heyneker et al., 1976; Scheller et al., 1977b; Shine et al., 1977; Ullrich et al., 1977). In each case, the linkers were first blunt-end ligated to the DNA of interest and then digested with the appropriate restriction enzyme to generate molecules with cohesive termini that could be joined to a vector with complementary ends.

We attached Eco RI dodecameric linkers to in vitro methylated eucaryotic DNA using T4 ligase and assayed for blunt-end ligation by the formation

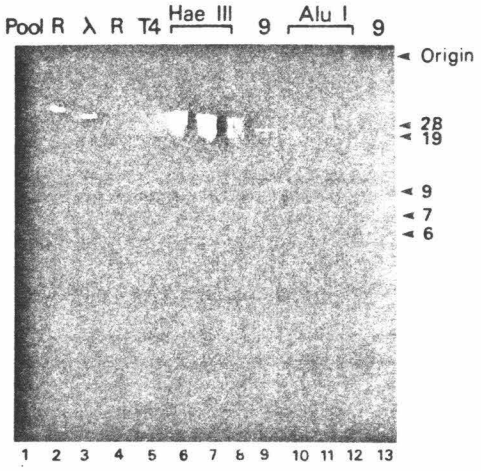


Figure 2. Agarose Gel Electrophoresis of Nonlimit Hae III and Alu I Digests of Rabbit Liver DNA

High molecular weight rabbit DNA was digested with different amounts of Hae III or Alu I and electrophoresed on a horizontal 0.5% agarose gel containing 0.5 μg/ml ethidium bromide, and the DNA was photographed on a short wave ultraviolet transilluminator (Sharp et al., 1973). (1, Pool) the pooled DNA from the Hae III and Alu I digests of slots 6-8 and 10-12; (2 and 4, R) undigested rabbit liver DNA; (3, λ) undigested wild-type λ DNA (49.4 kb) (Blattner et al., 1977); (5, T4) undigested T4 DNA (171 kb) (Kim and Davidson, 1974); (6-8, Hae III) Hae III-digested rabbit liver DNA; (9 and 13, 9) size markers, Eco RI-digested Charon 9 DNA; (10-12, Alu I) Alu I-digested rabbit DNA. The approximate sizes of the markers in kb are indicated on the side of the figure.

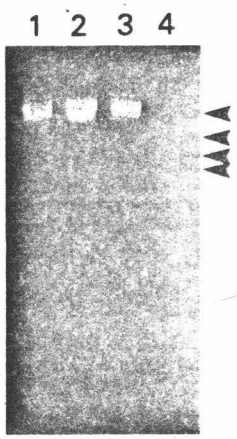


Figure 3. Assay for Methylation of Eco RI Sites in Eucaryotic DNA

A mixture of eucaryotic DNA and Charon 4A DNA was reacted with Eco RI methylase as described in Experimental Procedures and analyzed by agarose gel electrophoresis. (1) DNA mixture plus Eco RI methylase, minus Eco RI nuclease; (2) DNA mixture plus Eco RI methylase, plus Eco RI nuclease; (3) DNA mixture minus Eco RI methylase, minus Eco RI nuclease; (4) DNA mixture minus Eco RI methylase, plus Eco RI nuclease. The arrows indicate the positions of phage λ Eco RI fragments.

of linker oligomers. After ligation, the DNA was sedimented through a 10-40% sucrose gradient (or passed over a Sepharose 2B column) to remove unincorporated linker oligomers. Without this step, the digestion of linkers attached to eucaryotic DNA is difficult, since linker oligomers not incorporated into eucaryotic DNA compete for Eco RI. Eco RI digestion of the DNA thus purified provides large fragments with Eco RI cohesive ends that can be joined to the vector DNA.

Ligation of Eucaryotic DNA to Charon 4 DNA

Charon 4 contains three Eco RI cleavage sites (Figure 1). Digestion with Eco RI produces two internal fragments with genes nonessential for phage viability and two end fragments. To maximize the efficiency of in vitro recombination and to minimize the number of nonrecombinant phage in the library, we removed the internal fragments by sucrose gradient centrifugation. After annealing the cohesive ends of Charon 4 DNA and Eco RI digestion, excellent separation of the cohered λ DNA arms (31 kb) and the internal fragments (7 and 8 kb) can be achieved. To determine the ratio of vector DNA to eucaryotic DNA that produces the smallest number of background plaques without reducing the absolute yield of recombinants, we ligated varying amounts of eucaryotic DNA with a constant amount of purified λ DNA arms, packaged the DNA into phage and determined the number of plaque-forming units (pfu). Using the established optimal ratio, ligation reactions were performed at high DNA concentrations to minimize intramolecular joining and to maximize the formation of concatemeric DNA recombinants, the substrate for in vitro DNA packaging.

In Vitro Packaging of DNA into Phage Particles

The number of independently derived phage recombinants (library size) required for a 99% probability of finding any given single-copy sequence in the library can be calculated if the average size of the eucaryotic DNA inserts is known (Clarke and Carbon, 1976). Thus 7×10^5 recombinants are required for a mammalian DNA library of 20 kb DNA inserts. Using the $CaCl_2$ transfection procedure of Mandel and Higa (1970), approximately $2\text{-}10 \times 10^3$ pfu/μg of cleaved and religated λ vector DNA are obtained (in contrast to 10^6 pfu/μg of intact λ DNA). Thus approximately 100-400 μg of DNA fragments attached to synthetic linkers are needed for the construction of a complete library, an amount difficult to obtain. Fortunately, efficiencies of 2 and 0.15×10^7 pfu/μg have been achieved for intact and religated λ DNAs, respectively, using in

vitro packaging procedures (Hohn and Murray, 1977; Sternberg et al., 1977).

To use this technique for mammalian DNA cloning, it was necessary to demonstrate that the procedure does not alter the biological containment features of Charon 4A. Most in vitro packaging procedures involve the temperature induction of λ lysogens which carry amber mutations in different genes required for packaging. Each lysogen alone is incapable of producing viable phage particles, but mixed lysates of the two strains complement in vitro to convert λ DNA into a plaque-forming particle. If hybrid phage DNA carrying eucaryotic sequences is added to a mixed extract, there is a possibility that endogenous prophage DNA will recombine with the hybrid DNA in vitro or during subsequent in vivo amplification to produce DNA carrying the wild-type markers of the prophage.

Sternberg et al. (1977) have developed a packaging system that minimizes these problems. First, the prophage of their strains carry the λ*b2* mutation, which removes part of the attachment site and therefore prevents prophage excision after induction. Second, the lysogens are recombination-deficient (the prophage is red and the host is recA). To reduce further the chance of prophage DNA packaging and recombination we ultraviolet-irradiated the cells prior to their use in in vitro packaging reactions. Hohn and Murray (1977) found that both recombination and prophage packaging in their extracts could be suppressed by irradiation with ultraviolet light. To obtain EK2 certification for the system of Sternberg et al. (1977), we examined extracts derived from ultraviolet-irradiated cells for recombination and for the presence of in vitro packaged prophage. A procedure for the preparation and testing of in vitro packaging extracts for EK2 experiments is presented in Experimental Procedures.

Following in vitro packaging of recombinant DNA, the resulting phage particles were separated from cellular debris by sedimentation on a CsCl step gradient. This procedure concentrates the phage and removes material present in the extracts which inhibits the growth of bacterial cells.

Amplification of Libraries

An essential feature of our strategy for gene isolation is to establish a permanent library that can be repeatedly screened. To achieve this, it is necessary to amplify the in vitro packaged recombinant phage and to store the library in the form of a plate lysate. There is, of course, a risk that a particular recombinant phage will exhibit a growth disadvantage and will be eliminated from the library during amplification. It is therefore important to minimize competitive growth. We accomplished this by plating the in vitro packaged phage on agar plates at low density (10,000 pfu per 15 cm diameter plate). In this manner, the phage are amplified approximately 0.1 to 1 × 10^6 fold and recovered as a plate lysate.

Efficiency of Cloning Eucaryotic DNA

The efficiency of eucaryotic DNA cloning under our conditions depends primarily upon the quality of the in vitro packaging extracts (which varies between preparations) and the fraction of DNA fragments bearing Eco RI linkers. The latter is determined by the fraction of molecules with two blunt ends which, in turn, depends upon the method of preparation. The efficiency of our extracts prepared for EK1 experiments varies from 2–20 × 10^7 pfu/μg of intact λ DNA. Extracts prepared for EK2 experiments are consistently less efficient, varying from 0.4–5 × 10^7 pfu/μg of intact λ DNA. When using cleaved and religated DNA, this efficiency drops. Even the lowest efficiency observed (3.8 × 10^4 pfu/μg of DNA in the case of the rabbit library), however, is higher than that reported for cloning religated λ DNA by transfection (2–10 × 10^3 pfu/μg; Thomas, Cameron and Davis, 1974; Hohn and Murray, 1977).

Analysis of the data of Table 1, which describes the characterization of the libraries, reveals that DNA fragments produced by nonlimit endonuclease digestion (rabbit library) are cloned more efficiently than those produced by shearing followed by S1 nuclease treatment (Drosophila and silkmoth libraries). When the number of plaques formed per μg of eucaryotic DNA is normalized to the particular efficiency of the in vitro packaging extract used to construct each library, it becomes evident that the rabbit DNA was cloned 15.8 and 3.4 times more efficiently than the Drosophila and silkmoth DNAs, respectively. (We cannot, of course, rule out the unlikely possibility that the cloning efficiency is genome-specific).

Fraction of Clones Containing Eucaryotic DNA

To test whether a library containing the entire complement of genomic DNA can be constructed, we have measured the number of recombinants carrying eucaryotic sequences, the average size of eucaryotic DNA inserts and the single-copy DNA sequence representation in one of the libraries (Drosophila).

To estimate the number of recombinant phages in each library, it was necessary to determine the number of background (nonrecombinant) phage present. The number of background plaques was minimized by separating the annealed cohesive ends from the internal λ DNA fragments. When the purified cohesive ends of Charon 4 were ligated without the addition of eucaryotic DNA, however, a small number of pfu (on the order of a few percent

Table 1. Characterization of Libraries

Library	1 Efficiency of Extract	2 Plaques per μg of Eucaryotic DNA	3 Relative Efficiency of Cloning[a]	4 % Blue Plaques after Amplification	5 Total Number of Independent Recombinant Phage Recovered	6 Mean Length of Eucaryotic DNA	7 Number of Recombinant Phage Required for a "Complete" Library[b]
Drosophila	1×10^8	6.0×10^4	1	3.0	6.0×10^6	16 kb	4.8×10^4
Silkmoth	2×10^8	5.6×10^6	4.6	7.0	2.8×10^6	19 kb	2.4×10^6
Rabbit	4×10^8	3.8×10^4	15.8	2.6	7.8×10^6	17 kb	8.1×10^6

[a] This number was determined by dividing the number of column 2 by that in column 1 and normalizing to the value calculated for the Drosophila library.

[b] Calculated as described by Clarke and Carbon (1976) using the values in column 6 and assuming genome sizes of 1.65×10^8 bp for Drosophila (Rudkin, 1972), and 1×10^9 and 3×10^9 bp for silkmoth and rabbit, respectively (J. Yeh, L. Villa-Komaroff and A. Efstratiadis, unpublished results).

of the final number of plaques in the libraries) was recovered from the in vitro packaging reaction. The number of nonrecombinant phage in the libraries can also be estimated by an indicator plate assay. One of the internal Charon 4 fragments carries the E. coli lactose operator-promoter region and the gene for β-galactosidase. The presence of this fragment in library phage DNA can be detected by plating the phage on a lawn of lac⁻ E. coli grown on an Xgal indicator plate. Phage carrying an intact β-galactosidase gene produce blue plaques under these conditions (for a discussion of this assay, see Blattner et al., 1977). The number of blue (nonrecombinant) plaques observed after amplification is small (Table 1), indicating that most of the library phage carry eucaryotic DNA.

An independent, nonquantitative estimate of the degree of nonrecombinant phage contamination of the libraries can be obtained by determining the amount of internal Charon 4 DNA fragments in library DNA. This was accomplished by growing an aliquot of each library in liquid culture, purifying recombinant phage DNA and digesting the DNA with Eco RI. Figure 4 (lanes 2 and 4) shows the results of such an analysis of Drosophila and rabbit library DNAs. In addition to the left and right arms of the Charon 4 DNA, a characteristic smear of restriction endonuclease-digested eucaryotic DNA can be observed in the two library DNAs (compare to lanes 1 and 5 of Figure 4). The gel was intentionally overloaded to show that a small number of contaminating internal phage DNA fragments are present in both libraries. A similar result was obtained with the silkmoth library (not shown). All our assays together clearly demonstrate that most of the phage in each library contain eucaryotic DNA.

Mean Length of Eucaryotic DNA in the Libraries

The number of independent phage recombinants required for a complete library depends upon the average size of the cloned eucaryotic DNA inserts, which we estimated for each of the three libraries by CsCl sedimentation equilibrium analysis. Since the amount of protein in different λ phage is constant and the buoyant density depends upon the DNA/protein ratio, the distribution of DNA sizes in a λ phage population can be determined by measuring the distribution of phage in a CsCl density gradient (Weigle, Meselson and Paigen, 1959; Davidson and Szybalsky, 1971; Bellet, Busse and Baldwin, 1971). Figure 5 shows the results of a CsCl density gradient analysis of rabbit library phage. The density of each fraction was determined by its position in the gradient relative to two λ phage density markers. The average size of the rabbit DNA inserts calculated from the midpoint of the curve of Figure 5 is 17 kb, resulting in recombinant Charon DNA molecules whose average size is 97% of that of wild-type λ DNA. Similar analyses of the Drosophila (W. Bender and D. S. Hogness, personal communication) and silkmoth libraries yielded insert sizes of 16 and 19 kb, respectively.

Knowing the approximate size of the eucaryotic DNA fragments carried in each library and the complexity of each haploid genome (Table 1), we can calculate the number of independent recombinant phage needed to find any given single-copy sequence in the library with a probability of 0.99 (Clarke and Carbon, 1976), assuming that the entire genome consists of single-copy DNA sequences. Thus a 99% complete library of Drosophila, silkmoth or rabbit DNA would consist of 4.8×10^4, 2.4×10^5 or 8.1×10^5 recombinant phage, respectively. By comparing the theoretical library sizes to the actual ones (Table 1), we conclude that our libraries are "complete."

Sequence Representation in Library DNA

To determine whether a significant fraction of single-copy sequences is lost during amplification of

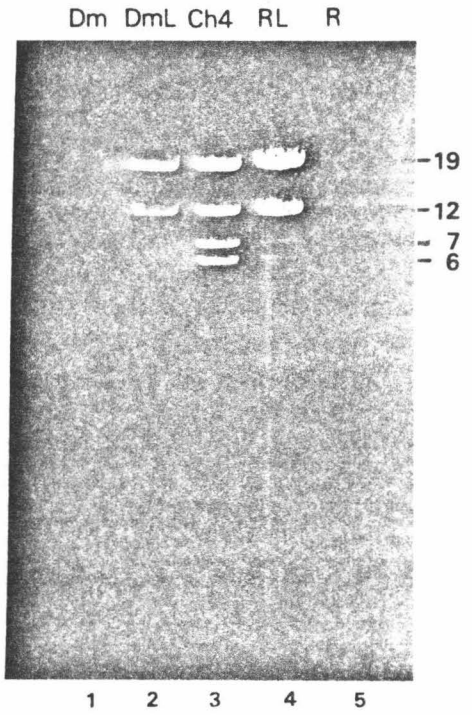


Figure 4. Presence of Internal Charon 4 DNA Fragments in the Drosophila and Rabbit Libraries

Aliquots of the Drosophila and rabbit libraries were each grown in liquid culture, and DNA was isolated as described in Experimental Procedures. The DNA samples were digested with Eco RI and analyzed on a 0.5% agarose gel. (1, Dm) Drosophila DNA; (2, DmL) Drosophila library DNA; (3, Ch4) Charon 4 DNA; (4, RL) rabbit library DNA; (5, R) rabbit liver DNA.

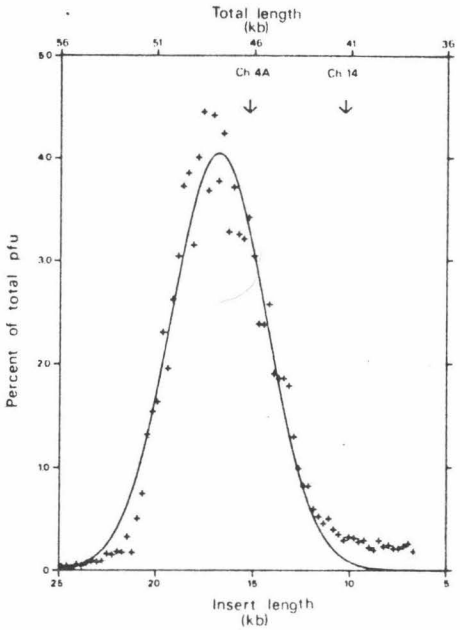


Figure 5. Average Size of Eucaryotic DNA Fragments in the Rabbit Library

The density distribution of phage from the rabbit library was determined by CsCl equilibrium centrifugation. Gradient fractions were titered, and the distribution of library phage (+) was computer fitted to a Gaussian curve (solid line). The arrows indicate the positions of the Charon 4A (93.6% λ) and Charon 14 (83.7% λ) phage density markers. Phage DNA length (upper abscissa) was calculated from the density relative to marker phage (Davidson and Szybalski, 1971). The insert length (lower abscissa) was determined by subtracting the length of the Charon 4A DNA arms (31 kb) from the total length of the phage DNA. The half-maximal band width of the curve is significantly greater than that observed for the marker phage (data not shown), indicating that a heterogeneous size population of rabbit DNA inserts is present in the library. This heterogeneity is not predicted by the relationship between insert size and packaging efficiency reported by Sternberg et al. (1977).

the libraries, we measured the single-copy complexity of library DNA using the procedure of Galau et al. (1976). Tritium-labeled single-copy tracer DNA was prepared from Drosophila DNA and driven with both sheared library and embryo DNAs. As shown in Figure 6, the reassociation rate and extent of reaction of the tracer are identical in both cases. We conclude that within the sensitivity of our measurements, the Drosophila library contains the entire complement of single-copy sequences present in genomic DNA. Phage libraries produced by nonlimit Eco RI digestion of sea urchin DNA (D. Anderson and E. Davidson, personal communication) and Drosophila DNA (R. Robinson and N. Davidson, personal communication) have been shown to be nearly complete by this criterion.

Screening Libraries for Structural Gene Sequences

Assuming an average size of 17 kb for the inserts in the rabbit library and a genome size of 3×10^9 bp, only 1 in 180,000 plaques will carry a particular single-copy sequence. Optimal screening conditions are therefore necessary to identify clones carrying such a sequence. We examined a number of variables in the plaque hybridization procedure of Benton and Davis (1977), including the type of medium used in the agar plates (L broth or NZCY), the strain of host bacteria (KH802 or DP50SupF), the concentration of plating bacteria and the method of preparing filters. Most of the variables had little effect on the intensity or the number of positive signals observed. The most significant differences resulted from varying the concentration of plating bacteria; the best signals were obtained

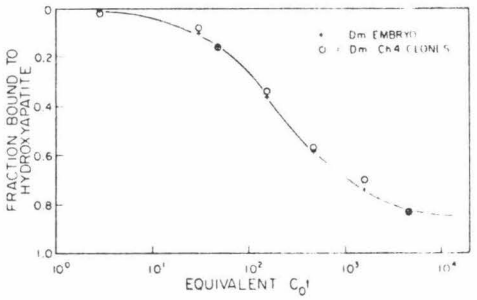


Figure 6. Single-Copy Sequence Representation in the Drosophila Library

A comparison of the reassociation kinetics of Drosophila embryo and Drosophila library DNAs with ^{3}H-labeled single-copy tracer (1×10^6 cpm/μg). Drosophila library DNA was prepared from phage grown in liquid culture (5×10^8 fold amplification). Single-copy tracer DNA was prepared from Drosophila pupae DNA, and reacted with sheared Drosophila embryo DNA and Drosophila library DNA according to the procedures of Galau et al. (1976). For each point, 3 μg of embryo DNA or 7.5 μg of library DNA were reacted with approximately 800 cpm of single-copy tracer. The library DNA was at 2.5 times the concentration of the embryo DNA on the assumption that 40% of each hybrid phage molecule consists of eucaryotic DNA. The solid line is a computer fit of the embryo data describing a single second-order component.

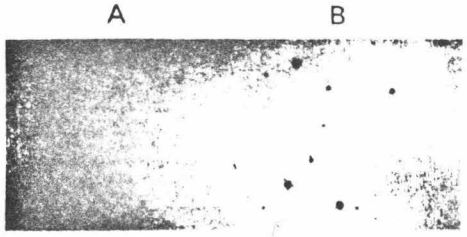


Figure 7. Screening of Silkmoth Library for Chorion Gene Sequences

(A and B) Autoradiograms of duplicate nitrocellulose filters showing specific hybridization of chorion cDNA to silkmoth library phage clones. 5000 phage were plated onto a 10 cm petri dish and incubated for 16 hr at 37°C, and the phage DNA was transferred to two nitrocellulose filters applied in succession. Agarose rather than agar was used in the 0.7% top agar layer. The filters were prepared for hybridization as described in Experimental Procedures and hybridized for 48 hr to 50 ng/ml ^{32}P-cDNA (spec. act. 2.5×10^7 cpm/μg) prepared from total chorion mRNA. Filters were washed, dried and exposed to X-ray film for 48 hr using a single intensifier screen. Five strong and two weak positive signals appear on both filters identifying phage clones bearing chorion gene sequences.

using 3.1×10^8 exponentially growing bacteria per 15 cm plate. The best correspondence between positive signals on duplicate filters occurred when the filters were sequentially applied to the plates; stacking filters often resulted in failure to observe duplicate signals.

To determine the optimal plaque density for screening, we plated various amounts of a mixture of Charon 4A and a Charon 16-pβG1 hybrid (pβG1 is a β-globin cDNA plasmid; Maniatis et al., 1976) at a constant ratio of 250:1. As many as 20,000 pfu per 15 cm plate could be screened for globin with no apparent loss of signal on autoradiograms. The most serious problem encountered in screening libraries is nonspecific background hybridization to nitrocellulose filters. Using conditions adapted from those developed for Southern transfer experiments (Jeffreys and Flavell, 1977a), however, we reproducibly observed low background.

All three libraries have been successfully screened using gene-specific hybridization probes. Several different phage recombinants that hybridize to different Drosophila cDNA plasmid clones have been selected from the Drosophila library (W. Bender and D. S. Hogness, personal communication). The frequencies at which these clones were detected and the single-copy complexity measurement of Figure 6 indicate that most if not all Drosophila structural gene sequences are present in the library.

The silkmoth library was screened for genes, sequentially expressed during oogenesis, which encode the approximately 100 eggshell (chorion) proteins of the developing oocyte (for a review of this system, see Kafatos et al., 1978). Although the exact number of chorion genes is not known, preliminary evidence suggests that each gene cannot be present in more than a few copies (J. Yeh, W. C. Jones and A. Efstratiadis, unpublished results). With an average insert size of 19 kb in the silkmoth library and a genome size of 10^9 bp, we expected to observe one positive signal per 530 plaques using cDNA transcribed from total chorion mRNA as probe, assuming that the chorion genes are unique. If some of the genes are closely linked, positive signals will be less frequent. Figure 7 (A and B) shows an autoradiogram of duplicate filters prepared from a plate containing 5000 plaques. 100% of the duplicate positive signals proved real when individual plaques were picked onto a bacterial lawn and rescreened. When only single (instead of duplicate) filters were prepared, 88% of the initial positives proved real upon rescreening. To date, screening of 350,000 plaques has yielded 350 independent isolates, about 53% of the number expected if the genes are unlinked.

The rabbit library was screened for globin sequences using cDNA prepared from total globin mRNA. Figure 8A shows an autoradiogram of a filter prepared from a 15 cm agar plate carrying 10,000 plaques. In the example shown, two positives were observed on one filter (a rare event), demonstrating two types of signals. One signal reflects the plaque morphology, while the other contains a head and a comet-like tail. The latter frequently observed morphology could result from

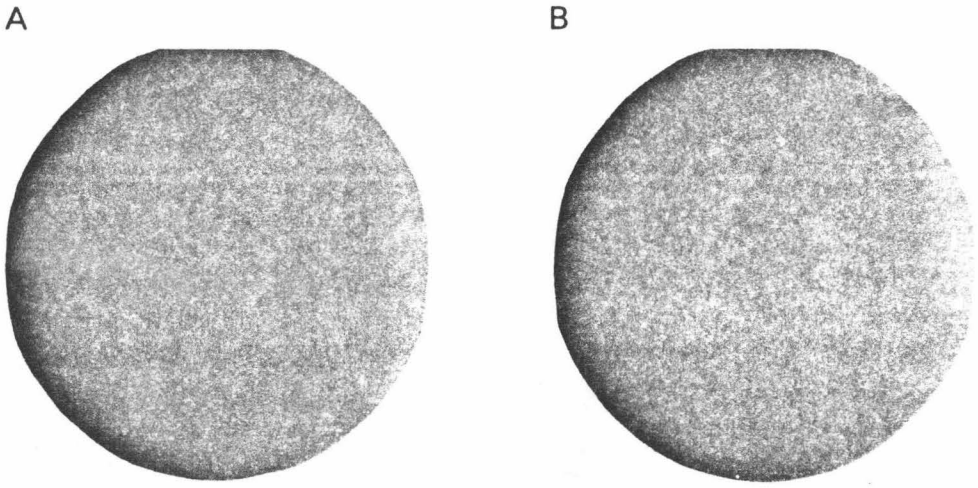


Figure 8. Screening the Rabbit Library for Globin Sequences

(A) Autoradiogram of a nitrocellulose filter prepared from a 15 cm plate containing 10,000 recombinant plaques. The filter was hybridized to 2.4 ng/ml ^{32}P-globin cDNA (5 × 10^8 cpm/μg) for 36 hr, washed as described in Experimental Procedures, dried and exposed to preflashed X-ray film for 48 hr at −70°C using a single intensifier screen. Arrows indicate the locations of two positive signals.

(B) Autoradiogram of a filter prepared from a plate containing 500 plaques obtained by plating a number of plaques from the area on the plate (A) corresponding to the location of one of the two positive signals shown in (A). The filter was hybridized with 25 ng/ml of nick-translated pβG1 DNA (5 × 10^7 cpm/μg) for 12 hr, and exposed to preflashed X-ray film for 24 hr at −70°C using two intensifier screens.

the spreading of phage DNA from the plaque during filter application or subsequent handling. Both spots were shown to be true positives by their appearance on duplicate filters and by rescreening (Figure 8B).

A total of four independent β-globin clones were recovered from 750, 000 plaques screened with globin cDNA. To identify clones carrying β-globin sequences, each clone was hybridized to in vitro labeled pβG1 DNA, a rabbit β-globin cDNA plasmid (Maniatis et al., 1976). With a genome size of 3 × 10^9 bp and cloned inserts of approximately 17 kb, we expected to recover four to five clones of the adult β-globin sequence from 750,000 plaques, a number close to that actually recovered. In this calculation, we assumed that cross-hybridization of adult β-globin probe to embryonic β-like rabbit globin genes is inadequate to allow detection of these genes in a total genome screen.

Characterization of Clones That Hybridize to Rabbit Globin Probes

To demonstrate that the four clones hybridizing to the β-globin plasmid actually carry the β-globin gene sequence, we digested DNA from each clone with Eco RI, fractionated the products on a 1.4% agarose gel, transferred the DNA to a nitrocellulose filter and hybridized them to in vitro labeled globin cDNA. Figure 9A shows the Eco RI cleavage pattern of DNAs from the four β-globin clones. As expected for DNA fragments generated by random cleavage, each clone contains common and unique Eco RI fragments. Presumably the common set contains the β-globin gene and its adjacent sequences, while the unique fragments lie further from the gene in the 5′ or 3′ direction. Figure 9B shows the resulting autoradiogram of the hybridization experiment. In all four clones, two fragments of approximately 2600 and 800 bp hybridize to the probe. These fragments are, respectively, the sizes of the 5′ and 3′ β-globin Eco RI fragments found in genomic DNA (Jeffreys and Flavell, 1977a, 1977b). The identification of these bands is confirmed by detailed restriction mapping and DNA sequence analysis of the cloned DNA (our unpublished results). In addition to these two fragments, RβG2 (lane 2) contains one Eco RI fragment and RβG5 (lane 4) contains three fragments that hybridize weakly to globin cDNA. These data and the failure of these two clones to hybridize α-globin probe indicate that the additional Eco RI fragments correspond to β-*like* sequences closely linked to the adult β-globin gene.

The 6.3 kb Eco RI fragment of RβG2 and RβG5, which hybridizes weakly compared with the 2.6 kb fragment, may correspond to the faint 6.9 kb Eco RI fragment detected in genomic DNA (Jeffreys and Flavell, 1977a). Preliminary restriction mapping data from RβG2 indicate that the β-like sequence

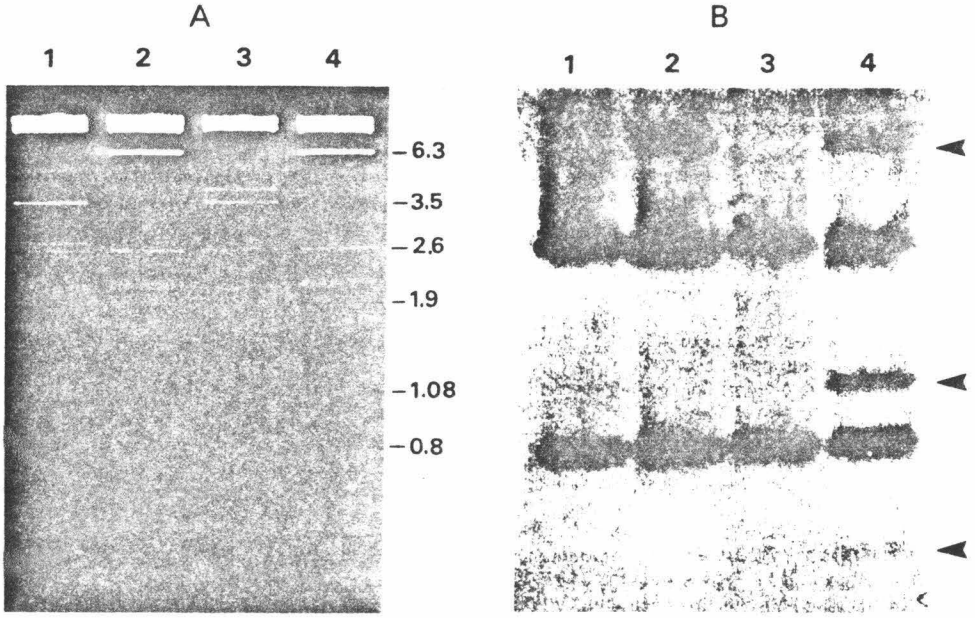


Figure 9. Eco RI Cleavage Patterns of DNAs from Rabbit β-Globin Phage Clones

DNAs from rabbit β-globin phage clones were digested with Eco RI, fractionated on a 1.4% agarose gel, transferred to a nitrocellulose filter (Southern, 1975) and hybridized to ^{32}P-globin cDNA (1 × 10^8 cpm/μg).
(A) Ethidium bromide-stained gel: (A) λCH4A-RβG1; (2) λCh4A-RβG2; (3) λCh4A-RβG3; (4) λCh4A-RβG5.
(B) Autoradiogram of the nitrocellulose filter prepared from the gel shown in (A). The sizes of various Eco RI fragments are indicated in kb. The arrows indicate DNA fragments bearing β-like globin sequences that are not a part of the adult β-globin gene.

carried on the 6.3 kb fragment is approximately 9.1 kb away from the adult β-globin gene in the 5′ direction (data not shown). The β-like sequences in RβG5 have not yet been mapped.

