A COMPARATIVE STUDY OF β-GLOBIN PSEUDOGENES IN MAN AND THE PRIMATES

BY

STEPHEN HARRIS

Thesis submitted for the degree of
Doctor of Philosophy
in the University of Leicester

1985

Summary

The human β-globin gene family, situated on chromosome 11, consists of five functional genes (ε, $^{G}\gamma$, $^{A}\gamma$, δ and β) and a non-processed pseudogene, Ψβ1. The major work undertaken in this thesis consisted of a phylogenetic analysis of the non-processed pseudogene of the human β-globin cluster in order to establish the tempo and mode of evolution of such sequences, their suitability as examples of more general non-coding DNA sequence evolution and their possible influences on eukaryotic multigene family evolution.

This study of contemporary Ψβ1 pseudogene sequences has shown that this gene has been a stable component of the β-globin gene cluster during the evolution of the primate, and other, mammalian orders. The pseudogene was probably functional early in primate evolution and silenced probably before the basal primate radiation ~70 million years ago. The presence of a functional orthologue to the human Ψβ1 gene in the goat (the ε^{II} gene) supports the view that the human Ψβ1 gene was functional prior to its silencing early in primate evolution.

After silencing, the primate Ψβ1 pseudogene has evolved randomly in terms of base substitution and insertion/deletion at a mean rate thought to be representative of non-functional non-coding DNA sequences throughout the primates. These conclusions are supported by the mode and tempo of non-coding DNA sequence evolution observed within the functional brown lemur β-globin gene. However, the tempo of primate Ψβ1 gene evolution conflicts with views concerning the universal constant rate of neutral evolution, the rate of non-coding DNA evolution having apparently decreased to varying extents within the different lineages of this mammalian order. The consequences for primate β-globin gene cluster evolution of the presence of a non-processed pseudogene are discussed.

The distinct nature of the Ψβ1 gene in the human β-globin gene cluster, the history of the Ψβ1 gene in the primates and the presence of sequences related to Ψβ1 in various other mammalian orders suggests an additional ancient genetic locus was present in the ancestral β-globin gene cluster prior to the mammalian radiation. In order to acknowledge the distinct nature of this locus the human Ψβ1 gene has therefore been renamed η and other contemporary η-related sequences renamed accordingly. The evolution and gene orthologies of the other non-primate β-globin gene families are discussed in the light of the phylogenetic analysis of the human Ψη pseudogene. The simplest interpretation of the evolution of contemporary mammalian β-globin gene clusters is that they resulted from a common minimal ancestral cluster composed of proto ε-, γ-, η-, δ- and β-like sequences.

The generality of the conclusions drawn from this work concerning pseudogene longevity and sequence evolution after silencing await the phylogenetic analysis of other pseudogene sequences. It is apparent however that pseudogenes may constitute another potential source of genetic variation on which the processes of natural selection can act in the evolution of both eukaryotic multigene families and the genome in general.

Publications

Some of this work has already been published,

"Isolation and sequence analysis of a hybrid δ-globin pseudogene from the brown lemur." by Jeffreys,A.J., Barrie,P.A., Harris,S., Fawcett,D.H., Nugent,Z.J. and Boyd,A.C. (1982). J.Mol.Biol.,156, 487-503.

"The primate Ψβ1 gene: an ancient β-globin pseudogene." by Harris,S. Barrie,P.A., Weiss,M.L. and Jeffreys,A.J. (1984). J.Mol.Biol.,180,785-801.

and reviewed elsewhere,

"Processes of gene duplication." by Jeffreys,A.J. and Harris,S. (1982). Nature 296, 9-10.

"Evolution of gene families: the globin genes." by Jeffreys,A.J. Harris,S., Barrie,P.A., Wood,D., Blanchetot,A and Adams,S.M. (1983). in "Evolution from Molecules to Men." (eds.Bendall,D.S.), pp175-195, Cambridge University Press.

"Pseudogenes." by Jeffreys,A.J. and Harris,S. (1984). Bioessays 1,253-258.

Acknowledgments

I am indebted to my supervisor Dr A.J.Jeffreys for his continuous enthusiasm and support throughout this project and for the excellent insight he has given me into good scientific technique. I hope that in the future I can continue to justify his efforts during the past 3 years.

Special thanks also to Dr P.A.Barrie and Mrs V.Wilson who introduced me to laboratory techniques; Dr J.Brookfield for his patience in discussions of the statistical aspects of the project and to Dr Z.Nugent for her efforts in compiling a series of DNA manipulation programmes, in particular the dot-matrix plotting programme.

I am also grateful to all members of Dr Jeffreys laboratory and the Genetic Dept, University of Leicester, past and present, for providing an enjoyable and stimulating enviroment in which to work. In particular, Mr T.Lymn and the technical and workshop staff who provided the infrastructure so that I could concentrate on the science and to other postgraduate students and friends for social distractions in times of need.

Finally, I acknowledge the receipt of a M.R.C. studentship.

DEDICATION

To family and friends

Abbreviations

bp	base pairs
BSA	bovine serum albumen
Ci	Curies
cpm	counts per minute
DMSO	dimethyl sulphoxide
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
DTT	dithiothreitol
EDTA	ethylenediaminetetra-acetic acid
hrs	hours
kb	kilobases (1000x bp)
mins	minutes
MY(s)	million year(s)
mRNA	messenger ribonucleic acid
OD_n	optical density wavelength in nanometres
o/n	overnight
PEG	polyethylene glycol
pfu	plaque formimg units
RNase	ribonuclease
SDS	sodium dodecyl sulphate
tris	tris(hydroxymethyl)aminomethane
TEMED	N,N,N',N' tetramethylethylenediamine
dNTP	2'-deoxy(N) 5'-triphosphate N=adenosine, cytidine, guanosine
dTTP	Thymidine 5'-triphosphate
ddNTP	2',3'-dideoxy(N) 5'-triphosphate N=adenosine, cytidine, guanosine or thymidine

Contents

Chapter 1

INTRODUCTION

1.1 General introduction

The advent of recombinant DNA technology has resulted in a tremendous increase in fundamental scientific understanding of the structural complexities of eukaryotic genomes and is beginning to provide insights into eukaryotic gene expression, regulation and ultimately cellular differentiation into complex organisms such as man.

One feature common to many eukaryotic genomes is the arrangement of genes with related function into clustered multigene families that have apparently evolved by a series of gene duplications from a single primordial gene. There are two basic types of multigene family, the specialised gene family, such as the haemoglobins, where each member of the gene family has a distinct functional role, and the larger homogeneous gene family, such as the rRNA genes, where all the members of the gene family perform the same function. In general, both types of multigene family also exhibit some degree of gene dispersal around the genome; in some cases the same multigene family may exhibit different topologies in different species and may even have a different topology within the same species, depending on the stage at which the genes are expressed during development (for example, the histone genes of the sea urchin). The globin gene families of higher vertebrates are representative of the specialised type of gene family with a limited degree of gene dispersal, the α- and β-like globin genes being clustered on different chromosmes (for example, chromosomes 16 and 11 respectively in the human genome).

The expression of the members of a multigene family is generally regulated by some form of cellular/developmental "clock" or in response to a physiological stimuli or a combination of both. Several different multigene families have therefore been extensively studied as model systems of gene arrangement, regulation and evolution on the assumption they exhibit features representative of the genome as a whole.

Several approaches have been employed in the study of the various multigene families. The gene clusters themselves have been isolated using recombinant DNA technology and individual genes subjected to detailed structural and functional analysis involving DNA sequencing and in vitro mutagenesis followed by in vitro and in vivo expression. Naturally occuring mutants can also be characterised in order to determine the primary defect and infer a function from the disruption to the normal phenotype, either of a single gene or regulation of a number of genes. The principles of evolutionary biology and population genetics have also been applied to the study of the same multigene family in a variety of different organisms in order to determine the evolutionary forces and structural requirements which have resulted in contemporary gene family arrangements.

This introduction concentrates on the contribution advances in biochemical and molecular genetics have made to our understanding of genome organisation and evolution at the molecular level with particular reference to the haemoglobin genes, by far the best characterised multigene family in a number of different species. While references to the globin system are extensive, it is worth stressing that this multigene family embodies many of the features found in the other eukaryotic multigene families examined, for example, the immunoglobulin supergene family (Hood _et_ _al_., 1985), histones (Maxson _et_ _al_., 1984), tubulins (Cowen and Dudley, 1983), actins (Firtel, 1981), rRNA genes (Long and Dawid, 1980), and snRNAs (Westin _et_ _al_., 1984; Busch _et_ _al_., 1982).

1.2 Molecular Evolution

a) Constancy in the rate of molecular evolution

Molecular evolution is a recent branch of evolutionary biology concerned with the processes of evolution within the fundamental informational macromolecules of the cell; proteins, RNA and DNA. The realisation that homologous protein and DNA sequences apparently evolve at near constant rates within different lineages has been one remarkable finding of molecular evolution.

Proteins such as fibrinopeptide, cytochrome c, insulin, haemoglobin, and histone each evolve at an apparently constant overall rate (in terms of amino acid substitutions/ site/ year [aa sub./site/yr]) when compared between species of known evolutionary age (Dickerson, 1971). The substitution rate may differ within the polypeptide chain itself. For example, preproinsulin can be divided into three functional domains (polypeptide chains A,B,C); for polypeptides A and B to achieve the correct configuration relative to each other before post-translational modification, polypeptide C must be present and of correct length. As polypeptide C acts primarily as a spacer polypeptide between polypeptides A and B, and is lost by cleavage after modification of A and B, there is little functional constraint on its amino acid sequence which evolves rapidly (2.4 x 10^{-9} aa sub./site/yr) compared to a rate of 0.4 x 10^{-9} aa sub./site/yr for polypeptides A and B, which form the physiologically active protein (see Kimura, 1983a). The overall rate of amino acid substitution is characteristic of a particular protein and reflects the degree of functional constraint imposed by natural selection on the protein. RNA encoding genes also exhibit different substitution rates within the transcribed region which are apparently related to the importance of a particular base to the final higher order structure and function of the RNA (Curtiss and Vournakis, 1984).

Similarly, DNA sequences are also believed to evolve at near constant rates in different lineages (in terms of nucleotide substitutions/ site/ year [nuc.sub./site/yr]), see Kimura (1983b). Within DNA sequences of the genome there are apparently several different rates of change. DNA coding regions (the exons) of protein encoding genes exhibit two different rates of nucleotide substitution. The replacement site rate is variable, gene specific and subject to purifying selection as the nucleotide substitutions, which occur predominantly at the 1st and 2nd position of the triplet codon, directly effect the amino acid composition of the active protein. In contrast, silent site substitutions (those which cause synonymous codon changes that do not directly effect the amino acid sequence, principally at the 3rd position of the triplet codon) evolve at a higher rate that is apparently common to different genes in different lineages.

The relatively high silent site substitution rate is thought to reflect a lack of constraint on substitutions at certain codon positions due to redundancy in the genetic code which allows several different triplet codons (up to 6) to encode the same amino acid position within the polypeptide chain of a protein. It is unclear however whether some form of weak purifying selection may act on certain silent site substitutions, for example, due to altered patterns of synonymous codon usage in different species (see Grantham _et al._, 1981; Ikemura, 1985), or selection for RNA secondary structure (Curtiss and Vournakis, 1984). The relatively high level of silent site substitutions has proved difficult to reconcile with mechanisms of molecular evolution based on Darwinian principles of natural selection (see 1.2(c)). An alternative interpretation (the neutral theory, see Kimura, 1983b) suggests that silent site substitutions may be neutral, or near neutral, mutations which are uncoupled from natural selection acting on the protein and accumulate by a process of random genetic drift driven by a constant mutation rate. The silent site substitution rate, according to the neutral theory, can therefore be considered a conservative approximation of the mutation rate per gamete per generation.

An early prediction of the neutral theory was that the tempo and mode of mutation would most accurately be estimated from non-functional non-coding DNA sequences such as introns and extragenic spacer DNA. In theory, the divergence between two genes within these sequences should more closely reflect the mutation process due to the absence of any selective constraint upon the nucleotide substitution rate _per se_. However, in order for rate comparisons to be made two non-coding sequences first have to be aligned. This requirement initially precluded the comparison of the majority of non-coding DNA sequences due to alterations in DNA sequence other than base substitutions that made alignment difficult; such alterations include micro and macro duplications/ deletions, transposon insertion/deletion and alterations in copy number of repeat sequences.

Where alignment between two homologous non-coding DNA sequences can be obtained (for example, between small introns, Efstratiadis et al., 1980) the relatively high rate of change in these regions results in the possibility of multiple base substitution at a single site such that the 'true' divergence is underestimated. This potential error in divergence estimates, due to multiple substitution at a single site, is also encountered when calculating the silent site subs~~i~~titution rate within coding regions. While general mathematical formulae to compensate for multiple base substitutions have been proposed, formulae that take into account all potential modes of base substitution (Gojobori, 1983; Kimura, 1983; Tajuma and Nei, 1984), the validity of any particular correction has yet to be tested by the phylogenetic analysis of the sequence evolution of a single gene.

Also, while the majority of non-coding DNA sequence within and in close proximity to genes is thought to be non-functional and therefore evolving as "junk" DNA, it is possible that certain mutation events within these regions may ~~e~~affect gene function and therefore be selected against (see 1.5(b)). The non-coding sequences in close proximity to important flanking and coding DNA sequences would therefore not necessarily reflect the mode and tempo of evolution of other such sequences in the genome. For example, a new splicing site within an intron may result from a base substitution that leads to aborted or incorrect mRNA maturation. Several reasons therefore reduce the value of observations of non-coding DNA evolution based on the analysis of non-coding DNA regions within and immediately surrounding functional genes.

The detection, isolation and sequence analysis of pseudogenes within the genome (see 1.4) has provided an ideal opportunity for analysing the effects of the mutation process on sequences that are believed, after silencing, to be representative of other non-functional "junk" DNA in the genome (see Kimura, 1983b). Detailed analytical comparisons of sequence evolution in a functional gene compared with that in a related pseudogene suggest that pseudogene sequences, after silencing, evolve freely in the apparent absence of any selective constraint (Li et al., 1981; Miyata and Yasunaga, 1981). However, these estimates are based on molecular phylogenies that assume constant rates of DNA sequence evolution in different lineages and the absence of recombinational exchange with other members of the gene family of which the pseudogene and functional gene are members. Neither of these two assumptions are necessarily correct. The phylogenetic analysis of the β-globin gene family in the primates (Barrie et al., 1981) has shown the potential effect recombinational exchange between different members of a gene family can have on phylogenetic reconstructions based solely on rates of protein or DNA evolution (see 1.3 and 1.7).

The phylogenetic analysis of a pseudogene is therefore needed to answer fundemental questions concerning the evolution of such a sequence within a gene cluster, to reveal the potential influences and consequences that a pseudogene may have on gene cluster evolution and any effects the other members of a gene family may have on the evolution of the pseudogene. Detailed sequence analysis of a pseudogene would reveal the evolutionary history of sequence change in different regions of the pseudogene after silencing and provide a clearer understanding of non-coding sequence evolution in general.

b) Molecular polymorphism

One early observation concerning evolution at the molecular level was the degree of polymorphism between individuals of a population (Harris, 1966; Lewontin and Hubby, 1966). Studies based upon protein polymorphism in various animal species suggest that in general about 2/3rds of all loci are polymorphic and that an individual is heterozygous for about a 1/3rd of all loci (Lewontin, 1974). In proteins these polymorphic substitutions are generally between amino acid residues with very similar physico-chemical properties which do not disrupt protein function (Sneath, 1966). DNA coding region polymorphisms are generally due to base substitutions which result in a synonymous codon being generated, or to a conservative amino acid substitution within the protein. Polymorphism within non-functional non-coding DNA sequences, as estimated from restriction site length polymorphisms and DNA sequencing, can potentially result from micro/macro insertions and deletions, inversions, movement of transposons and changes in copy number within repetitive DNA sequences as well as nucleotide base substitutions.

c) Opposing interpretations of molecular evolution and evolutionary change

The degree of observed polymorphism at the molecular level and the apparent "clock" rate of evolutionary change has resulted in a debate between two different interpretations, the so called "Neutralist-Selectionist controversy" (Lewontin, 1974; Harris, 1976). The debate centres around the relative contribution each side attributes to the mutation process and natural selection in determining molecular evolution.