Discussion

This paper shows that it is possible to isolate structural genes directly from large eucaryotic genomes by screening libraries of DNA fragments cloned in phage λ. The overall efficiency of the procedure described yields a collection of recombinants large enough to represent the entire genome of a mammalian cell. Such a collection can be amplified by a factor of 10^6, with no apparent loss of sequence complexity, to produce a library of eucaryotic DNA that can be screened repeatedly using different probes. This rapid method of gene isolation provides many advantages over existing techniques. For example, all the members of a family of evolutionarily or developmentally related genes can be isolated in a single step by screening a library with a mixed probe. Furthermore, isolation of a set of overlapping clones, all of which contain a given gene, permits the study of sequences extending many kilobases from the gene in the 5′ and 3′ directions. Moreover, even more distant regions along the chromosome can be obtained by rescreening the library using terminal fragments of the initially selected clones, allowing the isolation of linked genes.

The power of this approach is clearly illustrated by the isolation of globin and chorion genes. The mammalian globin genes constitute a relatively simple family comprised of at least two α-like and at least four β-like embryonic and adult genes, which are expressed during erythropoiesis in different cell populations at different developmental times (Clissold, Arnstein and Chesterton, 1974; Melderis, Steinheider and Ostertag, 1974; Steinheider, Melderis and Ostertag, 1975, 1977). The gene family that codes for the chorion proteins of the silkmoth Antheraea polyphemus is considerably more complex (Kafatos et al., 1978). Approximately 100 different chorion genes are sequentially expressed during oogenesis. We have used the procedure of cDNA cloning to purify to homogeneity sequences that correspond to individual members of the chorion gene family (Maniatis et al., 1977; Sim et al., 1978; G. K. Sim, unpublished results).

Some of these cDNA plasmids have been classified according to the time of their expression during development (Kafatos et al., 1978). Using these cDNA plasmids to rescreen the phage recombinants that hybridize to total chorion cDNA, it will be possible to isolate and study genes that are coordinately and/or sequentially expressed during choriogenesis.

Since the libraries were prepared from random DNA fragments, independent isolates of a given gene will carry DNA sequences that extend for various distances away from the gene in both directions. This is clearly illustrated by the analysis of the β-globin clones (Figure 9). In some cases, the cloned DNA extends far enough in one direction to include a linked gene.

Although the possibility of linkage between different members of the rabbit α- or β-globin gene families has not been studied, evidence exists that such linkage occurs in other mammals, including mouse (Gilman and Smithies, 1968) and human (Clegg and Weatherall, 1976). For example, in humans the β-, δ- and γ-globin are genetically linked (Huisman et al., 1972; Clegg and Weatherall, 1976) on chromosome 11 (Deisseroth et al., 1978). In most human populations, the α-globin genes are present in two copies per haploid genome (Lehmann, 1970; Hollan et al., 1972) and are located on chromosome 16 (Deisseroth et al., 1977). Thus the close physical linkage between rabbit globin genes reported here is probably a general characteristic of mammalian globin genes. Clones bearing such genes can be used to study the precise organization of linked genes and the possible relation between linkage and control of gene expression. Moreover, a permanent library of clones bearing overlapping sequences will facilitate the isolation of the many linked genes that constitute a complex genetic locus.

Cloned segments of eucaryotic DNA can also be used to study the fine structure of genes. Most current methods of mammalian gene isolation involve partial purification of genomic DNA fragments generated by a limit restriction endonuclease digestion prior to cloning (Tilghman et al., 1977; Tonegawa et al., 1977). If the restriction endonuclease used to fragment genomic DNA cleaves within the coding sequence or noncoding intervening sequences of the gene of interest, the gene must be cloned in pieces. The rabbit α- and β-globin genes, which carry a single Eco RI site within the coding sequence (Maniatis et al., 1976; Salser et al., 1976; Liu et al., 1977), and the chicken ovalbumin gene, which carries at least one Eco RI site within each of three intervening sequences (Breathnach, Mandel and Chambon, 1977; Weinstock et al., 1977; Lai et al., 1978), illustrate this problem. If more than one Eco RI site is located within an intervening sequence, a portion of the chromosomal gene structure will remain unidentified. The procedure described here avoids these problems by cloning large pieces of randomly fragmented DNA that should carry intact genes including their intervening sequences.

Experimental Procedures

Materials

DNA polymerase I and T4 polynucleotide kinase were purchased from Boehringer Mannheim. T4 ligase was purchased from Bethesda Research Laboratories. Eco RI methylase prepared according to Greene et al. (1975) was provided by John Rosenberg. DNA polymerase from avian myeloblastosis virus (AMV reverse transcriptase) was provided by Dr. J. W. Beard and the Office of Program Resources and Logistics (Viral Cancer Program, NIH). Eco RI was prepared according to the procedure of Greene et al. (1975). Hae III and Alu I were prepared as described by Roberts et al. (1976) and Roberts (1976), respectively. Proteinase K was purchased from EM Labs. Pancreatic DNAase I was purchased from Worthington Biochemicals. NZ amine was purchased from Humko-Sheffield (Linnhurst, New Jersey). Nitrocellulose filters were purchased from Millipore. S-adenosyl-L-methionine was purchased from Sigma. α-^{32}P-deoxynucleoside triphosphates were purchased from New England Nuclear (~300 Ci/mmole) or ICN (120–200 Ci/mmole). Oligo$(T)_{12-18}$ was purchased from Collaborative Research.

Preparation of Bacteriophage λ DNA

Charon phage were grown essentially as described by F. R. Blattner in the detailed protocol that accompanies the Charon λ phages. Phage were purified as described in the above protocol and by Yamamoto et al. (1970).

For preparation of phage DNA, purified phage were dialyzed against 10 mM Tris-Cl (pH 8), 25 mM NaCl, 1 mM $MgSO_4$, brought to 0.2% SDS, 10 mM EDTA, heated to 65°C for 15 min and digested for 1 hr at 37°C with 50 μg/ml Proteinase K. The DNA was extracted several times with phenol, ether-extracted and dialyzed extensively against TSE [5 mM NaCl, 10 mM Tris-Cl (pH 8), 1 mM EDTA].

To prepare the end fragments of Charon 4 (see Figure 1), the cohesive ends were annealed by incubation for 1 hr in 0.1 M Tris-Cl (pH 8.0), 10 mM $MgCl_2$ at 42°C. Dithiothreitol (DTT) was added to 1 mM along with an excess of Eco RI and the reaction mix was incubated for 3 hr at 37°C. An aliquot was run on a 0.5% agarose gel to verify that digestion was complete. The DNA was extracted with phenol and then with ether. 50–70 μg of Eco RI-cleaved phage DNA were layered onto a 10–40% linear sucrose gradient [1 M NaCl, 20 mM Tris-Cl (pH 8.0), 10 mM EDTA] in a Beckman SW27 centrifuge tube. The gradient was centrifuged (at 27,000 rpm for 24 hr at 20°C) (Neal and Florini, 1972), and 0.5 ml fractions were collected using an ISCO ultraviolet flow cell. Fractions were analyzed on an agarose gel and those containing the 31 kb annealed end fragments were pooled.

To examine the restriction endonuclease cleavage patterns of DNAs from individual plaques, DNA was prepared as described above from phage grown in 4 ml cultures. Enough DNA was obtained to perform several restriction endonuclease digestions.

Drosophila DNA

Drosophila embryos (Canton S wild-type) aged 6–16 hr were collected, washed, frozen on dry ice and stored at −70°C. DNA was prepared according to Brutlag et al. (1977) with modifications. DNA from the CsCl gradient was dialyzed, digested with Proteinase K and phenol-extracted. The DNA was then brought to 5 M NaCl and chilled to 0°C to minimize the generation of molecules with single-stranded tails during shearing (Pyeritz, Schlegel and Thomas, 1972). The DNA was sheared by slowly

drawing it into a chilled 5 ml plastic syringe with a 20 gauge 1 in needle and expelling it as hard as possible into a 50 ml conical polyethylene tube on ice. The number of passes through the needle prior to the addition of S1 required to generate 20 kb fragments following the nuclease treatment varied from preparation to preparation. For this particular DNA preparation, three passes through a 20 gauge needle produced a mean size of 30 kb. The DNA was dialyzed against 0.5 M NaCl, 10 mM Tris-Cl (pH 8.0), 1 mM EDTA. Sodium acetate (pH 4.5) and $ZnSO_4$ were added to final concentrations of 50 mM and 2 mM, respectively. An amount of S1 nuclease sufficient to convert an equivalent amount of single-stranded λ DNA to small fragments as assayed by agarose gel electrophoresis was added. Following incubation for 1 hr at 37°C, the reaction mixture was extracted repeatedly with phenol and then with ether.

Silkmoth DNA

DNA was isolated from silkmoth Antheraea polyphemus pupae as previously described (Efstratiadis et al., 1976). Shearing and S1 nuclease digestion were performed as described above.

Rabbit Liver DNA

The liver of the New Zealand rabbit was removed and frozen in small pieces in liquid nitrogen. DNA was isolated using a modification of the Blin and Stafford (1976) procedure. After Proteinase K digestion, phenol extraction and dialysis, solid CsCl (0.95 g/ml) and ethidium bromide (1/10 vol of a 5 mg/ml solution) were added (final density 1.65 g/cm³). The solution was centrifuged (in a Ti60 rotor at 45,000 rpm for 60 hr at 20°C) to separate DNA from RNA and polysaccharides, and the DNA was collected as a viscous band by puncturing the side of the tube with a needle. Ethidium bromide was removed by several extractions with isopropanol equilibrated with saturated CsCl, followed by exhaustive dialysis against TSE. The molecular weight of the DNA was estimated by electrophoresis on neutral (Sharp, Sugden and Sambrook, 1973) and alkaline (McDonnel, Simon and Studier, 1977) 0.5% agarose gels, using bacteriophage λ Charon 4 DNA (46,200 bp; Blattner et al., 1977) and bacteriophage T4 DNA (171,000 bp; Kim and Davidson, 1974) as molecular weight standards. Both the duplex and single-stranded lengths of the rabbit DNA molecules were estimated to be >100,000 bp or nucleotides.

Partial endonuclease digestion conditions were established for the restriction enzymes Hae III and Alu I by performing a serial dilution of each enzyme in the presence of 1 μg of rabbit liver DNA in 1X restriction enzyme buffer [6 mM Tris-Cl (pH 7.5), 6 mM $MgCl_2$, 6 mM β-mercaptoethanol]. Reactions were incubated for 1 hr at 37°C, and the extent of digestion was estimated by electrophoresis on a 0.5% neutral agarose gel using Eco RI-digested Charon 4 DNA as a molecular weight standard. On the basis of this information, six large scale digests (330 μg DNA per reaction) were performed with 0.5, 1 and 2 times the estimated amount of enzyme yielding the maximum proportion of 20 kb fragments. The six digests were pooled, phenol-extracted and concentrated by ethanol precipitation.

Isolation of 20 kb Eucaryotic DNA

250–300 μg of sheared or enzymatically cleaved DNA in 0.5 ml of 10 mM Tris–Cl (pH 8.0), 10 mM EDTA were heated at 68°C for 20 min and sedimented through a 10–40% linear sucrose gradient as described above. Aliquots of fractions were analyzed by electrophoresis on a 0.5% agarose gel using Eco RI-digested Charon 4A DNA as a molecular weight standard. The fractions containing 19–20 kb DNA were pooled, dialyzed against TSE, concentrated by ethanol precipitation and resuspended in TSE.

Eco RI Methylation of Eucaryotic DNA

115 μg of 20 kb DNA were brought to a volume of 1 ml in 0.1 M Tris–Cl (pH 8.0), 10 mM EDTA, 6 μm S-adenosyl-L-methionine. Eco RI methylase (20 units in 1 ml) was added. Two 10 μl aliquots were taken before and after the addition of the enzyme and mixed with 0.5 μg of λ DNA. The eucaryotic DNA and the four 10 μl control reactions were incubated for 1 hr at 37°C. Each 10 μl aliquot containing λ DNA was mixed with 25 μl of a buffer containing 0.2 M Tris-Cl (pH 7.5), 0.1 M NaCl, 20 mM $MgCl_2$, and 2 mM DTT. Two of the aliquots (one each withdrawn before and after the addition of methylase) were mixed with Eco RI and incubated for 1 hr at 37°C. The other two corresponding aliquots were incubated without the addition of Eco RI (see Figure 3). The methylated DNA was phenol-extracted, ether-extracted, ethanol-precipitated, redissolved in 100 μl of 5 mM Tris-Cl (pH 7.5), dialyzed against the same buffer in a Schleicher and Schuell collodion bag, and evaporated to 40 μl under nitrogen.

Covalent Joining of Eco RI Linkers to Eucaryotic DNA

The synthesis of dodecamer linkers produces molecules with 5′ hydroxyl ends (Scheller et al., 1977a). Since the ligase requires 5′ phosphate ends, the first step in the joining reaction is to phosphorylate the linker. 5 μg of dodecamer linker in 10 μl of 66 mM Tris-Cl (pH 7.6), 10 mM $MgCl_2$, 1.0 mM ATP, 1.0 mM spermidine, 15 mM DTT, 200 μg/ml gelatin were added to 2 μl (10 units) of T4 kinase and incubated at 37°C for 1 hr. This reaction mixture was then added directly to 100 μl of eucaryotic DNA (~100 μg) in the same buffer. 5 μl (5 units) of T4 ligase were added and the reaction was incubated at room temperature for 6 hr. A 5 μl aliquot of the reaction mixture was electrophoresed on a 12% polyacrylamide Tris-borate-EDTA gel (Maniatis, Jeffrey and van de Sande, 1975) and the DNA was visualized by ethidium bromide staining. A successful ligation was evidenced by the presence of a series of linker oligomers, from dimers up to 14-mers.

The ligation reaction mixture was diluted to 500 μl with 10 mM EDTA, incubated for 15 min at 68°C, layered onto a 10–40% linear sucrose gradient and centrifuged as described above. The gradient fractions containing 20 kb DNA were identified by agarose gel electrophoresis, pooled, dialyzed against TSE and ethanol-precipitated. The DNA (25 μg) was resuspended in 100 μl of 5 mM NaCl and brought to 1X Eco RI buffer [0.1 M Tris-Cl (pH 7.5), 50 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT], and 150 units of Eco RI were added. 3 μl of the reaction were mixed with 0.2 μg of λ DNA and both tubes were incubated for 1 hr at 37°C. The small scale reaction containing λ DNA was electrophoresed on a 0.5% agarose gel, in which a complete digest of the linkers attached to rabbit DNA was evidenced by a limit digestion pattern of λ DNA.

The Eco RI linkers were attached to 20 kb Drosophila and silkmoth DNAs using similar procedures, except that linker oligomers were removed on a Sepharose 2B column rather than on a sucrose gradient.

Ligation of Cohesive Ends

High cloning efficiencies were obtained using a 2 fold molar excess of Charon 4 arms to 20 kb eucaryotic DNA fragments. For example, in the case of the rabbit library, the reaction mixture contained 20.5 μg of eucaryotic DNA and 55 μg of purified Charon 4A arms in 300 μl of ligase buffer. The cohesive ends of the phage were annealed for a second time in $MgCl_2$, Tris and gelatin for 1 hr at 42°C before adding ATP, DTT, eucaryotic DNA and 19 μl (19 units) ligase. The mixture was incubated at 12°C for 12 hr. An aliquot was heated to 68°C, cooled and electrophoresed on a 0.3% agarose gel with Eco RI-digested Charon 4A DNA as a molecular weight standard. Successful ligation was evidenced by the absence of Charon 4A end fragments (~12 and 19 kb) and the presence of concatemeric DNA molecules larger than intact Charon 4A DNA.

In Vitro Packaging of Recombinant DNA into Phage Particles

In vitro packaging extracts were prepared as described by Sternberg et al. (1977), except that both types (A and B) of extracts which they describe were handled as 80 μl aliquots in 1.5 ml polypropylene tubes, which were frozen in liquid nitrogen and stored at −70°C. Proportionate amounts of lysozyme, then buffer B (Sternberg et al., 1977) and glycerol were added to each tube of

resuspended B extract cells and thoroughly stirred into the extremely viscous suspension using a glass micropipet.

Preparation and Testing of in Vitro Packaging Extracts for EK2 Experiments

NIH regulations require that in vitro packaging extracts "be irradiated with ultraviolet light to a dose of 40 phage lethal hits" before they can be used in EK2 level recombinant DNA experiments. Wavelengths between 250 and 280 nanometers are the most effective for photoinactivation (Hollaender, 1955). We use as an ultraviolet light source four 18 in 15 watt germicidal lamps (GE G15T8), which emit 1.3×10^4 erg/cm^2 - n (2200 μW/cm^2/sec) at a distance of 20 cm. The killing efficiency of this light source was calibrated as follows.

1 liter of NZCYM broth (1% NZ amine, 0.5% NaCl, 0.5% yeast extract, 0.1% casamino acids, 10 mM $MgSO_4$) was inoculated with 20 ml of an overnight culture of $\lambda C_I857Sam7$ (a heat-inducible λ lysogen). The culture was grown at 32°C to an OD_{600} of 0.3, transferred to a 42°C shaking water bath for 20 min to induce the lysogen and then incubated with shaking for 2.5 hr at 37°C. The cells were transferred to an enamel dish (32 cm × 18 cm × 5.5 cm) and irradiated at a distance of 20 cm with constant mixing on a rotary shaker platform. 1 ml aliquots were taken after 5, 10, 20, 30, 40 and 60 min of irradiation. The cells were lysed by the addition of two drops of chloroform and the DNA was removed by adding 10 μl of a 1 mg/ml DNAase I solution. The mixture was centrifuged for 2 min in a Brinkman Eppendorf centrifuge, and the number of surviving phage was determined by titering the supernatant on DP50SupF at 42°C under yellow light to avoid photoreactivation. Our yellow light source consists of a commercial "Gold" fluorescent lamp (GE) with a gold plexiglass filter. A plot of the log of survival versus time of irradiation extrapolates to 40 lethal hits at approximately 30 min of ultraviolet irradiation. Cells used in preparing EK2 extracts were ultraviolet-irradiated (as above) after heat induction and incubation at 37°C and prior to pelleting (see Sternberg et al., 1977).

Before any recombinant DNA was packaged, every preparation of extracts derived from ultraviolet-irradiated cells was tested for recombination and the presence of prophage. All of the steps described below were performed under yellow light. Extracts were tested for prophage excision and packaging by performing "mock" packaging reactions without exogenous DNA and plating on DP50SupF, which provides the appropriate suppressor for the S amber 7 mutation of the prophage (Sternberg et al., 1977). Plates were incubated at 42°C to inactivate the prophage repressor. No more than one tenth (30 μl) of one packaging reaction could be assayed on a single 10cm plate because the concentrated packaging mixture kills bacterial cells and thus masks the presence of prophage. According to NIH regulations, "the ratio of plaque-forming units without addition of exogenous λ DNA (endogenous virus) to plaque-forming units with exogenous DNA (exogenous virus) must be less than 10^{-6}." Exogenous viral DNA means recombinant DNA. We consistently find that this ratio is $< 10^{-6}$ for recombinant λ DNA and $< 10^{-8}$ using intact Charon 4A DNA in the assay. Without ultraviolet irradiation, the above ratio is $> 10^{-6}$.

NIH regulations further require that when the EK2 vector is packaged, "the ratio of am$^+$ phage (recombinants) to total phage must be less than 10^{-6}." We added Charon 4A DNA to the packaging extract and plated the packaged DNA on Su$^-$ and Su$^+$ strains. The ratio of pfu on Su$^-$ to pfu on Su$^+$ is a measure of the frequency of recombination with prophage DNA. We find that this ratio is $< 10^{-8}$; in fact, we have never observed a plaque resulting from recombination. On the basis of these and similar data obtained by N. Sternberg and L. Enquist (unpublished observations), the NIH has approved in vitro packaging for EK2 level experiments.

Packaging DNA for Libraries

To construct, for example, the rabbit library, 26 packaging reactions were performed, each containing 2.5 μg of recombinant DNA. After DNAase digestion and chloroform treatment, the packaged phage were purified and concentrated on a CsCl step gradient. The reactions were pooled, mixed with solid CsCl (0.5 g/ml), brought to a final volume of 30 ml with 0.5 g/ml CsCl in SM [0.1 M NaCl, 0.05 M Tris-Cl (pH 7.5), 10 mM $MgSO_4$, 0.01% gelatin], layered onto CsCl gradients (each one composed of 1.5 ml steps of 1.45, 1.5, 1.7 g/ml CsCl in SM, in an SW41 tube) and centrifuged (at 32,000 rpm for 1.5 hr at 4°C). 200 μl fractions were collected and phage were located by spotting dilutions of 10^{-3} to 10^{-4} onto a lawn of DP50SupF.

Library Amplification

The fractions containing phage were pooled and dialyzed against 0.1M NaCl, 50 mM Tris-Cl (pH 7.5), 10 mM $MgSO_4$. Gelatin was added to the dialysis bag to a concentration of 0.02% to stabilize the phage. In vitro packaged phage from the rabbit library were plated onto a fresh overnight of DP50SupF at a density of 10,000 phage per 15 cm diameter plate (75 plates total). The phage were preadsorbed with bacteria for 20 min at 37°C, mixed with 6.5 ml of NZCYM-DT (NZCYM medium with 0.01% diaminopimelic acid and 0.004% thymidine; Blattner et al., 1977), 0.5% top agar and plated. The plates were incubated for 14 hr at 37°C. (Plating was carried out in yellow light and the plates were wrapped in aluminum foil while growing to prevent photoreactivation.) The top agar was scraped into a sterile beaker and the plates were rinsed once with 3.75 ml of SM. The 285 ml lysate from 38 plates was mixed with 10 ml of chloroform and stirred slowly at room temperature for 20 min. The lysate was transferred to a 1 liter centrifuge bottle and centrifuged (in a Beckman J6 centrifuge at 52,000 rpm for 20 min at 4°C) to remove the top agar.

Screening the Libraries

Amplified libraries were screened using the in situ plaque hybridization technique of Benton and Davis (1977). 10,000 recombinant phage were plated on 3.1×10^4 exponential phase bacterial cells on 15 cm NZCYM petri dishes. To prevent top agar from adhering to the nitrocellulose filter when it was lifted from the plate (which tends to increase the background hybridization), plates were dried in a 37°C incubator for several hours or set on edge overnight to drain excess liquid. The use of 0.7% agarose rather than agar in the top agar layer also minimized this problem. The plates were incubated at 37°C for 14-16 hr, at which time the plaques were in contact but lysis was not confluent. Plates were refrigerated for an hour or longer before the filters were applied. Nitrocellulose filters were cut from rolls (HAMP 000 10) or circles (HAWP 142 50) of Millipore filter paper (type HA, pore size 0.45 μm) to fit easily over the agar plate. Phage and DNA were adsorbed to these filters in duplicate by placing two filters on each plate sequentially, 2 min for the first filter and 3 min for the second, at room temperature. The DNA was denatured and bound to the filters as described by Benton and Davis (1977).

To prepare the filters for hybridization to a labeled probe, they were wetted in about 10 ml per filter of 4X SET [1X SET = 0.15 M NaCl, 0.03 M Tris-Cl (pH 8), 2 mM EDTA] at room temperature for 30 min; washed for 3 hr at 68°C in about 10 ml per filter of 4X SET, 10X Denhardt's solution (10X Denhardt's solution = 0.2% bovine serum albumin, 0.2% polyvinylpyrolidone, 0.2% Ficoll, Denhardt, 1966) and 0.1% SDS, and prehybridized with continuous agitation for at least 1 hr at 68°C in about 4 ml per filter of 4X SET, 10X Denhardt's solution, 50 μg/ml denatured salmon sperm DNA, 10 μg/ml poly(A), 0.1% SDS and 0.1% sodium pyrophosphate. Denatured E. coli DNA (10-50 μg/ml) was included in the prehybridization mix when using labeled plasmid DNA as probe. The prehybridization and subsequent hybridization were carried out in a thermally sealed plastic bag. The filters were hybridized with agitation at 68°C in prehybridization solution containing ^{32}P-labeled hybridization probe at the concentrations and for the times indicated in the figure legends. After hybridization, the filters were washed once with agitation in about 15 ml per filter of 4X SET, 10X Denhardt's solution, 10 μg/ml poly(A), 0.1% SDS, 0.1% sodium pyrophosphate at 68°C for 1 hr, 3 times in a similar volume

of 3X SET, 0.1% SDS, 0.1% sodium pyrophosphate at 68°C for 15-30 min; twice in 1X SET, 0.1% SDS, 0.1% sodium pyrophosphate at 68°C for 15-30 min; and once in 4X SET at room temperature. In some cases, more stringent washing conditions (0.5X SET) were used as a final step. The filters were blotted dry, mounted on cardboard and exposed to preflashed Kodak XR5 X-ray film with Dupont Cronex:11R Xtra Life Lightning-plus intensifying screens at −70°C for 1-2 days.

Plaque Purification of Recombinant Phage

Plaques from the region of a plate corresponding to a positive on the autoradiogram were picked and suspended in 0.5 ml SM. The phage suspension was titered and the plate containing about 100 plaques was rescreened. The process of picking positives and rescreening was repeated until ~ 90% of the plaques on a plate gave positive signals after screening.

Plate lysates were prepared using 10^8 phage from an individual plaque on a 10 cm plate, and the lysates were harvested as described in Library Amplification. 10^{10}-10^{11} phage were usually recovered.

Hybridization Probes

cDNA plasmids were grown, purified and labeled in vitro by nick translation as previously described (Maniatis et al., 1976). Complementary DNA to globin and chorion mRNAs was synthesized as described by Efstratiadis et al. (1975) and Friedman and Rosbash (1977).

CsCl Sedimentation Equilibrium Analysis of Recombinant Phage

Phage from the amplified rabbit library (4×10^7 pfu) were mixed with 3×10^8 pfu of Charon 14 phage (Blattner et al., 1977) in a solution of CsCl, density = 1.514 g/ml, 10 mM Tris-Cl (pH 7.5) and 1.0 mM $MgSO_4$. The phage were banded by centrifugation to equilibrium in a Ti50 rotor, 5.38×10^8 g_{av} hr at 4°C. One drop fractions were collected from the bottom of the gradient into 0.5 ml SM for a total of 128 fractions. The fractions containing phage were titered on DP50SupF, a Su II + Su III strain, to determine the distribution of densities of phage in the library and also on Cla (SupO) to determine the position of the marker Charon 14, a nonamber phage. The position of a second marker, Charon 4A (representing the background phage in the library), was determined by counting the number of blue plaques formed on KH802, a Su II *lac*$^-$ strain, using Xgal plates. From the known sizes of Charon 14 and Charon 4A, a calibration curve was constructed relating fraction number, or buoyant density, to length of insert DNA. The average density of the recombinant phage provides a minimum estimate for the size of rabbit DNA inserts, since the empirical relationship between phage density and DNA molecular weight was established using λ DNA (50% G+C) (Davidson and Szybalski, 1971), while the base composition of rabbit DNA is 44.2% G+C (Kritskii, Arends and Nikolaev, 1967).

Recombinant DNA Safety

Experiments involving the cloning or propagation of bacteriophage λ carrying eucaryotic DNA were performed in accordance with the NIH Guidelines for recombinant DNA research. Drosophila and silkmoth cloning experiments were performed using EK1 vectors in P2 facilities located at the California Institute of Technology and Harvard University. Rabbit DNA cloning experiments were performed using the EK2 vector-host system λ Charon 4A/DP50SupF in a P3 facility at the California Institute of Technology. The relevant genotypes of *EK2* vectors were verified prior to their use in recombinant DNA experiments.

Acknowledgments

The procedures for making and screening libraries were established concurrently in the laboratories of E. Davidson, N. Davidson, J. Bonner and L. Hood at Caltech. We benefited from the free exchange of information and ideas among these labs. We are grateful to D. Goldberg, B. Seed, D. Engel, R. Scheller and D. Anderson for suggestions and discussions, and to N. Davidson for critical readings of the manuscript. We thank F. Blattner for providing Charon phage strains, N. Sternberg and L. Enquist for providing in vitro packaging strains, C. K. Itakura and R. Scheller for providing Eco RI linkers, J. Rosenberg for providing Eco RI methylase, R. Robinson for providing labeled single-copy Drosophila DNA tracer, and A. Cortenbach for preparing media and materials. We thank W. Bender and D. Hogness for discussions and for communicating unpublished data on the Drosophila library characterization, and F. C. Kafatos for the use of facilities and his support. We thank J. Maniatis for help with the figures. This work was supported by an NSF grant to F. C. Kafatos and T. M., an NIH grant to F. C. Kafatos, an American Cancer Society grant to T. M. and funds from an NIH Biomedical Research Support grant to the California Institute of Technology. R. C. H. was supported by a fellowship from the Jane Coffin Childs Memorial Fund for Medical Research. T. M. is the recipient of a Rita Allen Foundation career development award.

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

Received June 16, 1978

References

Bahl, C. P., Marians, K. J., Wu, R., Stawinsky, J. and Narang, S. (1977). A general method for inserting specific DNA sequences into cloning vehicles. Gene *1*, 81-92.

Bellet, A., Busse, H. and Baldwin, R. (1971). Tandem duplications in a derivative of phage lambda. In The Bacteriophage Lambda, A. J. Hershey, ed. (New York: Cold Spring Harbor Laboratory), pp. 501-513.

Benton, W. D. and Davis, R. W. (1977). Screening λgt recombinant clones by hybridization to single plaques in situ. Science *196*, 180-182.

Blattner, F. R., Williams, B. G., Blechl, A. E., Thompson, K. D., Faber, H. E., Furlong, L. A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schumm, J. W., Sheldon, E. L. and Smithies, O. (1977). Charon phages: safer derivatives of bacteriophage lambda for DNA cloning. Science *196*, 161-169.

Blin, N. and Stafford, D. W. (1976). A general method for isolation of high molecular weight DNA from eukaryotes. Nucl. Acids Res. *3*, 2303-2308.

Breathnach, R., Mandel, J. L. and Chambon, P. (1977). Ovalbumin gene is split in chicken DNA. Nature *270*, 314-319.

Brutlag, D., Appels, R., Dennis, E. S. and Peacock, W. J. (1977). Highly repeated DNA in Drosophila melanogaster. J. Mol. Biol. *112*, 31-47.

Carbon, J., Clarke, L., Ilgen, C. and Ratzkin, B. (1977). The construction and use of a hybrid plasmid gene bank in E. coli. In Recombinant Molecules: Impact on Science and Society, R. F. Beers and E. G. Bassatt, eds. (New York: Raven Press), pp. 335-378.