The "neutral mutation-random drift" hypothesis or "neutral theory" is based upon the premise that the majority of mutational changes that arise during molecular evolution are selectively neutral, or near neutral, changes that become fixed (or extinct) due to random drift within the population (for review see Kimura, 1983a). The observed rate of molecular evolution, in terms of mutant substitutions per site per year, is attributed primarily to the fixation of such neutral, or near neutral, variants by the process of random drift, a process that is apparently constant, lineage-independent and not influenced by such factors as generation time. Observed polymorphisms amongst individuals of a population are a result of the evolutionary "snapshots" taken when sampling contemporary protein and DNA sequences which contain neutral mutations at varying stages of fixation/extinction due to random drift. This hypothesis (first proposed by Kimura, 1968 and King and Jukes, 1969, and since championed by Kimura, 1983b) emphasises the role of mutation and random drift in driving molecular evolution and is based on quantitative population genetic principles which allow general predictions to be formulated as to how sequences within the genome should evolve.

The selectionists primary challenge of the neutral theory rests on the need to invoke "non-Darwinian" (neutral) evolution to explain high levels of polymorphism. They maintain that the primary determinant of molecular evolution is the same as that at the phenotypic level, that is, natural selection. By this criterion all polymorphic variants are maintained by some form of positive or balanced selection such as heterozygous advantage, frequency-dependent selection or a heterogeneous environment (Wills, 1978; Falconer, 1981). While undoubtedly some polymorphisms are maintained by processes of positive or balanced selection, no general mechanism of Darwinian selection has been proposed which can alleviate the genetic load associated with the maintenance of the estimated level of genetic polymorphism within populations (see Kimura, 1983b).

Other arguments put forward as evidence against the neutral theory are the apparently different evolutionary rates between and within different lineages (Goodman *et al.*, 1975; Goodman, 1981; Czelusniak *et al.*, 1982), implying that the mutation rate is not constant, and the larger than expected variance between observed rate observations (Langley and Fitch, 1974). The presence of different evolutionary rates implying possible alterations in the mutation rate would not falsify the neutral theory yet the neutralists defend rate constancy in the following manner. While accepting there are "local" fluctuations in the near constant rate of molecular evolution which exceed the statistical limits expected by chance alone, neutralists note that these fluctuations are small (observed variance 1.5-2.5x those expected by chance) and should not detract from the general agreement of most rate comparisons for rate constancy (Kimura, 1983b). Initial phylogenetic analysis of restriction endonuclease site variation in the β-globin gene cluster of the primate lineage may constitute the first indication that the mutation rate may genuinely be variable within different lineages (Barrie *et al.*, 1981).

While the principles of the neutral theory are generally accepted, many of the predictions concerning the tempo and mode of sequence evolution in different lineages have yet to be tested by the detailed phylogenetic analysis of homologous DNA or RNA sequences. Particularly important in this respect is the phylogenetic analysis of pseudogene evolution as this would test the predictions of the neutral theory which has championed pseudogenes as paradigms of neutral evolution (Li et al., 1981).

1.3 Molecular evolution of multigene families

a) Molecular "clocks" and the timing of gene duplications

Contemporary multigene families are thought to have evolved by a series of gene duplication events followed by sequence divergence to acquire new functions. Evolutionary relationships between members of a multigene family can be estimated in several ways. a) The taxonomic distribution of related proteins can be used to determine when related proteins arose relative to one another. This method requires a large body of protein data and an established phylogeny (see Dayhoff, 1972). b) The accumulated sequence divergence between homologous sequences from different species, or within a single species, can be used to construct molecular phylogenies based on the principal of "minimum evolution" or "maximum parsimony" (Fitch and Margoliash, 1966; Farris, 1970, 1972; Dayhoff, 1972; Sneath and Sokal, 1973). In the absence of an established phylogeny or other palaeontological evidence, these molecular phylogenies (often depicted as a branching tree, the root of which is the presumed common ancestor of all contemporary sequences) represent the similarity between individual sequences (or groups of sequences) but do not establish

when, on an evolutionary time scale, sequences diverged from one another.
c) The apparent molecular "clock" of homologous protein and DNA sequences can be utilised to construct molecular phylogenies relative to an evolutionary time scale. This approach requires the initial calibration of the "clock" by comparison of accumulated sequence divergence with the time of common ancestry of two species, as estimated from the palaeontological record. This comparison provides an estimate of the rate of change during a given period of evolution, a rate which appears to be fairly constant for a particular protein irrespective of the evolutionary lineage (see 1.2(a)), that can be used to estimate the divergence time of different species or of the members of a multigene family within a single species. The molecular phylogenies obtained in this manner generally accord well with the classical phylogenetic data and taxonomic distribution of the genes/proteins where known.

Extensive molecular phylogenies have been constructed by these methods from protein sequence data (Dayhoff, 1976; Goodman et al., 1974, 1975) and more recently this type of analysis has been extended to the accumulating DNA sequence database (Efstratiadis et al., 1980; Czelusniak et al., 1982). Replacement site substitutions within the exons of genes have been the most widely used in determining molecular phylogenies from DNA sequences; however, by definition, these replacement site substitutions result in molecular phylogenies equivalent to those determined from protein amino acid sequence comparisons.

As mentioned above, silent site substitutions are limited in their use as molecular "clocks" due to the relatively high rate of substitution at these positions. Similarly, non-coding DNA sequences such as introns and extragenic spacer DNA also accumulate substitutions at an even higher rate and they are also often difficult to align due to multiple micro and macro duplications/deletions, transposon insertions and changes in repeat element copy number. Analytical computer algorithms (for example see Goodman _et al._, 1984) and dot-matrix analyses have been used to aid the alignment of non-coding regions (Konkel _et al._, 1979), however it is still not always possible to distinguish the 'best' alignment on which to base divergence comparisons.

b) Concerted evolution in multigene families

There is growing evidence that members of a multigene family do not evolve independently of each other but can be involved in recombinational exchanges such as unequal exchange and gene conversion (for review see Arnheim, 1983). In particular, the non-reciprocal exchange or "concerted evolution" of related DNA sequences can have a profound effect upon the apparent evolutionary history of a multigene family.

For example, the single amino acid difference between the human foetal globin proteins ($^{G}\gamma$ and $^{A}\gamma$) suggests a recent duplication event some 6-7 MY ago, that is some time before the great ape-human divergence in the primate lineage. However, analysis of the arrangement of the β-globin gene cluster in the primate lineage demonstrated the presence of duplicate γ-loci in Old World monkeys, suggesting the presence of duplicate γ-loci in the common ancestor of the Old World monkeys and the hominoids some 20+ MY ago (Barrie _et al._, 1981). DNA sequencing of the linked $^{G}\gamma$ and $^{A}\gamma$ genes from a single polymorphic human individual has shown that this discrepency probably results from a gene conversion event between the duplicated genes (Slightom et al., 1981). The DNA sequences involved in the proposed gene conversion tract (1.5 kb) exhibit a reduced level of sequence divergence compared to the other DNA sequences thought to have been involved in the duplication event (a total of 5 kb) and lying outside the conversion tract (Shen et al., 1981). Since the proposed gene conversion event only one amino acid replacement substitution event has been fixed, which led to the incorrect estimate of the duplication time from protein sequence comparisons. This may not however have been the only gene conversion event during the evolutionary history of these two genes (Scott et al., 1984).

Further examples of gene conversion between members of the globin multigene family during their evolution have been described within the α-globin gene families of the goat (Schon et al., 1982), human (Zimmer et al., 1980, Liebhaber et al., 1981; Proudfoot et al., 1982) and mouse (Hill et al., 1984); the β-globin gene families of the rabbit (Hardison and Margot, 1984), mouse (Konkel et al., 1979) and man (this thesis, see Discussion). Gene conversion events have also been inferred within the following multigene families, rRNA genes (Arnheim et al., 1983), immunoglobulin V_H genes (Hood et al., 1975 however see Gojobori and Nei, 1984) and the heat shock genes of Drosophila (Leigh-Brown and Ish-Horowicz, 1981).

Detailed DNA sequence comparisons can detect recent gene conversion events as a non-random distribution of the accumulated divergence between two gene sequences (see Scott et al., 1984), or as regions of unexpected homology within non-coding regions on dot-matrix analysis of two sequences (Hill *et al.*, 1984; Hardison, 1984, ~~see,~~ for example, see Figure 3.2). However, while gene conversion appears to be a common feature of multigene family evolution, the mechanism(s) of gene conversion in higher eukaryotes is unknown; models are based on events characterised in the lower eukaryotes (see Radding, 1978), with reference to the enzymology of recombination in bacteria (see Radding, 1982). For example, it is unclear how a conversion event is initiated, though sequence homology and the possible involvement of DNA signal sequences within introns have been implicated (Slightom *et al.*, 198~~0~~1).

Those gene conversion events so far described appear to occur between duplicate genes of closely related function or between a functional gene and a pseudogene within the gene family. Mechanisms of biased concerted evolution or 'molecular drive' have been proposed to account for the homogenisation of the members of large identical multigene families (Dover, 1982; Arnheim, 1983). However, some mechanism must restrain the indiscriminate homogenisation of functionally specialised members within a multigene family such as the globins. Natural selection will act against conversion events which disrupt essential gene functions. In addition, it may be possible that structural features of gene families prevent certain combinations of genes being involved in concerted evolution.

1.4 Pseudogenes

a) Identification and characterisation

Multigene families have appar~~a~~ently evolved by a series of gene duplication events followed by adaptive divergence to produce new functions. DNA analysis has shown that families of related functional genes are larger than expected due to the presence within the genome of additional non-functional sequences called pseudogenes. Some of these extra gene copies appear to have arisen by the silencing of a previously functional gene, the non-processed pseudogenes, while others are apparently reintegrated DNA copies of mature RNA transcripts of their functional progenitor gene, the processed pseudogenes. Non-processed pseudogenes are generally linked to their parent gene cluster whereas processed pseudogenes are dispersed in the genome. Processed pseudogenes share many characteristics with other classes of mobile genetic element thought to disperse throughout the genome via a RNA intermediate (see below).

The first 'pseudogene' described formed part of the 5S DNA repeat unit of *Xenopus laevis* (Jacq *et al*., 1977). Since then other pseudogene sequences have been detected by nucleic acid hybridisation to related functional gene sequences. This experimental method limits the detection of pseudogene sequences to those less than 30-40% diverged from their functional counterpart, although in principle computer analysis of DNA sequence data banks may detect more diverged pseudogenes. Detailed DNA sequence analysis is required to confirm the authenticity and dysfunction of a pseudogene to ensure, for example, that the related sequence does not in fact encode a minor but previously unidentified essential product (Phillips *et al.*, 1984), or a spurious sequence with homology to the hybridisation probe (Shen and Smithies, 1982). Pseudogenes derived from protein encoding genes generally show several defects which would prevent production of a functional protein, though in some cases the defect apparently involves non-translation of the mRNA rather than absence of transcription (see below).

Pseudogenes differ from silent alleles of a functional gene in that they all appear to be fixed components of the genome, all individuals in a species possessing a copy of a particular pseudogene. A silent allele may however be an intermediate in the transformation of a functional gene into a pseudogene during evolution. Silent alleles of functional genes may attain a low frequency in the population in heterozygotes; if the functional allele was subsequently released from selective constraint due, for example, to a change in gene regulation making the gene superfluous, such a silent allele may become fixed thoughout the population by neutral genetic drift. The following discussion is based upon the limited number of such sequences so far described; their evolutionary origins, mode of evolution and possible contribution to genome organisation.

b) Non-processed pseudogenes

Most non-processed pseudogenes apparently result from the silencing of a functional gene after a tandem gene duplication event. They exhibit most or all the structural features of their related functional gene(s), including introns and the flanking signal sequences implicated in eukaryotic gene expression. As a result of the gene duplication process, non-processed pseudogenes are generally closely linked to their functional relatives, though an example of a dispersed non-processed pseudogene has been described (Leder et al., 1981).

Many potential defects can be envisaged which could disrupt a functional gene; these include abnormal initiation and termination codons, altered normally invariant codons, deletions/insertions and nonsense mutations, as well as defects in signal sequences implicated in transcription, RNA maturation and translation. The presence of defects within pseudogene coding sequences does not necessarily imply that the abnormal gene is transcriptionally inactive in vivo as there are pseudogene sequences, such as a human leukocyte interferon gene, which are transcribed to produce non-translated mRNA (Goeddel et al., 1981).

Contemporary non-processed pseudogenes generally show various combinations of these defects, a typical example being the human Ψβ1 pseudogene present within the human β-globin gene cluster (Fritsch et al., 1980; Jagadeeswaran et al., 1982; Chang and Slightom, 1984). The defects present in a contemporary pseudogene such as this do not tell us which was responsible for the initial silencing of the gene. Pseudogene sequences will accumulate many additional defects after the silencing event and the original defect may even 'revert' and therefore appear normal in the contemporary pseudogene. A detailed phylogenetic analysis of the defects in a pseudogene is essential if the initial defect is to be determined and also to give some insight into the evolution of the DNA sequence after silencing. In most cases however this is not possible as a suitable phylogeny is not available.

Non-processed pseudogenes appear to be common wherever multiple identical or diverged copies of functional genes exist: they have been found within mammalian globin gene families (Proudfoot, 1980; Little, 1982); the human and mouse immunoglobulin genes (Max *et al.*, 1982; Flanagan and Rabbits, 1982; Takashashi *et al.*, 1982; Selsing *et al.*, 1982; Hiroshi *et al.*, 1983; Cohen and Givol, 1983; Gideon *et al.*, 1983); in the rDNA genes of several species (Glover, 1981; Brownell *et al.*, 1983); amoungst the human leukocyte interferon genes (Goeddel *et al.*, 1981; Ullrich *et al.*, 1982); and within the cuticle protein gene cluster of *Drosophila* (Snyder *et al.*, 1982); the leghaemoglobin gene cluster of soybean (Lee *et al.*, 1983); and even organelle genes may produce pseudogene copies (see Lewin, 1983).

These examples include families of related genes transcribed in the nucleus by RNA polymerase I, II or III as well as organelle genes that may generate defective relatives which become fixed within the nuclear genome. The full extent of any particular non-processed gene family has not been established. There may be an upper limit to the number of such sequences a multigene family can tolerate without disruption of the functional genes or, alternatively, the genome may contain many pseudogenes in various stages of sequence 'decay', from recent derivatives to those so ancient and diverged to be indistinguishable as such.

Most non-processed pseudogenes are apparently generated by a tandem gene duplication event which may result in a concommitant gene silencing or later silencing after a period of divergent evolution. A single mutation event such as a base substitution or micro duplication/deletion within the functional gene sequence is the simplest possibility;

however several other mechanisms are possible. For example, a mobile genetic element may insert within the gene as is common in *Drosophila* (though not apparantly in man), and such an event may have resulted in the interrupted rRNA pseudogenes present in this species (Glover, 1981). One of a duplicate gene pair may occupy a non-transcribed "silent" chromatin domain or become separated from essential transcriptional enhancer elements; an example might be the human $\Psi\zeta$ gene which is the same in the promoter region as the neighbouring functional gene yet is not transcribed *in vivo* (Proudfoot *et al.*, 1982, 1984). Replicative transposition of a duplicate to an unlinked non-transcribed chromatin domain might also silence a gene; a few dispersed non-processed pseudogenes have been described which may have arisen by this process, for example, the mouse $\Psi\alpha4$ globin pseudogene (Leder *et al.*, 1981) and also some of the sea urchin dispersed "orphon" histone pseudogenes (Maxson *et al.*, 1983). Finally, one other process which may generate the wholesale appearance of non-processed pseudogenes is polyploidisation as found in some fish (see Li, 1980). Tetraploidization followed by a return to disomic segragation could in some instances result in the silencing of one of the duplicate loci to produce a pseudogene, or alternately, sequence divergence could occur to produce duplicate genes with different specialised functions. So far, fixed silent alleles generated by this method have not been analysed at the DNA sequence level.

For a non-processed pseudogene to become fixed within the population the silent allele must replace the original functional allele. This implies no selective constraint on the duplicate functional allele and also that the silent allele be essentially selectively neutral compared to the functional allele. The absence of selection for duplicate functional genes raises the problem of how the duplicate genes became established. One possible scenario is that normally the original gene duplication is advantageous and results in duplicate gene divergence to give two specialised functional genes. An evolutionary shift in gene regulation during this process may allow the accumulation of defects in one of the duplicate genes which ultimately becomes fixed within the population by neutral genetic drift.

Such a duplicate gene freed from selective constraint would be expected to accumulate genetic mutations relatively rapidly and behave as "junk DNA", a perfect model for neutral DNA evolution (Li et al., 1981, Kimura, 1983b). Occasionally however a pseudogene may have some active effect upon the parent gene cluster, perhaps through maintainance of chromatin domains or provision of regulatory elements. No evidence for such non-processed pseudogene function has yet been found. Theoretically a pseudogene may even become 'reactivated' after a period of neutral genetic drift (Ohno, 1970), strong selective pressure resulting in fixation of a new functional allele of a pseudogene. There is evidence for reversion of "cryptic" genes (pseudogenes with a single revertible defect) under strong selective pressure in bacteria (Hall et al., 1983; Li, 1984). It is difficult to envisage the independent reversion of several potentially silencing defects accumulated throughout a pseudogene. However genetic recombination, in the form of unequal exchange or gene conversion, may result in the reactivation of a pseudogene if the silencing defects were effectively replaced by sequences from a neighbouring functional gene. For example, the history of the contemporary human δ-globin gene may have included such an event during primate evolution (Martin *et al.*, 1983).

Present analyses of contemporary non-processed pseudogene histories estimate the time of the silencing event and subsequent sequence divergence by comparing the pseudogene sequence with that of functional relatives (Proudfoot and Maniatis, 1980; Lacy and Maniatis, 1980; Li *et al.*, 1981, Miyata and Yasunaga, 1981). These analyses, assuming "clock" rates of DNA sequence change within exons, suggest pseudogenes evolve rapidly at all positions after silencing and that the level of accumulated sequence divergence is generally consistent with the silencing event occurring after the initial duplication event. However, several factors may influence these conclusions. The estimates of the gene duplication and silencing times are subject to high statistically uncertainty such that in many cases the timing of these events is also consistant with concommitant duplication and silencing. Constant "clock" rates of DNA sequence change are assumed, as is the independent evolution of the pseudogene within the gene family, both as a functional gene after the initial gene duplication and subsequent to silencing. It is known however that members of a related multigene family do not necessarily evolve independently after duplication (see 1.3(b)).

Pseudogene sequences, due to their homology to related functional sequences, can potentially be involved in genetic exchanges with other members within the related gene cluster. Where described, pseudogenes involved in genetic recombinational events are within large homogeneous gene families where such events are unlikely to result in any, or very slight, selective disadvantage (Takashashi *et al.*, 1982; Hiroshi *et al.*, 1983). So far there is no information concerning the effect the presence of a closely related pseudogene such as the human Ψβ1 pseudogene may have had on the evolutionary history of a specialised multigene family such as the human β-globin genes.

c) Processed pseudogenes

First described in 1980, processed pseudogenes form a quite unexpected component of eukaryote genomes and contradict the central dogma in that they appear to result from the reintegration of cellular RNA sequences into the DNA of the germ line, a process previously thought possible in only a specialised group of RNA retroviruses and mobile genetic elements (see Sharp, 1983; Rogers, 1984; Vanin, 1984).

Processed pseudogenes are silent copies of functional genes which have undergone structural changes characteristic of RNA maturation, that is, they resemble DNA copies of mature RNA in that they lack introns and usually have a DNA copy of the RNA 3' poly(A) sequence. Processed pseudogenes are dispersed around the genome and are generally, though not always, flanked by direct repeats characteristic of the target site duplications associated with transposon insertion. Many processed pseudogenes contain defects within their "coding sequences". These defects probably accumulated subsequent to cDNA integration by neutral genetic drift. A processed pseudogene with no apparent coding sequence defects has been described (Karin and Richards, 1982). This is presumably a recent processed pseudogene insertion~~s~~, although absence of defects may have resulted from the recent correction of the pseudogene sequence by a functional processed mRNA.

Most processed pseudogenes that have been analysed originate in higher vertebrate genomes and include those of the mouse α-globin gene family (Nishioka *et al.*, 1980), the rat α-tubulin (Lemischka and Sharp, 1982), snRNA (Hiroshi and Kornberg, 1983) and cytochrome c (Scarpulla *et al.*, 1983) gene families, the human β-tubulin (Wilde and Cowen, 1982; Lee *et al.*, 1983), β-actin (Moos and Gallwitz, 1982, 1983), constant region immunoglobulin (Battey *et al.*, 1982; Ueda *et al.*, 1982; Hollis *et al.*, 1982), dihydrofolate reductase (Masters *et al.*, 1983), metallothionein (Karin and Richard, 1982), argininosuccinate synthetase (Freytag *et al.*, 1984), snRNA (Berstein *et al.*, 1983; Monstein *et al.*, 1983) and c-ras oncogene (McGrath *et al.*, 1983) families and a chicken calmodulin pseudogene (Stein *et al.*, 1983), though invertebrate genomes may contain such elements as some of the dispersed histone "orphons" of the sea urchin (Maxson *et al.*, 1983) and F-elements of *Drosophila* (Di Nocera and Dawid, 1983) also show signs of processing.

In mammals, the processed pseudogene component of a gene family can vary considerably from almost the entire complement of the gene family (*eg*, small nuclear RNA (snRNA)) to very few or no known members (*eg*, mouse α-globin family one $\Psi\alpha3$; human β-globin family none). Extensive processed pseudogene families are apparently restricted to higher vertebrates and may reflect the taxonomic host range distribution of retroviral like elements in the germ line (see below). The reason for the variation in processed pseudogene number is not understood; the relative level of germ line RNA transcription, efficiency of reverse transcription and reintegration of the cDNA of different RNA templates, plus the random genetic drift of such sequences through the population once integrated, probably all contribute to the observed variation in number.

As yet, the detailed mechanism that gives rise to processed pseudogenes remains unclear, although reverse transcriptase derived from endogenous proretroviruses is strongly implicated in the process of cDNA formation (Van Ardsell *et al*. 1981). Endogenous proretroviruses are common components of higher vertebrate genomes and reverse transcriptase activity has been demonstrated during early embryogenesis of mice (see Rogers, 1984). The generally accepted origin of processed pseudogenes is therefore via the reverse transcription of a mature RNA to produce a complementary DNA sequence (cDNA). This cDNA then reintegrates into the genome by a mechanism(s) that usually leads to the formation of a target site direct repeat.

The fidelity of RNA reverse transcription and integration is illustrated by the presence in the human genome of large numbers of processed snRNA pseudogenes corresponding to 3' truncated self-primed reverse transcripts produced by snRNA *in vitro* (Bernstein *et al*., 1983). In general full length cDNA copies are produced corresponding to the different mature RNAs (Scarpulla and Wu, 1983; Lee *et al*., 1983), however homology to the related functional gene sequence can extend 5' of the usual transcription initiation position into the promoter sequences of the gene and beyond (see Vanin, 1984). These 5' extended processed pseudogenes are thought to originate from minor RNA species whose transcription was initiated upstream of the usual position from a weak promoter, for example, a RNA polymerase III promoter.

It is unclear how the reverse transcriptase primes the cDNA synthesis of the different RNA species. Poly(A)$^+$ mRNA may be primed by oligo(dT) elements present in the nucleus or alternatively (Rogers, 1984) single stranded nicks in mammalian DNA might be extended by terminal transferase to produce oligo(dT)-rich single strand projections (analogous to the growth of teleomeres, see Blackburn and Szostak, 1983) which could prime cDNA synthesis at the site in the chromosome where ultimately integration occurs. As mentioned above, poly(A)$^-$ RNA can apparently undergo self-primed cDNA synthesis that results in 3' truncated processed pseudogenes when integrated. Hypothetical mechanisms have also been proposed for integration of the cDNAs involving DNA cleavage by topoisomerase enzymes that may/may not lead to a target site direct repeat (Bernstein et al., 1983), or the formation of a covalent RNA-DNA bond at a second staggered nick in the chromosome (Rogers, 1984).

The limited processed pseudogene sequence data allows few conclusions to be drawn about the rate of generation and evolution of these elements. However, most or all of the human arginosuccinate synthetase processed pseudogenes are also found in the chimpanzee, suggesting that at least 7 MY have elapsed since the last processed pseudogene was fixed (Freytag et al., 1984). The fixation of such sequences is therefore a rare germ line event. One interesting question is whether processed pseudogenes are produced continuously or if they are generated in a "big bang" due, for example, to the expression of reverse transcriptase during a transient germline retroviral infection. Unfortunately there is insufficient sequence data available to distinguish between these alternatives and further sequencing of large families of processed pseudogenes is needed to resolve this question.

As most processed pseudogenes lack the 5' flanking promoter sequences implicated in eukaryotic transcription it is unlikely that they are transcribed. Similarly, those abnormally initiated processed pseudogenes, derived from upstream promoters which contain their own promoter elements, are unlikely to be reintegrated at a position in the genome compatible with correctly regulated gene expression. Occasionally however a processed pseudogene may integrate adjacent to transcription promoter elements or an abnormally initiated processed pseudogene may integrate where it can be expressed. An example of such an expressed processed pseudogene may be a chicken calmodulin derived processed pseudogene which has been shown to be transcribed in a tissue specific manner (Stein *et al.*, 1983). Further analysis is required to determine whether the RNA produced is translated to give a functional protein or how the expression of the pseudogene is achieved.

Processed pseudogenes have several structural features in common with other classes of interspersed repetitive elements (see Rogers,1984). Like processed pseudogenes, these other mobile genetic elements are generally flanked by a genomic direct repeat at the point of insertion, have a DNA equivalent of a 3' oligo(A)-tract and are believed to reintegrate into the genome after reverse transcription of a RNA intermediate. Unlike processed pseudogenes, many of these sequences can initiate their own transcription and therefore replicate and disperse autonomously within the genome. An example of such an element are the Alu sequences found in the human genome which have equivalent related sequences in other species (see Singer, 1982).

The human genome contains approximately 300,000 copies of the 300 bp Alu sequence which show various degrees of homology to the 7SL RNA component of the signal recognition particle involved in export of proteins across membranes from several different species (Ullu and Tschudi, 1984). It is thought therefore that Alu sequences constitute an abundant processed 7SL RNA pseudogene, some of which have the ability to initiate their own transcription from an internal RNA polymerase III promoter (see Brown, 1984). The ability to replicate within the genome may account for the abundance of these Alu sequences in the human genome, the upper limit of which may only be restricted by the number of integration sites available within the genome which engender no selective disadvantage when occupied. Within the genome one might therefore expect some competition between different retroposon variants for the available "niches", that is, intragenomic selection for the most "selfish DNA" (see Rogers, 1984). The speed with which a new mobile genetic element (or variant) can spread through the genome and related populations is illustrated by the apparently recent spread of P-elements in wild populations of Drosophila (see Engels, 1983; Kidwell, 1983; Daniels et al., 1984).

The potential for such sequences to generate genetic variation within a population is illustrated by mobile genetic elements in Drosophila, yeast, bacteria and, to a lesser extent, mammalian cells (see Syvanen, 1984). Many of these mutational events will be disadvantageous and eliminated from the population by purifying selection; however some insertion events may result in shifts in gene regulation and expression which have a positive selective advantage and will therefore spread throughout the population. For example, it has been proposed that insertion of a member of the Kpn family of mobile genetic elements 5' of the duplicated human γ genes may have resulted in the shift to foetal expression of these genes during primate evolution, the orthologues of the human γ genes in other mammalian species being expressed during embryonic development (see Collins and Weissman, 1984). Direct evidence has also been obtained that specific retroposons at defined locations can be transcribed in a tissue specific manner (Allan and Paul, 1984) and that certain transposon classes are associated with developmentally co-ordinated gene expression, perhaps serving as "identifier sequences" in developmental regulation (Sutcliffe et al., 1983).

1.5 The globin gene family

a) Introduction

The globins are an extremely well characterised group of related respiratory proteins found in vertebrates (myoglobin and the haemoglobins, Jeffreys et al., 1983), invertebrates (haemoglobin, haemocyanin, chlorocruorin and haemerythrin, Mill, 1972) and leguminous plants (leghaemoglobin, Lee et al., 1983). The presence of leghaemoglobin in the genome of higher plants implies the ancestral globin gene is potentially extremely ancient. The globins are pigmented proteins involved in oxygen transport, diffusion and storage and are possibly distantly related to other haem binding proteins such as cytochromes (Dickerson, 1971). In vertebrates, haemoglobin is the major constituent of the blood involved in oxygen transport while myoglobin is involved in oxygen storage and facilitated diffusion in the muscle.

The elucidation of the relationship between protein structure and function and the physiological role of myoglobin and the different developmentally regulated haemoglobins has been one of the highlights of recent advances in biochemistry and molecular biology. Recombinant DNA analysis of the globins has revealed many of the structural features and molecular phenomenon thought to be of general importance in the organisation, expression, regulation and evolution of eukaryotic genomes. The following discussion outlines some of the more recent features of the molecular biology of the globins with particular reference to the structure, organisation and evolution of this group of related genes in the eukaryotic genome (for a recent comprehensive review and references see Collins and Weissman, 1984).

b) Globin gene structure

The archetypal globin gene structure is an ancient and apparently very stable one consisting of three coding regions (exons) interrupted by two non-coding regions (introns). Mammalian β-like globin genes have intron sizes from 116-132 bp and 628-906 bp respectively while the α-like globins have two similar sized introns of 103-140 bp in length (Blanchetot et al., 1983). Initially it was suggested that this stability of intron length in mammalian α and β globin genes may refect a more general functionally imposed constraint, for example, on RNA processing (Engel and Dodgson, 1980). However exceptions have emerged, the human ζ and Ψζ genes both have introns larger than characteristic of the α-like globin genes due to the presence of tandemly repeated sequences within the introns (Proudfoot et al., 1982). A similar situation is also found in chicken π' gene (Engel *et al.*, 1983). *Xenopus* globin introns, especially those of the larval genes, are also somewhat larger than those of the higher vertebrates (Hosbach *et al.*, 1983), while mammalian myoglobin genes contain the largest introns yet described for a gene that is a member of the globin gene family (Blanchetot *et al.*, 1983; Weller *et al.*, 1984). Apart from the intron/exon boundaries there therefore appears to be little obvious selective constraint on the overall size or nucleotide sequence of globin introns, though a minimum intron size has recently been shown to be important for efficient splicing during globin mRNA maturation (Wieringa *et al.*, 1984).