Clarke, L. and Carbon, J. (1976). A colony bank containing synthetic Col E1 hybrid plasmids representative of the entire E. coli genome. Cell *9*, 91-99.

Clegg, J. B. and Weatherall, D. J. (1976). Molecular basis of thalassaemia. Br. Med. Bull. *32*, 262-269.

Clissold, P. M., Arnstein, H. R. V. and Chesterton, C. J. (1977). Quantitation of globin mRNA levels during erythroid development in the rabbit and discovery of a new β-related species in immature erythroblasts. Cell *11*, 353-361.

Davidson, N. and Szybalski, W. (1971). Physical and chemical characteristics of lambda DNA. In The Bacteriophage Lambda, A. D. Hershey, ed. (New York: Cold Spring Harbor Laboratory), pp. 45-82.

Deisseroth, A., Nienhuis, A., Turner, P., Velez, R., Anderson, W. F., Ruddle, F., Lawrence, J., Creagen, R. and Kucherlapati, R. (1977). Localization of the human α-globin structural gene to chromosome 16 in somatic cell hybrids by molecular hybridization assay. Cell *12*, 205-218.

Deisseroth, A., Nienhuis, A., Lawrence, J., Giles, R., Turner, P. and Ruddle, F. (1978). Chromosomal localization of human β-globin gene on human chromosome 11 in somatic cell hybrids. Proc. Nat. Acad. Sci. USA *75*, 1456-1460.

Denhardt, D. T. (1966). A membrane-filter technique for the detection of complementary DNA. Biochem. Biophys. Res. Commun. *23*, 641-646.

Efstratiadis, A., Maniatis, T., Kafatos, F. C., Jeffrey, A. and Vournakis, J. N. (1975). Full length and discrete partial reverse transcripts of globin and chorion mRNAs. Cell *4*, 367-378.

Efstratiadis, A., Crain, W. R., Britten, R. J., Davidson, E. H. and Kafatos, F. (1976). DNA sequence organization in the lepidopteran Antherea pernyi. Proc. Nat. Acad. Sci. USA *73*, 2289-2293.

Friedman, E. Y. and Rosbash, M. (1977). The synthesis of high yields of full-length reverse transcripts of globin mRNA. Nucl. Acids Res. *4*, 3455-3471.

Galau, G. A., Klein, W. H., Davis, M. M., Wold, B. J., Britten, R. J. and Davidson, E. H. (1976). Structural gene sets active in embryos and adult tissues of the sea urchin. Cell *7*, 487-505.

Gilman, J. G. and Smithies, O. (1968). Fetal hemoglobin variants in mice. Science *160*, 885-886.

Glover, D. M., White, R. L., Finnegan, D. J. and Hogness, D. S. (1975). Characterization of six cloned DNAs from Drosophila melanogaster, including one that contains the genes for rRNA. Cell *5*, 149-157.

Greene, P. J., Poonian, M. S., Nussbaum, A. L., Tobias, L., Garfen, D. E., Boyer, H. W. and Goodman, H. M. (1975). Restriction and modification of a self complementary octanucleotide containing the EcoRI substrate. J. Mol. Biol. *99*, 237-261.

Heyneker, H. L., Shine, J., Goodman, H. J., Boyer, H., Rosenberg, J., Dickerson, R. E., Narang, D. A. and Riggs, A. D. (1976). Synthetic lac operator DNA is functional in vivo. Nature *263*, 748-752.

Hohn, B. and Murray, K. (1977). Packaging recombinant DNA molecules into bacteriophage particles in vitro. Proc. Nat. Acad. Sci. USA *74*, 3259-3263.

Hollaender, A. (1955). Radiation Biology, *II* (New York: McGraw-Hill).

Hollan, S. R., Szelenyii, J. G., Brimhall, B., Duerst, M., Jones, R. T., Koler, R. D. and Stocklen, I. (1972). Multiple alpha chain loci for human haemoglobins: Hb J-Buda and Hb G-Pest. Nature *235*, 47-50.

Huisman, T. H. J., Wrightstone, R. M., Wilson, J. B., Schroeder, W. A. and Kendall, A. G. (1972). Hemoglobin Kenya, the product of a fusion of γ and β polypeptide chains. Arch. Biochem. Biophys. *153*, 850-853.

Jeffreys, A. J. and Flavell, R. A. (1977a). A physical map of the DNA regions flanking the rabbit β-globin gene. Cell *12*, 429-439.

Jeffreys, A. J. and Flavell, R. A. (1977b). The rabbit β-globin gene contains a large insert in the coding sequence. Cell *12*, 1097-1108.

Kafatos, F. C., Efstratiadis, A., Goldsmith, M. R., Jones, C. W., Oaniatis, T., Reiger, J. C., Rodakis, G., Rosenthal, N., Sim, G. K., Thireos, G. and Villa-Komaroff, L. (1978). The developmentally regulated multigene families encoding chorion proteins in silkmoth. In Structure and Organization of Developmentally Regulated Genes, J. Schultz and F. Ahmad, eds. (New York: Academic Press), in press.

Kim, J. S. and Davidson, N. (1974). Electron microscope heteroduplex study of sequence relations of T2, T4 and T6 bacteriophage DNAs. Virology *57*, 93-111.

Kritskii, G., Arends, I. and Nikolaev, Y. (1967). Nucleotide composition of DNA fractions of bone marrow in normal and X-irradiated animals. Biochemistry (USSR) *32*, 372-376.

Lai, E. C., Woo, S. L. C., Dugaiczyk, A., Catteral, J. F. and O'Malley, B. W. (1978). The ovalbumin gene: structural sequences in native chick DNA are not contiguous. Proc. Nat. Acad. Sci. USA *75*, 2205-2209.

Leder, P., Tiemeier, D. and Enquist, L. (1977). EK2 derivatives of bacteriophage lambda useful in the cloning of DNA from higher organisms: the λgtWES system. Science *196*, 175-177.

Lehmann, H. (1970). Different types of alpha thalassemia and significance of haemoglobin Bart's in neonates. Lancet *2*, 78-80.

Liu, A. Y., Paddock, G. V., Heindell, H. C. and Salser, W. (1977). Nucleotide sequences from a rabbit alpha globin gene inserted in a chimeric plasmid. Science *196*, 192-197.

Mandel, M. and Higa, A. (1970). Calcium-dependent bacteriophage DNA interaction. J. Mol. Biol. *53*, 159-162.

Maniatis, T., Jeffrey, A. and van de Sande, H. (1975). Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis. Biochemistry *14*, 3787-3794.

Maniatis, T., Sim, G. K., Efstratiadis, A. and Kafatos, F. C. (1976). Amplification and characterization of a β-globin gene synthesized in vitro. Cell *8*, 163-182.

Maniatis, T., Sim, G. K., Kafatos, F. C., Villa-Komaroff, L. and Efstratiadis, A. (1977). An approach to the study of developmentally regulated genes. In Molecular Cloning of Recombinant DNA (New York: Academic Press), pp. 173-203.

McDonnel, M., Simon, M. N. and Studier, W. F. (1977). Analysis of restriction fragments of T7 DNA and determination of molecular weights by electrophoresis in neutral and alkaline gels. J. Mol. Biol. *110*, 119-146.

Melderis, H., Steinheider, G. and Ostertag, W. (1974). Evidence for a unique kind of α type globin chain in early mammalian embryos. Nature *250*, 774-776.

Neal, M. W. and Florini, J. R. (1972). Effect of sucrose gradient composition on resolution of RNA species. Anal. Biochem. *45*, 271-276.

Pyeritz, R., Schlegel, R. and Thomas, C., Jr. (1972). Hydrodynamic shear breakage of DNA may produce single-chained terminals. Biochim. Biophys. Acta *272*, 504-509.

Roberts, R. J. (1976). Restriction endonucleases. CRC Crit. Rev. Biochem. *4*, 123-164.

Roberts, R. J., Myers, P. A., Morrison, A. and Murray, K. (1976). A specific endonuclease from Arthobacter luteus. J. Mol. Biol. *102*, 157-165.

Rudkin, G. T. (1972). Replication in polytene chromosomes. In Results and Problems in Cell Differentiation, *4*, W. Beerman, ed. (New York: Springer-Verlag), pp. 59-85.

Salser, W., Bowen, S., Brown, D., Adli, F. E., Federoff, N., Fry, K., Heindell, H., Paddock, G., Poor, R., Wallace, B. and Whitcome, P. (1976). Investigation of the organization of mammalian chromosomes at the DNA sequence level. Fed. Proc. *35*, 23-35.

Scheller, R. H., Dickerson, R. E., Boyer, H. W., Riggs, A. D. and Itakura, K. (1977a). Chemical synthesis of restriction enzyme recognition sites useful for cloning. Science *196*, 177-180.

Scheller, R. H., Thomas, T. L., Lee, A. S., Klein, W. H., Niles, W. D., Britten, R. J. and Davidson, E. H. (1977b). Clones of individual repetitive sequences from sea urchin DNA constructed with synthetic eco RI sites. Science *196*, 197-200.

Seeburg, P. H., Shine, J., Martial, J. A., Baxter, J. D. and Goodman, H. M. (1977). Nucleotide sequence and amplification in bacteria of structural gene for rat growth hormone. Nature *270*, 486-494.

Sharp, P. A., Sugden, B. and Sambrook, J. (1973). Detection of two restriction endonuclease activities in Haempophilus parain-

fluenzae using analytical agarose ethidium bromide electrophoresis. Biochemistry *12*, 3055-3063.

Shine, J., Seeburg, P. H., Martial, J. A., Baxter, J. D. and Goodman, H. M. (1977). Construction and analysis of recombinant DNA for human chorionic somatomammotropin. Nature *270*, 494-499.

Sim, G. K., Efstratiadis, A., Jones, C. W., Kafatos, F. C., Koehler, M., Kronenberg, H. M., Maniatis, T., Regier, J. C., Roberts, B. F. and Rosenthal, N. (1978). Studies on the structure of genes expressed during development. Cold Spring Harbor Symp. Quant. Biol. *42*, 933-945.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. *98*, 503-517.

Steinheider, G., Melderis, H. and Ostertag, W. (1975). Embryonic ϵ chains of mice and rabbits. Nature *257*, 714-716.

Steinheider, G., Melderis, H. and Ostertag, W. (1977). In International Symposium on the Synthesis, Structure and Function of Hemoglobin, H. Martin and L. Novicke, eds. (Munich: Lehmans), pp. 222-235.

Sternberg, N., Tiemeier, D. and Enquist, L. (1977). In vitro packaging of a λ *Dam* vector containing Eco RI DNA fragments of E. and phage P1. Gene *1*, 255-280.

Thomas, M., Cameron, J. R. and Davis, R. W. (1974). Viable molecular hybrids of bacteriophage lambda and eukaryotic DNA. Proc. Nat. Acad. Sci. USA *71*, 4579-4583.

Tilghman, S. M., Tiemeier, D. C., Polsky, F., Edgell, M. H., Seidman, J. G., Leder, A., Enquist, L. W., Norman, B. and Leder, P. (1977). Cloning specific segments of the mammalian genome: bacteriophage λ containing mouse globin and surrounding gene sequences. Proc. Nat. Acad. Sci. USA *74*, 4406-4410.

Tilghman, S. M., Tiemeier, D. C., Seidman, J. G., Peterlin, B. M., Sullivan, M., Maizel, J. and Leder, P. (1978). Intervening sequence of DNA identified in the structural portion of a mouse β-globin gene. Proc. Nat. Acad. Sci. USA *75*, 725-729.

Tonegawa, S., Brack, C., Hozumi, N. and Scholler, R. (1977). Cloning of an immunoglobulin variable region gene from mouse embryo. Proc. Nat. Acad. Sci. USA *74*, 3518-3522.

Ullrich, A., Shine, J., Chirgwin, J., Dictet, R., Tischer, E., Roltes, W. J. and Goodman, H. J. (1977). Rat insulin genes: construction of plasmids containing the coding sequences. Science *196*, 1313-1319.

Weigle, J., Meselson, M. and Paigen, K. (1959). Density alterations associated with transducing ability in bacteriophage lambda. J. Mol. Biol. *1*, 379-386.

Weinstock, R., Sweet, R., Weiss, M., Cedar, H. and Axel, R. (1977). Intragenic spacers interrupt the ovalbumin gene. Proc. Nat. Acad. Sci. USA *75*, 1299-1303.

Wensink, P. C., Finnegan, D. J., Donelson, J. E. and Hogness, D. S. (1974). A system for mapping DNA sequences in the chromosomes of Drosophila melanogaster. Cell *3*, 315-325.

Yamamoto, K. R., Alberts, B. M., Benzinger, R., Lawhorne, L. and Treiber, C. (1970). Rapid bacteriophage sedimentation in the presence of polyethylene glycol and its application to large scale virus purification. Virology *40*, 734-744.

Young, M. and Hogness, D. (1977). A new approach for identifying and mapping structural genes in Drosophila melanogaster. In Eucaroytic Genetic Systems, ICN-UCLA Symposia on Molecular and Cellular Biology, *VIII* (New York: Academic Press), pp. 315-331.

Note Added in Proof

The NIH has recently revised the rules for in vitro packaging in EK2 systems using phage λ vectors. In particular, ultraviolet irradiation of extracts to a dose of 40 phage-lethal hits and assaying extracts without the addition of exogenous λ DNA are no longer required (for current rules, consult Office of Recombinant DNA Activities).

Chapter 2

The chromosomal arrangement of human α-like globin genes: Sequence homology and α-globin gene deletions

Cell, Vol. 20, 119–130, May 1980, Copyright © 1980 by MIT

The Chromosomal Arrangement of Human α-Like Globin Genes: Sequence Homology and α-Globin Gene Deletions

Joyce Lauer, Che-Kun James Shen
and Tom Maniatis
Division of Biology
California Institute of Technology
Pasadena, California 91125

Summary

We report the isolation of a cluster of four α-like globin genes from a bacteriophage λ library of human DNA (Lawn et al., 1978). Analysis of the cloned DNA confirms the linkage arrangement of the two adult α-globin genes (α1 and α2) previously derived from genomic blotting experiments (Orkin, 1978) and identifies two additional closely linked α-like genes. The nucleotide sequence of a portion of each of these α-like genes was determined. One of these sequences is tentatively identified as an embryonic ζ-globin gene (ζ1) by comparison with structural data derived from purified ζ-globin protein (J. Clegg, personal communication), while the other sequence cannot be matched with any known α-like polypeptide sequence (we designate this sequence ψα1). Localization of the four α-like sequences on a restriction map of the gene cluster indicates that the genes have the same transcriptional orientation and are arranged in the order 5′-ζ1-ψα1-α2-α1-3′. Genomic blotting experiments identified a second, nonallelic ζ-like globin gene (ζ2) located 10–12 kb 5′ to the cloned ζ-globin gene. Comparison of the locations of restriction sites within α1 and α2 and heteroduplex studies reveal extensive sequence homology within and flanking the two genes. The homologous sequences, which are interrupted by two blocks of nonhomology, span a region of approximately 4 kb. This extensive sequence homology between two genes which are thought to be the products of an ancient duplication event suggests the existence of a mechanism for sequence matching during evolution. One consequence of this arrangement of homologous sequences is the occurrence of two types of deletions in recombinant phage DNA during propagation in E. coli. The locations and sizes of the two types of deletions are indistinguishable from those of the two types of deletions associated with α-thalassemia 2 (Embury et al., 1979; Orkin et al., 1979; S. Embury et al., manuscript submitted). This information strongly suggests that the genetic disease is a consequence of unequal crossing over between homologous sequences within and/or surrounding the two adult α-globin genes.

Introduction

The α-like and β-like subunits of hemoglobin are encoded by a small group of genes which are expressed sequentially during development. The earliest embryonic hemoglobin consists of ε (β-like) and ζ (α-like) polypeptide chains. The ε polypeptide is gradually replaced by γ (fetal β) and this, in turn, is followed by the adult globins δ and β. In contrast, the ζ chain is succeeded immediately by α, which functions during both fetal and adult life (Weatherall and Clegg, 1979). Genetic studies have demonstrated linkage between the human fetal and adult β-like genes (Weatherall and Clegg, 1979), but the adult β-like and α-like genes are located on separate chromosomes (Deisseroth et al., 1977, 1978). Physical linkage between the β-like genes was demonstrated by genomic blotting experiments (Flavell et al., 1978; Mears et al., 1978; Fritsch, Lawn and Maniatis, 1979; Little et al., 1979a; Tuan et al., 1979) and examination of cloned DNA (Lawn et al., 1978; Fritsch et al., 1979; Fritsch, Lawn and Maniatis, 1980). All the β-like genes have the same transcriptional orientation and are arranged in the order 5′-ε-^{G}γ-^{A}γ-δ-β-3′. The 5′ to 3′ organization of the genes reflects the order of their expression during development. It is not known whether this developmentally correlated gene arrangement plays a role in the mechanism of sequential activation of globin genes or is simply a consequence of the duplication events which gave rise to the gene cluster.

Genetic and structural studies of α-chain variants suggest that most individuals carry at least two linked α-globin genes per haploid genome (see Bunn, Forget and Ranney, 1977; Weatherall and Clegg, 1979). It appears that both of these genes are expressed since individuals heterozygous for two different α-globin variants produce normal α-globin in addition to both of the variant polypeptides. Since only one α-globin amino acid sequence is found in most individuals (Dayhoff, 1972), the two linked genes encode identical polypeptide chains.

Physical linkage between the two adult α-globin genes was demonstrated by genomic blotting experiments. These experiments also revealed that both genes are transcribed from the same DNA strand and are separated by approximately 3.7 kb (Orkin, 1978; Embury et al., 1979). The location of the embryonic ζ-globin gene with respect to the α genes was previously unknown. In this study we used molecular cloning procedures to demonstrate physical linkage between the two α-globin genes and two additional α-like sequences, one of which appears to be ζ. We also used genomic blotting procedures to identify a second ζ-like gene located 5′ to the cloned ζ-globin gene. Finally, we present an analysis of homologous regions within the coding, intervening and flanking sequences of the two adult α-globin genes, and characterize two types of deletions which occur in recombinant phage DNA during propagation in E. coli.

Results

Localization of α-Like Sequences in Cloned DNA

Two clones denoted λHαG1 and λHαG2 containing α-globin genes were isolated from a library of human DNA (Lawn et al., 1978) by screening with a ^{32}P-labeled human α-globin cDNA plasmid, JW101 (Wilson et al., 1978). α-Globin sequences were located on the restriction map of each clone using the blot hybridization procedure (Southern, 1975). The adult α-globin genes were identified by comparison with the restriction map derived from genomic blotting experiments (Orkin, 1978; Embury et al., 1979). The two genes have the same transcriptional orientation. For convenience we will refer to the 5′ and 3′ members of the α-globin gene pair as α2 and α1, respectively. One intensely hybridizing band and two fainter bands are detected when λHαG1 DNA is digested with Hpa I plus Bam HI and hybridized to ^{32}P-labeled α-globin cDNA plasmid (Figure 1, lanes a and b). The intense band at 4.3 kb corresponds in size to a Hpa I fragment containing the α2 gene. The identification of the two fainter bands of Figure 1b will be discussed below.

The clone λHαG2 contains both adult α-globin genes. λHαG2 DNA digested with Hind III plus Bgl II displays four strongly hybridizing bands (Figure 1, lanes c and d). Hind III cuts within each α-globin gene to generate a 3.8 kb intergene fragment (Orkin, 1978) which is cleaved by Bgl II to produce 1.8 and 2.0 kb fragments. Correlation of the blot of Figure 1d with the restriction map of λHαG2 (Figure 1) indicates that the 1.8 and 2.0 kb fragments contain, respectively, the 3′ half of α2 and the 5′ half of α1. Similarly, the 10.2 kb Bgl II plus Hind III hybridizing fragment contains the 5′ half of α2 (Figure 1). The 3′ half of α1 is located within the 3.8 kb hybridizing fragment, which includes part of the lambda vector (Figure 1). These data confirm the arrangement of the adult genes derived from genomic blotting experiments (Orkin, 1978; Embury et al., 1979). A detailed restriction map of λHαG1 and λHαG2 DNA is presented in Figure 2. The precise localization of the four α-like sequences on this map will be discussed below.

Comparison of the Positions of Restriction Sites in Adult α-Globin Genes

To determine the extent of similarity between the two adult α-globin genes and to establish whether these genes contain intervening sequences, the distances between many restriction enzyme sites within and near α1 and α2 were compared with the distances between corresponding sites predicted by the nucleotide sequence of α-globin mRNA (Forget et al., 1979). This comparison was accomplished by labeling the single Hind III site within α1, α2 and the cDNA clone JW101 (Wilson et al., 1978) with ^{32}P, digesting with a second restriction enzyme and purifying the end-labeled fragments containing the 5′ and 3′ halves of the three α-like sequences. These fragments were then digested with a variety of restriction enzymes and the sizes of the products were determined by polyacrylamide gel electrophoresis to give the distance be-

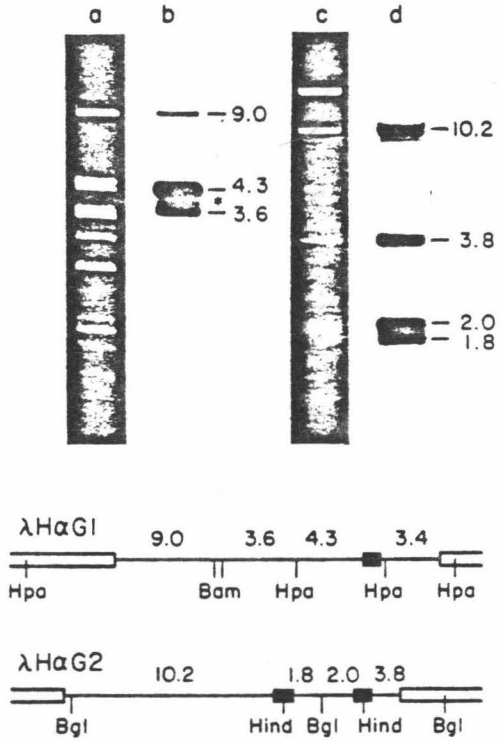


Figure 1. Localization of α-Like Globin Sequences in λHαG1 and λHαG2 DNAs

Purified λHαG1 and λHαG2 DNAs were digested with restriction endonucleases as indicated below, fractionated on 0.7% agarose gels, transferred to nitrocellulose filters (Southern, 1975) and hybridized to the ^{32}P-labeled α-globin cDNA plasmid JW101 (Wilson et al., 1978).

(a) Ethidium bromide-stained gel of λHαG1 DNA digested with Hpa I plus Bam HI.

(b) Autoradiogram of the blot of the gel shown in (a). The band labeled 4.3 corresponds to a Hpa I fragment containing the α2 gene (see λHαG1 map below autoradiogram). The fainter bands at 3.6 and 9.0 kb correspond to Bam HI plus Hpa I fragments containing α-like sequences. The very faint band marked with an asterisk does not align with a fluorescent band in (a) and therefore corresponds to a minor contaminant, the nature of which has not been determined.

(c) Ethidium bromide-stained gel of λHαG2 DNA digested with Bgl II plus Hind III.

(d) Autoradiogram of the blot of the gel shown in (c). The four bands correspond respectively to a 10.2 kb Bgl II plus Hind III fragment containing two α-like sequences plus the 5′ half of the 5′ adult α-globin gene, a 3.8 kb Hind III plus Bgl II fragment containing the 3′ half of the 3′ α-globin gene, a 2.0 kb Bgl II plus Hind III fragment containing the 5′ half of the 3′ α-globin gene and a 1.8 kb Hind III plus Bgl II fragment containing the 3′ half of the 5′ α-globin gene. These hybridizing fragments are indicated on the map of λHαG2 DNA below the autoradiogram.

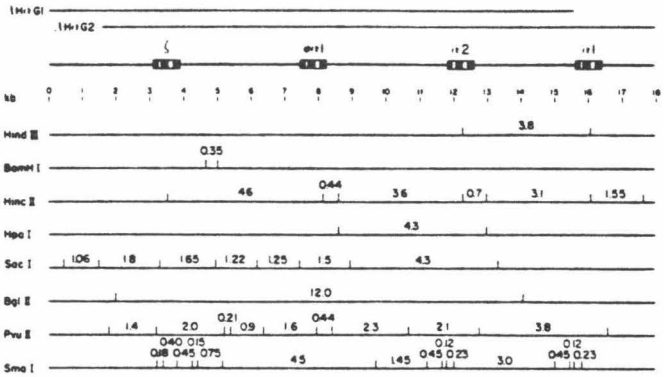


Figure 2. Map of Restriction Endonuclease Cleavage Sites in λHαG1 and λHαG2 DNAs

Restriction sites are indicated by vertical lines, and the size of each fragment is given in kilobase pairs (kb). The positions of α1, α2, ψα1 and ζ are shown. Black boxes represent mRNA coding regions and white boxes represent noncoding intervening sequences. The region of human DNA included within the insert of λHαG1 or λHαG2 is indicated by the horizontal lines at the top of the figure.

tween the Hind III site and the nearest recognition site for each enzyme.

The results of this analysis (Figure 3) indicate that both of these genes contain two intervening sequences (IVS). The 5′ half of each gene digested with Mbo II or Hinf I generates fragments 95 bp longer than the corresponding fragments of the cDNA, indicating the presence of a 95 bp IVS (IVS I). Because there is no Sma I or Hpa II site in the 5′ coding region, those sites in each gene must lie within IVS I. Hph I cleaves the 3′ half of each gene to yield a fragment 120 (α2) or 128 bp (α1) longer than the corresponding fragment of the cDNA, indicating the presence of a second intervening sequence (IVS II) in each gene. A number of sites (Hae III, Taq I, Hha I, Tha I, Bst NI, Mbo II, Alu I) are found in the 3′ half of both cloned adult genes but are not present in the corresponding positions in the cDNA clone, suggesting that these sites are within IVS II. Two IVSs have been located in the mouse α-globin gene between codons 31 and 32 and 99 and 100 (Nishioka and Leder, 1979). These locations are consistent with the approximate positions of the human α-globin IVSs I and II (Figure 3). It seems likely that the human and mouse α-globin IVS positions are identical, since the locations of two IVSs in all mammalian β-globin genes examined thus far are identical (for references see Hardison et al., 1979) and are analogous to the positions of the IVSs in the mouse α-globin gene (Nishioka and Leder, 1979).

The results presented in Figure 3 indicate that the coding, intervening and flanking sequences of α1 and α2 share considerable homology. The locations of restriction sites in the two genes can be aligned by assuming that IVS II of α1 contains a block of approximately seven nucleotides which is not found in α2. This unshared region can be located precisely as an insertion within the few nucleotides separating the Bst NI and Mbo II sites of α2 (or a deletion from α1). The homology between the coding sequences of α1 and α2 is consistent with the fact that only one α-globin protein (Dayhoff, 1972) and mRNA (for references see Forget et al., 1979) sequence have been reported. However, we were surprised to observe the homology within intervening and flanking sequences which is indicated by the similar positions in α1 and α2 of restriction sites in those regions, since other globin gene pairs show little homology in those regions (Konkel, Maizels and Leder, 1979; Hardison et al., 1979).

Heteroduplex Analysis of α1 and α2

The homology indicated by the results of Figure 3 was demonstrated directly by examining a heteroduplex between 4.3 and 5.7 kb Hpa I fragments of λHαG2 which contain α2 and α1, respectively (Figure 4A, 1–3). Alignment of the restriction map of Figure 2 with length measurements of α1–α2 heteroduplexes indicates that the region of homology begins 100 bp 3′ to the mRNA coding sequence of each gene and continues through the genes into the 5′ flanking sequence for a total of 1.8 kb (Figure 4B, 1 and 3). Beyond this, we observe a small bubble of nonhomologous DNA, a short duplex region, a second, larger nonhomology bubble and finally a 1.0 kb stretch of homologous sequence which extends to the 5′ end of the Hpa I fragments.

The large nonhomology bubble contains at least one inverted repeat sequence which is evident in the alternative structures shown in Figure 4A, 2 and 3. Since homoduplexes formed by the α2 Hpa I fragment also contain the inverted repeats (Figure 4A, 4 and 5) and no inverted repeats are observed in α1 homoduplexes, we conclude that the inverted repeats in the α1–α2 heteroduplex (Figure 4A, 2 and 3) are contributed by the α2 strand. Formation of this hairpin loop results in a reduction in the length of the duplex region 5′ to the large nonhomology bubble and a concomitant increase in the length of the shorter strand of the bubble. This suggests that part of the sequence in the hairpin stem is also present in the 5.7 kb Hpa I fragment.

To determine whether there are additional regions of sequence homology 3′ to α1 and α2, a heteroduplex

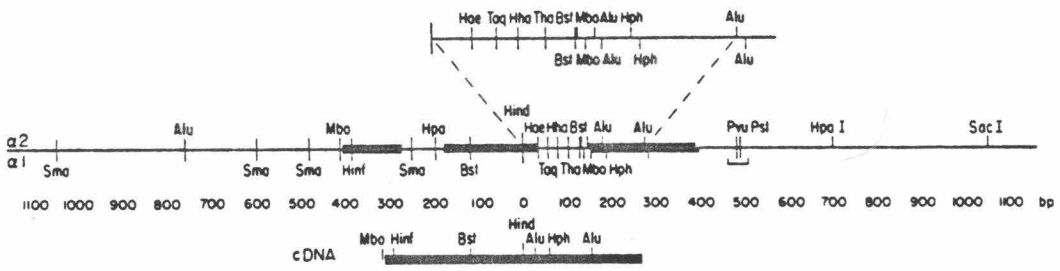


Figure 3. Comparison of the Locations of Restriction Sites in $\alpha 1$, $\alpha 2$ and α–Globin cDNA

Positions of restriction endonuclease cleavage sites within $\alpha 1$ and $\alpha 2$ were determined as described in Experimental Procedures, except for a few sites which were determined by partial or double digestion experiments. Sites within the cDNA were determined by Forget et al. (1979). Black boxes represent sequences present in mRNA, and the unshaded regions between these boxes represent intervening sequences. Abbreviations shown in the figure correspond to the following enzymes: (Alu) Alu I; (Bst) Bst NI; (Hae) Hae III; (Hha) Hha I; (Hind) Hind III; (Hinf) Hinf I; (Hpa) Hpa II; (Hph) Hph I; (Mbo) Mbo II; (Pst) Pst I; (Pvu) Pvu II; (Sma) Sma I; (Taq) Taq I; (Tha) Tha I. In cases where the cleavage site differs from the recognition site for an enzyme (for example, Mbo II) the position of the cleavage site is indicated. The number of base pairs between each site and the Hind III site within each gene was determined using labeled fragments derived from the cDNA clone as standards.

between DNA of subclones containing the 3′ half of $\alpha 1$ or $\alpha 2$ plus adjacent 3′ sequences was examined in the electron microscope. This experiment confirms that the $\alpha 1$–$\alpha 2$ homology extends only 100 bp beyond the sequences encoding the 3′ ends of the two mRNAs (Figure 4A, 6 and Figure 4B, 2 and 3). No other sequence homology is detected except for a 260 bp duplex region located about 1 kb 3′ to each gene. This 260 bp region may correspond to a sequence which is highly repeated in the human genome and is found at several locations within the β–like globin gene cluster (Fritsch et al., 1980). This is suggested by the fact that a hybridization probe containing this repeat derived from the β–like gene cluster hybridizes to four different restriction fragments within the α–like globin gene cluster (E. Fritsch and R. Lawn, personal communication). One of these fragments contains only 400 bp of human DNA and maps to the same position as the 260 bp region of homology which is located 3′ to $\alpha 1$.