The origin and maintainance of introns within eukaryotic genes remains one of the fundamental unanswered questions of eukaryotic molecular biology. Introns cannot be an absolute requirement for eukaryotic gene expression as some genes lack any introns (*eg.* interferon types α and β, and the histones) while others contain a large number of introns (*eg.* collagen, ovalbumin), the total sequence of which can greatly exceed the total coding sequence (*eg.* myoglobin and dihydrofolate reductase). Introns may even be lost without any apparent effect on gene expression (Lomedico *et al.*, 1979; Ng *et al.*, 1985)). Within the globin gene family the numerous haemoglobinopathies caused by mutations within introns illustrate that there must be some evolutionary disadvantage associated with the presence of introns (see Orkin and Kazazian, 1984), yet they have been maintained over millions of years of independent globin gene evolution, since at least the myoglobin-haemoglobin duplication some 500-800 MY ago (Blanchetot *et al.*, 1983).

By their very presence introns impose both a metabolic and genetic load on eukaryotes due to the requirement for their efficient excision during gene expression and the potential consequences of mutations within these regions on the protein encoded by a gene (see below). What advantage therefore, if any, is conferred by their continued presence ?. It has been suggested that the interrupted nature of eukaryotic genes may facilitate the evolution of new proteins (Gilbert, 1978). Blake (1978) extended this view by suggesting exons may in fact correspond to protein domains or modules (the secondary and tertiary structures encoded by various primary amino acid sequences) such that exon rearrangment may, by chance, create new gene functions and therefore confer an evolutionary advantage during evolution. The correlation between exon structure and functional protein domains is particularly striking in the immunoglobulin genes (Sakano _et_ _al_., 1979) but has also been proposed for the lysozyme gene (Artymiuk _et_ _al_., 1981) and haemoglobin gene (Go, 1981), see Blake (1983).

The correlation proposed between globin gene exon structure and protein modules (Go, 1981) has been supported by the discovery of an additional intron in the leghaemoglobin genes (Jensen _et_ _al_., 1981). This results in a four exon structure equivalent to the four protein modules deduced from the protein structure; the central two modules being fused into one exon in all vertebrate globins. While there is no evidence for recent exon exchange with other genes in the evolution of the globin genes, the specialised class switching of immunoglobulin heavy chain genes during the immune response illustrates the potential for such exon switching to generate new genes with differing functional potential (see Tonegawa, 1983).

What of the origin of introns ? One proposal is that introns are the result of insertion of transposable elements into contiguous genes during eukaryotic evolution (Darnell, 1978). This model requires the presence of primitive RNA splicing enzymes to remove the transposon from the RNA prior to translation otherwise the transposon insertion would be lethal. An alternative view is that split genes are the primitive form of gene structure predating the prokaryotic/eukaryotic divergence (Doolittle, 1978); introns have subsequently been lost from prokaryotic genes as their genomes became more "streamlined" but have been retained in eukaryotes due to some selective advantage, such as that put forward by Gilbert (1978), or due to the lack of pressure for genome streamlining and the removal of introns. The recent discovery of an intron in a tRNA gene of an archaebacterium (Kaine et al., 1984) and in the prokaryotic T4 coliphage thymidylate synthetase gene (Chu et al., 1984) support the second view, that is, the common ancestor, or 'progenote', of prokaryotes and eukaryotes initially contained introns.

Whatever the origin of introns it is possibly not surprising to find that their presence has influenced eukaryotic genome evolution. For example, introns may play a role in concerted evolution between members of a multigene family. While the precise role of introns in this process is unclear they have been implicated both as the instigators of the conversion event, due to the presence of specific sequences in the intron (Slightom et al., 1980), and in inhibiting the progress of a conversion tract due to either a) accumulated sequence divergence between the two genes since the last conversion event, or gene duplication (Hill et al., 1984) or b) insertion of a mobile genetic element into one gene causing a region of non-homology between the genes across which a conversion tract cannot migrate (Hess et al., 1983; Schimenti and Duncan, 1983).

A further consequence of the presence of introns within genes is the generation of alternative 'cryptic' splicing signals that may disrupt normal mRNA maturation. For example, several of the human β-thalassemias result from incorrect mRNA splicing due to nucleotide substitutions that result in the use of cryptic splice sites (see Orkin and Kazazain, 1984). There are however several examples where alternate RNA splicing has been exploited as a means of generating different proteins from a single gene due to the presence, or absence, of potential coding sequence in the translated mRNA. For example, the three polyoma virus T antigen proteins are produced from a single pre-mRNA, the 'intron' of which is spliced at three different positions (see Sharp, 1979). Other examples include the Drosophila myosin gene (Rozek and Davidson, 1983), murine α-crystallin gene (King and Piatigorsky, 1983), secreted and non-secreted forms of immunoglobulins (Early et al., 1980), the different forms of the rat fibronectin gene (Tamkun et al., 1984) and the rat fibrinogen gene (Crabtree and Kant, 1982).

Recombination between intron sequences can apparently result in the intragenic duplication of regions of a gene. For example, the intragenic duplication that is thought to have gaven rise to the human Hp^2 gene apparently arose due to a non-homologous, probably random, cross-over within different introns of two Hp^1 genes (Maeda et al., 1984). Similarly, exon duplication is thought responsible for the structure of the collagen (Wozeny et al., 1981), α-fetoprotein (Eiferman et al., 1981), ovalbumin (Cochet et al., 1982), ovamucoid (Stein et al., 1980), and the immunoglobulin genes (Sakano *et al.*,1979)

Only one example of an apparently essential intron encoded function has so far been described and is found in the mitochondrial genome of certain strains of yeast. The second intron of the long form of the yeast mitochondrial apocytochrome b gene encodes a maturase protein involved in the excision of the intron thereby removing the mRNA that directed its own synthesis during maturation of the RNA (see Borst and Grivell, 1981).

While it is unclear what general role, if any, intron sequences play in the eukaryotic genome they have proved very useful in studies of the molecular evolution of related gene families in different species. As this thesis will show, non-coding DNA sequences are extremely useful in determining gene orthologies between different mammalian β-globin genes (see Chapter 3 and Discussion). Improvements in this technique (White *et al.*, 1984) have been employed during this thesis and by several other groups in the analysis of the evolutionary histories of several different mammalian β-globin gene clusters.

c) Globin gene expression

The globin genes have been studied extensively as one of the model systems of eukaryotic RNA polymerase II gene transcription, RNA maturation, translation and developmental gene regulation. While the mechanism(s) of co-ordinate developmental gene expression await a suitable animal or cellular model in which the process can be analysed, the sequence elements involved in transcription, RNA maturation, and translation of avian and mammalian globin genes have been increasingly well characterised during the course of this work (see Collins and Weissman, 1984). The importance of many of these sequences for globin gene expression in situ in the chromosome has been confirmed by analysis of the defects which give rise to many human haemoglobinopathies (see Orkin and Kazazian, 1984). The comparison of functional globin genes from different species has also shown these and additional sequences to be conserved during evolution and therefore likely to be functionally important (Hardison, 1983).

Vertebrate globin genes are expressed at different times during development, for example, the six specialised tetrameric human globins ($\zeta_2\varepsilon_2$, $\zeta_2\gamma_2$, $\alpha_2\varepsilon_2$, $\alpha_2\gamma_2$, $\alpha_2\delta_2$, and $\alpha_2\beta_2$) are produced from the eight functional human globin genes; the clustered α- and β-like globin genes being expressed in a co-ordinate developmental order. The switch in gene expression during development from foetal (γ) to adult (δ and β) globin gene expression has been particularly well studied with relation to the structure of the human β-globin cluster; however analysis of haemoglobinopathies which disrupt the normal switch during development have failed to establish a connection between primary DNA sequence and developmental regulation (see Orkin and Kazazian, 1984). Similarly, comparison of β-globin genes expressed at the same stage of development in different mammalian species, in order to find specific conserved promoter elements which determine their developmental expression, have also proved inconclusive (Hardison, 1983). It seems likely that the developmental regulation of globin gene expression will not reside entirely within the DNA sequence immediately encompassing a particular gene but that gene cluster arrangement, possibly reflected as chromatin domain structure (see

Orkin and Kazazian, 1984), and/or trans-acting regulatory factors (Papayannopoulou et al., 1984) will both play an important role in this process.

d) Vertebrate Globin gene cluster organisation

The structural organisation of the globin genes has been determined for several other species as well as in man (see Collins and Weissman, 1984). There are several features common to all these globin gene clusters even though the number of genes and the size of the gene clusters varies considerably. All the globin genes have the archetype exon/intron structure (see 1.5(b)) and are orientated in the same transcriptional direction (5'-3'). The gene clusters generally have embryonic gene(s) at the head (5' end) of the cluster and the adult gene(s) at the opposite end, the exception being the chicken β-globin gene cluster. Even the triplicated goat β-globin gene cluster consists of a basic four gene unit consisting of genes expressed early (5') then later (3') in development. The mammalian β-globin gene clusters also contain a non-processed pseudogene component. The relative position of these pseudogenes in the various mammalian gene families, 5' to the functional adult gene(s), led some to propose some functional role for these sequences (Vanin et al., 1980). However, while in man the intergenic region in which the Ψβ1 gene resides is implicated in the developmental regulation of β-globin gene expression there is no evidence for the involvement of this, or any other non-processed pseudogene sequence, in developmental regulation of the gene cluster to which they belong.

The clustered arrangement of the globin genes suggests that tandem gene duplication is an important mechanism in the generation of new specialised gene functions within related gene families during evolution. In all cases except the amphibian Xenopus (Jeffreys et al., 1980), the vertebrate α and β globin gene clusters are dispersed to different chromosomal locations in the genome. For example, the human α and β globin gene clusters are on chromosome 16 and 11 respectively. The simplest evolutionary explanation for the dispersal of the α and β globin gene clusters is that the ancient α and β globin genes were initially linked following the tandem duplication of their common ancestor. The two loci subsequently became dispersed early in vertebrate evolution, in an early reptilian ancestor of mammals and birds, while remaining linked in the amphibians (see Jeffreys et al., 1983).

Similarly, gene dispersal since the ancient tandem duplication which gave rise to the related myoglobin and haemoglobin genes would account for the presence of the human myoglobin gene on chromosome 22 (Jeffreys et al., 1984). Several hypothetical mechanisms by which dispersal could occur have been proposed; these include non-replicative transposition, chromosome translocation with the break point between the duplicated genes, and polyploidisation followed by silencing of one or other of the duplicated loci (Jeffreys and Harris, 1982). Subsequent successive duplications of the dispersed functional genes would result in contemporary gene family arrangements.

The organisation of the different mammalian β-globin genes is consistant with an evolutionary history consisting of a series of mainly adaptive tandem gene duplications from a common ancestral cluster.

However, the organisation of the contemporary mammalian β-globin gene clusters suggests their independent evolutionary histories differed substantially, with changes in gene number and cluster length implying that gene duplication and cluster contraction are not infrequent events during evolution. For example, the goat β-globin gene cluster consists of a triplicated four gene cluster (Townes et al., 1984) whereas the rabbit has a single set of four β-globin genes (Hardison, 1984).

The contemporary mammalian β-globin gene clusters are too far diverged for any detailed analysis of cluster evolution; in particularly changes in the organisation and rate of sequence evolution of the non-coding component of the genome. For example, does the contemporary cluster reflect gradual or sudden alterations in structure ?. How are alterations in developmental gene expression achieved and are intergenic non-coding DNA sequences involved ?. What, if any, constraints are there upon the organisation of the intergenic non-coding sequences of the cluster ?. Do intergenic sequences in the cluster evolve as neutral "junk" DNA, in terms of gross alterations in intergenic distance and rate of nucleotide substitution ?. Phylogenetic analysis of the primate β-globin gene cluster has therefore been undertaken in order to help establish the tempo of cluster evolution and possible molecular phenomenon which have contributed to the evolutionary history of the contemporary human β-globin gene cluster (see 1.7).

1.6 The primate phylogeny

There are 183 species of contemporary primates, including man, that form the six generally accepted groups of living primates. These are

(1) the tarsiers, of which there are three species found in South East Asia; (2) the lemurs of Madagascar of which there are 28 species; (3) the loris group consisting of 10 species found in Africa and India; (4) the New World monkeys of Central and South America with 77 species; (5) the Owl World monkeys of Africa and Asia with 51 species and (6) the Hominoids, which includes the great apes (13 species) found in Africa and Asia and Man, whose domain encompasses the whole globe.

The lemur and loris groups taxonomically comprise the primate suborder prosimii while the remaining groups (including the tarsier group) are all part of the suborder anthropoidea. The following simplified classification is adapted from Kavanagh (1983). Primate species referred to and used in the course of this thesis are shown in the brackets after the classification.

```
Order: Primates
 Suborder: Strepsirhini (prosimians).
  Superfamily: Lemuroidea.
   Family: 1.  Cheirogaleidae         (dwarf lemur)
           2.  Lemuridae              (brown lemur, ruffed lemur)
           3.  Lepilemuridae
           4.  Indriidae
  Superfamily: Daubentonioidea.
   Family: 1.  Daubentoniidae
  Superfamily: Lorisoidea.
   Family: 1.  Lorisidae
 Suborder: Haplorhini (anthropoids).
  Superfamily: Tarsioidea
   Family: 1.  Tarsiidae
  Superfamily: Ceboidea.  (Plattyrhini, New World monkeys)
   Family: 1.  Callitrichidae         (marmosets and tamarins)
           2.  Callimiconidae
           3.  Cebidae                (squirrel and owl monkey)
  Superfamily: Cercopithecidea.  (Catarrhini, Old World monkeys)
   Family: Cercopithecidae
    Subfamily: 1.  Colobinae
               2.  Cercopithecinae    (yellow baboon)
  Superfamily: Hominoidea.  (apes and man)
   Family: 1.  Hylobatidae
           2.  Pongidae (gorilla, chimpanzee and orang-utan)
           3.  Hominidae              (man)
```

While the division of the primates into six major groups is generally accepted it is often difficult discriminating between different phylogenetic relationships within the groups using the sparse and incomplete fossil record (see Gingerich and Schoeninger, 1977). Even between the major groups divergence times derived from the fossil data tend to cover large evolutionary periods and are constantly being revised in the light of new fossil data and more recently from the analysis of biochemical and DNA sequence evolution in the primates (see Wilson et al., 1977; Sibley and Ahlquist, 1984). As well as the fossil record primate relationships have been reconstructed from morphological (Andrews and Cronin, 1982; Eaglen, 1983), karyotypic (Yunis and Prakash, 1982), DNA-DNA hybridisation (Sibley and Ahlquist, 1984), and immunological data (Sarich and Wilson, 1967), and also by the analysis of protein sequences (Wilson et al., 1977), mitochondrial DNA sequences (Brown et al., 1982), rDNA sequences (Wilson et al., 1984) and repeated DNA families (Gillespie, 1977). However, biochemical and DNA sequence comparisons also depend to some extent on some accurate divergence times from the fossil record, in order to calibrate observed differences against an evolutionary time scale, and therefore these methods are also subject to some degree of ambiquity.

The divergence times employed in this thesis are based on several criteria (including fossil and biochemical data) as discussed previously by Barrie (1982). This places the basal primate radiation, that is, the separation of the prosimains from the simians, at about 52-70 MY ago. This was followed by the divergence of the New World monkey lineage from that which gave rise to the Old World monkeys, great apes and man some 35-50 MY ago. The subsequent divergence of the Old World monkey lineage from the hominoids is thought to have occurred 20-30 MY ago. Finally, the lineages leading to the great apes and man probably separated some 7 MYs ago.

1.7 Phylogenetic analysis of the primate β-globin gene cluster

The organisation of the β-globin gene cluster has previously been characterised by restriction endonuclease mapping and the cross-hybridisation of human and rabbit β-globin gene probes to genomic DNA from representative species from each of the major primate groups (Barrie _et_ _al_., 1981; Barrie, 1982). This analysis shows that the β-globin gene cluster has evolved in a discontinuous manner with long periods of stable organisation. The human, great ape (gorilla and chimpanzee) and Old World monkey (yellow baboon) β-globin gene clusters are indistinguishable in functional gene number and the organisation of intergenic DNA. In contrast, while the β-globin gene clusters of the New World monkey (owl monkey) and prosimians (brown and ruffed lemur) also contain ε-, γ- and β-related globin genes, in a similar gene order and orientation 5'-3', there is limited similarity in intergenic distances when compared to the higher primates (see Discussion). The observed stability in overall gene cluster organisation, particularly in the higher primates, seems incompatable with the continuous gross rearrangements of intergenic DNA that might be expected of non-functional "junk" DNA sequences (Barrie _et_ _al_., 1981).

The rate of sequence divergence within the primate β-globin gene cluster (1 x 10^{-9} nuc.sub./site/yr, estimated from restriction endonuclease site variation, Barrie et al., 1981) is much lower than that currently accepted for non-functional DNA (5 x 10^{-9} nuc.sub/site/yr; estimated from rates of silent site substitution within functional genes, and in all positions of a silent pseudogene, and taken as an approximation of the rate of neutral evolution, see Kimura, 1983b). The overall stability and apparently reduced rate of intergenic sequence divergence suggests that this arrangement may be functionally significant and that the cluster may be evolving as a complete unit, not as individual genes within non-coding "junk" DNA. Alternatively, the intrinsic rate of non-coding sequence evolution may be much lower in the primates (see Discussion). In the absence of phylogenetic analysis of other contemporary mammalian β-globin gene clusters it is not possible to say whether this pattern of evolution (long periods of organisational stasis and/or different rates of sequence evolution) is a general feature of β-globin gene cluster evolution.

The analysis of the primate β-globin gene clusters provides a direct means of testing the timing of gene duplication events deduced from accumulated amino-acid and DNA sequence divergence between the different human globins and thought to have occurred since the mammalian radiation 80 MY ago (Dayhoff, 1972; Efstratiadis et al., 1980). For example, as mentioned previously, the duplication time estimated for the human γ genes is inconsistant with the phylogenetic data due to the apparent concerted evolution of these two genes (Slightom et al., 1980). From the analysis of primate β-globin gene cluster organisation it is clear that a duplication of the ancestral γ gene prior to the divergence of the Old World monkey from the hominoids, but after the divergence of New World monkeys from their common ancestor with the higher primates, would be the simplest evolutionary explanation for the presence of two γ genes in the yellow baboon and other higher primates but not in the owl monkey, a New World monkey (Barrie et al., 1981). Further examples of the benefits of the phylogenetic approach in determining gene histories arise from the work conducted in this thesis (see Discussion).

The intergenic non-coding DNA sequences of the different mammalian β-globin gene clusters have not been as well characterised, in evolutionary terms, as the functional genes. As well as stretches of simple repetitive and unique sequence DNA that compose the majority of the different mammalian β-globin gene clusters the intergenic DNA also contains silenced supernumary gene copies (non-processed pseudogenes) and interspersed repetitive elements (transposons such as Alu and Kpn in man and B1/B2 and Bam-HI-R in mouse). These recognisable non-coding DNA features of the various gene clusters are of interest not only due to their potential influence on the evolution of the gene family but also as representatives of more general non-coding DNA sequence evolution. The analysis of the β-globin gene cluster in the primate phylogeny provides an opportunity to investigate phylogenetically the evolution of such sequences and to evaluate their possible contribution to the evolution of a multigene family.

The potential mobility of transposon elements (while of considerable interest) suggests that these sequences may have complex evolutionary histories and will not be representative of other non-coding DNA sequence evolution in gene clusters or the genome. In contrast, after silencing, the sequence evolution of non-processed pseudogenes is already thought to represent that of other non-coding DNA sequences, to be paradigms of neutral evolution (see 1.4b), and therefore ideal sequences for such an analysis.

The presence of a non-processed pseudogene ($\Psi\beta1$) in the human β-globin gene cluster (Fritsch et al., 1980) provides an ideal opportunity to investigate phylogenetically the history of such a sequence and to test the predictions of the neutral theory concerning their evolution. The intergenic distance, restriction endonuclease site map and cross-hybridisation of an adult β-globin gene probe from the rabbit suggest the region between the $^{A}\gamma$ and δ genes in great ape and Old World monkey species is very similar to that of man (Barrie et al., 1981), indicating the potential presence of a $\Psi\beta1$-related sequence in this region. As linkage over this region of the owl monkey (a New World monkey) β-globin gene cluster has not yet been established it is unclear whether the additional fragment detected by the rabbit adult β-globin gene probe in this species is also a $\Psi\beta1$-related sequence. The functional status of these potential $\Psi\beta1$-related sequences is unknown. There is apparently no $\Psi\beta1$-related sequence in the contracted brown lemur β-globin gene cluster, the four β-like globin gene sequences having previously been identified from hybridisation analysis. The history of the human $\Psi\beta1$ pseudogene may therefore reside within the primate phylogeny.

1.8 Object of research

The work in this thesis is mainly concerned with the evolutionary history of the human Ψβ1 gene, a non-processed pseudogene found in the human β-globin gene cluster between the functional $^{A}\gamma$ and δ genes.

The main questions addressed are 1) can this sequence be found in the same relative position in other primates and what is the functional status of the sequences if present ?, 2) when was this pseudogene silenced and can the initial defect be distinguished ?, 3) did this pseudogene have a functional history between the duplication and silencing events ?, 4) has the pseudogene behaved as neutral "junk" DNA after silencing and if so what has been the mode and tempo of accumulated sequence change in this non-coding DNA sequence and can it be considered representative of other non-coding DNA sequences in the primates ?, 5) what, if any, effects does the presence of a non-processed pseudogene have on the evolutionary history of a specialised gene family such as the globins ?.

Chapter 2

MATERIALS AND METHODS

1 DNA AND TISSUES

DNA was prepared (or had previously been prepared) from human blood and from liver taken from a yellow baboon (*Papio cynocephalus*), owl monkey (*Aotus trivirgatus*), squirrel monkey (*Saimiri sciureus*), red-mantled tamarin (*Saguinus illigeri*), brown lemur (*Lemur macaco* (*fulvus*) *mayottensis*), ruffed lemur (*Lemur variegatus*), lion (*Panthera leo*), dog (*Canis familiaris*), blackbuck (*Antilope cervicapra*) and flying fox (*Pteropus lastat*). Other DNAs were prepared from grey seal (*Halichoerus grypus*) muscle and from cow (*Bos taurus*) thymus. DNA from the whole carcass of a dwarf lemur (*Cheirogaleus major*) was kindly provided by Dr M.Weiss (Wayne State University, Michigan, U.S.A). Yellow baboon (*Papio cynocephalus*), Western lowland gorilla (*Gorilla gorilla gorilla*), common chimpanzee (*Pan troglodytes verus*) and orang-utan (*Pongo pygmaeus*) blood DNA samples were generously provided by Dr A.F. Scott (John Hopkins University School of Medicine, Baltimore, U.S.A.).

Tissues from which DNAs were prepared were kindly supplied by the following people and Institutions:

Dr Ian Craig (Genetics Department, University of Oxford), lung tissue from one male western lowland gorilla (*Gorilla gorilla gorilla*). National Institute for Medical Research (London), an entire owl monkey (*Aotus trivirgatus*) corpse. Lynne Walters (Jersey Wildlife Preservation Trust), a one week old infant brown lemur (*Lemur macaco* (*fulvus*) *mayottensis*). Dr Rachel Fisher (Zoological Society of London), the livers of a ruffed lemur (*Lemur variegatus*), squirrel monkey (*Saimiri sciureus*) and a red-mantled tamarin (*Saquinus illigeri*). Mr J.Prime (British Antarctic Survey), grey seal (*Halichoerus grypus*) muscle. Miss Brancker (Twycross Zoo), dog (*Canis familiaris*) and blackbuck (*Antilope cervicapra*) tissue. Dr P.Little (St Marys Hospital Medical School, London), lion (*Panthera leo*) tissue. Dr M.Weiss (Wayne State University, Michigan, U.S.A), flying fox (*Pteropus lastat*) tissue.

2 RECOMBINANT PLASMIDS AND BACTERIOPHAGE LAMBDA

pβG1 and λHYG4 DNAs were gifts from Prof. C.Weissman (University of Zurich, Switzerland) and Dr T.Maniatis (Harvard University, U.S.A.) respectively.

3 NONRECOMBINANT PLASMIDS, BACTERIOPHAGE AND BACTERIAL STRAINS

The lambda replacement vector λL47.1 (Loenen and Brammer, 1980) was used to make the human and owl monkey genomic libraries. Nonrecombinant plasmids pAT153 (Twigg and Sherratt, 1980) and pUC13 (Messing, 1983) were used in subcloning DNA from recombinant bacteriophage; and the M13 sequencing vectors M13mp8 and mp9 (Messing and Vieira, 1982) were used in shotgun cloning of DNAs for M13 sequencing.

E.coli bacterial strains used are listed below :

HB101 (*rec*A, *hsd*R$_k$, *hsd*M$_k$, *leu*, *thi*1, *lac*Y, *end*A, *rpsl*20, *ara*14, *gal*K2, *xyl*-5, *mtl*-1, *sup*E44, *trp*)

ED8910 (*sup*E44, *sup*F58, *rec*B21, *rec*C22, *hsd*S, *met*B, *lac*Y1, *gal*K2, *gal*T22)

JM83	(ara, Δ(lac-pro), strA, thi1, Φ80dlacI^q, ZΔM15)
JM101	(Δ(lac-pro), supE44, thi1. F'traD36, proAB, lacI^q, ZΔM15)
JM103	(Δ(lac-pro), thi1, strA, supE44, endA, sbcB15, hsdR4. F'traD36, proAB, lacI^q,ZΔM15)

4 ENZYMES, ANTIBIOTICS, CHEMICALS AND REAGENTS

Unless otherwise stated all restriction enzymes were obtained from Bethesda Research Laboratories Inc, Rockville, Maryland, U.S.A., as were deoxyribonucleoside triphosphates and M13mp8 and mp9 RF DNA. Bovine pancreatic ribonuclease A, lysozyme, dextran sulphate (sodium salt), dithiothreitol, spermidine trichloride, Ficoll 400, salmon sperm DNA (sodium salt), bovine serum albumin, dimethyldichlorosilane, piperidine, isopropyl-β-D-galactopyranoside(IPTG), ampicillin (sodium salt), and N,N,N',N'-tetramethylethylenediamine (TEMED) were obtained from Sigma, London, England. Hydrazine was from Pierce Chemical Company, Rockford, Illinois, U.S.A. Proteinase K, DNA Polymerase I (large fragment,Klenow enzyme), calf intestinal phosphatase and dideoxyribonucleoside triphosphates were from Boehringer Corporation, London, England. DNase I came from Worthington Biochemical Corporation, Freehold, New Jersey, U.S.A. Restriction endonuclease MboII and RsaI, T4-polynucleotide kinase and T4-DNA ligase were obtained from New England Biolabs, Beverly, Massachusetts, U.S.A. Avian myeloblastosis virus reverse transcriptase was from Dr J.W.Beard, Life Sciences Incorporated, St Petersburg, Florida, U.S.A. Polyvinylpyrrolidine and phenol (AR) were from Fisons, Loughborough, England; acrylamide from Unisciences Ltd, London, England; Bisacrylamide Bio-rad Laboratories Ltd, Watford, England, and dimethylsulphate was obtained from Aldrich Co Ltd, Gillingham, England. Agarose was from F.M.C. Corporation, Rockland, Maine, U.S.A.; Streptomycin sulphate from Glaxo Laboratories; E.coli DNA polymerase I (for nick-translations) plus all radionucleotides were obtained from The Radiochemical Centre, Amersham, England. Bachem Inc, Torrance, California, U.S.A. supplied 5-Bromo-4-chloro-3-indolyl-β-D-galactopyranoside (Xgal). The 17mer primer (Duckworth et al. 1981) used for M13 sequencing was kindly provided by Dr P.Meacock (Biocentre, Leicester)

All other chemicals were analytical grade.

5 MEDIA AND GENERAL DNA HANDLING TECHNIQUES

i) Media

The following liquid and solid media were used:

Luria Broth (10g Difco Bacto Tryptone, 5g Difco Bacto Yeast Extract, 5g NaCl per litre of distilled water). Luria agar plates were prepared by solidifying liquid media with 15g Difco bacto-agar per litre; 6g agar per litre was used to prepare soft agar overlays.

BBL Agar, used for phage assays and initial growth, contained 10g Trypticase peptone (Becton Dickerson and Company), 5g NaCl and 5g of $MgSO_4$ per litre of distilled water solidified with 15g or 6g of agar as above; soft agar overlays were supplemented with 10mM $MgCl_2$.

Glucose supplemented minimal medium plates, used to maintain E.coli strains JM101 and JM103, contained per litre; 500mls of 4% BBL agar plus 500mls of M56 salts (61.1 ml 0.5M Na_2HPO_4, 19.3ml 1M KH_2PO_4, 0.5ml 10% $(NH_4)_2SO_4$, 1ml 0.05% $FeSO_4$, 1ml 10% $MgSO_4.7H_2O$, 0.5ml 1% $Ca(NO_3)_2$

supplemented with 0.2ml of a 0.1% B_1 (thiamine) solution and 10ml of a 20% glucose solution.

ii) Phenol Extraction

DNA solutions were mixed with 0.5 vol of phenol: chloroform: isoamyl alcohol: 8-hydroxyquinoline (100:100:4:0.1,w:v:v:w) saturated with 10mM Tris-HCl,pH 7.5 and briefly centrifuged to separate the phases. The upper aqueous phase containing the DNA was removed and the phenol layer re-extracted with an equal volume of 10mM Tris-HCl,pH 7.5. The phenol used was AR grade and not redistilled.

iii) Ethanol Extraction

DNA was precipitated from solution by the addition of 0.1 vol of 2M sodium acetate,pH 5.6, and 2 volumes of ethanol. After mixing, the solution was chilled for 10 minutes in an I.M.S/dry-ice bath (I.M.S= Industrial Methylated Spirits). DNA precipitates were pelleted by centrification at 16300xg, -15°C for 10 minutes (or in a bench Eppendorf centrifuge for 5 minutes at maximum speed), the supernatent discarded and the pellet rinsed with 70% ethanol, centrifuged for 2 minutes and the 70% ethanol removed. DNA pellets were then vacuum dried and resuspended in an appropriate solution for further manipulation.

iv) Methoxyethanol/phosphate Extraction to removal carbohydrates

This method was adapted from Kirby, 1957.

DNA solutions were mixed with equal volumes of 2.5M potassium phosphate,pH 8.0, and 2-methoxyethanol and the resulting turbid solution centrifuged. The upper aqueous phase was then precipitated by the addition of 2 volumes of I.M.S and the cloudy liquid and 'oily beads' poured off. The DNA was washed in 70% alcohol, redissolved in 10mM

Tris-HCl,pH 7.5 and dialysed o/n against 2mM Tris-HCl, 0.1mM EDTA,pH 7.5 at 4°C.

v) Butanol Concentration

DNA solutions were concentrated by extraction of water with butan-2-ol. Solutions were mixed with butan-2-ol and briefly centrifuged. The top phase was discarded and the lower phase extracted 3 times with diethyl ether to remove remaining butan-2-ol. Traces of diethyl ether were removed by gentle passage of air over the solution.