The locations of homologous sequences determined by heteroduplex analysis are aligned with a restriction map of the gene cluster in Figure 4B, 3. This alignment reveals that $\alpha 1$ and $\alpha 2$ are each located within an approximately 4 kb homology unit interrupted by two regions of nonhomology. Under the conditions used in the formation of the heteroduplexes described above, a sequence with as much as 25–30% base mismatch would be scored as duplex (Davis, Simon and Davidson, 1971). The actual degree of sequence homology between the $\alpha 1$ and $\alpha 2$ regions is therefore not known.

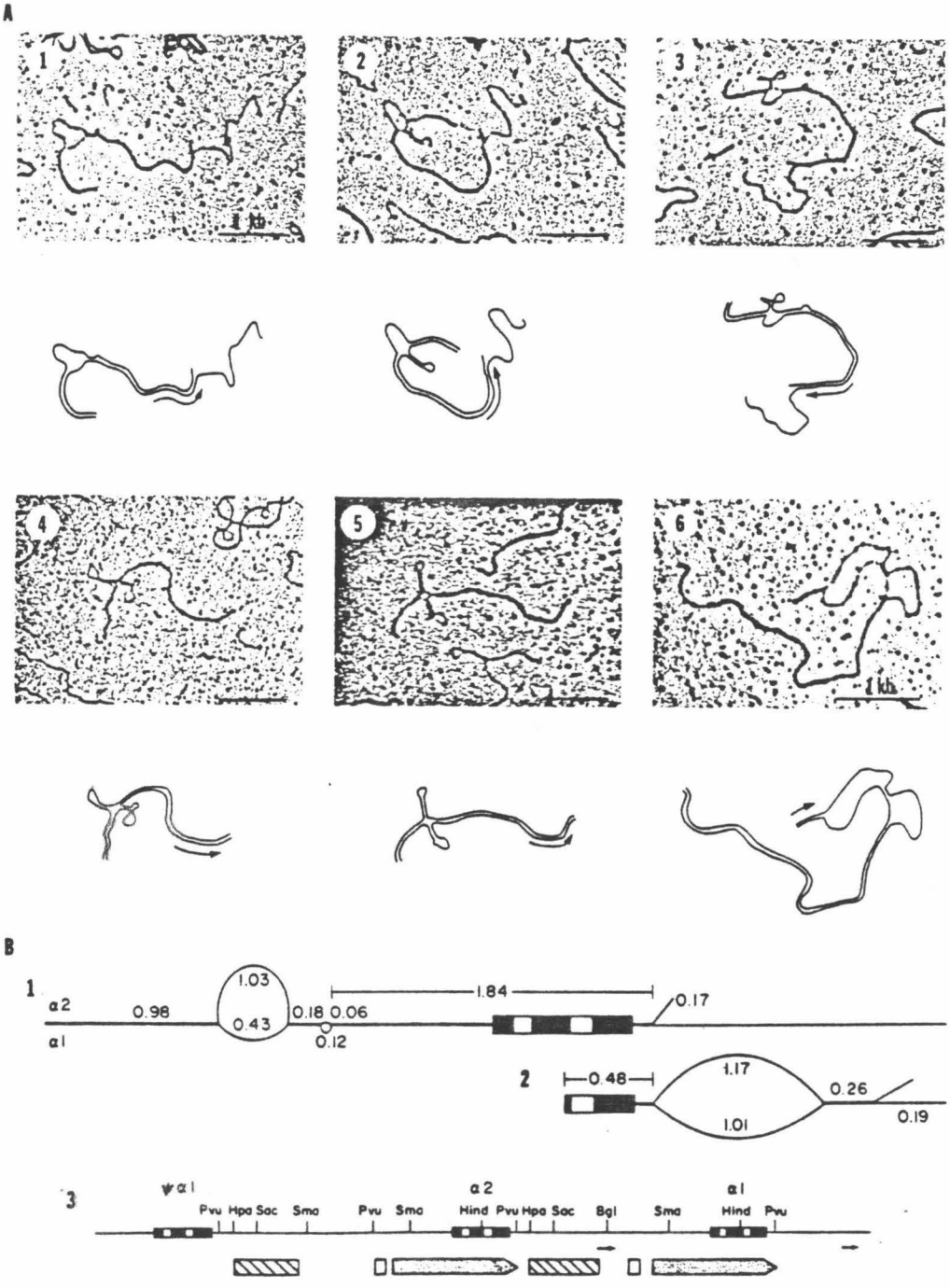
Specific Deletions in λHαG1 and λHαG2

The restriction map of Figure 2 is consistent with data obtained from genomic blotting experiments (Orkin, 1978; Embury et al., 1979; S. Embury et al., manuscript submitted; our unpublished data), indicating that the α–globin gene region can be cloned without detectable sequence rearrangements. However, both λHαG1 and λHαG2 give rise to deletions at high frequency. When these phage are purified by CsCl equilibrium sedimentation, two bands are observed, with the majority of the phage in the upper band. The

Figure 4. Heteroduplex Analysis of the Adult α–Globin Gene Regions

(A) Electron micrographs and interpretive drawings of hetero- and homoduplexes formed by restriction fragments or plasmid DNAs containing $\alpha 1$ and $\alpha 2$. The arrow in each tracing indicates the position of the mRNA coding region and the direction of transcription. (1–3) Heteroduplexes formed between the 4.3 and 5.7 kb Hpa I fragments of λHαG2 DNA (these fragments contain $\alpha 2$ and $\alpha 1$, respectively). 60% of the molecules are of the type shown in (1). Comparison of the summed lengths of single- and double-stranded regions with the known lengths of the two Hpa I fragments indicates that the 1.03 kb strand of the larger nonhomology bubble is contributed by the 4.3 kb Hpa I fragment. Frequently this single-stranded DNA forms a hairpin loop (2 and 3). (4 and 5) Homoduplexes formed by denaturation and reassociation of the 4.3 kb Hpa I fragment. The length of duplex DNA 5′ to the hairpin loops is the same as that 5′ to the large nonhomology bubble in the heteroduplexes of (2) and (3). (6) Heteroduplexes formed between the plasmid pRH$\alpha 2$ cut with Hind III (this molecule contains the 3′ half of $\alpha 2$ plus 3′ flanking sequences) and the plasmid pBR$\alpha 1$ cut with Hind III plus Sal I (this molecule contains the 3′ half of $\alpha 1$ plus 3′ flanking sequences). Comparisons of the summed lengths of single- and double-stranded regions with the lengths of the human DNA inserts in pRH and pRB indicate that the 1.17 and 1.01 kb strands are from the regions 3′ to $\alpha 2$ and $\alpha 1$, respectively.

(B) Schematic drawings showing the locations of homologous and nonhomologous sequences between the $\alpha 1$ and $\alpha 2$ regions. (1) A heteroduplex exemplified by the molecule of (A1). The double- and single-stranded DNA regions are represented by thick and thin lines, respectively, and the lengths are given in kilobase pairs and kilonucleotides, respectively. (2) A heteroduplex exemplified by the molecule of (A6). (3) A map showing the positions of restriction enzyme sites in relation to the three regions of sequence homology represented by the crosshatched boxes, the small open boxes and the stippled arrows. The small arrows represent the 260 bp region of homology located on the 3′ side of $\alpha 1$ and $\alpha 2$. The open areas between the boxes and arrows correspond to regions of nonhomology.


A
1
1 kb
2
3
4
5
6
1 kb
B
1
α2
α1
0.98
1.03
0.43
0.18
0.06
0.12
1.84
0.17
2
0.48
1.17
1.01
0.26
0.19
3
ψα1
α2
α1
Pvu
Hpa
Sac
Sma
Pvu
Sma
Hind
Pvu
Hpa
Sac
Bgl
Sma
Hind
Pvu

restriction map of the human DNA insert in lower band phage DNA corresponds to the map derived by genomic blotting experiments, whereas two types of deletions are detected in upper band phage DNA of both λHαG1 and λHαG2.

The breakpoints of these deletions were mapped as described in the Appendix and are shown in Figure 5. The breakpoints of the leftward type of deletion which removes 4.3 kb of DNA map within regions of sequence homology located approximately 3–4 kb 5′ to α1 and α2 (Figure 5A). The distance between corresponding restriction sites within these regions (for example, Hpa I to Hpa I or Sac I to Sac I; Figure 2) is also 4.3 kb, strongly suggesting that the leftward type of deletion is generated by unequal crossing over between homologous sequences. Similarly, the breakpoints of the rightward type of deletion which removes 3.8 kb of DNA map within the regions of sequence homology within and immediately flanking α1 and α2. The corresponding restriction sites in these homologous regions are separated by 3.8 kb (for example, Hind III to Hind III or Pvu II to Pvu II; Figure 2), suggesting that the rightward type of deletion is generated by unequal crossing over between these homologous regions. Thus the occurrence of both types of deletions must be a consequence of the distribution of direct repeats around the α1 and α2 genes. We do not know whether this recombination is intra- or intermolecular. Since recombination can occur anywhere within the repeats, the breakpoints of the deletions cannot be mapped precisely. In fact, the leftward and rightward types of deletions might each consist of a number of different deletions generated by recombination at sites anywhere within the region of homology. This uncertainty is represented by the dashed arrows in Figure 5.

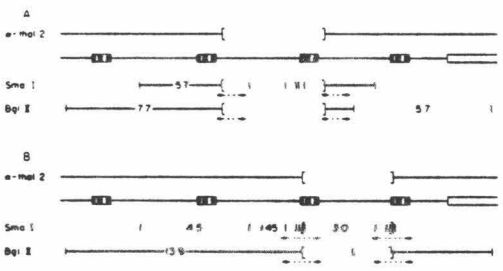


Figure 5. Locations of Deletions in λHαG1 and λHαG2 DNAs

Restriction enzyme cleavage sites are indicated by vertical lines, and the size of each fragment is given in kilobase pairs. Fragments detected only in upper band phage DNA are represented by horizontal lines interrupted by a bracketed region which indicates the approximate location of a deletion. The lines labeled "α–thal 2" indicate the positions of deletions in the DNA of individuals with α–thalassemia (Embury et al., 1979; Orkin et al., 1979; S. Embury et al., manuscript submitted). The dashed arrows below the brackets represent the limits over which the deletion breakpoints can occur in either the cloned DNA or the α–thal 2 deletions. The length of each arrow equals the length of the region of homology (Figure 4B). (See Appendix for a discussion of this point and a description of the mapping data). (A) Leftward-type deletion; (B) rightward-type deletion.

Characterization of Linked α–Like Globin Sequences

The restriction fragments which hybridize weakly to the α–globin cDNA plasmid (Figure 1b) were located on the map of Figure 2 by blot hybridization experiments and DNA sequence analysis. Table 1 shows a portion of the nucleotide sequence of the two non-adult α–like regions which can be aligned with codons 73–96 of the α–globin mRNA sequence. This sequence information precisely locates each α–like sequence on the map of Figure 2 and establishes the relative transcriptional orientation of each gene. In both non-adult α–like regions the mRNA sequence is read from left to right (5′ to 3′) in the map of Figure 2. α1 and α2 have the same transcriptional orientation.

An embryonic α–like polypeptide, ζ, has been identified (Huehns et al., 1961; Capp, Rigas and Jones, 1967, 1970; Weatherall, Clegg and Wong, 1970; Todd et al., 1970; Huehns and Farooqui, 1975; Kamuzora and Lehmann, 1975; Gale, Clegg and Huehns, 1979). The chromosomal location of the ζ–globin gene has not been established. To determine whether either of the α–like sequences linked to α1 and α2 encodes an embryonic ζ–globin, the amino acid sequence predicted by the nucleotide sequence corresponding to codons 73–96 of each gene was compared with amino acid sequence data for the ζ protein. Neither sequence could be aligned with a preliminary sequence derived by matching the amino acid composition of ζ and α chain peptides (Kamuzora and Lehmann, 1975). However, recent peptide composition data derived from a more purified preparation of ζ chain (J. Clegg, personal communication) provide a sequence for codons 73–89 which is identical to that predicted by the nucleotide sequence of the 5′-most gene. We therefore tentatively identify this sequence as a ζ–globin gene. The α–like sequence between ζ and α2 does not encode any known globin polypeptide (for convenience, we refer to this sequence as ψα1). The complete nucleotide sequence of this gene has been determined and will be reported elsewhere (N. Proudfoot and T. Maniatis, manuscript submitted).

In addition to the four α–like sequences contained in λHαG1 and λHαG2, we have detected one other α–like sequence in the human genome. DNA from four individuals was digested with the restriction enzyme Bgl II and probed with a labeled restriction fragment containing the 5′ half of the cloned ζ–globin gene. Comparison of the map of Figure 2 with the autoradiogram of Figure 6A indicates that each of the four DNAs contains a hybridizing fragment of 12 kb which corresponds in size to the λHαG1 Bgl II fragment containing a ζ–globin gene. In addition, at least one other Bgl II fragment not found in the cloned DNA is

Table 1. A Portion of the Nucleotide Sequences of the Non-adult α–Like Globin Regions of λHαG1

Codon		73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
nt	α	GUG	GAC	GAC	AUG	CCC	AAC	GCG	CUG	UCC	GCC	CUG	AGC	GAC	CUG	CAC	GCG	CAC	AAG	CUU	CGG	GUG	GAC	CCG	GUC
	ψα1	UUA	GAU	GAC	AUG	CCC	AAU	GAU	GUG	UCU	GAG	CUG	AGG	AAG	CUG	CAU	GUC	CAC	GAG	CUG	UGG	GUG	GAC	CCA	GGC
	ζ	AUC	GAC	GAC	AUC	GGC	GGC	GCC	CUG	UCC	AAG	CUG	AGC	GAG	CUG	CAC	GCC	UAC	AUC	CUG	CGC	GUG	GAC	CCG	GUC
aa	α	Val	Asp	Asp	Met	Pro	Asn	Ala	Leu	Ser	Ala	Leu	Ser	Asp	Leu	His	Ala	His	Lys	Leu	Arg	Val	Asp	Pro	Val
	ψα1	Leu	Asp	Asp	Met	Pro	Asn	Asp	Val	Ser	Glu	Leu	Arg	Lys	Leu	His	Val	His	Glu	Leu	Trp	Val	Asp	Pro	Gly
	ζ	Ile	Asp	Asp	Ile	Gly	Gly	Ala	Leu	Ser	Lys	Leu	Ser	Glu	Leu	His	Ala	Tyr	Ile	Leu	Arg	Val	Asp	Pro	Val

Corresponding regions of the ζ and ψα1 sequences determined as described in Experimental Procedures and the α–globin cDNA sequence (Forget et al., 1979) are aligned for comparison. The amino acid sequence encoded in the three types of α–like globin sequences is also shown. Nucleotides and amino acids are indicated by (nt) and (aa), respectively.

Figure 6. Identification of a Second, Linked ζ–Globin Gene

(A) DNA from four individuals (lanes 1–4) was digested with Bgl II and the products were fractionated on a 0.7% agarose gel, transferred to nitrocellulose (Southern, 1975) and hybridized with a ^{32}P-labeled Pvu II–Hinc II fragment containing the 5′ half of ζ1.

(B) DNA from individual 1 of (A) was digested with Bgl II plus Hpa I (lane 1), Bgl II plus Eco RI (lane 2) or Bgl II (lane 3), fractionated on a 0.7% agarose gel, transferred to nitrocellulose and probed with a ^{32}P-labeled Eco RI–Sac I fragment containing the terminal 0.49 kb of the human DNA insert of λHαG1. (See Figure 7 for the location of this fragment.)

detected in DNA from each individual. The size of this latter fragment varies between individuals, suggesting the presence of restriction site polymorphisms in the regions flanking the ζ–like sequence. Individual 4 appears to be heterozygous for polymorphisms in the Bgl II sites surrounding both ζ–globin sequences, since four different Bgl II fragments are detected (Figure 6A).

To determine whether the second ζ–like sequence is a nonallelic ζ–globin gene, we established a physical map of the restriction sites surrounding this sequence (Figure 7). The sizes of hybridizing fragments resulting from single or double digestion experiments are presented in Table 2. Analysis of these data made it possible to construct a restriction map distinct from that of the cloned ζ–globin gene, indicating that the hybridizing sequence does in fact correspond to a second ζ–globin gene.

The data in Table 2 also suggest that the two ζ sequences are linked, as indicated in the map of Figure 7. Linkage was definitively established by genomic blotting experiments using a terminal fragment of λHαG1 as a hybridization probe. As expected from the map of Figure 7, this probe detected the 10.9 kb Bgl II fragment containing the ζ–like sequence, as well as fragments of 6.8 and 5.2 kb resulting from double digests using Bgl II plus Eco RI and Bgl II plus Hpa I, respectively (Figure 6B). The sizes of these latter

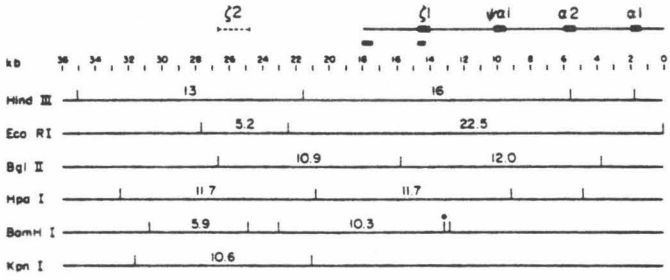


Figure 7. Linkage Arrangement of Embryonic and Adult α-Like Globin Genes

The sizes of fragments detected by a ζ1 probe (Table 2) were used to construct a map of restriction sites surrounding the ζ2 sequence. The solid line connecting ζ1, ψα1, α2 and α1 indicates the DNA region which has been cloned. The locations of the ζ1 probe, and of the λHαG1 terminal fragment probe used in the experiment of Figure 6B, are indicated by bars above the kb scale. The asterisk marks a Bam HI site present in the cloned DNA but not in the DNA of individual 1.

Table 2. Sizes of Fragments Detected by ζ1 Probe in Genomic Blots

Hind III	16, 13	Hpa I	11.7 doublet
Eco RI	22.5, 5.2	Bam HI	10.3, 5.9
Bgl II	12.0, 10.9	Kpn I	>40, 10.6
Bgl II/Hind III	10.2, 5.1	Bam HI/Eco RI	10, 2.8
Bgl II/Eco RI	12.0, 4.2	Bam HI/Bgl II	3.0, 1.9
Hpa I/Hind III	11.7, 11.1	Kpn I/Hind III	10.2, 15.5
Bam HI/Hind III	9.0, 5.8		
Pvu II	2.0, 1.5	Sal I	>40
		Xho I	>40

DNA from individual 1 of Figure 6A was digested with the indicated enzymes, fractionated on 0.7% agarose gels, transferred to nitrocellulose, and probed with a ^{32}P-labeled Pvu II–Hinc II fragment containing the 5′ half of ζ1. The sizes of hybridizing fragments resulting from each digest are given in kilobase pairs.

fragments are those predicted by the linkage map of Figure 7. We therefore conclude that the ζ-like sequence corresponds to a second ζ-globin gene located 10–12 kb 5′ to the cloned ζ-globin gene. For convenience, we will refer to the cloned ζ-globin gene and the ζ-globin gene identified by genomic blotting as ζ1 and ζ2, respectively.

ζ1 and ζ2 appear to share considerable sequence homology. Identical strips of Bgl II or Hind III blots were hybridized with ζ1 probe and washed in 0.33, 0.16 or 0.08 × SSC at 68°C. Both the ζ1 and ζ2 bands are still visible and are equally intense after washing in 0.08 × SSC (data not shown).

Discussion

Linkage Arrangement of Globin Genes

We have used molecular cloning and genomic blotting procedures to establish the linkage arrangement of the entire human α-like globin gene cluster. The cluster consists of two adult α-globin genes (α1 and α2), an apparently nonfunctional α-like gene (ψα1) and two embryonic ζ-globin genes (ζ1 and ζ2). An exact correspondence between the amino acid sequence encoded by a short region of ζ1 and the sequence of the corresponding region of the ζ-globin polypeptide suggests that ζ1 is a functional globin gene. This suggestion is strengthened by comparison of additional amino acid (J. Clegg, personal communication) and DNA sequence (C. O'Connell, personal communication) information. ζ2, on the other hand, has been demonstrated to be a functional ζ-globin gene by analysis of a deletion associated with a case of hydrops fetalis (L. Pressley, B. R. Higgs, J. B. Clegg and D. J. Weatherall, manuscript submitted). The ζ1 sequence information and the fact that this gene is indistinguishable from the functional ζ2 gene in hybridization-melting experiments suggest that there are two functional ζ-globin genes.

The chromosomal organization of the human α-like and β-like globin gene clusters is very similar. Within each cluster all the genes have the same transcriptional orientation. The α-like genes are arranged in the order 5′-embryonic-adult-3′ (this paper), while the β-like genes are arranged in the order 5′-embryonic-fetal-adult-3′ (Fritsch et al., 1980). A similar organization has been reported for the β-like globin gene cluster of rabbit (Hardison et al., 1979; Lacy et al., 1979) and the α-like gene cluster of chicken (Engel and Dodgson, 1980). An apparent exception to this pattern is the chicken β-like globin gene cluster, where a sequence which preferentially hybridizes embryonic globin mRNA is located 3′ to an adult globin gene. This sequence, however, has not been definitively identified as a functional embryonic gene (Dodgson, Strommer and Engel, 1979; D. Engel, personal communication). The possible significance of globin gene linkage has been discussed elsewhere (Fritsch et al., 1979).

Another feature common to globin gene clusters is the presence of globin-like sequences such as ψα1 which cannot be identified with known polypeptides. The complete nucleotide sequence of ψα1 indicates that it cannot encode a functional globin polypeptide (N. Proudfoot and T. Maniatis, manuscript submitted). Two globin-like sequences which have not been identified with polypeptides are found within the human β-like gene cluster. One of these sequences (ψβ2) is located 5′ to the ε-globin gene and the other (ψβ1) is located between the ^{A}γ- and δ-globin genes (Fritsch

et al., 1980). Similarly, an apparently unexpressed β-like sequence (β2) is found 5′ to the rabbit adult β-globin gene (Hardison et al., 1979; Lacy et al., 1979). These sequences may represent vestigial genes which resulted from gene duplication followed by sequence divergence.

In many cases, globin genes immediately adjacent to one another are expressed at the same stage of development. Examples of this pairwise arrangement are the human δ and β (Flavell et al., 1978; Lawn et al., 1978; Mears et al., 1978), $^{G}\gamma$ and $^{A}\gamma$ (Fritsch et al., 1979, 1980; Little et al., 1979; Tuan et al., 1979) and α1 and α2 genes (Orkin, 1978), the rabbit β3 and β4 genes (Hardison et al., 1979), the mouse β_{maj} and β_{min} genes (Konkel et al., 1979; M. Edgell, personal communication) and the chicken α_A and α_D genes (Engel and Dodgson, 1980). The single rabbit adult β gene (Hardison et al., 1979) and the human ε gene (Proudfoot and Baralle, 1979; Fritsch et al., 1980) are exceptions. However, each of these genes is immediately adjacent to a β-like sequence of unknown function. In many cases, comparison of the members of a gene pair indicates that the intervening and flanking sequences are more divergent than the coding sequences (Hardison et al., 1979; Konkel et al., 1979). The human $^{G}\gamma$ and $^{A}\gamma$ genes, however, are essentially identical throughout their coding, intervening and flanking sequences (J. Slightom, A. Blechl and O. Smithies, manuscript in preparation), as is the case for the human α-globin genes.

Maintenance of Sequence Homology between Adult α-Globin Genes during Evolution

Genetic studies and structural analyses of α-globin polypeptides from a variety of vertebrate species indicate that duplication of adult α-globin genes is not limited to humans (Clegg, 1974; Kitchen, 1974; Nute, 1974; Russell and McFarland, 1974; Dresler et al., 1974; Chapman, Tobin and Hood, 1980). Comparison of α-globin amino acid sequences between various primate species reveals differences of 1–7%, which are consistent with the expected rate of divergence of globins over evolutionary time (Dayhoff, 1972). However, the two adult α-globin proteins within a species show considerably less divergence. For example, human and gorilla α-globins differ by 6%, yet no differences are observed when the intraspecies comparisons are made. The observed similarity between the α-globin proteins within a species could result from recent independent duplication events or from a mechanism for gene correction. Comparison of α-globin genes from a number of primate species has revealed a remarkable similarity in the distribution of restriction sites surrounding the genes (Zimmer et al., 1980), suggesting that the α-globin gene duplication occurred prior to the time of primate divergence. This observation, in conjunction with the evidence for α-globin gene duplication in most vertebrates, indicates that the α-globin gene duplication is ancient, and therefore favors the existence of a process which leads to sequence matching between α-globin genes within a species.

Maintenance of homology among a family of evolving genes within a species has been termed coincidental evolution (Hood, Campbell and Elgin, 1975). Coincidental evolution can be imposed by natural selection, or by a mechanism for gene correction. Two such mechanisms are gene conversion and expansion and contraction of gene number by homologous but unequal crossing over (Smith, 1973; Tartof, 1973). Comparison of the complete nucleotide sequences of the $^{G}\gamma$- and $^{A}\gamma$-globin genes has led to a specific proposal of intrachromosomal gene conversion (J. Slightom, A. Blechl and O. Smithies, manuscript in preparation). Similarly, intrachromosomal recombination has been proposed as an explanation for the presence of an identical restriction enzyme site polymorphism within the intervening sequences of the two linked human γ-globin genes (Jeffreys, 1979).

Evidence for unequal crossing over between human α-globin genes is provided by analysis of deletion-type α-thalassemias, which are caused by deletion of one (or, less frequently, both) α-globin genes (Dozy et al., 1979; Embury et al., 1979; Orkin et al., 1979; S. Embury et al., manuscript submitted). Moreover, chromosomes containing three α-globin genes have recently been observed (Goossens et al., 1980; D. Higgs and J. Clegg, personal communication). Unequal crossing over has been proposed to explain the sequence homology between the human $^{G}\gamma$- and $^{A}\gamma$-globin genes (Little et al., 1979b). Coincidental evolution of duplicated α-globin genes resulting from gene conversion or unequal crossing over has been proposed independently on the basis of a comparison of the positions of restriction enzyme sites surrounding the α-globin genes in various primate species (Zimmer et al., 1980).

It is not known why intra- or interchromosomal recombination mechanisms would operate on the α- and γ-globin gene pairs and not on the other more divergent globin gene pairs mentioned above. It is possible that the α and γ genes lie adjacent to special sites which promote recombination (Stahl, 1979). Another possibility is that the homology units of the more divergent gene pairs were interrupted at some time after the duplication event by a random insertion of DNA into the noncoding region within or flanking one of the two genes. This region of nonhomology would be expected to reduce the frequency of recombination, as is the case in bacteriophage lambda (Sodergren and Fox, 1979) and in yeast (G. Fink, personal communication), and thus allow the flanking and intervening sequences of the two genes to diverge. Perhaps the large and small nonhomology bubbles

which we have detected 5′ to α1 and α2 are the results of such insertion events and the genes are in the early stages of divergence.

Globin Gene Deletions and α-Thalassemia

The occurrence of deletions in clones containing α-globin genes and their correlation with the locations of extensive regions of homology surrounding those genes strongly suggests that intra- or interchromosomal unequal crossing over occurs during propagation of these recombinant phage in E. coli. It is remarkable that the locations of the breakpoints of deletions which occur in cloned DNA are indistinguishable from those associated with a form of α-thalassemia designated α-thalassemia 2. α-Thalassemia 2 has been shown to result from a deletion of one of the two α-globin genes (Orkin et al., 1979). Two types of α-thalassemia 2 deletions have been characterized. One type corresponds to the leftward-type deletion described above (Figure 5A) and is found primarily among Asians, whereas the other type corresponds to the rightward-type deletion in cloned DNA (Figure 5B) and occurs frequently in Asian, black and Mediterranean populations (S. Embury et al., manuscript submitted). Thus the precise localization of regions of homology surrounding the α-globin genes in cloned DNA and the occurrence of deletions in E. coli which appear to involve these sequences provide an explanation for the occurrence of specific deletions associated with α-thalassemia 2.

The structural studies reported here in conjunction with studies of the molecular basis of α-thalassemia portray a dynamic gene family undergoing rapid evolutionary change. This family includes genes which have specialized to function at different developmental stages (ζ and α), a gene for which no globin polypeptide has been identified and which may therefore no longer be functional (ψα1), and genes which are presently undergoing deletion and duplication in human populations (α1 and α2).

Appendix: Mapping Deletions in λHαG1 and λHαG2 DNAs

Information regarding the locations and lengths of deleted regions in λHαG2 DNA is provided by comparing Bgl II digests of upper and lower band DNAs. Digestion of lower band DNA with Bgl II produces a 12.0 kb fragment which contains α2 and a 5.7 kb fragment which contains α1 and extends into the lambda vector (Figure 8A; Figure 5). The 12.0 kb Bgl II fragment is missing from upper band DNA and the relative amount of the 5.7 kb fragment is reduced (Figure 8B). This suggests that one type of deletion removes part of the 12.0 and 5.7 kb Bgl II fragments (rightward deletion) while another type of deletion (leftward) is contained entirely within the 12.0 kb fragment, sparing the 5.7 kb fragment. The missing 12.0 and 5.7 kb Bgl II fragments in upper band DNA are replaced by fragments of 13.9 and 7.7 kb (Figure 8B). The 13.9 kb fragment cannot be a deleted form of the 12.0 kb fragment and must therefore correspond to the rightward type of deletion which removes part of both the 12.0 and 5.7 kb fragments. The deletion event removes the Bgl II site which ordinarily separates these two fragments, resulting in a fusion of the remaining parts of these fragments to yield the 13.9 kb fragment (Figure 5). In intact DNA these two fragments total 17.7 kb (12.0 plus 5.7 kb). 3.8 kb (17.7 minus 13.9 kb) have been removed by the rightward type of deletion. The 7.7 kb fragment must result from the deletion which is contained entirely within the 12.0 kg fragment (Figure 5). 4.3 kb (12.0 minus 7.7 kb) are removed by the leftward type of deletion.

Sma I was used to locate further the breakpoints of these two types of deletions. Lower band DNA digested with Sma I displays fragments of 4.5 and 1.45 kb which lie to the left of α2, and a 3.0 kb fragment which contains most of α2 and most of the region between α2 and α1 (Figure 8C; Figure 5). In upper band DNA both the 4.5 and 1.45 kb Sma I fragments are reduced in amount relative to that of fragments unaffected by the deletions, and the 3.0 kb fragment is absent (Figure 8D). Both the leftward and rightward types of deletions must therefore remove part or all of the 3.0 kb fragment, which is consistent with the fact that no α2 coding sequence can be detected (data not shown). Because the 4.5 kb fragment is reduced in amount, the leftward type of deletion must extend from α2 to include the Sma I site which lies between the 1.45 and 4.5 kg fragments (Figure 5). The 5.7 kb fragment present in upper band DNA digested with Sma I (Figure 8D) is most likely a fusion fragment resulting from the leftward-type deletion (Figure 5). A small amount of the 1.45 and 4.5 kb fragments remains, however, so most or all of the rightward-type deletions must lie to the right of the 1.45 kb fragment, beginning within a region which includes the 0.45 kb Sma I fragment and the α2 gene (Figure 5). The presence of an approximately 4 kb deletion has been confirmed by examination of heteroduplexes between upper and lower band DNAs (data not shown).