6 AUTORADIOGRAPHY AND PHOTOGRAPHY

Hybridisation filters and sequencing gels were autoradiographed using Kodak X-ray film (35x40 or 13x18cm X-Omat RP). Exposure times depended upon the dpm (disintegrations per minute) detected using a hand-held mini-monitor (Mini-Instruments Ltd, g-m monitor type 5.10). If an intensifying screen was required then exposures were performed in a -80°C freezer. In the absence of an intensifying screen they were performed at room temperature. Autoradiographs were photographed using a Nikon F camera with orange filter and Kodak AHU 35mm microfilm.

DNA in agarose gels was visualised by bound ethidium bromide fluoresence using a short wavelength ultraviolet transilluminator (Ultra-violet Products Inc, California, U.S.A) and photographed using a Polaroid MP-3 land camera and Polaroid 4x5 type-55 or type-57 film. Agarose gels to be autoradiographed were first dried down using a commercial hair drier before exposure to the X-ray film, usually overnight, in the absence of an intensifying screen.

Autoradiographic X-ray film was developed by immersion in developer (Kodak DX80) for 5 minutes followed by a rinse in water (containing a splash of acetic acid) then 5 minutes immersion in fixer (Kodak FX40 plus HX-40 hardener). After rinsing in tap water developed autoradiographs were dried at room temperature or in a 37°C hot room.

35mm photographs were developed according to the manufacturers recommendations using Kodak D19 developer and May and Baker "Amfix" fixer.

7 DNA PREPARATION

a) Genomic DNA

This method was based upon that of Kirby(1957). Upto 5g of tissue, previously stored at -80°C, was frozen in liquid nitrogen then pulverised before being homogenised in a DuPont Instruments "Sorvall" Omni-mixer containing 5 volumes of ice-cold S.E. buffer (S.E.= 150mM NaCl,100mM Na_2EDTA,pH 8.0) The cells were lysed by adding 0.1 vol of an ice-cold 10% S.D.S solution (S.D.S= sodium dodecyl sulphate) and left on ice for 5 minutes. The homogenate was then phenol extracted once (with gentle shaking to avoid shearing the high molecular weight DNA) without reextracting the phenol, and centrifuged at 16300xg, 4°C for 5 minutes to separate the two phases. The upper aqueous phase was then decanted off and the nucleic acids, remaining proteins and the carbohydrates precipitated by the addition of 2 volumes of I.M.S (I.M.S= Industrial Methylated Spirits). The flocculent precipitate formed was removed and washed in a 70% I.M.S solution then redissolved in a small volume of 0.1 x T.N.E (T.N.E= 50mM Tris-HCl,5mM Na_2 EDTA,100mM NaCl,pH 7.5) at 0-4°C. This solution was then incubated for 15 minutes at 37°C with 100µg/ml of pancreatic RNaseA (from a 20mg/ml in 0.15M NaCl stock solution which had previously been incubated at 80°C for 15 minutes to inactivate any DNases). The DNA solution was then proteinased (after addition of 0.1 vol 10% S.D.S and 0.05 vol 20 x TNE,pH 8.0) by incubation at 37°C for 15 minutes with 100μg/ml of proteinase K. The DNA solution was then phenol extracted once and precipitated in the presence of 0.1 vol of 2M sodium acetate,pH 5.6 by mixing with 2 volumes of ethanol (without chilling or centrifugation). The precipitate was washed with 70% ethanol then redissolved in 10mM Tris-HCl,pH 7.5 and methoxyethanol/phosphate extracted to remove carbohydrates (see 4(v)). The DNA was recovered from the upper aqueous phase by careful swirling after addition of 2 volumes of ethanol. The precipitate was washed 2-3x in 70% ethanol, redissolved in 10mM Tris-HCl,1mM Na_2EDTA,pH 7.5 and dialysed overnight at 4°C against 2 litres of 2mM Tris-HCl,0.1mM Na_2EDTA,pH 7.5 with one change.

DNA concentrations were determined by measurement of optical density at 260nm on a Cecil Instruments CE 202 ultraviolet spectrophotometer with CE 235 Micro-sipette control attachment. An OD of 20 at 260nm is equal to 1mg/ml of double stranded DNA.

The quality of the DNA preparation was determined by horizontal agarose gel electrophoresis of native (0.5μg) and denatured (2μg,see 8b) DNA samples were run against bacteriophage lambda marker DNA digested with restriction endonucleases _HindIII_ or _EcoRI_ in the native or denatured form. Molecular weights of these DNA markers were taken from Daniel _et al._ (1980). The DNA yielded by this method had a single stranded size of not less than 15kb (kb=1000 bp) and a double stranded size of greater than 23kb.

b) Plasmid DNA

i) Small scale plasmid preparations

This method is a modification of the small scale alkaline extraction method of Birnboim and Doly, (1979).

A 5ml culture of the plasmid containing strain of E.coli was grown to stationary phase overnight in Luria broth supplemented with a suitable selective antibiotic. 1.5 mls of culture were pelleted by a 15 second spin in an Eppendorf bench centrifuge. The cells were resuspended in 100µl of ice-cold lysis solution (25mM Tris-HCl,10mM Na_2EDTA,50mM sucrose,pH 8.0 containing freshly added lysozyme at a final concentration of 1mg/ml). This solution was left on ice for 10 minutes then 2 volumes of ice-cold alkaline/S.D.S (0.2M sodium hydroxide,1% sodium dodecyl sulphate) were added and the solution mixed (to disrupt the cells) and left on ice for a further 5 minutes. Chromosomal DNA and most of the proteins were precipitated by the addition and mixing of 150µl of ice-cold potassium acetate,pH 5.2 while on ice for 10 minutes. After a 5 minute spin in an Eppendorf centrifuge the supernatant was removed avoiding the precipitated material. Plasmid DNA was then precipitated by mixing 2 volumes of ice-cold ethanol with the supernatant and immersion in a -80°C I.M.S/dry-ice bath for 5 minutes followed by a 2 minute spin in an Eppendorf centrifuge. The DNA was redissolved in 0.2M sodium acetate,pH 5.6 and precipitated as before, rinsed in 70% ethanol and finally redissolved in 40µl of H_2O.

Approximately 5-10µg of plasmid DNA can be prepared by this method. The preparation contains very little E.coli chromosomal DNA or cellular protein, but does contain large amounts of RNA. The plasmid DNA can be used directly in restriction endonuclease digests, followed by RNase treatment to remove RNA, to confirm the presence of a required DNA insert within a plasmid.

ii) Large scale plasmid preparation

This is essentially a scaled up version of the previous method with the additional caesium chloride gradient centrifugation purification step.

The E.coli strain containing the plasmid was grown, with shaking, overnight at 37°C in 5mls of Luria broth containing 20µg/ml thymine plus suitable selective antibiotic. 1ml of this culture was used to inoculate 2 x 400ml of the same medium and grown, with shaking, overnight at 37°C. The cells were pelleted by centrifugation at 4200xg, 4°C for 5 minutes, the supernatant discarded and the cells resuspended in 40mls of ice-cold lyses solution (containing 2mg/ml of freshly added lysozyme) and kept on ice for 5 minutes. 80mls of ice-cold alkaline/S.D.S were then added, mixed, and left on ice a further 5 minutes. To this solution 60mls of ice-cold 3M potassium acetate,pH 5.2 was added and mixed to precipitate chromosomal DNA and proteins. After centrifugation at 6000xg,4°C for 10 minutes the supernatant was decanted from the precipitate through a polyallomer wool plug in a glass funnel. The nucleic acids were then precipitated by addition of 0.5 volumes of propan-2-ol followed by centrifugation at 4200xg, 4°C for 10 minutes. The pellet was gently rinsed with a small volume of 70% ethanol, followed by diethyl ether, blown dry and redissolved in 10mls of TE buffer (TE=10mM Tris-HCl,1mM Na_2EDTA,pH 7.5). The volume was adjusted gravimetrically to 20mls with more TE to which was then added 4mls of a 5mg/ml ethidium bromide solution plus 23.76gms of AR grade caesium chloride, to give a final density of ρ=1.392-1.394. This solution was split between two Beckman polyallomer 5/8 x 3in "Quickseal" tubes and centrifuged for either a) 110000xg, 15°C for 40 hours in a 50Ti rotor or b) 270000xg, 15°C overnight in a 75Ti rotor. The plasmid band was removed from the caesium chloride gradient using a 5ml disposable syringe fitted with polyvinyl flexible tubing. Ethidium bromide was removed by repeated extraction with propan-2-ol saturated with caesium chloride/H_2O. Plasmid DNA was then precipitated by the addition and mixing of 2 volumes of H_2O plus 2 (new) volumes of ethanol followed by centrifugation at 16300, 0°C for 10 minutes. The precipitate was washed with 70% ethanol, vacuum dried, then redissolved in 500µl of 10mM Tris-HCl,pH 7.5.

The concentration and quality of the plasmid preparation were determined by optical density at 260nm and agarose gel electrophoresis as mentioned previously.

c) Denatured Salmon Sperm DNA

1g of salmon sperm DNA (Sigma, TypeIII) plus 20mls of 0.5M Na_2EDTA was added to 500mls of H_2O. The DNA was dissolved while immersed in a boiling waterbath then 15mls of 10M NaOH were stirred into the solution. After checking the alkalinity the solution was returned to the boiling waterbath for a further 20 minutes, allowed to cool on ice, then 20mls of 1M Tris-HCl,pH 7.5 added. The DNA solution was then adjusted, with stirring, to pH 7-8 with concentrated HCl. The denatured DNA was phenol extracted, I.M.S precipitated, rinsed in 70% ethanol, and the last traces of ethanol allowed to evaporate off in a fume cupboard overnight. The DNA was finally redissolved in 50mls of H_2O and the concentration determined via optical density at 260nm.

d) High Molecular Weight Salmon Sperm DNA

200mg of salmon sperm DNA (Sigma, TypeIII) were dissolved in 200mls of 10mM Tris-HCl,pH 7.5 overnight at 4°C. The DNA was then phenol extracted, ethanol precipitated, vacuum dried, and redissolved in 70mls of 10mM Tris-HCl,pH 7.5. The DNA concentration was determined as before.

e) Human Competitor DNA

Sheared single stranded human competitor DNA used in hybridisations was produced from human muscle DNA by a scaled down version of (c).

8 <u>RESTRICTION ENDONUCLEASE DIGESTION</u>

DNAs at a final concentration of ≤0.5mg/ml were incubated in the manufacturers recommended buffer at 37°C for 1 hour unless otherwise stated. Spermidine trichloride was routinely added to a final concentration of 4mM as this is known to enhance the efficiency of many restriction endonucleases, especially if the DNA has previously been recovered from an agarose gel (Bouche, 1981). Complete digestion was checked by running an aliquot equivalent to 0.5μg of the DNA digested on a suitable horizontal agarose gel against marker DNAs of known molecular weight. If the digest was incomplete more restriction endonuclease was added, incubated for a further hour and another aliquot tested. After complete digestion Na_2EDTA was added to a final concentration of 20mM and the DNA phenol extracted, ethanol precipitated twice, vacuum dried and redissolved in 10mM Tris-HCl,pH 7.5. The DNA was ready for further manipulation in this solution.

9 AGAROSE GEL ELECTROPHORESIS

a) Test Gels

Horizontal agarose gels with 3-7mm loading slots were prepared and run in 40mM Tris-acetate, 2mM Na_2EDTA,pH 7.7 buffer containing ethidium bromide at 0.5µg/ml (Aaij and Borst, 1972). The gel size varied with the number of samples to be run from 5x7cm (minigels) to 20x20cm (mapping gels). The concentration of the agarose varied between 0.5%-2% (w/v) according to the anticipated molecular weight of the DNA sample(s) which were run against a suitable set of known molecular weight marker DNAs; either lambda x _Hin_dIII or pBR322 x _Sau_3A or both. DNA samples were mixed with 0.5-1.0 vol of a 0.2% suspension of agarose beads in 20mM EDTA containing 10% glycerol and a small amount of bromophenol blue dye as an electrophoresis marker; this suspension was prepared as described by Schaffner _et al._, (1976). Gels were run at room temperature at 120V for 1-2 hours, or overnight at 15-20V, until the marker dye had run approximately 2/3 the length of the gel.

b) Mapping Gels

Single standed DNA samples were run on 0.8% horizontal agarose gels with 5-7mm loading slots. DNA was denatured to the single stranded form by the addition of 0.1 vol of 1.5M NaOH,0.1M Na_2EDTA 5 minutes before loading. After electrophoresis the DNA was transferred directly to nitrocellulose by a modification of the method of Southern (1975), as described by Barrie (1982).

Native double stranded DNA samples were run on 0.5% horizontal agarose gels with 5-7mm loading slots, the DNA was denatured by acid/alkaline treatment (see 11) before transfer to nitrocellulose as above.

Gels to be transferred to nitrocellulose were electrophoresed until the dye front was 8cm or 15cm from the loading slots. The distance travelled from the loading slots depended on the length of hybridisation chambers to be used and the degree of resolution that was desired (the further the distance travelled the better the resolution between different sized fragments).

c) Preparative Gels

Preparation of samples, gel loading and electrophoresis were as described by Jeffreys et al. (1980). The amount of native DNA loaded was adjusted to ≤0.5μg/mm^2 of gel slot surface area to avoid overloading. Three different procedures were used to recover DNA from gels (see below).

10 RECOVERY OF DNA FROM AGAROSE GELS

i) DNA was electrophoresed onto a vertical dialysis membrane inserted into a slot cut into the gel as described by Yang et al. (1979). The DNA was rinsed off the membrane with sterile water.

ii) Gel slices were inserted into a sealed dialysis membrane bag, together with a small amount of electrophoresis buffer, and placed in shallow buffer such that the DNA could be electroeluted out of the gel slice into the bag (Smith, 1980).

DNA was recovered from solution in methods (i) and (ii) as follows: traces of agarose were removed by a 2 minute spin in a bench top M.S.E centrifuge at maximum speed through a polyallomer wool column. The solution volume was reduced by several rounds of butanol concentration then phenol extracted, ethanol precipitated twice, vacuum dried, and redissolved in 10mM Tris-HCl,pH 7.5.

iii) DNA was electrophoresed onto Whatman DE81 DEAE-cellulose paper and recovered by a modification of the method of Dretzen et al. (1981).

DE81 paper was presoaking in 2.5M NaCl for 15 minutes, rinsed 3 x 5 minutes in H_2O and stored in 1mM Na_2EDTA or used immediately. DNA was electrophoresed onto DE81 paper as follows; the DNA was visualised under a UV source and a slot cut immediatly in front and behind the desired DNA fragment(s). A strip of DE81 paper the correct size was placed in the slots, the current switched on and the DNA electrophoresed onto the paper. The DE81 paper "behind" the desired DNA fragment stopped contamination by any higher molecular weight DNA. DE81 paper onto which the DNA was electrophoresed was rinsed 3 x 5 minutes with sterile water and blotted dry on Whatman 3MM paper. The DE81 paper was placed in an Eppendorf tube and shredded by vortexing in high salt buffer (1M NaCl,50mM Tris-HCl,1mM Na_2EDTA,pH 7.5) then incubated for 1 hour at 37°C to release the DNA from the DE81 paper. Alternatively the incubation was shortened to 10 minutes at 37°C followed by 10 minutes at 65°C with no detectable loss in quantity or quality of recovered DNA. The DNA solution was separated from the DE81 paper by centrifugation through a small polyallomer wool column in a bench top M.S.E centrifuge at maximum speed for a few minutes. Small remaining traces of DE81 paper were removed by a further 5 minute centrifugation in an Eppendorf centrifuge and transfer of the supernatant to another tube avoiding any DE81 pellet. The DNA was then ethanol precipitated twice, rinsed with 70% ethanol, vacuum dried,and redissolved in 10mM Tris-HCl,pH 7.5 ready for further manipulation.

11 ACID/ALKALINE DENATURATION OF AGAROSE GELS

Mapping gels, with DNA run in the native double stranded form, were removed from the electrophoresis equipment, photographed, then soaked twice in 0.25M HCl for 15 minutes to reduce the size of the DNA by depurination. The gel was then neutralised and the DNA denatured by soaking twice in 0.5M NaOH,1M NaCl for 15 minutes each. A brief wash with water was followed by two 15 minute washes in 0.5M Tris-Hcl,3M NaCl,pH 7.5. DNA transfer to Sartorius nitrocellulose (0.45µ pore size) was performed by a modification of the method of Southern (1975) exactly as described by Barrie (1982).

12 ^{32}P-LABELLING OF DNA PROBES BY "NICK-TRANSLATION"

The method used to label DNA probes was essentially that of Weller et al. (1984); 50-100ng of DNA in 5µl of sterile H_2O were heated at 100°C for 3 minutes, chilled on ice, then added to the following reaction mixture

2.5µl 10x nick mix (500mM Tris-HCl,pH 7.5, 50mM $MgCl_2$, 100mM 2-mercaptoethanol)
2µl each of 50µM dGTP,dATP,and dTTP
1µl 0.1M spermidine
1µl 8ng/ml DNase I (freshly diluted from a 1mg/ml stock in H_2O)
1.5µl α-^{32}P-dCTP (10µCi/µl, ~3000Ci/mMol)
5 units of E.coli DNA polymerase I
H_2O up to 25µl

The solution was mixed then incubated at 15°C for 90 minutes. Samples may be removed to check incorporation and quality of the translation at this point via a) mobility on an agarose gel and autoradiography and/or b) DE81 binding (Maniatis et al. 1982). ≥50% incorporation was observed even with impure substrates which did not normally label very well. The reason for this increased incorporation over other methods is unknown but DNA labelled by this method behaves indistinguisably in filter hybridisations from DNA labelled by other methods. The reaction was stopped by the addition of 25μl of 0.5% S.D.S,12.5mM Na_2EDTA,10mM Tris-HCl,pH 7.5, phenol extracted and the aqueous phase recovered. 100μg of high molecular weight salmon sperm DNA was added as carrier and the DNA precipitated by the addition of 0.1 vol of sodium acetate and 2 volumes of ethanol (without chilling or centrifugation). The aqueous phase was removed and the DNA redissolved in 0.5ml of Tris-HCl,pH 7.5 before being precipitated as before. The DNA was rinsed with 70% ethanol before being redissolved in 500μl of 10mM Tris-HCl,pH 7.5. Specific activities of 10^7-10^8 dpm/μg were generally achieved.

13 HYBRIDISATIONS

Filter hybridisations were carried out as described in detail by Barrie (1982). DNA was bound to nitrocellulose after transfer from agarose gels by baking in an oven at 80°C for 2-5 hours. Filters were then cut into strips and prehybridised for 30 minutes in a hybridisation box in a gently rocking waterbath at the appropriate hybridisation temperature in the following changes of degassed solution;

3 x SSC	(not degassed)
1 x Denhardts	(0.2% ficoll, 0.2% polyvinylpyrrolidone 0.2% bovine serum albumin in 3 x SSC)
1 x CFHM	(1 x Denhardts plus denatured salmon sperm DNA at 50μg/ml and 0.1% S.D.S)
1 x CFHM	(+/- 9%(w/v) dextran sulphate, +/- competitor DNA 30-50μg/ml)

The final hybridisation solution contained ^{32}P-labelled DNA at no more than 10~~μ~~ng/ml which had been denatured by heating to 100°C for 5 minutes prior to adding to the hybridisation solution. The presence of dextran sulphate in the hybridisation solution greatly increases the hybridisation kinetics. Where used competitor DNA was added to the labelled probe DNA before the denaturation step in order to reduce any background hybridisation of repetitive DNA sequences present in the probe to similar sequences in the DNA on the filters. Genomic DNA filters were hybridised overnight in the presence of dextran sulphate while lambda recombinant DNA filters were either hybridised 3-5 hours in the presence of dextran sulphate or overnight without dextran sulphate.

Unbound labelled DNA was washed off the filters by the following changes of solution (preheated to the hybridisation temperature);

1 x CFHM	(repeated changes till unbound ^{32}P-labelled DNA present in the wash solution after each change was at a low level)
1 x CFHM	(2-3 changes of 15 minutes each)
Final wash	(60 minutes, 2x 30 minutes with CFHM at desired SSC concentration)
3xSSC	(quick rinse)

The filters were blotted dry on Whatman 3MM paper, allowed to dry completely at room temperature, reconstructed and then autoradiographed for 1-14 days (see 5). Filters to be hybridised with other probes were washed repeatedly in H_2O at 65°C to remove the previously hybridised ^{32}P-labelled DNA.

14 GENOMIC LIBRARIES

The human and owl monkey genomic libraries were prepared by P.Weller and Dr P.A.Barrie respectively; detailed protocols for the methods involved are presented by Barrie (1982).

Genomic high molecular weight DNA was partially digested with the restriction endonuclease *Sau*3A, recovered after phenol extraction by ethanol precipitation, vacuum dried and redissolved in H_2O. Size selection between 11-23kb was achieved using a preparative agarose gel and electrophoresis of DNA in this size bracket onto a dialysis membrane. The DNA was recovered off the membrane as described before, 2.10(i).

λL47.1 vector arms were prepared by digestion of λL47.1 DNA by the restriction endonuclease *Bam*HI. This enzyme was used because the "sticky" ends produced allow ligation to the complementary ends produced by *Sau*3A digestion of genomic DNA. The right and left arms of λL47.1 were separated from the internal "inessential" region on a preparative agarose gel, electrophoresed onto dialysis membrane and recovered as before.

The genomic *Sau*3A partials and λL47.1 arms were then ligated together, *in vitro* packaged, transfected into the *E.coli* strain ED8910 and plated out on BBL agar plates (supplemented with 10mM $Mg\overset{SO_4}{\cancel{Cl_2}}$ and 0.2% maltose) exactly as described by Barrie (1982).

15 SCREENING OF GENOMIC LIBRARIES

The method used was a modification of that of Benton and Davis, (1977). After phage growth the agar was hardened by cooling at 4°C for 15 minutes. Nitrocellulose filters (88mm diameter, Schleicher and Schüll BA 85/20) were placed directly onto the agar. Phage were allowed to transfer onto the filters for 5 minutes while the filters and plates were being uniquely marked for future reference. The filters were removed and layered for 1 minute on a 0.1M NaOH,1.5M NaCl solution, neutralised by transfer on to a 1xSSC,0.2M Tris-HCl,pH 7.5 solution for another minute. Filters were then blotted dry and baked for 2-5 hours at 80°C. The filters were hybridised overnight with an appropriate ^{32}P labelled probe (in the presence of dextran sulphate). Autoradiography of the washed filters was generally for 1-4 days at -80°C with an intensifying screen.

16 PURIFICATION OF POSITIVE RECOMBINANT CLONES AND PHAGE AMPLIFICATION

Positively hybridising recombinant plaques or regions were picked and resuspended in 0.5ml of phage buffer (6mM Tris-HCl,10mM $MgSO_4.7H_2O$,0.005% gelatin,pH 7.5), serially diluted in the same, and 100µl of each dilution allowed to absorb to 100µl of ED8910 cells in Luria broth (supplemented with 10mM $MgSO_4$) for 20-30 minutes at room temperature. Cells and phage were layered onto BBL agar plates in 3ml of BBL top agar containing 20µg/ml thymine and 10mM $MgSO_4$ and grown overnight at 37°C. This purification procedure was repeated three times until a single plate contained only positively hybridising plaques. Single plaques were then picked from such a plate for phage amplification and storage at 4°C exactly as described by Barrie (1982).

17 LARGE SCALE RECOMBINANT LAMBDA PHAGE PREPARATION

Phage lysates were prepared by a modification of the method of Blattner _et_ _al_., (1977).

An aliquot of recombinant phage from an amplified phage stock solution was serially diluted in phage buffer. 100µl of each dilution was added to 100µl of an ED8910 overnight previously diluted 1/10 in Luria broth (containing 10mM $MgSO_4$) and allowed to absorb for 30 minutes. Each solution was layered onto a Luria agar plate in 3ml of Luria top agar (containing 10mM$MgSO_4$) then grown overnight at 37°C. 3-5 well separated phage plaques were picked, together with surrounding bacteria, as 1.5-3mm diameter plugs and used to inoculate 200ml of Luria broth (containing 20µg/ml thymine and 10mM $MgSO_4$) in a 2l unbaffled flask. After growth overnight at 37°C, with gentle shaking, successful phage growth was indicated by the presence of cellular debris in an otherwise almost clear solution. Chloroform was added to the culture to 0.5%(v/v) and then left to stand for 10 minutes to lyse the remaining cells. Lysates were cleared by centrifugation at 13000xg, 4°C for 10 minutes and the phage then harvested by centrifugation at 160000xg, 4°C for 1 hour. Phage pellets were resuspended in 1.2 mls of lambda buffer, by gentle shaking overnight, then cleared again by centrifugation at 16000xg, 4°C for 10 minutes. The supernatant (in a total volume of 4.8ml of lambda buffer) was layered onto the top of a caesium chloride block gradient composed of 2ml at density ρ= 1.7, 3ml at ρ=1.5, and 2ml at ρ=1.3 followed by centrifuged at 220000xg, 20°C for 1 hour. Each CsCl solution was diluted in phage buffer to the correct density from a 65%(w/v) stock solution made up in water. The phage band in the ρ=1.5 region was removed and dialysed overnight against 10mM Tris-HCl,1mM Na_2EDTA,pH 7.5 to remove CsCl. While still in the dialysis bag the DNA was RNased by addition of heat treated pancreatic RNase A to 20µg/ml and dialysis for a further hour against 10mM Tris-HCl,1mM Na_2EDTA,pH 7.5. The DNA was then proteinased by addition of proteinase K, to 1mg/ml, whilst dialysing against 20mM Tris-HCl,pH 8.0,1mM Na_2EDTA,0.1M NaCl,0.01% Triton X-100 for 2 hours. The DNA solution was then removed from the dialysis bag, phenol extracted twice, ethanol precipitated twice (without centrifugation and cooling), and finally redissolved in 10mM Tris-HCl,pH 7.5. Upto 1mg of recombinant phage DNA could be obtained by this method.

18 MAPPING OF LAMBDA RECOMBINANTS

2-3μg of recombinant phage DNA were single and double digested with a series of restriction endonucleases (BamHI, BglII, EcoRI, and HindIII). Digests were run unrecovered on 0.5% horizontal agarose gels against marker DNAs of known molecular weight until the bromophenol blue dye had travelled ≈8cm from the loading slots. The agarose gel was photographed and then the DNA in the gel acid/alkaline denatured before transfer to Sartorius nitrocellulose membrane. Recombinant DNA fragments containing regions homologous to globin DNA probes were visualised by hybridisation of the filters overnight (usually in the absence of dextran sulphate) and autoradiography as described before.

19 PREPARATION OF PLASMID VECTOR DNA FOR SUBCLONING

20μg of plasmid DNA was digested at suitable cloning sites with the appropriate restriction endonuclease(s). The linearised DNA was recovered after phenol extraction by two ethanol precipitations, vacuum dried and redissolved in 10mM Tris-HCL,pH 7.5. 15μg of linearised plasmid DNA was then phosphatased with 0.15 units of calf intestinal phosphatase at 37°C for 1 hour. The DNA was then phenol extracted twice, ethanol precipitated, vacuum dried and redissolved in 10mM Tris-HCl,pH 7.5 ready for use.

20 PREPARATION OF LAMBDA RECOMBINANT DNA FOR SUBCLONING

2-5μg of lambda recombinant DNA were digested with a suitable restriction endonuclease, recovered after phenol extraction by two ethanol precipitations then resuspended in 10mM Tris-HCl,pH 7.5. Where a single or limited population of fragments could be isolated this was achieved by preparative horizontal agarose gel electrophoresis and DNA recovery from the gel using the DE81 paper method.

21 LIGATION OF DNA FRAGMENTS INTO PLASMID VECTORS

Fragments produced by restriction endonuclease digestion were ligated to plasmid vector DNAs in a 4:1 ratio of fragment to vector DNA. Ligations containing a total of ≤2.5µg of DNA were performed in a total volume of 25µl of ligase buffer (50mM Tris-HCl,10mM $MgCl_2$,20mM DTT,1mM ATP 50µg/ml BSA,pH 7.8) and incubated with 2µl of T4 DNA ligase (400u/µl) at 4°C overnight. Successful ligation was tested by horizontal agarose gel electrophoresis.

22 TRANSFORMATION

Transformations were performed by a modification of the method of Cohen _et al_. (1972). _E.coli_ strains JM83 and HB101 were grown in Luria broth plus streptomycin at 200µg/ml and Luria broth plus 20µg/ml thymine respectively.

The desired strain was grown, with gentle shaking, overnight at 37°C in Luria broth plus supplements then diluted 1/100 in identical media and grown at 37°C to an OD_{600} of 0.2(HB101) or 0.4(JM83). In the case of HB101 this step was repeated. 80 mls of cells were pelleted by centrifugation at 4000xg, 4°C for 5 minutes then resuspended in 40mls of ice-cold 0.1M $MgCl_2$ and kept on ice for 5 minutes. The cells were repelleted and resuspended in 40mls of ice-cold 0.1M $CaCl_2$ and kept on ice for 20 minutes. Finally the cells were repelleted and resuspended in 4mls of ice-cold 0.1M $CaCl_2$ and kept on ice until required.

Typical transformation mixtures contained 200µl of competent cells, 100µl of 1xSSC(0.15M NaCl,15mM trisodium citrate,pH 7.0) and 0.1, 0.2 and 0.3µg of ligated DNA in a total volume of 10µl of H_2O. Each transformation mixture was kept on ice for 30 minutes with occassional shaking, transferred to a 42°C waterbath for 2 minutes then returned to ice for a further 20 minutes. The competent cells were then allowed to recover by growing at 37°C with gentle shaking in 1.2mls of Luria broth (plus 20µg/ml thymine for HB101) for 60-90 minutes. Cells were then pelleted in an Eppendorf centrifuge and resuspended in 0.1mls of Luria broth before plating on selective media. Control transformations were performed with phosphatased vector, either religated or linear, and with native plasmid DNA.

HB101/pAT153-recombinant transformants were selected on Luria agar plates containing 20µg/ml thymine plus 25µg/ml Na_2ampicillin. JM83/pUC13-recombinant transformants were selected on Luria agar plates containing 200µg/ml streptomycin, 25µg/ml Na_2ampicillin and 40µg/ml Xgal. Plates were incubated overnight at 37°C. An efficiency corresponding to ~10^6 transformants/µg were generally achieved.

23 SCREENING OF TRANSFORMANTS

The general method used was a modification of the filter hybridisation procedure of Grunstein and Hogness, (1975) which has been described in detail by Barrie (1982). The differences were in i) the plates used to grow replicates, which depended on the strain used and ii) the radioactive probe used to detect the clones of interest, which depended on the insert. Transformants derived from recombinant pUC13 plasmids were readily identified by their inability to develop the blue colony colouration associated with nonrecombinant plasmid transformants. This reduced the number of colonies to be screened by the filter hybridisation procedure. DNA from the clones of interest was prepared by the miniplasmid preparation method (see 7(bi)), restricted and the insert identified on an agarose gel before a full scale plasmid preparation was performed.