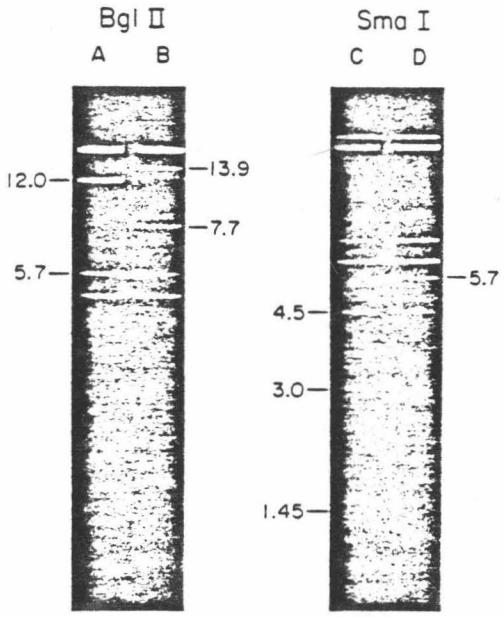


Figure 8. Comparison of Restriction Endonuclease Digests of Intact and Deleted λHαG2 DNA

The recombinant phage λHαG2 was grown in liquid culture and purified by CsCl sedimentation equilibrium as described in Experimental Procedures. The two phage bands were collected separately and rerun on separate equilibrium gradients to obtain purified upper and lower band DNAs. These DNAs were digested with Bgl II or Sma I and the products were fractionated by electrophoresis in a 0.7% agarose gel. (A) Lower band DNA digested with Bgl II; (B) upper band DNA digested with Bgl II; (C) lower band DNA digested with Sma I; (D) upper band DNA digested with Sma I. The faint, high molecular weight fragment in lanes C and D is due to reassociation of vector fragments at the cos site of lambda.

Experimental Procedures

Isolation and Characterization of Human α-Like Globin Genes

A bacteriophage λ library of human fetal liver DNA (Lawn et al., 1978) was screened with ^{32}P-labeled α-globin cDNA plasmid (JW101; Wilson et al., 1978) using procedures described by Fritsch et al. (1980). The isolation of recombinant bacteriophage DNA, the recovery of specific fragments from agarose or acrylamide gels, the mapping of restriction endonuclease cleavage sites and blotting and hybridization experiments were carried out using published procedures (Maniatis et al., 1978; Lacy et al., 1979; Fritsch et al., 1980).

Construction of Plasmid Subclones

Restriction endonuclease fragments of human DNA were isolated from λHαG1 or λHαG2 and cloned in pBR322 (Rodriguez et al., 1976) digested with the appropriate enzymes, essentially as described by Lacy et al. (1979). pRBα1 extends from the right-hand λHαG2 artificial Eco RI site (this site was created by the addition of Eco RI linkers during construction of the human library) to the Bgl II site between α1 and α2. pRHα2 extends from the λHαG1 right-hand artificial Eco RI site to the Hind III site within α2. pHBα2 ψα1 extends from the Hind III site within α2 to the right-hand Bam HI site of the pair of sites located at about 5 kb in the map of Figure 2. pBRζ extends from the left-hand Bam HI site (Figure 2) to the artificial Eco RI site which constitutes the left-hand boundary of λHαG1.

Fine-Structure Restriction Mapping

In those cases where single or double restriction endonuclease digestion yielded numerous small DNA fragments, the order of these fragments was determined using the double-stranded exonuclease of Alteromonas espejiana BAL31, which was provided by H. Gray. A plasmid subclone was linearized by cleavage at one of the sites bounding the human DNA insert, and then exonuclease and restriction endonuclease digestions were carried out as described (Legerski, Hodnett and Gray, 1978).

The locations of restriction sites in α1, α2 and the α-globin cDNA were compared by digesting pHBα2ψα1, pRHα2, pRBα1 and JW101 with Hind III and end-labeling with ^{32}P. pHBα2ψα1 and pRHα2 were then redigested with Pvu II, pRBα1 was redigested with Pst I and JW101 was redigested with Hinf I, and end-labeled gene-containing fragments were isolated by gel purification. These fragments were then digested with a variety of enzymes and the sizes of resulting fragments were determined by electrophoresis in parallel with labeled markers in an acrylamide gel.

DNA Sequence Analysis

pBRζ DNA was digested with Hinc II, end-labeled with ^{32}P, redigested with Sac I, eluted from an acrylamide gel and sequenced by the procedure of Maxam and Gilbert (1977). pHBα2ψα1 DNA was digested with Pvu II, end-labeled, redigested with Sac I and sequenced as above.

Electron Microscopy

To examine the extent of homology between sequences within and 5′ to α1 and α2, λHαG2 DNA was digested with Hpa I and the gene-containing fragments (5.7 and 4.3 kb, respectively) were purified by agarose gel electrophoresis. Heteroduplexes were formed in 0.18 M $NaPO_4$ at 60°C, as described previously (Shen and Maniatis, 1980). To examine the homology between the 3′ flanking sequences of the two genes, the plasmids pRHα2 and pRBα1 were digested with Hind III and Hind III plus Sal I, respectively. The DNA was phenol-extracted, ether-extracted and ethanol-precipitated, and heteroduplexes were formed as described above. The DNA samples were spread in 50% formamide, 0.1 M Tris–HCl, 0.01 M Na EDTA (Davis et al., 1971). Single- and double-stranded φX DNA were used as length markers. The grids were shadowed and examined in a Philips 300 electron microscope.

Acknowledgments

We thank J. Clegg, S. Embury, E. Fritsch, Y. W. Kan, R. Lawn, N. Proudfoot, O. Smithies and A. Wilson for permission to cite their unpublished results. We are grateful to L. Hood, N. Proudfoot and B. Seed for discussions. We thank A. Cortenbach for preparing media and materials. J. L. and C.-K. J. S. were supported by NIH graduate and postdoctoral training grants, respectively, to the California Institute of Technology, and T. M. was supported by the Rita Allen Foundation. This research was funded by a grant from the NIH.

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

Received January 22, 1980; revised February 25, 1980

References

Bunn, H. F., Forget, B. G. and Ranney, H. M. (1977). Human Hemoglobins (Philadelphia: W. B. Saunders).

Capp, G., Rigas, D. and Jones, R. (1967). Science *157*, 65–66.

Capp, G., Rigas, D. and Jones, R. (1970). Nature *228*, 278–290.

Chapman, B., Tobin, A. and Hood, L. (1980). Dev. Biol., in press.

Clegg, J. B. (1974). Ann. NY Acad. Sci. *241*, 61–69.

Davis, R. W., Simon, M. and Davidson, N. (1971). Meth. Enzymol. *210*, 413–428.

Dayhoff, M. O. (1972). Atlas of Protein Sequence and Structure, 5 (Washington, D.C.: National Biomedical Research Foundation).

Deisseroth, A., Nienhuis, A., Turner, P., Velez, R., Anderson, W. F., Ruddle, F., Lawrence, J., Creagan, R. and Kucherlapati, R. (1977). Cell *12*, 205–218.

Deisseroth, A., Nienhuis, A., Lawrence, J., Giles, R., Turner, P. and Ruddle, F. H. (1978). Proc. Nat. Acad. Sci. USA *75*, 1456–1460.

Dodgson, J. B., Strommer, J. and Engel, J. D. (1979). Cell *17*, 879–887.

Dozy, A., Kan, Y. W., Embury, S., Mentzer, W., Wang, W., Lubin, B., Davis, J. and Koenig, H. (1979). Nature *280*, 605–607.

Dresler, S. L., Runkel, D., Strenzel, P., Brimhall, B. and Jones, R. T. (1974). Ann. NY Acad. Sci. *241*, 411–415.

Embury, S., Lebo, R., Dozy, A. and Kan, Y. W. (1979). J. Clin. Invest. *63*, 1307–1310.

Engel, J. D. and Dodgson, J. B. (1980). Proc. Nat. Acad. Sci. USA, in press.

Flavell, R. A., Kooter, J. M., De Boer, E., Little, P. F. R. and Williamson, R. (1978). Cell *15*, 25–41.

Forget, B. G., Davallesco, C., de Riel, J. K., Spritz, R. A., Choudary, P. V., Wilson, J. T., Wilson, L. B., Reddy, V. B. and Weissman, S. M. (1979). In Eukaryotic Gene Regulation, R. Axel, T. Maniatis and C. F. Fox, eds. (New York: Academic Press).

Fritsch, E. F., Lawn, R. M. and Maniatis, T. (1979). Nature *279*, 598–603.

Fritsch, E. F., Lawn, R. M. and Maniatis, T. (1980). Cell *19*, 959–972.

Gale, R., Clegg, J. and Huehns, E. (1979). Nature *280*, 162–164.

Goossens, M., Dozy, A., Embury, S., Zacharides, Z., Hadjiminis, M., Stamatoyannopoulos, M. and Kan, W. Y. (1980). Proc. Nat. Acad. Sci. USA *77*, 518–521.

Hardison, R. C., Butler, E. T., III, Lacy, E., Maniatis, T., Rosenthal, N. and Efstratiadis, A. (1979). Cell *18*, 1285–1297.

Hood, L., Campbell, J. H. and Elgin, S. C. R. (1975). Ann. Rev. Genet. *9*, 305–353.

Huehns, E. and Farooqui, A. (1975). Nature *254*, 335–337.

Huehns, E., Flynn, F., Butler, E. and Beaven, G. (1961). Nature *189*, 496–497.

Jeffreys, A. J. (1979). Cell *18*, 1–10.

Kamuzora, H. and Lehmann, H. (1975). Nature *256*, 511.

Kitchen, H. (1974). Ann. NY Acad. Sci. *241*, 12–24.

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

Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. and Maniatis, T. (1978). Cell *15*, 1157–1174.

Lacy, E., Hardison, R. C., Quon, D. and Maniatis, T. (1979). Cell *18*, 1273–1283.

Legerski, R., Hodnett, J. and Gray, H. B., Jr. (1978). Nucl. Acids Res. *5*, 1445–1464.

Little, P. F. R., Flavell, R. A., Kooter, J. M., Annison, G. and Williamson, R. (1979a). Nature *278*, 227–231.

Little, P. F. R., Williamson, R., Annison, G., Flavell, R. A., De Boer, E., Bernini, L. F., Ottolenghi, S., Saglio, G. and Mazza U. (1979b). Nature *282*, 316–318.

Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. and Efstratiadis, A. (1978). Cell *15*, 687–701.

Maxam, A. M. and Gilbert, W. (1977). Proc. Nat. Acad. Sci. USA *74*, 560–564.

Mears, J. G., Ramirez, F., Leibowitz, D. and Bank, A. (1978). Cell *15*, 15–23.

Nishioka, Y. and Leder, P. (1979). Cell *18*, 875–882.

Nute, P. (1974). Ann. NY Acad. Sci. *241*, 39–70.

Orkin, S. H. (1978). Proc. Nat. Acad. Sci. USA *75*, 5950–5954.

Orkin, S. H., Old, J., Lazarus, H., Altay, C., Gurgey, A., Weatherall, D. J. and Nathan, D. G. (1979). Cell *17*, 33–42.

Proudfoot, N. J. and Baralle, F. E. (1979). Proc. Nat. Acad. Sci. USA *76*, 5435–5439.

Rodriguez, R. L., Bolivar, F., Goodman, H. M., Boyer, H. W. and Betlach, M. (1976). In Molecular Mechanisms in the Control of Gene Expression, D. P. Nierlich, W. J. Rutter and C. F. Fox, eds. (New York: Academic Press), pp. 471–478.

Russell, E. S. and McFarland, E. C. (1974). Ann. NY Acad. Sci. *241*, 25–38.

Shen, C.-K. J. and Maniatis, T. (1980). Cell *19*, 379–391.

Smith, G. P. (1973). Cold Spring Harbor Symp. Quant. Biol. *38*, 507–514.

Sodergren, E. and Fox, M. (1979). J. Mol. Biol. *130*, 357–377.

Southern, E. M. (1975). J. Mol. Biol. *98*, 503–517.

Stahl, F. (1979). Ann. Rev. Genet. *13*, 7–24.

Tartof, K. D. (1973). Cold Spring Harbor Symp. Quant. Biol. *38*, 491–500.

Todd, D., Lai, M. C. S., Beaven, G. and Huehns, E. (1970). Brit. J. Haematol. *19*, 27–31.

Tuan, D., Biro, P. A., De Riel, J. K., Lazarus, H. and Forget, B. G. (1979). Nucl. Acids Res. *6*, 2519–2544.

Weatherall, D. J. and Clegg, J. B. (1979). Cell *16*, 467–479.

Weatherall, D., Clegg, J. and Wong, H. (1970). Brit. J. Haematol. *18*, 357–367.

Wilson, J. T., Wilson, L. B., de Riel, J. K., Villa-Komaroff, L., Efstratiadis, A., Forget, B. G. and Weissman, S. M. (1978). Nucl. Acids Res. *5*, 563–581.

Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W. and Wilson, A. C. (1980). Proc. Nat. Acad. Sci. USA, in press.

Note Added in Proof

The existence of the embryonic ζ2-globin gene and its linkage to the other α-like genes was recently confirmed by analysis of a bacteriophage recombinant bearing a 10.9 kb Bgl II fragment which overlaps λHαG1 and contains the ζ2-globin gene (J. Lauer, unpublished results).

After this manuscript was submitted, we learned of a study of cloned mouse ribosomal DNA which revealed that deletions which occur in the ribosomal gene inserts during propagation in E. coli are similar to those which occur in mice (N. Arnheim and M. Kuehn, 1979, J. Mol. Biol. *134*, 743–765). As in the case of the human α-globin gene deletions reported here, the deletions in mouse ribosomal DNA can be correlated with the presence of direct repeat sequences.

Appendix: Isolation of the ζ2 globin gene

The linkage of the ζ2 globin gene to the other α-like genes was confirmed by analysis of a bacteriophage recombinant (λHζG1) bearing a 10.9 kb Bgl II fragment which overlaps λHαG1 and contains the ζ2 gene. The human fetal liver DNA used in isolation of this clone was from the same individual as the DNA that was used in constructing the human DNA library (Lawn et al., 1978) from which λHαG1 and λHαG2 were isolated. Human DNA was digested with Bgl II and DNA fragments of 9-12 kb were isolated by preparative agarose gel electrophoresis. The bacteriophage lambda vector, Charon 28 (F. Blattner, personal communication), was digested with BamH I and the annealed left and right arms were isolated by preparative agarose gel electrophoresis. The human DNA Bgl II fragments and the Charon 28 vector arms were ligated and packaged in vitro. Approximately 350,000 recombinant phage were screened using a ^{32}P-labeled Eco RI-Sac I fragment containing the terminal 0.49 kb of the human DNA insert of λHαG1 (see Figure 7 of Lauer et al., 1980 for the location of this fragment). Ten positive phage were isolated.

The restriction map of one of these phage (λHζG1) is shown in Figure 1a. λHζG1 DNA was digested with various restriction enzymes and the products were fractionated on an agarose gel, transferred to nitrocellulose, and hybridized with a ^{32}P-labeled Pvu II-Hinc II fragment containing part of ζ1 (Figure 1c). The approximate location of the ζ2 gene as determined in this manner agrees with that previously determined by genomic blotting. The precise position of the middle exon of ζ2 was determined by alignment of the restriction maps of ζ2 (Figure 1b) and ζ1 (Figure 1c).

References

Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. and Maniatis, T. (1978). The isolation and characterization of linked δ- and β-globin genes from a cloned library of human DNA. Cell **15**, 1157-1174.

Lauer, J., Shen, C.-K. J. and Maniatis, T. (1980). The chromosomal arrangement of human α-like globin genes: sequence homology and α-globin gene deletions. Cell **20**, 119-130.

Figure 1. Map of Restriction Endonuclease Cleavage Sites in λH ζG1 DNA.

a) The locations of cleavage sites in λH ζG1 DNA for the enzymes Bgl II (Bg), BamH I (Ba), Eco RI (RI), Hind III (H) and Sac I (S) are indicated. Sac I cleaves at additional sites in the vector which are not shown on this map. Charon 28 vector sequences are represented by the heavy bars, and the 10.9 kb Bgl II fragment of human DNA which contains ζ2 is represented by the thin line.

b) and c) Selected restriction endonuclease cleavage sites within and adjacent to the coding region of ζ2 (b) and ζ1 (c) are indicated. Black boxes represent coding sequences (exons). The approximate position of the first exon of ζ2 and ζ1 has been determined by in vitro transcription experiments (N. Proudfoot and M. Shander, personal communciation). The positions of the second and third exons of ζ1 have been determined by DNA sequence analysis (Lauer et al., 1980; C. O'Connell and N. Proudfoot, personal communication). The position of the second exon of ζ2 is assumed by alignment of the Hinc II, Sac I, and Pvu II sites of ζ1 and ζ2. The position of the third exon of ζ2 is unknown and it is therefore tentatively indicated by an open rather than solid box. The hatched box indicates the location of the Pvu II-Hinc II ζ1 probe fragment.

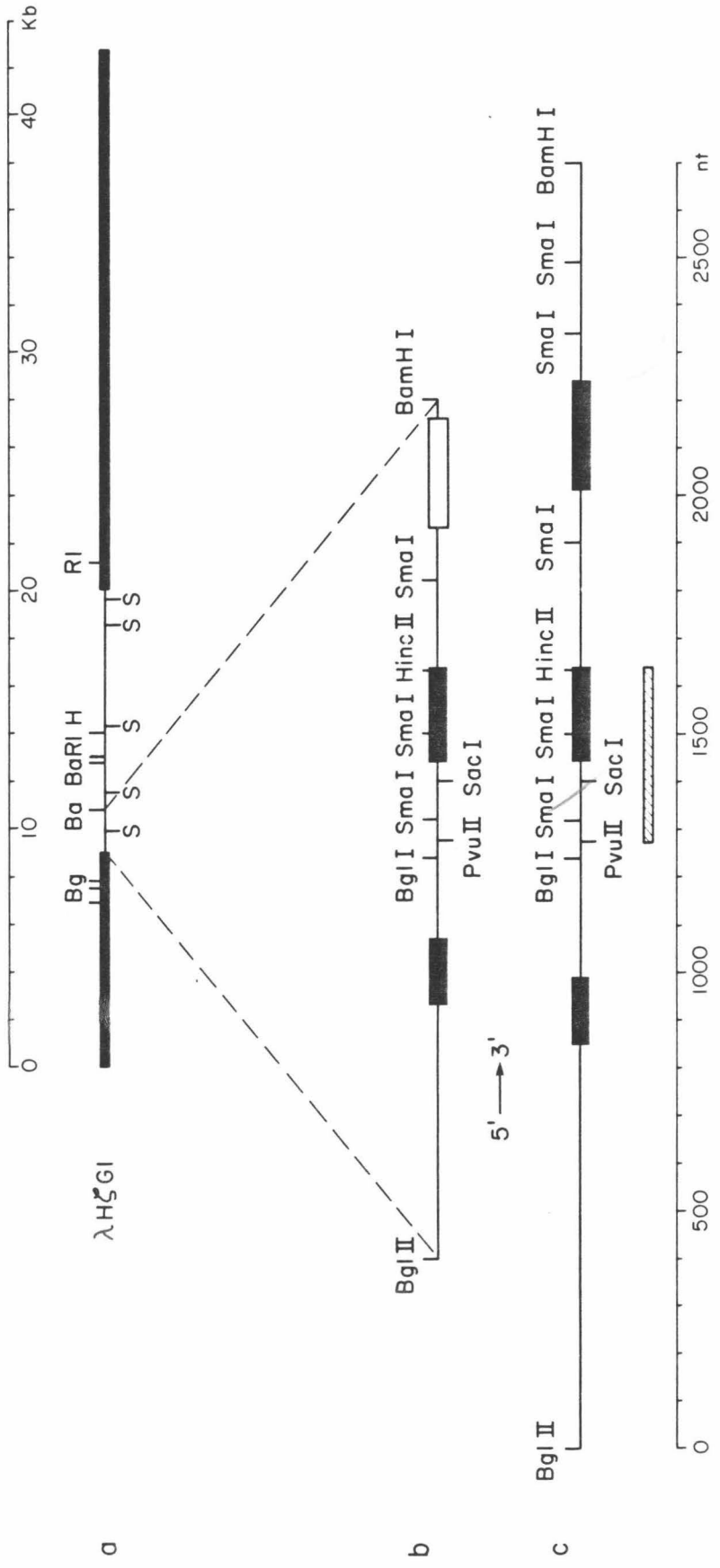

Kb
0
10
20
30
40
Bg
Ba
BaRI H
RI
S
λ HSG1
Bgl II
Bgl I Sma I Sma I Hinc II Sma I
BamH I
Pvu II Sac I
5' → 3'
Bgl II
Bgl I Sma I Sma I Hinc II Sma I
Sma I Sma I BamH I
Pvu II Sac I
nt
0
500
1000
1500
2000
2500
a
b
c

Figure 1

Chapter 3

The molecular genetics of human hemoglobins

THE MOLECULAR GENETICS OF HUMAN HEMOGLOBINS

Tom Maniatis, Edward F. Fritsch[‡], Joyce Lauer[+], and Richard M. Lawn[*]

Division of Biology

California Institute of Technology

Pasadena, California 91125

Present Addresses:

[‡]Department of Biochemistry, Michigan State University, East Lansing, Michigan 48824.

[+]Department of Molecular Biology, University of Edinburgh, Edinburgh, Scotland.

[*]Genentech, Inc., 460 Pt. San Bruno Blvd., South San Francisco, California 94080.

Shortened Title: MOLECULAR GENETICS OF HUMAN HEMOGLOBINS

Send Proofs to: Tom Maniatis, Division of Biology, California Institute of Technology, Pasadena, California 91125.

CONTENTS PAGE

INTRODUCTION

The human globin gene family is a paradigm for studying differential gene activity during development and the molecular basis of genetic disorders in gene expression (see Bunn et al 1977; Weatherall and Clegg 1979a; Weatherall et al 1979 for recent reviews). Structural characterization of globin polypeptides and messenger RNAs and extensive clinical investigations of inherited disorders in hemoglobin expression have provided information which is not available for any other eukaryotic gene system. The primary objective of this review is to discuss our current understanding of the structure, chromosomal arrangement, and expression of normal and abnormal globin genes, emphasizing new information derived from gene mapping and molecular cloning studies.

The Ontogeny of Globin Gene Expression

The α-like and β-like subunits of hemoglobin (Hb) are encoded by a small group of genes which are expressed sequentially during development. The earliest embryonic hemoglobin tetramer, Gower 1, consists of ε (β-like) and ζ (α-like) polypeptide chains (Huehns and Farooqui 1975; Gale et al 1979). Beginning at approximately eight weeks of gestation the embryonic chains are gradually replaced by the adult α-globin chain and two different fetal β-like chains, designated $^{G}\gamma$ and $^{A}\gamma$. The γ chains differ only in the presence of glycine or alanine at position 136, respectively (Schroeder et al 1968). During the transition period between embryonic and fetal development, Hb Gower 2 ($\alpha_2\varepsilon_2$) and Hb Portland ($\zeta_2\gamma_2$) are detected. Hb F ($\alpha_2\gamma_2$) eventually becomes the predominant Hb tetramer throughout the remainder of fetal life. Beginning just prior to birth, the γ-globin chains are gradually replaced by the adult β- and δ-globin polypeptides. At six months after birth 97-98% of the hemoglobin is Hb A ($\alpha_2\beta_2$), while Hb A_2 ($\alpha_2\delta_2$) accounts for approximately 2%. Small amounts of Hb F (1%) are also found in adult peripheral blood. The site of erythropoiesis changes from the yolk sac in the early embryo,

to the developing liver, spleen and bone marrow in the fetus, and finally to the bone marrow in adults. Because the ratio of Hb A to Hb F is the same in all fetal erythroid tissues, the switch from fetal to adult globin production is not correlated with the site of erythropoiesis (Wood and Weatherall 1973).

In addition to the gene switching described above, globin gene expression is regulated within a particular developmental stage. For example, the patterns of expression of the δ and β genes during adult red cell maturation are quite different (Roberts et al 1972). δ-globin mRNA can be detected in nucleated erythroid precursor cells but not in mature reticulocytes (Wood et al 1978). Thus, the small amount of δ-globin polypeptide found in circulating reticulocytes was synthesized in more immature cells in the bone marrow. Furthermore, bone marrow nuclei contain tenfold less δ-globin mRNA precursor than β-globin precursor (Kantor et al 1980), suggesting that the difference in expression of the two genes is at the level of transcription or RNA processing. The biological significance of the restriction of δ gene expression to immature cells is unknown.

The information summarized above indicates that the α-like and β-like globin gene families have coordinated programs for differential gene expression. The primary difference between the two gene families is that two switches in gene expression (embryonic to fetal to adult) are observed for the β-like genes, while a single switch results in activation of adult α-globin production early in fetal life.

Globin Gene Mapping and Molecular Cloning

Initial advances in globin gene mapping and isolation were made possible by the development of procedures for synthesizing and cloning double-stranded DNA copies of poly(A)-containing mRNAs (see Efstratiadis and Villa-Komaroff 1979; Maniatis 1980, for reviews). The successful introduction of double-stranded mouse (Rougeon et al 1975) and rabbit (Maniatis et al 1976; Higuchi et al 1976)

β-globin cDNA into bacterial plasmids provided homogeneous hybridization probes for gene mapping experiments. More recently cDNA plasmids carrying human α-, β- and γ-globin mRNA sequences were cloned and characterized (Wilson et al 1978; Little et al 1978).

The availability of cloned hybridization probes and the successful application of the Southern blotting procedure (Southern 1975) to mapping single copy sequences in mammalian genomes (genomic blotting; Botchan et al 1976) made it possible to construct a physical map of restriction endonuclease cleavage sites within and flanking the rabbit adult β-globin gene in chromosomal DNA (Jeffreys and Flavell 1977a). The interrupted colinearity between the restriction map of a rabbit β-globin cDNA plasmid (Maniatis et al 1976) and the corresponding map of the rabbit β-globin gene (Jeffreys and Flavell 1977a,b) provided evidence for the presence of noncoding intervening sequences (introns) in cellular genes (Jeffreys and Flavell 1977b). An intron within the mouse β-globin gene was independently discovered using molecular cloning techniques (Tilghman et al 1978a). The presence of similar sequences in the human globin genes will be discussed below.

Primarily two procedures have been used for cloning mammalian globin genes (see Maniatis 1980 for review). The mouse adult β-globin genes were isolated by gene enrichment followed by molecular cloning in a bacteriophage λ vector (Tilghman et al 1977). Globin genes were also isolated without a pre-enrichment step (Maniatis et al 1978; Blattner et al 1978) by constructing and screening collections (libraries) of recombinant phage containing the products of a limit restriction endonuclease digestion of mammalian DNA (Blattner et al 1978) or containing random, high molecular weight fragments of mammalian DNA (Maniatis et al 1978). The advantage of libraries of high molecular weight DNA is that the isolation of a set of overlapping clones, all of which contain a given gene, permits study of the sequences extending many kilobases (kb) from the gene.

Moreover, even more distant regions along the chromosome can be obtained by rescreening a library using terminal fragments of the initially selected clones as hybridization probes. Thus, this approach to gene isolation makes it possible to study the organization of closely linked genes. The entire human β-like (Lawn et al 1978; Fritsch et al 1979, 1980; Ramirez et al 1979) and α-like (Lauer et al 1980; J. Lauer, unpublished results) globin gene clusters including the intergene sequences have been isolated by this approach.

In the following sections, the human α-like and β-like globin gene clusters will be described separately in terms of the structure, chromosomal arrangement and expression of normal and abnormal genes within each cluster. Some of the information presented has been previously reviewed (Bunn et al 1977; Weatherall and Clegg 1979a) and is included for background. This review will focus on the application of molecular cloning and gene mapping procedures to the study of human globin genes.

THE HUMAN β-LIKE GLOBIN GENES

β-Like Globin Gene Fine Structure

The structure and organization of the human β-like globin genes have been studied by genomic blotting using globin cDNA or cDNA plasmids as hybridization probes, and by cloning segments of the genome containing globin genes. Restriction sites within and surrounding the δ and β genes were mapped using the genomic blotting procedure (Flavell et al 1978; Mears et al 1978). The existence of a large intron within each gene was demonstrated by comparing the restriction endonuclease cleavage map of the β-globin gene in genomic DNA with the map of a β-globin cDNA plasmid (Flavell et al 1978; Mears et al 1978). These conclusions were independently reached by analysis of bacteriophage recombinants which carry both the δ and β genes (Lawn et al 1978). The identities of the two genes

and the precise location of the large intron within each gene were established by DNA sequence analysis. In addition, a second smaller intron similar to that found in the mouse β gene (Konkel et al 1978) was identified by comparing the fine structure maps of a human β-globin cDNA plasmid (Wilson et al 1978) and of the chromosomal gene (Lawn et al 1978).

All five of the known human β-like globin genes have been obtained from recombinant clones of genomic DNA: β and δ (Lawn et al 1978), $^{G}\gamma$ and $^{A}\gamma$ (Smithies et al 1978; Fritsch et al 1979, 1980; Ramirez et al 1979) and ε (Proudfoot and Baralle 1979; Fritsch et al 1980). Identification of the ε gene was made possible by the recent determination of the partial amino acid sequence of ε-globin polypeptide (Gale et al 1979). All five genes are interrupted by two introns at identical locations: the first, of 125-150 base pairs (bp) in length is located between codons 30 and 31, and the second, of 700-900 bp is between codons 104 and 105 (Figure 1). Since the human α-globin genes (Figure 1; see below) as well as the mouse α- and β-globin genes, the chicken β-like globin genes and the four known rabbit β-like globin genes also contain two introns at homologous locations, it is reasonable to assume that these interruptions antedate the emergence of separate α- and β-globin genes about 500 million years ago (Tilghman et al 1977; Jeffreys and Flavell 1977b; Van den Berg et al 1978; Dodgson et al 1979; Hardison et al 1979; Konkel et al 1979; Nishioka and Leder 1979; Lauer et al 1980; Baralle et al 1980). As was shown for the mouse (Tilghman et al 1978b; Kinniburgh et al 1978; Kinniburgh and Ross 1979) and rabbit (Flavell et al 1979a; Hardison et al 1979) β-globin genes, the introns of the human β-like globin genes are transcribed as part of a nuclear mRNA precursor and are removed in steps by splicing to produce the mature globin mRNA (Maquat et al 1980; Kantor et al 1980). RNA-DNA hybridization mapping experiments have shown that the 5' ends of the rabbit (Flavell et al 1979a) and mouse (Weaver and Weissman 1979) β-globin nuclear mRNA precursors are coterminal

with their respective cytoplasmic mRNAs.

Although the function of introns is presently unknown, it has been proposed that they play a role in evolution by joining different combinations of DNA sequences encoding protein structural domains (Gilbert 1978, 1979; see Wahli et al 1979; Crick 1979 for discussion). Evidence for this hypothesis is provided by the observation that introns separate the three constant region domains within heavy chain immunoglobulin genes (Sakano et al 1979; Early et al 1979). In addition, the distribution of functional amino acid residues in the α- and β-globin polypeptides can be correlated with the arrangement of coding (exon) and noncoding (intron) sequences in these genes (Gilbert 1978, 1979; Blake 1979; Eaton 1980). For example, the middle exon encodes the region of the protein containing most of the heme contacts, while the α_1-β_1 protein-protein contacts are located predominantly in the region of the globin chain encoded by the third exon. Further evidence that the middle exon encodes a functional domain is provided by the demonstration that the polypeptide fragment encoded by the middle exon can bind heme tightly and specifically (Craik et al 1980).

The nucleotide sequences of human β- and γ-globin mRNAs have been determined (see Forget et al 1979 for discussion and references). The complete nucleotide sequences of the five human β-like globin genes and flanking regions have been determined (Lawn et al 1980; Spritz et al 1980; Slightom et al 1980; Baralle et al 1980) and a detailed comparison of these and other mammalian globin gene sequences has been presented (Efstratiadis et al 1980). These comparisons made it possible to establish accurate phylogenetic relationships among the human β-like globin genes and revealed interesting sequence homologies in regions which are potentially involved in globin gene transcription and splicing (Efstratiadis et al 1980). Alignment of the sequences 5' to the human β-like globin genes provides an example of these homologies. The most notable feature of this alignment is the presence of two blocks of sequence homology which are present in analogous positions adjacent

to most eukaryotic genes. The first homology block (designated the ATAA box) is found 29-30 bp 5' to each gene and the second block (designated the CCAAT box) is located 70-78 bp 5' to each gene. A possible role of these sequences in transcriptional initiation and/or RNA processing has been discussed (Efstratiadis et al 1980).

In addition to the ATAA and CCAAT box sequences which are found in many eukaryotic genes, other regions of homology which are shared among the β-like globin genes are scattered throughout the 5' flanking sequence. The conservation of these sequences during the 200 million years since the time of ε,γ-δ,β divergence (Efstratiadis et al 1980) suggests that they are functionally significant. The identification by sequence comparisons of regions of possible functional importance may be useful in the search for structural differences between normal and mutant globin genes.

Comparison of the complete nucleotide sequences of the $^{G}\gamma$ and $^{A}\gamma$ genes has revealed that the coding, intervening and flanking sequences of these two genes are virtually identical (Slightom et al 1980). The distribution of the infrequent sequence differences has led to formulation of a model for gene correction (see below).

β-Like Globin Gene Linkage

Cell fusion studies have shown that the human adult β-globin gene is located on chromosome 11 (Deisseroth et al 1978). More recent experiments have localized the β-like globin gene cluster to the distal portion of the short arm of chromosome 11 (Jeffreys et al 1979; Gusella et al 1979; Lebo et al 1979).

Genetic studies (Weatherall and Clegg 1979b) and structural analysis of hemoglobin fusion proteins (Baglioni 1962; Huisman et al 1972) predicted linkage of the γ, δ and β genes. These predictions were recently confirmed by genomic blotting and molecular cloning experiments (Figure 2). Physical linkage between the δ and β genes was demonstrated by genomic blotting experiments (Flavell et al

1978) and by analysis of a recombinant bacteriophage containing both the δ and β genes (Lawn et al 1978). Linkage of the $^{G}\gamma$ and $^{A}\gamma$ genes was determined by genomic blotting experiments (Little et al 1979a) and was confirmed by analysis of clones containing both of these genes (Fritsch et al 1979, 1980; Ramirez et al 1979). Linkage between the $^{A}\gamma$ and δ genes was established using a cloned hybridization probe from the $^{A}\gamma$-δ intergenic region and a well-defined deletion located in this region (Fritsch et al 1980; Tuan et al 1979), by mapping DNA fragments from the $^{A}\gamma$-δ intergenic region generated by non-limit restriction enzyme digests of genomic DNA (Bernards et al 1979a), and ultimately by cloning overlapping DNA fragments which span the $^{A}\gamma$-δ intergenic region (Fritsch et al 1980). Analysis of overlapping cloned DNA fragments also made it possible to demonstrate physical linkage between the ε gene, the isolation of which was initially reported by Proudfoot and Baralle (1979), and the remainder of the gene cluster (Fritsch et al 1980).

The linkage arrangement of the human β-like globin gene cluster is shown in Figure 2. All of the genes are transcribed from the same DNA strand and are arranged on the chromosome in the order of their expression during development: 5'-ε-$^{G}\gamma$-$^{A}\gamma$-δ-β-3'. A similar pattern of gene organization has been found in the human α-like (Lauer et al 1980; see below) and rabbit (Lacy et al 1979; Hardison et al 1979) and mouse (Jahn et al 1980) β-like globin gene clusters. Linkage between four chicken β-like globin genes has also been demonstrated (Dodgson et al 1979), but the identities of the genes have not been unambiguously determined (J. D. Engel, personal communication). Two β-like sequences, ψβ1 and ψβ2, which cannot be identified with known polypeptides, are also shown on the map of Figure 2. These sequences probably correspond to pseudogenes, which have been studied in detail in other globin gene clusters (see below).

Repetitive Sequence Elements within Globin Gene Clusters

The universal occurrence of repetitive sequences in eukaryotic DNA and

their interspersion with single-copy sequences have led to the proposal that repetitive sequences play a role in the regulation of gene expression. Previous studies of repetitive sequence elements have described their genomic distribution and their differential transcription during development (see Davidson and Britten 1979 for review). The isolation of globin gene clusters by molecular cloning provides an opportunity to study repeat sequences in relation to well-defined sets of developmentally regulated genes.

A detailed study of the rabbit β-like globin gene cluster revealed a complex array of sequences which are repeated within the gene cluster as well as throughout the genome (Shen and Maniatis 1980). A sequence which is repeated within the human β-like globin gene cluster (Fritsch et al 1980) is also interspersed with the human α-like globin genes (E. Fritsch, J. Shen, R. Lawn and T. Maniatis, unpublished results), is highly repeated elsewhere in the genome (Fritsch et al 1980), and is homologous to a sequence which is found in the rabbit β-like globin gene cluster (Shen and Maniatis 1980; J. Shen, unpublished results). Nucleotide sequence analysis of the copy of this repeat located 5' to the $^{G}\gamma$ gene indicates that the repeat is a member of a family of sequences which is reiterated approximately 300,000 times in the human genome (Houck et al 1979; Jelinek et al 1980).

This family of repeated sequences is characterized by the following diverse set of properties: 1) the repeats are transcribed by RNA polymerase III in vitro (Duncan et al 1980; Fritsch et al in preparation); 2) the repeats share sequence homology with an abundant small nuclear RNA molecule (Jelinek and Leinwand 1978; Jelinek et al 1980) and with double-stranded heterogeneous nuclear RNA (Jelinek et al 1980; Fritsch et al 1980); 3) the repeats contain a sequence which is homologous to a sequence found near the replication origins of SV40, polyoma and BK DNA tumor viruses (Jelinek et al 1980). Whether such repeat sequences are involved in differential gene expression or play a more general role in DNA replication or chromosome structure remains to be determined.

β-Globin Variants

The most common structural variants of the β-globin polypeptide are simple amino acid substitutions, presumably resulting from a single base change in the nucleotide sequence (see Bunn et al 1977 for review). The majority of such substitutions have no demonstrable effect on globin chain activity or stability. However, some amino acid substitutions result in severe clinical effects such as those associated with sickle cell anemia (Weatherall and Clegg 1979b).

The widespread occurrence of a variant, Hb $F^{Sardinia}$ ($^{T}\gamma$; Ricco et al 1976) and differences in the $^{A}\gamma/^{G}\gamma$ ratio during fetal development (Huisman et al 1970; Schroeder et al 1973) led to the proposal that there are more than two γ-globin genes per haploid genome. However, only two γ genes are detected by genomic blotting in both normal individuals and individuals with Hb $F^{Sardinia}$ (Little et al 1979a,b). Moreover, structural analysis of the $^{T}\gamma$ chain of Hb $F^{Sardinia}$ indicates that $^{T}\gamma$ is a variant of the $^{A}\gamma$ chain (Saglio et al 1979).

Some β-globin structural variants lack one to eight amino acids (Bunn et al 1977). Since the amino acid sequence is normal on either side of the deleted residues the genes which encode these variants must contain small deletions which do not alter the translational reading frame. Examination of the nucleotide sequences in normal DNA within and adjacent to the codons for the amino acids which are deleted in the variants reveals the presence of small (2-8 bp) direct repeat sequences (Marotta et al 1977). Small repeats are also associated with deletions in the lac i gene of E. coli (Farabaugh and Miller 1978). These deletions are thought to result from base mispairing and slippage during DNA replication (Streisinger et al 1966; Farabaugh and Miller 1978). As noted by Marotta et al (1977) and Efstratiadis et al (1980), a similar mechanism may have generated the deletions which are associated with the β-globin variants described above.

Another type of hemoglobin variant is the β-globin fusion proteins. Hb Lepore is a fused N-terminal δ- and C-terminal β-globin protein (Baglioni 1962) which is the result of a fusion between the δ and β genes (Flavell et al 1978; Mears et al 1978) presumably caused by unequal crossing over during meiosis (Baglioni 1962). Another well-characterized fusion protein is Hb Kenya which must result from fusion of the 5' portion of the $^{A}\gamma$- and the 3' portion of the β-globin genes (Huisman et al 1972). As mentioned above, the existence of Hb Lepore and Hb Kenya provided the first strong evidence for linkage of the γ, δ and β genes (Baglioni 1962; Huisman et al 1972).

DNA Sequence Polymorphisms within the β-Like Globin Gene Cluster

Genetic polymorphisms of restriction endonuclease cleavage sites have been detected in the human β-like globin gene cluster. The first example of such a linked genetic polymorphism was an alteration in an Hpa I recognition site located 5 kb 3' to the β gene (Kan and Dozy 1978). The frequent association of the absence of this Hpa I site with the allele for sickle cell anemia was exploited by Kan and Dozy (1978) as a tool for prenatal diagnosis. Rarely, absence of this Hpa I site is associated with the normal allele, leading to uncertainties in diagnosis (Kan et al 1980). β°-thalassemia in Sardinia may be diagnosed by the presence of a polymorphic BamH I site located 9.3 kb 3' to the β gene (Kan et al 1980). Although the frequent occurrence of this BamH I site in normal individuals leads to considerable uncertainty in diagnosis, genomic blotting information is a valuable complement to existing procedures for prenatal diagnosis. Sequence polymorphisms have also been detected within the large introns of the δ (Lawn et al 1978; Jeffreys 1979) and γ (Jeffreys 1979; Tuan et al 1979) genes. These studies indicate that restriction site polymorphisms are frequent within the β-like globin gene cluster. Analysis of linked polymorphisms associated with other genetic disorders may provide an important tool for studying the genetics of human disease (Jeffreys 1979; Kan et al 1980; Little et al 1980).

GENETIC DISORDERS IN β-GLOBIN GENE EXPRESSION

β°- and β^{+}-thalassemia are the most frequently occurring mutations in β-globin gene expression. β^{+}-thalassemia is characterized by reduced levels of β-globin production while in β°-thalassemia there is no detectable β-chain synthesis (Weatherall and Clegg 1979b). β-globin gene deletions, which affect the expression of other globin genes in addition to β, constitute a third class. In the discussion which follows we will briefly summarize the information currently available regarding the molecular basis of various types of β-thalassemia.

β^{+}-Thalassemia

In β^{+}-thalassemia a small amount of normal β-globin protein is made suggesting that the gene and mRNA are intact. Globin cDNA/DNA solution hybridization (Old et al 1978; Benz et al 1978), in vitro translation (Benz et al 1978), and genomic blotting (Flavell et al 1979b; Orkin et al 1979c) experiments have confirmed these conclusions. Thus, the primary defect in β^{+}-thalassemias appears to be a reduced level of transcription or inefficient processing of mRNA precursors. Comparison of the ratios of globin RNA sequences in bone marrow nuclear ($\alpha/\beta = 1$) and cytoplasmic ($\alpha/\beta = 15\text{-}20$) RNA obtained from individuals homozygous for β^{+}-thalassemia suggested that the molecular defect in these individuals is abnormal processing or decreased stability of β-globin mRNA precursor (Nienhuis et al 1977).

Evidence for the former possibility was recently provided by analysis of pulse-labeled α- and β-globin mRNA precursors from nucleated bone marrow cells (Maquat et al 1980). Following pulse-labeling with tritiated nucleosides a series of β-globin mRNA precursors is observed. In normal individuals these precursors are quantitatively processed to mature mRNA during a chase period with unlabeled nucleosides. By contrast, in β^{+}-thalassemic individuals discrete intermediates accumulate and only a fraction of the labeled RNA is chased into β-globin mRNA. Processing of α-globin mRNA precursors is identical in normal and β^{+}-thalassemic

individuals. Similar conclusions were obtained by Kantor et al (1980) who used a hybridization probe specific for the large intron of the β gene to quantitate the level of β-globin RNA and to distinguish between δ- and β-globin mRNA precursors. Furthermore, using an RNA blotting procedure (Alwine et al 1977), an abnormal 650 nucleotide (nt) intermediate was observed in one patient while a normal 1300 nt intermediate was found to accumulate in another (Kantor et al 1980). These studies strongly suggest that some β^+-thalassemias result from mutations which alter RNA processing.

β°-Thalassemia

β°-thalassemia is characterized by complete absence of β-globin polypeptide synthesis (Weatherall and Clegg 1979b). The molecular defects in β°-thalassemia appear to be quite heterogeneous. In the majority of cases no deletions or alterations of the β gene can be detected by genomic blotting (Flavell et al 1979b; Orkin et al 1979c). Rarely, this phenotype is associated with a 600 bp deletion which removes part of the large intron, the third exon, and 150 bp beyond the 3' end of the normal gene (Figure 2; Orkin et al 1979c, 1980).

Three classes of β°-thalassemia have been defined on the basis of cDNA/mRNA hybridization experiments (Old et al 1978; Benz et al 1978). In the first class, no β-globin mRNA can be detected although in at least one case, β-globin sequences were detected in the nuclear RNA (Comi et al 1977). This class of β°-thalassemia probably results from a block in RNA transcription or processing.

In the second class of β°-thalassemia up to 30% of the normal level of β-globin mRNA is present although no β-globin protein is synthesized in vivo or in vitro (Benz et al 1978; see Conconi et al 1972; Conconi and Del Senno 1974 for a possible exception). In one case of this type Temple et al (1977) showed by RNA fingerprinting that an apparently normal β-globin mRNA was present. Sequence analysis of this RNA revealed, however, the presence of a single base change at

codon position 17 resulting in a nonsense mutation (Chang and Kan 1979). The fact that this nonsense mutation can be suppressed in vitro demonstrates that it is the primary defect (Chang et al 1979). In another case of β°-thalassemia which appears to contain intact β-globin mRNA the initiation codon may be defective (Old et al 1978). In a third class of β°-thalassemia, a partial RNA transcript is produced (Old et al 1978).

β-Globin Gene Deletions

Most cases of β°- and β^{+}-thalassemia result in moderate to severe anemia and are not associated with detectable deletions in the β-like globin gene cluster. In these cases, the level of fetal globin expression in adults is normal or only slightly increased, suggesting that the switch from fetal to adult globin synthesis is operative. Surprisingly, in other classes of genetic disorder known as δβ-thalassemia and Hereditary Persistence of Fetal Hemoglobin (HPFH), the absence of β-globin chains is compensated for by continued expression of γ-globin chains in the adult. The degree of fetal globin compensation differs among these disorders (see Wood et al 1979 for review).

As classically defined, HPFH and δβ-thalassemia are distinguished on the basis of three criteria: 1) the mean level of Hb F is higher in HPFH than in δβ-thalassemia; 2) Hb F as measured by the acid-elution technique is uniformly distributed throughout the red cell population in HPFH and non-uniformly distributed in δβ-thalassemia (more recently, however, the greater sensitivity of immunofluorescent techniques indicates that the cellular distribution is uniform in both HPFH and δβ-thalassemia; see Wood et al 1979 for critical discussion); 3) HPFH is not associated with clinical symptoms whereas δβ-thalassemia is characterized by red cell abnormalities and mild to severe anemia. Studies of individuals heterozygous for either of these disorders indicated that these mutations which abolish β-globin production act in cis to affect γ gene expression (Weatherall and Clegg 1979b; Huisman

et al 1974). On the basis of this information, Huisman et al (1974) predicted that both HPFH and δβ-thalassemia are associated with β gene deletions and that the HPFH deletion, but not the δβ-thalassemia deletion, removes regulatory sequences involved in the normal suppression of γ gene expression in adults. This model has received considerable attention because it offered a genetic approach to understanding the mechanism of differential globin gene expression.

β-globin cDNA/DNA solution hybridization experiments confirmed that both HPFH and δβ-thalassemia are associated with deletions which remove the β and possibly the δ genes (Kan et al 1975b; Forget et al 1976; Ottolenghi et al 1976; Ramirez et al 1976). With the introduction of the genomic blotting procedure and the availability of specific hybridization probes, it became possible to map the endpoints of these and other deletions within the β-like globin gene cluster. The locations of the deletions which have been mapped to date, and the level of Hb F associated with each syndrome, are presented in Figure 2. A unified molecular interpretation of this information is not possible at present, due primarily to the heterogeneity of these genetic disorders. As discussed by Wood et al (1979), the wide spectrum of phenotypes observed in HPFH and δβ-thalassemia leads to overlap between these syndromes. In addition, types of HPFH have been described which show genuinely non-uniform distribution of Hb F in the red cell population (British or Swiss-type HPFH; Weatherall et al 1979) or which express different relative amounts of $^{G}\gamma$ and $^{A}\gamma$ chains (Wood et al 1979). Nonetheless, genomic blotting mapping studies have been very important in demonstrating the variety of molecular rearrangements which can alter γ-globin gene expression, and have led to the formulation of new models for the mechanism of hemoglobin switching.

For example, the locations of some of these deletions can be interpreted in the context of the Huisman et al (1974) deletion model. Comparison of $^{G}\gamma$-$^{A}\gamma$-δβ-thalassemia (see legend to Figure 2 for an explanation of the nomenclature) with

four cases of HPFH is consistent with the possibility that a sequence 5' to the δ gene exerts some influence on the level of Hb F production (Fritsch et al 1979; Bernards et al 1979b; Tuan et al 1979; Ottolenghi et al 1979). Similarly, comparison of Hb Lepore with $^{G}\gamma$-$^{A}\gamma$-δβ-thalassemia implicates a second region located 3' to the β gene (Bernards et al 1979b). Other deletions are apparently inconsistent with these interpretations. For example, the $^{G}\gamma$-δβ-thalassemia deletion removes the regions 5' to the δ gene and 3' to the β gene but does not result in an HPFH phenotype (Fritsch et al 1979; Orkin et al 1979a). However, because a large portion of the γ-globin gene region is removed in $^{G}\gamma$-δβ-thalassemia, the significance of this syndrome with respect to the deletion model cannot be assessed (Fritsch et al 1979).

An alternative and equally plausible explanation of the deletion data is that the human β-like globin gene cluster consists of one or more functional chromosomal domains and that deletions within the cluster alter the chromosome structure and thereby affect the normal pattern of differential globin gene expression (see Fritsch et al 1979; Bernards et al 1979b; and Van der Ploeg et al 1980 for discussion). The concept that higher order chromosome structure is involved in regulating gene activity is central to certain hypotheses of eukaryotic differentiation (for example, see Cook 1973; Weintraub et al 1978). A correlation between alterations in chromosome structure and changes in gene activity is suggested by the observation that actively-transcribed genes are more sensitive to DNase I digestion than non-transcribed genes (Weintraub and Groudine 1976). Genomic blotting analysis of DNase I-digested chromatin has shown that moderate DNase I sensitivity extends for many kilobases on either side of an active gene (Wu et al 1979a,b; Stalder et al 1980), suggesting that differential gene activity involves a structural alteration of a large region of the chromosome.

A third explanation of the data is that a second mutation, unrelated to the β gene deletion, might be responsible for the altered γ gene regulation which

is observed in HPFH and/or $\delta\beta$-thalassemia. The occurrence of different deletion events in independent cases of HPFH and the absence of the HPFH phenotype in non-deletion β-thalassemias (β^0- or β^+-thalassemia) argues against the general occurrence of a second-site compensating mutation. A rare type of HPFH ($^A\gamma$-HPFH, Greek type) appears not to be associated with a detectable deletion. In $^A\gamma$-HPFH, 10-20% Hb F is produced (9:1 ratio of $^A\gamma$ to $^G\gamma$; Clegg et al 1980), and the δ and β genes are expressed at low levels in some individuals. No alterations of the γ-δ-β-globin gene region have been detected (Tuan et al 1980).

In the syndromes described above, deletions in the δ-β-globin gene region are associated with changes in the pattern of γ gene expression. In γ-β-thalassemia, suppression of γ and β gene expression occurs, resulting in severe anemia in newborns and thalassemia in adults. No homozygous case has been reported. A heterozygous case of this syndrome is characterized by a deletion of at least 40 kb removing both γ genes and the δ gene (Van der Ploeg et al 1980). However, the β gene and approximately 2.5 kb of 5' flanking sequence are intact. The lack of expression of the β gene on the mutant chromosome is not understood. Because this is the only example of γ-β-thalassemia which has been characterized at the DNA level, the possibility that a second mutation has inactivated the β gene (similar to β^0- or β^+-thalassemia) must be considered seriously. Alternatively, Van der Ploeg et al (1980) have suggested that within the β-like globin gene cluster the DNA is organized into $^G\gamma$-$^A\gamma$ and δ-β domains and that the γ-β-thalassemia deletion removes sequences necessary for activation of the δ-β domain.

In conclusion, analysis of deletion types of thalassemia has suggested that deletions may act in cis over considerable distances to influence differential gene expression within the β-like globin gene cluster. A similar association of deletions with alterations in globin gene switching has been described in the human α-like globin gene cluster (see below).

THE HUMAN α-LIKE GLOBIN GENES

α-Like Globin Gene Fine Structure

The existence of two α-globin genes per haploid genome was established on the basis of genetic studies (see Weatherall and Clegg 1979a for discussion). Although only one α-globin polypeptide sequence has been identified (Dayhoff 1972), both genes are expressed. A partial amino acid sequence of ζ-globin polypeptide has been determined (J. Clegg, personal communication).

The nucleotide sequence of human α-globin mRNA has been determined by analysis of α-globin cDNA and of a cDNA plasmid (Wilson et al 1978; see Forget et al 1979 for discussion and references). The structure and organization of the human α-like globin genes have been examined by genomic blotting experiments (Orkin 1978; Embury et al 1979) and by analysis of cloned DNA segments (Lauer et al 1980). For convenience, the two α genes are referred to as α1 and α2 (Figure 3). Restriction endonuclease analysis demonstrated that the α1 and α2 genes are each interrupted by two introns of approximately 95 and 125 bp (Lauer et al 1980; see above for a discussion of intron function). Determination of the complete nucleotide sequence of an independently isolated α2 gene indicated that the introns are located between the codons for amino acids 31 and 32 and 99 and 100 (S. Liebhaber, personal communication). The locations of introns in the human and mouse (Nishioka and Leder 1979) α-globin genes are identical, and are analogous to the positions of introns in β-like globin genes. Fine structure mapping of restriction sites in the coding, intervening and flanking sequences of the α1 and α2 genes indicated that the two genes are virtually identical (Lauer et al 1980). Restriction sites could be aligned by assuming the insertion of a block of approximately seven nucleotides into the large intron of the α1 gene (or a deletion from the α2 gene). The α1-α2 homology is discussed further below.

A cloned ζ-globin gene (ζ1) was identified by the exact correspondence between the DNA sequence of this gene (Lauer et al 1980; C. O'Connell, N. Proudfoot and T. Maniatis, unpublished results) and the partial amino acid sequence of ζ-globin polypeptide (J. Clegg, personal communication). The ζ1 gene contains two introns at locations identical to the positions of introns in the adult α-globin genes.

α-Like Globin Gene Linkage

Cell fusion studies have shown that the human adult α-globin genes are located on chromosome 16 (Deisseroth et al 1977). Genomic blotting experiments demonstrated that the linked adult α-globin genes are transcribed from the same DNA strand and are 3.7 kb apart (Figure 3; Orkin 1978; Embury et al 1979). α-globin gene linkage was confirmed by isolation of recombinant bacteriophage clones (Lauer et al 1980) containing the two adult α-globin genes, the embryonic ζ1 gene and an α-globin pseudogene (ψα1; see below). Nucleotide sequence analysis of a portion of the ψα1 and ζ1 genes indicated that they are oriented in the same direction as the adult α-globin genes (Lauer et al 1980).

A second ζ-globin gene, ζ2, was identified by genomic blotting experiments using a labeled restriction fragment containing part of the ζ1 gene as a hybridization probe (Lauer et al 1980). Physical linkage of the ζ2 gene to the other α-like globin genes was demonstrated by genomic blotting using a cloned restriction fragment located 5' to the ζ1 gene as a hybridization probe. The ζ2 gene is located approximately 12 kb 5' to the ζ1 gene (Figure 3). The ζ1-ζ2 gene linkage was confirmed by analysis of overlapping clones which contain the ζ1 and ζ2 genes (J. Lauer, unpublished results).

Evidence suggests that both ζ-globin genes are functional. Detection of ζ-globin protein in an infant with a homozygous deletion which removes the ζ1 gene but spares the ζ2 gene (see below) indicates that ζ2 is a functional gene. Furthermore, since the nucleotide sequence of the ζ1 gene (C. O'Connell, N. Proudfoot

and T. Maniatis, unpublished results) agrees with the amino acid sequence of ζ-globin polypeptide (J. Clegg, personal communication), the ζ1 gene is probably also functional. Genomic blotting-melting experiments suggest that the two genes are very homologous (Lauer et al 1980). Since only one ζ-globin polypeptide sequence has been identified, the ζ1 and ζ2 genes may encode identical polypeptide chains. Alternatively, the two genes may be expressed at different times or at different levels during embryonic development.

Globin Gene Duplication

A common feature of globin gene clusters is the occurrence of two immediately-adjacent genes which are coordinately expressed during a given developmental stage. Examples of this are the human δ-β, $^{G}\gamma$-$^{A}\gamma$, α1-α2, and ζ1-ζ2 globin gene pairs. The δ and β genes are highly homologous in the coding region, but the noncoding sequences within and surrounding the two genes have diverged considerably (Lawn et al 1978; Efstratiadis et al 1980). Extensive divergence of noncoding regions has also been observed in some other closely-linked, coordinately-expressed globin gene pairs (Konkel et al 1979; Hardison et al 1979; Jahn 1980). In contrast, the two members of the $^{G}\gamma$-$^{A}\gamma$ gene pair are virtually identical to one another throughout their coding, intervening and flanking sequences (Slightom et al 1980). Although the nucleotide sequences of linked human α-globin genes have not yet been determined, restriction mapping and heteroduplex analysis of the α1 and α2 genes indicate that the sequences within and flanking these two genes are virtually identical. Each α-globin gene is located within an approximately 4 kb region of homology interrupted by two small regions of non-homology (Figure 4; Lauer et al 1980).

The virtual identity of sequences within the $^{G}\gamma$-$^{A}\gamma$ and α1-α2 gene pairs appears to be the product of a mechanism for gene matching during evolution. Based on the nearly identical distribution of restriction sites surrounding the α-globin genes in a number of primate species, it has been suggested that the α-globin gene duplication

occurred prior to the time of primate divergence (Zimmer et al 1980). Differences between the α-globin amino acid sequences of various primate species are consistent with sequence drift following primate divergence (Dayhoff 1972). However, intra-species comparisons show much less divergence, indicating that the α-globin genes within a species have been corrected against one another. Maintenance of homology among a family of evolving genes within a species has been termed "horizontal" (Brown et al 1972) or "coincidental" (Hood et al 1975) or "concerted" (Zimmer et al 1980) evolution. Gene conversion and expansion/contraction of gene number by homologous but unequal crossing-over have been proposed as mechanisms for concerted evolution (see Hood et al 1975 for review).

Analysis of the complete nucleotide sequences of cloned $^{G}\gamma$ and $^{A}\gamma$ genes has led to formulation of a specific intrachromosomal gene conversion model to explain sequence matching between linked genes (Slightom et al 1980). The $^{G}\gamma$ and $^{A}\gamma$ genes on one chromosome are identical in the region 5' to the center of the large intron, yet show greater divergence 3' to that position. Examination of the boundary between the conserved and divergent regions revealed a block of "simple sequence" DNA ($[TG]_n$) (Slightom et al 1980). Slightom et al (1980) have proposed that this simple sequence is a "hot spot" for initiation of recombination events which lead to unidirectional gene conversion.

As described above, an active mechanism for gene matching must also be invoked to explain the observed homology between the human α-globin genes. Although α-globin gene matching could occur by gene conversion, sequence data which would address this possibility are not available. Evidence that α-globin gene sequence matching could occur by expansion and contraction of gene number by unequal crossing-over is provided by the frequent occurrence of one-gene (Orkin et al 1979b; Embury et al 1979) and three-gene (Goosens et al 1980; Higgs et al 1980) chromosomes in some human populations (see Zimmer et al 1980 and Lauer et al 1980 for discussion).

GENETIC DISORDERS IN α-GLOBIN GENE EXPRESSION

The α-thalassemias are characterized by a reduced rate of α-globin synthesis (Weatherall and Clegg 1979b) resulting in the presence of excess β-like chains which associate to form the abnormal tetramers Hb Bart's ($\gamma 4$) and Hb H ($\beta 4$). Four α-thalassemia syndromes of increasing clinical severity occur in the Asian population: 1) the silent carrier state (α-thalassemia 2) which is asymptomatic; 1-2% Hb Bart's at birth; 2) α-thalassemia trait (α-thalassemia 1) which is associated with red cell abnormalities but little or no anemia; 5-6% Hb Bart's at birth; 3) Hb H disease which is associated with anemia; Hb Bart's at birth and Hb H in adulthood; 4) hydrops fetalis which is fatal at or before birth; 80-90% Hb Bart's, 10-20% Hb Portland ($\zeta_2 \gamma_2$). As will be discussed below, the occurrence and frequency of these syndromes differ in non-Asian populations.

Homozygosity for the "α-thalassemia trait", leading to complete absence of α-globin synthesis, was postulated to be the cause of hydrops fetalis (Lie-Injo et al 1962). Subsequently, genetic analysis of Hb H disease led Wasi et al (1964) to propose the existence of two α-thalassemia alleles: a "severe" allele corresponding to the previously-identified α-thalassemia trait, and a "mild" allele recognizable only when present along with the severe allele thereby causing Hb H disease. When it became apparent that human α-globin genes are duplicated, it was proposed that the mild and severe alleles correspond not to differing degrees of impairment of a single α-globin gene, but rather to a lack of function of one or both genes, respectively, on a chromosome, whether due to deletion or to other gene defects (Lehmann 1970).

The four syndromes described above might thus result from involvement of one, two, three, or four α-globin structural genes, respectively (Lehmann 1970). cDNA/DNA solution hybridization analysis indicated that persons with Hb H disease have only one-quarter of the normal number of α-globin genes, supporting this

hypothesis (Kan et al 1975a). In addition, solution hybridization analysis of hydrops fetalis (which is characterized by complete absence of α-globin synthesis) indicated that the α-globin genes are deleted in this syndrome (Ottolenghi et al 1974; Taylor et al 1974). Although the precision of solution hybridization experiments is limited, these studies provided the first direct physical evidence for the association of gene deletion with α-thalassemia. The genotypes resulting in each α-thalassemia syndrome are diagrammed in Figure 5, where the presence of a functional α-globin gene is indicated by a black rectangle.

Although the majority of α-thalassemias appear to be due to gene deletion, some non-deletion types of defects were recently reported (Kan et al 1977, 1979; Orkin et al 1979b). An α-thalassemia phenotype also results from the unstable α-globin variant, Hb Constant Spring, which is synthesized at a normal level but is rapidly degraded (Weatherall and Clegg 1975). In the discussion which follows we will present the information currently available regarding the molecular basis of various forms of α-thalassemia.

α-Globin Gene Deletions

The association of gene deletions with α-thalassemia was confirmed and extended by genomic blotting. In normal DNA the duplicated α-globin genes on each chromosome are detected in a 23 kb Eco RI fragment (Orkin 1978; Embury et al 1979). Chromosomes containing a single functional α gene can arise by inactivation of one gene (23 kb Eco RI fragment; Orkin et al 1979b; Kan et al 1979) or by deletion of one gene (19 kb Eco RI fragment; Embury et al 1979; Orkin et al 1979b). Chromosomes without any functional α-globin genes are associated with a 2.6 kb Eco RI fragment containing a partial α gene (Orkin and Michelson 1980) or with deletion of both genes. Different combinations of these chromosomes are associated with each of the α-thalassemia syndromes, as summarized in Table 1.

α-thalassemia 1 is common in Asians, blacks and Mediterraneans, yet hydrops

fetalis is extremely rare in the latter two populations. The rarity of hydrops fetalis could be due to rarity of the zero-gene chromosome in these populations (Lehmann 1970). Genomic blotting studies of α-thalassemia 1 in blacks confirmed that this syndrome arises from homozygosity for the single gene chromosome (Dozy et al 1979; Table 1).

Restriction endonuclease mapping analysis indicated that chromosomes containing a single α-globin gene could have resulted from deletion of the α2 gene or from unequal crossing over between the α2 and α1 genes to form a single hybrid gene (Orkin et al 1979b; Embury et al 1979). More detailed mapping of single-gene chromosomes from Asians, blacks and Mediterraneans revealed that these chromosomes are of two types (Embury et al 1980). One type of deletion (termed leftward) removes 4.2 kb of DNA containing the α2 gene, while the other type of deletion (rightward) removes 3.7 kb of DNA, apparently by unequal crossing over between the α1 and α2 genes (Figure 3). The rightward type of deletion predominates in all three populations. The leftward type has been found only in several Asian cases.

Deletions associated with α-thalassemia 2 (Embury et al 1980) are indistinguishable in size and position from two types of deletions which occur during propagation of bacteriophage λ clones containing the α-globin genes (Lauer et al 1980). The breakpoints of these deletions (Figure 3) are located within the blocks of α1-α2 homology described above (Figure 4). The precise lengths of the leftward (4.3 kb) and rightward (3.8 kb) types of deletions indicate that both types of deletions in cloned DNA occur by homologous but unequal crossing-over between corresponding sequences in the α1 and α2 gene regions.

Homologous but unequal crossing-over at the α-globin locus should produce a chromosome with three α-globin genes in addition to the chromosome with a single α-globin gene. An analogous crossing-over at the β-globin locus produces Hb Lepore (δ-β fusion) on one chromosome and anti-Lepore (β-δ fusion; e.g., Hb P Congo or Miyada) along with the β and δ genes on the other chromosome (Weatherall and

Clegg 1979a). Three-gene chromosomes have been found in black and Greek Cypriot populations where α-thalassemia 2 is common (Goosens et al 1980; Higgs et al 1980).

Recently, the breakpoints of deletions associated with two cases of hydrops fetalis have been mapped (Pressley et al 1980). In a Greek case, the deletion removes the α1, α2 and ζ1 genes but leaves the ζ2 gene intact (Figure 3). The presence of Hb Portland ($\zeta_2\gamma_2$) in the Greek infant establishes that ζ2 is a functional gene. In a Thai case the deletion removes the α1 and α2 genes but spares both the ζ1 and ζ2 genes (Figure 3).

It is noteworthy that deletions associated with hydrops fetalis may be similar to those associated with HPFH in their effect on differential globin gene expression (Weatherall et al 1970; Pressley et al 1980). During normal development Hb Portland is found in significant amount only until about ten weeks gestation (Gale et al 1979), whereas in infants with hydrops fetalis 10-20% Hb Portland is found at birth (Weatherall et al 1970; Todd et al 1970). Thus, in hydrops fetalis, ζ-globin gene expression continues beyond the time at which it is normally switched off. In both hydrops fetalis and HPFH, a deletion in one region of a gene cluster affects the expression of a distant gene within the cluster.

α- AND β-GLOBIN PSEUDOGENES

α- and β-globin sequences which cannot be identified with known globin polypeptide chains (peuedogenes) have been detected in several mammalian species (Hardison et al 1979; Fritsch et al 1980; Lauer et al 1980; Jahn et al 1980; Nishioka and Leder 1980; Vannin and Smithies 1980). Nucleotide sequence analysis of a rabbit β pseudogene (β2; Hardison et al 1979; L. Lacy and T. Maniatis, manuscript submitted), a human α pseudogene (ψα1; Figure 3; Lauer et al 1980; N. Proudfoot and T. Maniatis, manuscript submitted), a mouse β pseudogene (waw-a; Jahn et al 1980), and a mouse α pseudogene (α-3 as designated by Nishioka and Leder 1980 or α-30.5 as designated

by Vannin and Smithies 1980) has demonstrated a variety of structural differences between each gene and its functional counterpart. The human β-like sequences (ψβ1 and ψβ2; Figure 2; Fritsch et al 1980) have not yet been extensively characterized.

Each of the pseudogenes which has been analyzed exhibits 75-80% sequence homology when compared with its corresponding normal gene. None of these pseudogenes can encode a functional globin polypeptide, due to the presence of small deletions or insertions which result in alterations of the translational reading frame. In addition, one or more of the intron/exon junctions of β2, waw-a and ψα1 are different from the "consensus" sequence common to splicing junctions in globin genes and all expressed genes studied to date (Breathnach et al 1978; Lerner et al 1980). The mouse α pseudogene differs from the other pseudogenes in that both introns of the mouse α pseudogene have been precisely removed, leaving an uninterrupted coding sequence. The location of the mouse α pseudogene with respect to the functional mouse α-globin genes is unknown.

It is interesting to note that in all of the mammalian globin gene clusters thus far characterized, a pseudogene is found between the embryonic (or fetal) genes and the adult genes (for example, see Figures 2 and 3). It is possible that pseudogenes have some as yet unidentified function in globin gene clusters. Alternatively, pseudogenes may be the products of gene duplication and subsequent sequence divergence (Ohno 1970). The variation in human α-globin gene number observed in present day populations and the location of ψα1 within the α-like globin gene cluster are consistent with the latter possibility. As shown in Figure 3, ψα1, α2 and α1 are separated from each other by approximately 4 kb, which is the size of the α1-α2 duplication unit noted above (Figure 4). The nucleotide sequence of ψα1 indicates that it is α-like rather than ζ-like (N. Proudfoot and T. Maniatis, manuscript submitted). It therefore seems possible that ψα1 was once part of a set of three functional α-globin genes.

CONCLUDING REMARKS

The application of gene mapping and molecular cloning procedures to the study of human globin genes has led to significant advances in our understanding of their structure and chromosomal organization. The linkage arrangement of the known α- and β -like genes has been established and the nucleotide sequence determination for each of these genes has been completed or is in progress. This structural information and the information provided by clinical investigations of inherited disorders in globin gene expression can now be combined to study the molecular genetics of globin gene regulation. Recent improvements in molecular cloning techniques make it possible to isolate individual globin genes and their flanking sequences from the DNA which can be obtained from a small quantity of blood. Structural comparisons between mutant globin genes and their normal counterparts may lead to the identification of sequences involved in globin gene expression. However, in vivo or in vitro assays for globin gene expression will be required to discriminate between functionally significant sequence differences and random sequence polymorphisms.

An in vitro approach to studying globin gene expression is now possible because of the recent development of cell-free extracts for RNA polymerase II-dependent transcription of cloned eukaryotic genes. Two in vitro transcription systems have been described. One system consists of a cytoplasmic extract which requires the addition of purified RNA polymerase II for activity (Weil et al 1979), while the other system consists of a concentrated whole cell extract with endogenous RNA polymerase II activity (Manley et al 1980; C. Parker, personal communication). In both systems, specific transcription of adenovirus genes (Weil et al 1979; Manley et al 1980) was shown by the fact that the capped 5' terminus of the in vitro transcript is indistinguishable from that found in vivo. The general applicability of these in vitro

transcription systems was recently demonstrated by specific transcription of mouse (Luse and Roeder 1980), human (Manley et al 1980; N. Proudfoot, M. Shander and T. Maniatis, unpublished results), and rabbit (N. Proudfoot, M. Shander and T. Maniatis, unpublished results) globin genes. These in vitro transcription systems may provide a rapid assay for transcriptional defects in globin genes. Moreover, comparison of extracts prepared from a variety of erythroid and non-erythroid cells may lead to the identification of factors necessary for tissue-specific globin gene expression.

In vivo assays for globin gene expression utilize DNA-mediated gene transfer procedures ("convection" procedures) and SV40 tumor virus vectors. Transient introduction of globin genes into mammalian cells can be achieved by replacing the genes for SV40 viral capsid proteins with foreign DNA and growing the recombinants in the presence of an SV40 helper virus (Mulligen et al 1979; Hamer and Leder 1979a,b). The recombinant virus is amplified many thousandfold, facilitating the study of globin gene expression. The disadvantage of this approach is that only small DNA fragments ($\leq$ 2 kb) can be introduced into the vector because of size constraints on viral DNA packaging. Furthermore, transcriptional initiation from a globin gene promoter has not yet been reported.

Large DNA fragments containing rabbit or human globin genes have been stably introduced into mouse L cells in culture using the calcium phosphate DNA transformation procedure and the herpes virus thymidine kinase gene as a selectable marker (Wigler et al 1977, 1978; Mantei et al 1979). Usually, 1-10 copies of the gene are integrated into high molecular weight DNA and one or more of the copies are expressed at a relatively low level (5-2000 copies of globin mRNA per cell compared with 40,000-50,000 globin mRNA molecules in an erythroid cell). Analysis of the human and rabbit β-globin gene transcripts in mouse L cells indicated that both introns are removed by splicing. However, in at least one cell line, the mature globin mRNA lacks 46 nt at the 5' end (Wold et al 1979). Thus, transcriptional

initiation or RNA processing must be abnormal in this cell line. Direct microinjection of DNA into nuclei may provide an alternative and more efficient means of introducing DNA into cells (W. F. Anderson, personal communication; C. Lo and A. Efstratiadis, personal communication). Both convection procedures (calcium phosphate DNA transformation and microinjection) suffer from difficulties in detecting small amounts of globin transcripts because of the low levels of expression of integrated genes. In addition, there is significant heterogeneity in the level of expression of the foreign gene within a cloned cell population (Hanahan et al 1980). In spite of the difficulties inherent in the SV40 and convection procedures, it should be possible to study the basic mechanisms of globin gene transcription and processing.

In addition to using these in vitro and in vivo assays for studying the expression of normal and mutant globin genes (e.g., β°- and β^{+}-thalassemias), it will also be possible to study genes which have been altered by site-directed in vitro mutagenesis. Procedures are available for introducing point mutations, deletions, or rearrangements at specific sites in DNA (Carbon et al 1975; Domingo et al 1976; Shortle and Nathans 1978; Weissman et al 1979; Shortle et al 1979). A study of the consequences of alterations within globin gene clusters may lead to the identification of regions important for globin gene expression.

It may also be possible to identify sequences necessary for tissue-specific gene expression by introducing globin genes into erythroid cells in culture. The feasibility of this approach is suggested by cell fusion studies which demonstrated that mouse erythroleukemia cells carrying human chromosome 16 or 11 can be induced by DMSO to express high levels of human α- or β-globin mRNA and protein (Deisseroth et al 1978). By introducing normal and in vitro altered globin genes into these cells, it may be possible to define the minimal genetic unit necessary for developmentally regulated gene expression. This unit may be an individual globin gene and its flanking sequences or the entire gene cluster.

Acknowledgements

We thank P. F. R. Little for his critical reading of the manuscript. The work cited from the authors' laboratory was supported by grants from the National Science Foundation and the National Institutes of Health.

REFERENCES

Alwine, J. C., Kemp, D. J., Stark, G. R. 1977. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc. Natl. Acad. Sci. USA 74:5350-5354

Baglioni, C. 1962. The fusion of two polypeptide chains in hemoglobin Lepore and its interpretation as a genetic deletion. Proc. Natl. Acad. Sci. USA 48:1880-1885

Baralle, F. E., Shoulders, C., Proudfoot, N. J. 1980. The primary structure of the human epsilon gene. Cell, submitted for publication.

Benz, E. J., Jr., Forget, B. G., Hillman, D. G., Cohen-Solal, M., Pritchard, J., Cavallesco, C., Prensky, W., Housman, D. 1978. Variability in the amount of β-globin mRNA in β° thalassemia. Cell 14L299-312

Bernards, R., Little, P. F. R., Annison, G., Williamson, R., Flavell, R. A. 1979a. Structure of the human $^{G}\gamma$-$^{A}\gamma$-δ-β-globin gene locus. Proc. Natl. Acad. Sci. USA 76:4827-4831

Bernards, R., Kooter, J. M., Flavell, R. A. 1979b. Physical mapping of the globin gene deletion in (δβ)°-thalassemia. Gene 6:265-280

Blake, C. C. F. 1980. Exons encode protein functional units. Nature 277:598

Blattner, F. R., Blechl, A. E., Denniston-Thompson, K., Faber, H. E., Richards, J. E., Slightom, J. L., Tucker, P. W., Smithies, O. 1978. Cloning human fetal γ-globin and mouse α-type globin DNA: Preparation and screening of shotgun collections. Science 202:1279-1283

Botchan, M., Topp, W., Sambrook, J. 1976. The arrangement of SV40 sequences in the DNA of transformed cells. Cell 9:269-287

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., Chambon, P. 1978. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc. Natl. Acad. Sci. USA 75:4853-4857

Brown, D., Wensink, P., Jordan, E. 1972. A comparison of the ribosomal DNAs of Xenopus laevis and Xenopus mulleri: the evolution of tandem genes. J. Mol. Biol. 63:57-74

Bunn, H. F., Forget, B. G., Ranney, H. M. 1977. Human Hemoglobins. Philadelphia: W. B. Saunders Co.

Carbon, J., Shenk, T., Berg, P. 1975. Biochemical procedure for production of small deletions in SV40 DNA. Proc. Natl. Acad. Sci. USA 72:1392-1396

Chang, J. C., Kan, Y. W. 1979. β° Thalassemia, a nonsense mutation in man. Proc. Natl. Acad. Sci. USA 76:2886-2889

Chang, J. C., Temple, G. F., Trecartin, R. F., Kan, Y. W. 1979. Suppression of the nonsense mutation in homozygous β° thalassemia. Nature 281:602

Clegg, J. B., Metaxatou-Mavromati, A., Kattamis, C., Sofroniadou, K., Wood, W. G., Weatherall, D. J. 1979. Occurrence of $^{G}\gamma$ Hb F in Greek HPFH: analysis of heterozygotes and compound heterozygotes with β-thalassemia. Brit. J. Hematology 43:521-536

Comi, P., Giglioni, B., Barbarano, L., Ottolenghi, S., Williamson, R., Novakova, M., Masera, G. 1977. Transcriptional and post-transcriptional defects in β° thalassemia. Eur. J. Biochem. 79:617-622

Conconi, F., Rowley, P. T., Del Senno, L., Pontremoli, S. 1972. Induction of β-globin synthesis in the β-thalassemia of Ferrara. Nature New Biol. 238:83-85

Conconi, F., Del Senno, 1974. L. The molecular defect of Ferrara β-thalassemia. Ann. N.Y. Acad. Sci. 232:54

Cook, P. R. 1973. Hypothesis on differentiation and the inheritance of gene superstructure. Nature 245:23-25

Craik, C. S., Buchman, S. R., Beychok, S. 1980. Characterization of globin domains: Heme binding to the central exon product. Proc. Natl. Acad. Sci. USA 77: 1384-1388

Crick, F. 1979. Split genes and RNA splicing. Science 204:264-271

Davidson, E., Britten, R. 1979. Regulation of gene expression: possible role of repetitive sequences. Science 204:1052-1059

Dawid, I., Wahli, W. 1979. Application of recombinant DNA technology to questions of developmental biology: a review. Dev. Biol. 69:305-328

Dayhoff, M. O. 1972. Atlas of Protein Sequence and Structure. Washington, D.C.: National Biomedical Research Foundation.

Deisseroth, A., Hendrick, D. 1978. Human α-globin gene expression following chromosomal dependent gene transfer into mouse erythroleukemia cells. Cell 15:55-63

Deisseroth, A., Nienhuis, A., Turner, P., Velez, R., Anderson, W. F., Lawrence, J., Creagan, R., Kucherlapati, R. 1977. Localization of the human α-globin structural gene to chromosome 16 in somatic cell hybrids by molecular hybridization assay. Cell 12:205-218

Deisseroth, A., Nienhuis, A., Lawrence, J., Giles, R., Turner, P., Ruddle, F. H. 1978. Chromosomal localization of human β-globin gene on human chromosome 11 in somatic cell hybrids. Proc. Natl. Acad. Sci. USA 75:1456-1460

Dodgson, J. B., Strommer, J., Engel, J. D. 1979. Isolation of the chicken β-globin gene and a linked embryonic β-like globin gene from a chicken DNA recombinant library. Cell 17:879-887

Domingo, E., Flavell, R. A., Weissman, C. 1976. In vitro site-directed mutagenesis: generation and properties of an infectious extracistronic mutant of bacteriophage Qβ. Gene 1:3-25

Dozy, A., Kan, Y. W., Embury, S., Mentzer, W., Wang, W., Lubin, B., Davis, J., Koenig, H. 1979. α-Globin gene organization in blacks precludes the severe form of α-thalassemia. Nature 280:605-607

Duncan, C., Biro, P. A., Chowdary, P. V., Elder, J. T., Wang, R. R. C., Forget, B. G., De Riel, J. K., Weissman, S. M. 1980. RNA polymerase III transcriptional units are interspersed among human non-α-globin genes. Proc. Natl. Acad. Sci. USA 76:5095-5099

Early, P. E., Davis, M., Kaback, D., Davidson, N., Hood,L. 1979. Immunoglobulin heavy chain gene organization in mice: Analysis of a myeloma genomic clone containing variable and α constant regions. Proc. Natl. Acad. Sci. USA 76:857-861

Eaton, W. A. 1980. The relationship between coding sequences and function in hemoglobin. Nature 284:183-185

Efstratiadis, A., Posakony, J., Maniatis, T., Baralle, F., Blechl, A., De Riel, J., Forget, B., Lawn, R., O'Connell, C., Proudfoot, N., Shoulders, C., Slightom, J., Smithies, O., Spritz, R., Weissman, S. 1980. DNA sequence comparison of the human β-like globin genes. Cell, submitted for publication.

Efstratiadis, A., Villa-Komaroff, L. 1979. Cloning of double-stranded cDNA. In Genetic Engineering: Principles and Methods, ed. J. Setlow. New York: Plenum Press.

Embury, S., Lebo, R., Dozy, A., Kan, Y. W. 1979. Organization of the α-globin genes in the Chinese α-thalassemia syndromes. J. Clin. Invest. 63:1307-1310

Embury, S. H., Miller, J. A., Dozy, A. M., Kan, Y. W., Chan, V., Todd, D. 1980. Two different molecular organizations account for the single α-globin gene of the α-thalassemia 2 genotype. Submitted for publication.

Farabaugh, P. J., Miller, J. H. 1978. Genetic studies of the lac repressor. VII. On the molecular nature of spontaneous hotspots in the lac i gene of Escherichia coli. J. Mol. Biol. 126:847-863

Flavell, R. A., Kooter, J. M., De Boer, E., Little, P. F. R., Williamson, R. 1978. Analysis of the human β-δ-globin gene loci in normal and Hb Lepore DNA: direct determination of gene linkage and intergene distance. Cell 15:25-41

Flavell, R. A., Bernards, R., Grosveld, G. C., Hooijmakers-VanDommelen, H. A. M., Kooter, J. M., De Boer, E. 1979a. The structure and expression of globin genes in rabbit and man. In Eukaryotic Gene Regulation, ICN-UCLA Symposium on Molecular and Cellular Biology, XIV, ed. Axel, R., Maniatis, T., Fox, C.F. New York: Academic Press.

Flavell, R. A., Bernards, R., Kooter, J. M., De Boer, E. 1979b. The structure of the human β-globin gene in β-thalassemia. Nucl. Acids Res. 6:2749-2760.

Forget, B. G., Hillman, D. G., Lazarus, H., Barell, E. F., Benz, E. J., Jr., Caskey, C. T., Huisman, T. H. J., Schroeder, W. A., Housman, D. 1976. Absence of messenger RNA and gene DNA for β globin chains in hereditary persistence of fetal hemoglobin. Cell 7:323-329.

Forget, B. G., Cavallesco, C., de Riel, J. K., Spritz, R. A., Choudary, P. V., Wilson, J. T., Wilson, L. B., Reddy, V. B., Weissman, S. M. 1979. Structure of the human globin genes. In Eukaryotic Gene Regulation, ICN-UCLA Symposium on Molecular and Cellular Biology, XIV ed. Axel, R., Maniatis, T., Fox, C.F. New York: Academic Press.

Fritsch, E. F., Lawn, R. M., Maniatis, T. 1979. Characterization of deletions which affect the expression of fetal globin genes in man. Nature 279:598-603.

Fritsch, E. F., Lawn, R. M., Maniatis, T. 1980. Molecular cloning and characterization of the human β-like globin gene cluster. Cell 19:959-972.

Gale, R., Clegg, J., Huehns, E. 1979. Human embryonic hemoglobins Gower 1 and Gower 2. Nature 280:162-164.

Gilbert, W. 1978. Why genes in pieces? Nature 271:501.

Gilbert, W. 1979. Introns and exons: playgrounds of evolution. In Eukaryotic Gene Regulation, ICN-UCLA Symposium on Molecular and Cellular Biology, XIV, ed. Axel, R., Maniatis, T., Fox, C.F. New York: Academic Press.

Goosens, M., Dozy, A., Embury, S., Zacharides, Z., Hadjiminas, M., Stamatoyannopoulos, G., Kan, Y. W. 1980. Triplicated α-globin loci in humans. Proc. Nat. Acad. Sci. USA 77:518-521

Gusella, J., Varsanyi-Brelner, A., Kao, F. T., Jones, C., Puck, T., Keys, C., Orkin, S., Housman, D. 1979. Precise localization of human β-globin gene complex on chromosome 11. Proc. Natl. Acad. Sci. USA 76:5239-5243

Hardison, R. C., Butler, E. T., Lacy, E., Maniatis, T., Rosenthal, N., Efstratiadis, A. 1979. The structure and transcription of four linked rabbit β-like globin genes. Cell 18:1285-1297

Hamer, D. H., Leder, P. 1979. Expression of the chromosomal mouse β^{maj}-globin gene cloned in SV40. Nature 281:35-40

Hamer, D. H., Smith, K. D., Boyer, S. H., Leder, P. 1979. SV40 recombinants carrying rabbit β-globin gene coding sequences. Cell 17:725-735

Hanahan, D., Lane, D., Lipsich, L., Wigler, M., Botchan, M. 1980. Characteristics of an SV40-plasmid recombinant and its movement into and out of the genome of a murine cell. Submitted for publication.

Higgs, D. R., Old, J. M., Pressley, L., Clegg, J. B., Weatherall, D. J. 1980. A novel α-globin gene arrangement in man. Nature 284:632-635

Higuchi, R., Paddock, G. V., Wall, R., Salser, W. 1976. A general method for cloning eukaryotic structural gene sequences. Proc. Natl. Acad. Sci. USA 73:3146-3150

Hood, L., Campbell, J. H., Elgin, S. C. R. 1975. The organization, expression and evolution of antibody genes and other multigene families. Ann. Rev. Genet. 9:305-353

Houck, C. M., Rinehart, F. P., Schmid, C. W. 1979. A ubiquitous family of repeated DNA sequences in the human genome. J. Mol. Biol. 132:289-306

Huehns, E., Farooqui, A. 1975. Oxygen dissociation properties of human embryonic red cells. Nature 254:335-337

Huisman, T. H. J., Schroeder, W. A., Bannister, W. H., Grech, J. L. 1972. Evidence for four nonallelic structural genes for the γ chain of human fetal hemoglobin. Biochem. Genet. 7:131-139

Huisman, T. H. J., Wrightstone, R. N., Wilson, J. B., Schroeder, W. A., Kendall, A. G. 1972. Hemoglobin Kenya, the product of fusion of γ and β polypeptide chains. Arch. Biochem. Biophys. 153:850-853

Huisman, T. H. J., Schroeder, W. A., Efremov, G. D., Duma, H., Meadenovski, B., Hyman, C. B., Rachmilewitz, E. A., Bouver, N., Miller, A., Brodie, A. R., Shelton, J. R., Appel, G. 1974. The present status of the heterogeneity of fetal hemoglobinin β-thalassemia: an attempt to unify some observations in thalassemia and related conditions. Ann. N.Y. Acad. Sci. 232:107-124

Jahn, C. L., Hutchinson, C. A. III, Phillips, S. J., Weaver, S., Haigwood, N. L., Voliva, C. F., Edgell, M. H. 1980. DNA sequence organization of the β-globin complex in the BALB/c mouse. Submitted for publication.

Jeffreys, A. J. 1979. DNA sequence variants in the $^{G}\gamma$-, $^{A}\gamma$-, δ- and β-globin genes of man. Cell 18:1-10

Jeffreys, A., Craig, I., Francke, U. 1979. Localization of the $^{G}\gamma$-, $^{A}\gamma$-, δ- and β-globin genes on the short arm of human chromosome 11. Nature 281:606-608

Jeffreys, A. J., Flavell, R. A. 1977a. A physical map of the DNA regions flanking the rabbit β-globin gene. Cell 12:429-439

Jeffreys, A. J., Flavell, R. A. 1977b. The rabbit β-globin gene contains a large insert in the coding sequence. Cell 12:1097-1108

Jelinek, W. G., Leinwand, L. 1978. Low molecular weight RNAs hydrogen-bonded to nuclear and cytoplasmic poly(A)-terminated RNA from cultured Chinese hamster ovary cells. Cell 15:205-214

Jelinek, W. R., Toomey, T. P., Leinwand, L., Duncan, C. H., Biro, P. A., Choudary, P. N., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L., Schmid, C. W. 1980. Ubiquitous, interspersed repeated sequences in mammalian genomes. Proc. Natl. Acad. Sci. USA 77:1398-1402

Kan, Y. W., Dozy, A., Varmus, H., Taylor, J., Holland, J., Lie-Injo, L., Ganesan, J., Todd, D. 1975a. Deletion of α-globin genes in hemoglobin H disease demonstrates multiple α-globin structural loci. Nature 255:255-256

Kan, Y. W., Holland, J. P., Dozy, A. M., Charache, S., Kazazian, H. H. 1975b. Deletion of the β globin structural gene in hereditary persistence of fetal hemoglobin. Nature 258:162-163

Kan, Y. W., Dozy, A., Trecartin, R., Todd, D. 1977. Identification of a nondeletion defect in α-thalassemia. N. Engl. J. Med. 297:1081-1084

Kan, Y. W., Dozy, A. M. 1978. Polymorphism of DNA sequence adjacent to human β-globin structural gene: relationship to sickle mutation. Proc. Natl. Acad. Sci. USA 75:5631-5635

Kan, Y. W., Dozy, A., Stamatoyannopoulos, G., Hadjiminas, M., Zacharides, Z., Furbetta, M., Cao, A. 1979. Molecular basis of hemoglobin H disease in the Mediterranean population. Blood 54:1434-1438

Kan, Y. W., Lee, K. Y., Furbetta, M., Angius, A., Cao, A. 1980. Polymorphism of DNA sequence in the β-globin gene region. New Engl. J. Med. 302:185-188

Kantor, J. A., Turner, P. H., Nienhuis, A. W. 1980. Beta$^+$ thalassemia: mutations which affect processing of the β-globin mRNA precursor. Cell, submitted.

Kinniburgh, A. J., Mertz, J. E., Ross, J. 1978. The precursor of mouse β-globin messenger RNA contains two intervening RNA sequences. Cell 14:681-693

Kinniburgh, A. J., Ross, J. 1979. Processing of the mouse β-globin mRNA precursor: at least two cleavage-ligation reactions are necessary to excise the large intervening sequence. Cell 17:915-921

Konkel, D. A., Maizel, J. V., Leder, P. 1979. The evolution and sequence comparison of two recently diverged mouse chromosomal β-globin genes. Cell 18:865-873

Konkel, D. A., Tilghman, S. M., Leder, P. 1978. The sequence of the chromosomal mouse β-globin major gene: homologies in capping, splicing and poly(A) sites. Cell 15:1125-1132

Lacy, E., Hardison, R. C., Quon, D., Maniatis, T. 1979. The linkage arrangement of four rabbit β-like globin genes. Cell 18:1273-1283

Lauer, J., Shen, C.-K. J., Maniatis, T. 1980. The chromosomal arrangement of human α-like globin genes: sequence homology and α-globin gene deletions. Cell 20:119-130

Lawn, R. M., Efstratiadis, A., O'Connell, C., Maniatis, T. 1980. Cell, submitted for publication.

Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G., Maniatis, T. 1978. The isolation and characterization of linked δ- and β-globin genes from a cloned library of human DNA. Cell 15:1157-1174

Lebo, R., Carrano, A., Burkhart-Schultz, K., Dozy, A., Yu, L. C., Kan, Y. W. 1979. Assignment of human β-, γ-, and δ-globin genes to the short arm of chromosome 11 by chromosome sorting and DNA restriction enzyme analysis. Proc. Natl. Acad. Sci. USA 76:5804-5808

Lehmann, H. 1970. Different types of α-thalassemia and significance of hemoglobin Barts in neonates. Lancet ii:73

Lerner, M., Boyle, J., Mount, S., Wolin, S., Steitz, J. 1980. Are snRNPs involved in splicing? Nature 283:220-224

Lie-Injo Luan Eng, Lie Hong Gie, Ager, J. A. M., Lehmann, H. 1962. α-thalassemia as a cause of hydrops fetalis. Brit. J. Haematol. 8:1

Little, P. F. R., Annison, G., Darling, S., Williamson, R., Camba, L., Modell, B. 1980. A model for antenatal diagnosis of β-thalassemia and other monogenic disorders using molecular analysis of linked DNA polymorphisms. Nature, in press.

Little, P., Curtis, P., Coutelle, C., Van Den Berg, J., Dalgleish, R., Malcolm, S., Courtney, M., Weatway, D., Williamson, R. 1978. Isolation and partial sequence of recombinant plasmids containing human α-, β-, and γ-globin cDNA fragments. Nature 273:640-643

Little, P. F. R., Flavell, R. A., Kooter, J. M., Annison, G., Williamson, R. 1979a. Structure of the human fetal globin gene locus. Nature 278:227-231

Little, P. F. R., Williamson, R., Annison, G., Flavell, R. A., De Boer, E., Bernini, L. G., Ottolenghi, S., Saglio, G., Mazza, U. 1979b. Polymorphisms of human γ-globin genes in Mediterranean populations. Nature 282:316-318

Luse, D. S., Roeder, R. G. 1980. Accurate transcription initiation on a purified mouse β-globin DNA fragment in a cell-free system. Cell, submitted for publication.

Maniatis, T. 1980. Recombinant DNA procedures in the study of eukaryotic genes. In Comprehensive Cell Biology, Vol. 3, ed. L. Goldstein, D. Prescott. New York: Academic Press.

Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K., Efstratiadis, A. 1978. The isolation of structural genes from libraries of eucaryotic DNA. Cell 15:687-701

Maniatis, T., Sim, G.-K., Efstratiadis, A., Kafatos, F. 1976. Amplification and characterization of a β-globin gene synthesized in vitro. Cell 8:163-182

Manley, J. L., Fire, A., Cano, A., Sharp, P. A., Gefter, M. L. 1980. DNA-dependent transcription of adenovirus genes in a soluble whole-cell extract. Proc. Natl. Acad. Sci. USA, in press.

Mantei, N., Boll, W., Weissman, C. 1979. Rabbit β-globin mRNA production in mouse L cells transformed with cloned rabbit β-globin chromosomal DNA. Nature 281:40-46.

Maquat, L. E., Kinniburgh, A. J., Beach, L. R., Honig, G. R., Lazerson, J., Ershler, W. B., Ross, J. 1980. Processing of the human β-globin mRNA precursor to mRNA is defective in three patients with β^+ thalassemia. Proc. Natl. Acad. Sci. USA, submitted for publication.

Marotta, C. A., Wilson, J. T., Forget, B. G., Weissman, S. M. 1977. Human β-globin messenger RNA. III. Nucleotide sequences derived from complementary DNA. J. Biol. Chem. 252:5040-5053.

Mears, J. G., Ramirez, F., Leibowitz, D., Bank, A. 1978. Organization of human δ- and β-globin genes in cellular DNA and the presence of intragenic inserts. Cell 15:15-23

Mulligan, R. C., Howard, B. H., Berg, P. 1979. Synthesis of rabbit β-globin in cultured monkey kidney cells following infection with a SV40 β-globin recombinant clone. Nature 277:108-114

Nienhuis, A. W., Turner, P., Benz, E. J., Jr. 1977. Relative stability of α- and β-globin messenger RNAs in homozygous β^+ thalassemia. Proc. Natl. Acad. Sci. USA 74:3960-3964.

Nishioka, Y., Leder, P. 1979. The complete sequence of a chromosomal mouse α-globin gene reveals elements conserved throughout vertebrate evolution. Cell 18:875-882.

Ohno, S. 1970. Evolution by Gene Duplication, pp. 76-77. New York: Springer-Verlag.

Old, J. M., Proudfoot, N. J., Wood, W. G., Longley, J. I., Clegg, J. B., Weatherall, D. J. 1978. Characterization of β-globin mRNA in β° thalassemias. Cell 14:289-298

Orkin, S. H. 1978. The duplicated human α globin genes lie close together in cellular DNA. Proc. Natl. Acad. Sci. USA 75:5950-5954

Orkin, S. H., Alter, B. P., Altay, C. 1979a. Deletion of the $^A\gamma$-globin gene in $^G\gamma$-δβ-thalassemia. J. Clin. Invest. 64:866-869

Orkin, S. H., Kolodner, R., Michelson, A., Husson, R. 1980. Cloning and direct examination of a structurally abnormal human beta°-thalassemia globin gene. Proc. Natl. Acad. Sci. USA, submitted for publication.

Orkin, S., Michelson, A. 1980. Submitted for publication.

Orkin, S., Old, J., Lazarus, H., Altay, C., Gurgey, A., Weatherall, D., Nathan, D. 1979b. The molecular basis of α-thalassemias: frequent occurrence of dysfunctional α loci among non-Asians with Hb H disease. Cell 17:33-42

Orkin, S. H., Old, J. M., Weatherall, D. J., Nathan, D. G. 1979c. Partial deletion of β-globin gene DNA in certain patients with β°-thalassemia. Proc. Natl. Acad. Sci. USA 76:2400-2404

Ottolenghi, S., Lanyon, W. G., Paul, J., Williamson, R., Weatherall, D. J., Clegg, J. B., Pritchard, J., Pootrakul, S., Wong, H. B. 1974. The severe form of α thalassemia is caused by a hemoglobin gene deletion. Nature 251:389-392

Ottolenghi, S., Comi, P., Giglioni, B., Tolstoshev, P., Lanyon, W. G., Mitchell, G. J., Williamson, R., Russo, G., Musumeci, S., Schiliro, G., Tsistrakis, G. A., Charache, S., Wood, W. G., Clegg, J. B., Weatherall, D. J. 1976. δβ-thalassemia is due to a gene deletion. Cell 9:71-80

Ottolenghi, S., Giglions, B., Come, P., Gianni, A. M., Polli, E., Acquaye, C. T. A., Oldhom, J. H., Masera, G. 1979. Globin gene deletion in HPFH, δ°β° thalassemia and Hb Lepore disease. Nature 278:654-659

Pressley, L., Higgs, D., Clegg, J., Weatherall, D. 1980. Gene deletions in α-thalassemia prove that the 5' ζ locus is functional. Proc. Natl. Acad. Sci. USA, in press.

Proudfoot, N., Baralle, F. 1979. Molecular cloning of the human ε-globin gene. Proc. Natl. Acad. Sci. USA 76:5435-5439

Ramirez, F., O'Donnell, J. V., Marks, P. A., Bank, A., Musumeci, S., Schiliro, G., Pizzarelli, G., Russo, G., Luppis, B., Gambino, R. 1976. Abnormal or absent beta mRNA in beta° Ferrara and gene deletion in delta-beta thalassemia. Nature 263:471-475

Ramirez, F., Burns, A. L., Mears, J. G., Spence, S., Starkman, D., Bank, A. 1979. Isolation and characterization of cloned human fetal globin genes. Nucl. Acid Res. 7:1147-1162

Ricco, G., Muzza, U., Turi, R. M., Pich, P. G., Camashella, C., Saglio, G., Bernini, L. F. 1976. Significance of a new type of human fetal hemoglobin carrying a replacement isoleucine→ threonine at position 75 (E19) of the γ chain. Hum. Genet. 32:305-313

Roberts, A. V., Weatherall, D. J., Clegg, J. B. 1972. The synthesis of human hemoglobin A2 during erythroid maturation. Biochem. Biophys. Res. Comm. 47:81-87

Rougeon, F., Kourilsky, P., Mach, B. 1975. Insertion of a rabbit β-globin gene sequence into an E. coli plasmid. Nucl. Acids Res. 2:2365-2378

Saglio, G., Ricco, G., Mazza, U., Camaschella, C., Pich, P. G., Gianni, A. M., Gianazza, E., Righette, P. G., Giglione, B., Comi, P., Gusmerole, M., Ottolenghi, S. 1979. Human $^{T}\gamma$ globin chain is a variant of $^{A}\gamma$ chain ($^{A}\gamma$ sardinia). Proc. Natl. Acad. Sci. USA 76:3420-3424

Sakano, H., Rogers, J., Huppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R., Tonegawa, S. 1979. Domains and the hinge region of an immunoglobulin heavy chain are encoded in separate DNA segments. Nature 277:627-633

Schroeder, W. A., Bannister, W. H., Grech, J., Brown, A., Weightstone, R., Huisman, T. 1973. Non-synchronized suppression of postnatal activity in non-allelic genes which synthesize the $^{G}\gamma$ chain in human fetal hemoglobin. Nature New Biol. 244: 89-90

Shen, C.-K. J., Maniatis, T. 1980. The organization of repetitive sequences in a cluster of rabbit β-like globin genes. Cell 19:379-391

Shortle, D., Nathans, D. 1978. Local mutagenesis: a method for generating viral mutants with base substitutions in preselected regions of the viral genome. Proc. Natl. Acad. Sci. USA 75:2170-2174

Shortle, D., Pipas, J., Lazarowitz, S., DiMaio, D., Nathans, D. 1979. Constructed mutants of simian virus 40. In Genetic Engineering, ed. J. K. Setlow, A. Hollaender, 1:73-92. New York: Plenum.

Slightom, J., Blechl, A., Smithies, O. 1980. Human fetal $^{G}\gamma$ and $^{A}\gamma$ globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell, in press.

Smithies, O., Blechl, A., Denniston-Thompson, K., Newell, N., Richards, J., Slightom, J., Tucker, D., Blattner, F. Cloning human fetal γ-globin and mouse α-type globin DNA: characterization and partial sequencing. Science 202:1284-1289.

Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98:503-517

Spritz, R., DeRiel, J., Forget, B., Weissman, S. 1980. Nucleotide sequence of the human δ globin gene. Cell. Submitted for publication

Stalder, J., Larsen, A., Groudine, M., Engel, J. D., Dodgson, J., Weintraub, H. 1980. Tissue-specific cuts in the globin chromatin domain introduced by DNase I. Submitted for publication.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., Inouye, M. 1966. Frameshift mutations and the genetic code. Cold Spring Harbor Symp. Quant. Biol. 31:77-84

Taylor, J. M., Dozy, A., Kan, Y. W., Varmus, H. E., Lie-Injo, L. E., Ganesan, J., Todd, D. 1974. Genetic lesion in homozygous α thalassemia (hydrops fetalis). Nature 251:392-393

Temple, G. F., Chang, J. C., Kan, Y. W. 1977. Authentic β-globin mRNA sequences in homozygous β°-thalassemia. Proc. Natl. Acad. Sci. USA 74:3047-3051.

Tilghman, S. M., Curtis, P. J., Tiemeier, D. C., Leder, P., Weissman, C. 1978b. The intervening sequence of a mouse β globin gene is transcribed within the 15 S β-globin mRNA precursor. Proc. Natl. Acad. Sci. USA 75:1309-1313

Tilghman, S. M., Tiemeier, D. C., Polsky, F., Edgell, M. H., Seidman, J. G., Leder, A., Enquist, L. W., Norman, B., Leder, P. 1977. Cloning specific segments of the mammalian genome: bacteriophage λ containing mouse globin and surrounding sequences. Proc. Natl. Acad. Sci. USA 74:4406-4410

Tilghman, S. M., Tiemeier, D. C., Seidman, J. G., Peterlin, B. M., Sullivan, M., Maizel, J., Leder, P. 1978a. Intervening sequence of DNA identified in the structural portion of a mouse β-globin gene. Proc. Natl. Acad. Sci. USA 75:1309-1313

Todd, D., Lai, M. C. S., Beaven, G., Huehns, E. 1970. The abnormal haemoglobins in homozygous α-thalassemia. Brit J. Haematol. 19:27-31

Tuan, D., Biro, P. A., de Riel, J. K., Lazarus, H., Forget, B. G. 1979. Restriction endonuclease mapping of the human γ globin gene loci. Nucl. Acids. Res. 6:2519-2544

Tuan, D., Murnane, M. J., de Riel, J. K., Forget, B. G. 1980. Heterogenity in the molecular basis of hereditary persistence of fetal hemoglobin. Nature, in press.

van den Berg, J., van Ooyen, A., Mantei, N., Shambock, A., Grosveld, G., Flavell, R. A., Weissmann, C. 1978. Comparison of cloned rabbit and mouse β-globin genes showing strong evolutionary divergence of two homologous pairs of introns. Nature 276:37-44

van der Ploeg, L. H. T., Konings, A., Oort, M., Roos, D., Bernini, L., Flavell, R. A. 1980. γ-β-thalassemia studies showing that deletion of the γ- and δ-genes influences β-globin gene expression in man. Nature 283:637-642

Vannin, E., Smithies, O. 1980. Submitted for publication.

Wasi, P., Na-Nakorn, S., Suingdumrong, A. 1964. Hemoglobin H disease in Thailand: a genetical study. Nature 204:907-908

Weatherall, D. J., Clegg, J. B. 1975. The α-chain-termination mutants and their relationship to α-thalassemia. Philos. Trans. R. Soc. Lond. 271:411-455.

Weatherall, D. J., Clegg, J. B. 1979a. Recent developments in the molecular genetics of human hemoglobin. Cell 16:467-479

Weatherall, D. J., Clegg, J. B. 1979b. The Thalassemia Syndromes, Blackwell Scientific Publications, 3rd ed.

Weatherall, D., Clegg, J., Wong, H. 1970. The haemoglobin constitution of infants with the haemoglobin Bart's hydrops fetalis syndrome. Brit. J. Haematol. 18:357-367

Weatherall, D. J., Clegg, J. B., Wood, W. G., Pasvol, G. 1979. Human haemoglobin genetics. In Human Genetics: Possibilities and Realities. Ciba Foundation Symp. 66 (New Series), pp. 147-186. New York: Exerptea Medica.

Weaver, R. F., Weissman, C. 1979. Mapping of RNA by a modification of the Berk-Sharp procedure: The 5' termini of 15 S β-globin mRNA precursor and mature 10 S β-globin mRNA have identical map coordinates. Nucl. Acids Res. 7:1175-1193

Weil, P. A., Luse, D. S., Segall, J., Roeder, R. G. 1979. Selective and accurate initiation of transcription at the Ad2 major late promotor in a soluble system dependent on purified RNA polymerase II and DNA. Cell 18:469-484

Weintraub, H., Flint, S. J., Leffak, I. M., Groudine, M., Grainger, R. M. 1978. The generation and propagation of variegated chromosome structures. Cold Spring Harbor Symp. Quant. Biol. 42:401-407

Weintraub, H., Groudine, M. 1976. Chromosomal subunits in active genes have an altered conformation. Science 193:848-853

Weissman, C., Nagota, S., Taniguchi, T., Weber, H., Meyer, F. 1979. The use of site directed mutagenesis in reversed genetics. In Genetic Engineering, ed. J. Setlow, A. Hollaender, 1:133-150. New York: Plenum.

Wigler, M., Silverstein, S., Lee, L. S., Pellicer, A., Cheng, Y. C., Axel, R. 1977. Transfer of purified herpes virus thymidine kinase gene to cultured mouse cells. Cell 11:223-232

Wigler, M., Sweet, R., Sim, G. K., Wold, B., Pellicer, A., Lacy, E., Maniatis, T., Silverstein, S., Axel, R. 1979. Transformation of mammalian cells with genes from procaryotes and eucaryotes. Cell 16:777-785.

Wilson, J. T., Wilson, L. B., de Riel, J. K., Villa-Komaroff, L., Efstratiadis, A., Forget, B. G., Weissman, S. M. 1978. Insertion of synthetic copies of human globin genes into bacterial plasmids. Nucl. Acids Res. 5:563-581

Wold, B., Wigler, M., Lacy, E., Maniatis, T., Silverstein, S., Axel, R. 1979. Expression of an adult rabbit β-globin gene stably inserted into the genome of mouse L cells. Proc. Natl. Acad. Sci. USA 76:5684-5688

Wood, W. G., Weatherall, D. J. 1973. Haemoglobin synthesis during human fetal development. Nature 244:162-165

Wood, W. G., Old, J. M., Roberts, A. V. S., Clegg, J. B., Weatherall, D. J. 1978. Human globin gene expression: control of β, δ, and δβ chain production. Cell 15:437-446

Wood, W. B., Clegg, J. B., Weatherall, D. J. 1979. Hereditary persistence of fetal hemoglobin (HPFH) and δβ thalassemia. Brit. J. Hematol. 43:509-520

Wu, C., Bingham, P. M., Livak, K. J., Holmgren, R., Elgin, S. C. R. 1979. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. Cell 16:797-806

Wu, C., Wong, Y.-C., Elgin, S. C. R. 1979. The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. Cell 16: 807-814

Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W., Wilson, A. C. 1980. Rapid duplication and loss of genes coding for the α chains of hemoglobin. Proc. Natl. Acad. Sci. USA 77 (in press).

Table 1 α-specific Eco RI fragments associated with α-thalasssemia syndromes

	Chromosome A		Chromosome B	
	Gene number	Fragment size	Gene number	Fragment size
Normal	2	23	2	23
α-thalassemia 2 (A,B,M)	2	23	1	19
α-thalassemia 2 (M)	2	23	1 (1)	23
α-thalassemia 1 (A)	0	-	2	23
α-thalassemia 1 (A,B,M)	1	19	1	19
α-thalassemia 1 (M)	1	19	1 (1)	23
Hb H (A,B,M)	0	-	1	19
Hb H (A,M)	0	-	1 (1)	23
Hb H (M)	1	19	0 (1)	2.6
Hb H (M)	1 (1)	23	0 (1)	2.6
hydrops fetalis (A)	0	-	0	-

The various α-thalassemia syndromes are listed in the left column. The size of the Eco RI fragment containing the α-globin genes is given in kilobase pairs. The number of functional α-globin genes on each chromosome is indicated, followed by a number in parentheses to indicate the number of nonfunctional genes on the same chromosome. A, B, and M in parentheses following the name of a syndrome signify that the indicated combination of Eco RI fragments has been observed in Asian, black, or Mediterranean populations, respectively.

Figure 1 Structure of human globin genes. The canonical structures for the human α-like and β-like globin genes are drawn to approximate scale. Solid and open boxes represent coding (exon) and noncoding (intron) sequences, respectively. The α-like globin genes contain introns of approximately 95 and 125 bp, located between codons 31 and 32 and 99 and 100, respectively. The β-like globin genes contain introns of approximately 125-150 and 800-900 bp, located between codons 30 and 31 and 104 and 105, respectively.

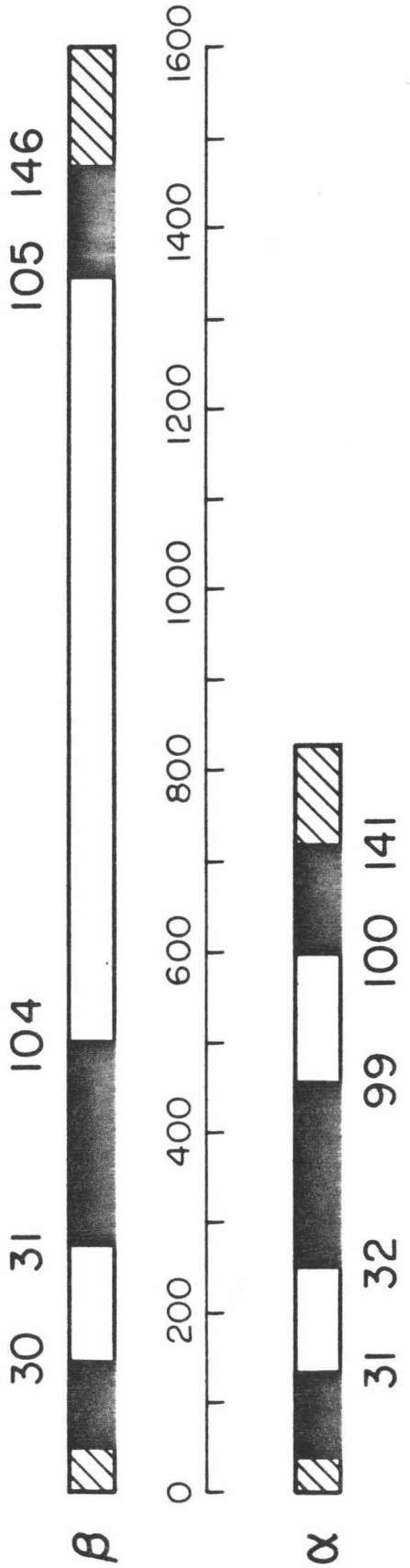

β
α
0
200
400
600
800
1000
1200
1400
1600
30
31
104
105
146
31
32
99
100
141

Figure 1

Figure 2 Linkage arrangement of the human β-like globin genes and locations of deletions within the β-like gene cluster. The positions of the embryonic (ε), fetal ($^{G}\gamma$, $^{A}\gamma$) and adult (δ, β) β-like globin genes and the two β-like pseudogenes (ψβ1, ψβ2) are shown. For each gene the black and white boxes represent the coding (exon) and noncoding (intron) sequences, respectively. The distribution of coding and noncoding sequences within ψβ1 and ψβ2 is not known. The locations of various deletions within the gene cluster are presented below the map. Open boxes represent areas known to be deleted; dashed lines indicate that the endpoint of the deletion has not been determined; and stippled boxes represent uncertainty in the extent of the deletions. For δβ-thalassemia and HPFH, the type of fetal globin chain which is produced ($^{G}\gamma$ and/or $^{A}\gamma$) is indicated in the name of each syndrome (for example, in $^{G}\gamma$-$^{A}\gamma$-δβ-thalassemia, the $^{G}\gamma$- and $^{A}\gamma$-globin chains are produced). The percentage of Hb F observed in heterozygotes is given to the right of each deletion. An asterisk (*) indicates that the Hb F is entirely of the $^{G}\gamma$-type. See the text for references and discussion.

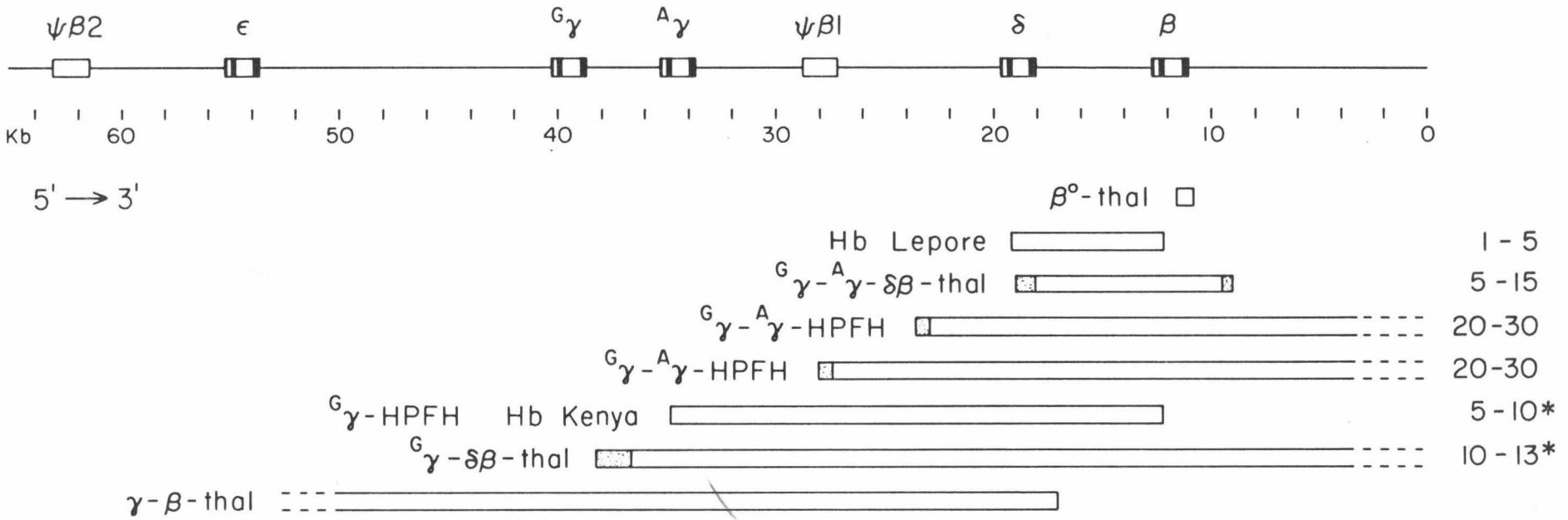

ψβ2
ε
Gγ
Aγ
ψβ1
δ
β
Kb
60
50
40
30
20
10
0
5' → 3'
β°-thal
Hb Lepore
1 - 5
Gγ-Aγ-δβ-thal
5 - 15
Gγ-Aγ-HPFH
20-30
Gγ-Aγ-HPFH
20-30
Gγ-HPFH Hb Kenya
5 - 10*
Gγ-δβ-thal
10 - 13*
γ-β-thal

Figure 2

Figure 3 Linkage arrangement of the human α-like globin genes and locations of deletions within the α-like gene cluster. The positions of the adult (α1, α2) and embryonic (ζ1, ζ2) α-like globin genes and the α-like pseudogene (ψα1) are shown. For each gene the black and white boxes represent coding (exon) and noncoding (intron) sequences. The introns in ζ2 are assumed to exist by analogy with the other α-like genes. The locations of deletions associated with the leftward and rightward types of α-thalassemia 2 are indicated by the rectangles labeled α-thal 2 L and α-thal 2 R. The cross-hatched boxes at the ends of these rectangles indicate regions of sequence homology. The breakpoints of each type of α-thalassemia 2 deletion can occur anywhere within the regions of homology. The locations of deletions associated with two cases of α-thalassemia 1 (α-thal 1 Thai and α-thal 1 Greek) are shown below the linkage map. The stippled boxes indicate uncertainty in the extent of each deletion. See the text for discussion and references.

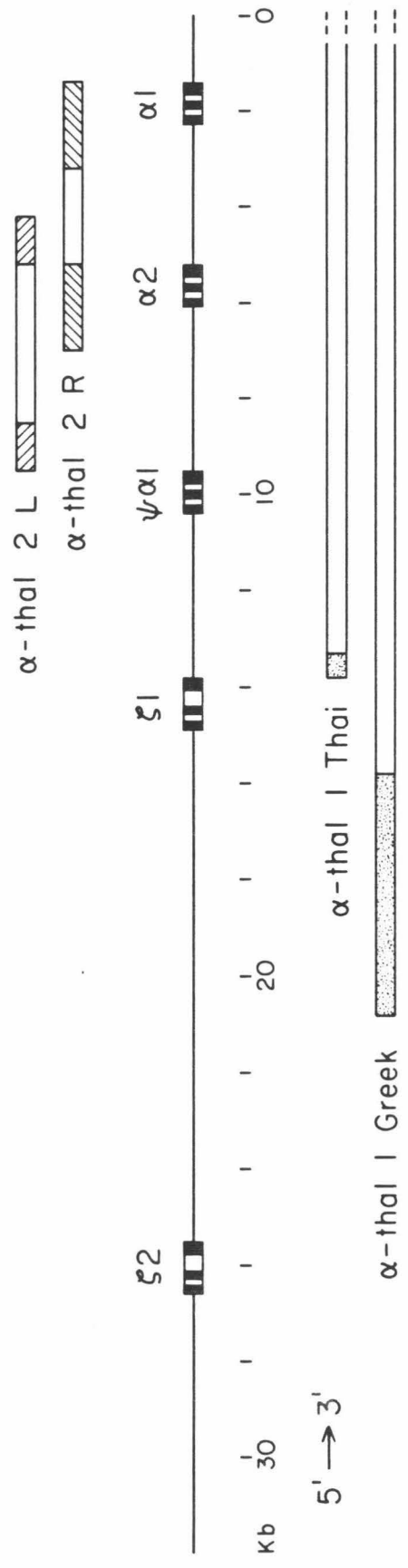

α-thal 2 L
α-thal 2 R
ζ1
ψα1
α2
α1
ζ2
α-thal 1 Thai
α-thal 1 Greek
Kb
30
20
10
0
5' → 3'

Figure 3

Figure 4 The distribution of sequence homologies within a region of the human α-like globin gene cluster. Regions of sequence homology are indicated by cross-hatched boxes, white boxes or stippled arrows. These homologies were detected by heteroduplex analysis of the cloned α-globin gene cluster (Lauer et al 1980). Recent nucleotide sequencing data indicate that the cross-hatched boxes extend to the left as far as the Pvu II sites (N. Proudfoot, unpublished results).

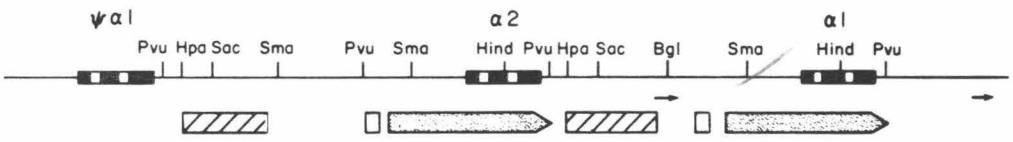

ψα1
α2
α1
Pvu Hpa Sac Sma
Pvu Sma
Hind Pvu Hpa Sac
Bgl
Sma
Hind Pvu

Figure 4

Figure 5 Schematic representation of genotypes associated with α-thalassemia syndromes. The pair of horizontal lines for each syndrome represents the chromosome 16 homologues. A black rectangle indicates the presence of a functional α-globin gene. Absence of a black rectangle signifies either gene deletion or a nondeletion defect leading to a nonfunctional α-globin gene. α-thal 1 or 2 signifies the α-thalassemia 1 or 2 syndromes.

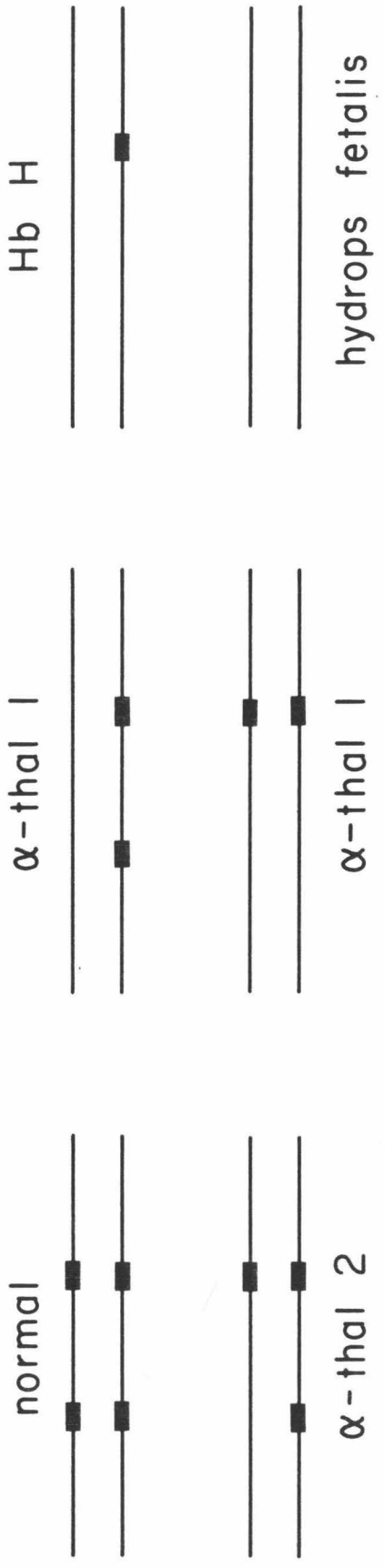


Figure 5