24 RESTRICTION MAPPING OF RECOMBINANT PLASMIDS

Detailed restrictions maps were produced by a modification of the method of Smith and Birnstiel, (1976).

a) Single digests

0.5μg samples of DNA were incubated with all the restriction endonucleases available for 1 hour in the manufacturers recommended buffer and at the optimum temperature for each enzyme. The ability to cut the DNA was determined by electrophoresis of the DNA on agarose gels. Only those restriction endonucleases which had digested the DNA were use in fine mapping of the DNA (see (c)).

b) Partial mapping

i) DNA 5'end-labelling with polynucleotide kinase and γ-^{32}P-ATP.

2μg of restriction endonuclease digested DNA was phosphatased and recovered into 10μl of 10mM Tris-HCl,pH 7.5 as described previously (19). The DNA was then incubated with 5 units of T4 polynucleotide kinase for 30 minutes at 37°C in kinase buffer(70mM Tris-HCl,10mM $MgCl_2$,5mM DTT,4mM spermidine,1mM Na_2EDTA,pH 7.6) plus 20μCi of γ-^{32}P-ATP. Incorporation was tested by running an aliquot on a horizontal agarose gel, cutting out the fragment and Cerenkov counting in a Packard 3255 liquid scintillation counter. The labelled DNA was recovered after two phenol extractions by two ethanol precipitations, vacuum dried, then redissolved in 10μl of 10mM Tris-HCl,pH 7.5. The DNA, labelled at both ends, was digested with a second restriction endonuclease to produce asymmetric uniquely end-labelled fragments. If required these fragments were separated on an agarose gel and recovered by method 10(ii) or 10(iii) ready for partial mapping.

ii) Restriction of labelled DNA

Several μg of unlabelled plasmid DNA was added as carrier to uniquely end-labelled fragment (equivilent to ~20000 dpm per digest) and the volume adjusted with 10mM Tris-HCl,pH7.5 such that there was sufficient volume to add 8μl to each restriction endonuclease digest. An array of restriction endonucleases (1μl) were used, in the manufacturers recommemnded buffer (1ul of a 10 x mix) and incubation temperature, to produce partial digestion of the DNA. For each enzyme 2.5μl samples of partially digested DNA were removed at 1, 2, 4 and 8 minutes after addition of the enzyme. These samples were pooled in an Eppendorf tube containing 1μl of 0.5M Na_2EDTA,pH 8.0 and frozen in an IMS/dry-ice bath on completion of the incubation period. Samples were then run on a 1% horizontal agarose gel which was dried down and then autoradiographed overnight without an intensifying screen.

25 Maxam and Gilbert sequencing

Recombinant plasmids containing the region of interest, or larger restriction fragments thereof, were digested by suitable restriction endonuclease(s) and recovered after phenol extraction by two ethanol precipitations into 10mM Tris-HCl,pH 7.5. DNA to be end-labelled at the 5' phosphate residue were phosphatased as described previously (19). DNA was then end-labelled at the 5' terminus using γ-^{32}P-ATP and polynucleotide kinase as described by Maxam and Gilbert, (1980). DNA was labelled at the 3' hydroxyl terminus using α-^{32}P-CTP and reverse transcriptase as described by Goodman, (1980). After phenol extraction and two ethanol precipitations the labelled DNA was digested with a second restriction endonuclease to produce uniquely end-labelled DNA fragments which were recovered from preparative horizontal agarose gels by method 10(ii) or 10(iii). All sequencing chemistry was performed by the procedures of Maxam and Gilbert, (1980) using the five chemical modification reactions (G, G+A, T+C, C, A>C).

26 Preparation of M13 recombinants

i) Sonication of Plasmid DNA

Sonication was performed in a sonicating waterbath (Kerry Ultrasonics Ltd) containing 1-2cm of water. A 1.5ml Eppendorf tube containing ≤15μg of recombinant plasmid DNA in a total volume of 30μl (made up with water) was placed on the bottom of the waterbath for 4 x 30 second bursts of sonication. Between each burst of sonication the solution was placed on ice and given a quick spin to bring the solution back to the bottom of the tube. The appearance of a "mist" on the sides of the tube was an indication of successful sonication. A 1µl aliquot was electrophoresed against pBR332 x Sau3A markers on an agarose gel to check the sonication (the majority of the DNA smear should be between 1.2 and 0.6 kb in size). Further 30 second bursts of sonication were employed until complete if incomplete first time. DNA was recovered after phenol extraction by two ethanol precipitations, vacuum dried then redissolved into 10µl of H_2O ready for end-repair.

ii) End-repair of sonicated DNA and size-selection

The sonicated DNA was end-repaired by incubation overnight at 15°C in the following reaction mixture;

DNA	(in H_2O)	20µl
10x ligase mix	(500mM Tris-HCl,100mM $MgCl_2$ 100mM DTT,pH 7.5)	3µl
TM buffer	(100mM Tris-HCl,100mM $MgCl_2$,pH 7.5)	3µl
0.1M spermidine		1.2µl
sequence chase mix	(0.25mM solution of each dNTP in TM buffer)	2µl
DNA polymerase I	(10 units of large fragment Klenow enzyme)	2µl

After end-repair sonicated DNA was electrophoresed in a 1.5% preparative agarose gel against pBR322 x Sau3A markers. DNA between 800-1200bps was collected on DE81 paper and recovered as described previously then redissolved in 20µl of H_2O.

iii) Preparation of stock M13 vector DNA

2µg of M13mp8 or M13mp9 RF DNA were cleaved at the desired cloning site by a restriction endonuclease(s) (for blunt ended substrates the enzyme used was SmaI) then recovered after phenol extraction by two ethanol precipitations, vacuum dried and redissolved into 10mM Tris-HCl,pH 7.5. The DNA was then phosphatased as previously described (19) and the DNA diluted to 20µg/ml with 10mM Tris-HCl,pH 7.5 ready for use.

iv) Ligation of sonicated DNA into M13 vector DNA

The following ligation reaction mixes were prepared;

Size selected DNA partials (ii)	1µl	2µl	4µl
Phosphatsed M13 vector DNA (iii)	1µl	1µl	1µl
10mM ATP	1µl	1µl	1µl
20mM Spermidine	2µl	2µl	2µl
10x Ligase buffer	1µl	1µl	1µl
H_2O	4µl	3µl	1µl

400 units of T4-DNA ligase were added to each and incubated overnight at 15°C. A further 0.5µl of 10mM ATP and 200 unit of T*4 DNA ligase were then added and incubated at 4°C for a further 1-4 days after which they could be stored at -80°C indefinitely.

v) Transformation of recombinant M13 into the E.coli strain JM101 or JM103

Competent cells were prepared by a modification of the method of Kushner (1978).

JM101 (or JM103) was grown, with shaking, overnight at 37°C in Luria broth containing thiamine (0.0002%). 0.5ml were diluted 1/100 in the same medium and grown to an OD_{600}=0.3. The culture was kept at room temperature (for later use) while 1.4 ml aliquots of cells were pelleted by a 30 second spin in an Eppendrof centrifuge (the number of tubes = number of ligation reactions). The supernatant was flicked off and the cells (gently) resuspended in 0.5ml of MR (MR= 10mM MOPS,pH 7.0,10mM RbCl), spun again for 30 seconds and the supernatant removed as before. The cells were resuspended in 0.5ml of MRC (MRC= 100mM MOPS,pH 6.5,10mM RbCl,50mM $CaCl_2$) and left on ice for 30 minutes. After another 30 second spin the cells were resuspended in 0.15ml of MRC and kept on ice until ready. To each tube of "competent" cells was added 3µl DMSO (dimethylsulphoxide) and 5µl of the ligation mix. The mixture was left on ice for 1 hour, placed at 55°C for 35 seconds, cooled on ice for 1 minute, then left at room temperature. Before plating out the cells were transferred to a glass tube containing 200µl of JM101 or JM103 log phase cells (those held at room temperature), 25µl of 25mg/ml BCIG (in dimethylformamide) and 25µl of IPTG (in H_2O). These tubes were then mixed with 3ml of LUB soft agar before plating out on LUB agar plates and incubated overnight at 37°C.

White and blue "plaques" develop overnight in the bacterial lawn corresponding to recombinant and nonrecombinant M13 transformants respectively. "Plaque" in this sense refers to an area of reduced bacterial lawn growth due to *in vivo* M13 replication rather than the cell death associated with the virulant replication cycle of lambda. White plaques were screened for recombinant sequences and single-stranded DNA sequencing templates prepared from positives as described by Weller *et al.* (1984).

27 M13 recombinant sequencing

Sequencing of M13 recombinant clones was based on the methods of Sanger *et al.* (1978) and Biggin *et al.* (198~~4~~3) for M13-dideoxyribonucleotide chain-termination using α-^{32}P-dATP and α-^{32}S-dATP respectively. Quantities quoted are for 15 sequencing templates which are a comfortable number to process at any one time. Reactions were performed in 1.5ml Eppendorf tubes and all centrifugations were done in an Eppendorf bench centrifuge.

To ensure the cloned template DNAs were fully redissolved after preparation they were incubated at 60°C for 10 minutes prior to annealing.

Each clone was annealed to the 17-mer universal primer by taking 5µl of clone DNA plus 5µl of primer mix (7.2µl of 2µg/ml 17-mer primer, 64µl H_2O, 8µl TM buffer (100mM Tris-HCl,pH 8.0, 100mM $MgCl_2$) and incubating at 60°C for 2 x 30 minutes with a quick centrifugation in between. The annealed mix can be held at room temperature until ready to start the sequencing reactions themselves. For each clone 4 reaction tubes were prepared containing 2µl of annealed clone plus 2µl of either a "T","C","G", or "A" NTP mix.

NTP mixes for sequencing

	"T"	"C"	"G"	"A"
0.5mM dTTP	12.5	250	250	250
0.5mM dGTP	250	250	12.5	250
0.5mM dCTP	250	12.5	250	250
10mM ddTTP	12.5(6.2)			
10mM ddCTP		2(4)		
10mM ddGTP			4(8)	
10mM ddATP				0.75(1.2)
TE buffer	500	500	500	250

TE buffer= 10mM Tris-HCl,pH 8.0, 0.1mM EDTA). 0.5mM dNTP and 10mM ddNTPs were prepared in TE buffer. The figure in brackets is the amount of ddNTP added to the mix when ^{35}S-dATP was the radiolabelled substrate.

To each reaction tube in turn was added 2µl of freshly prepared "Klenow" mix (^{32}P-mix= 117µl TE buffer, 3.3µl Klenow polymerase (5 units/µl), 10µl α-^{32}P-dATP, 0.7µl 50µM dATP; ^{35}S-mix= 114µl H_2O, 7µl Klenow polymerase, 10µl α-^{35}S-dATP). The tube was given a gentle mix then incubated at 37°C for 20 minutes. For 15 clones all 60 tubes can be completed at a steady pace in 20 minutes. After 20 minutes 2µl of sequence chase mix (0.25mM each of dATP, dGTP, dTTP, dCTP made up in TE buffer) was added to each tube in turn (in the same order as before), mixed and incubated a further 20 minutes at 37°C. At this point the reaction tubes can be prepared for loading onto a sequencing gel by addition of 4µl of formamide dye (stock solution containing 10ml deionised formamide, 10mg xylencyanol FF, 10mg bromophenol blue, 0.2ml 0.5M Na_2EDTA,pH 8.0). ^{32}P-labelled substrates were run as soon as possible while ^{32}-S labelled substrates were occasionally stored for several days at -20°C or -80°C as long as the formamide dye had not been added.

28 Sequencing gels

i) Maxam and Gilbert sequencing gels

Substrates were run on 40cm 8% or 6%(w/v) polyacrylamide gels 0.35mm thick prepared by the following method;

For 2 gels:	8%	6%
acrylamide	15.2g	11.4
bisacrylamide	0.8g	0.6g
urea	100g	100g
10 x TBE buffer (1MTris-Borate pH 8.3, 20mMEDTA)	10ml	10ml
H_2O	upto 200ml	

After all the ingredients had dissolved 1.4ml of a 10% APS (Ammonium persulphate) solution (made up in H_2O) was added and the solution filtered through 2 x 9cm diameter Whatman filters using a Buchner funnel and vacuum line. 37µl of TEMED was added to 100ml degassed aliquots of this solution immediately before the gel was poured.

Before loading onto the gel, sequencing substrates were first denatured by incubation at 90°C for 1 minute then placed in ice/water. Gels were run at 1.4-1.5KV for 3-8 hours then removed from the gel mould. The two glass plates were separated such that the gel remained attached to the larger unsiliconised plate. The gel was then covered with aluminium foil and autoradiographed at -70°C, in the presence of an intensifying screen if required, for 3-14 days.

ii) M13 sequencing gels

Preparation and running of 40cm 6%(w/v) polyacrylamide gels or buffer gradient gels were essentially as described in (i) or by Biggins _et al._ (1983) respectively. A "sharks" tooth comb was used which enabled 15-18 clones to be run on a single gel. The standard mixes used to prepare two buffer gradient gels are given overleaf.

For two gels:	0.5x	2.5x
acrylamide	17.1g	2.28g
bisacrylamide	0.9g	0.12g
urea	150g	20g
sucrose	---	2g
10 x TBE	15ml	10ml
H_2O upto final vol	300ml	40ml

The solutions were filtered as for Maxam and Gilbert gels then aliquots degassed before addition of;

	0.5x (150ml)	2.5x (20ml)
10% APS	1.05ml	0.14ml
TEMED (prior to ~~on~~ pouring)	72µl	9.6µl

1.5µl (^{32}P) or 2.5µl (^{35}S) samples of each sequencing reaction were boiled for 3 minutes before loading onto the gel. After electrophoresis for 3-8 hours at 1.4-1.7KV the gels were fixed in a 10% methanol, 10% acetic acid solution for 15 minutes then dried down using a Bio-rad Gel drier. Gels were autoradiographed at room temperature for 16 hours-4 days.

29 COMPUTING

Various computing facilities were used in the course of this work. The dot-matrices presented were prepared using a DNA manipulation package written in Fortran 77 by Dr Z.Nugent and run on University facilities including a Cyber 73 (Control Data Corporation) mainframe computer and

Calcomp 936 graph plotter. Alignment of sequences was also performed on the Cyber 73 using a Basic program written by Dr C.Boyd.

M13 clone sequences were aligned against reference sequences and each other using a version of the Staden (1980) programme modified to run on a Digital PDP 11/44 minicomputer. This minicomputer was also used in conjunction with the word-processing package Word 11 to format the DNA sequences as presented.

A BBC/Acorn microcomputer was used to run a package of DNA sequence manipulation programmes written in BBC Basic by Dr A.J.Jeffreys. The dot-matrix program in this package is based on the algorithum of White _et al._ (1984).

Finally, this thesis was compiled using a Fortune 32:16 dedicated word-processor running the multiuser version of Fortune:Word.

28 CONTAINMENT

All experiments undertaken in this thesis were conducted with reference to the Genetic Manipulation Advisory Groups guidelines on safety and containment conditions for such work.

Chapter 3

DETECTION OF HUMAN $\Psi\beta1$-RELATED SEQUENCES IN OTHER PRIMATES AND MAMMALS

3.1 Introduction

The human β-globin gene family consists of five functional genes and a single non-processed pseudogene arranged in a cluster covering 65 kb of genomic DNA on chromosome 11 (Figure 3.1*). The functional genes can be further classified, according to their period of developmental expression, into embryonic (ε), foetal ($^G\gamma$ and $^A\gamma$) and adult (δ and β) globin genes. In order to gain a clearer understanding of the evolutionary history of the contemporary human β-globin gene family the physical organisation of the β-globin gene cluster has also been determined for a representative primate species from each of the contemporary primate orders from prosimians to man (Figure 3.1). Using appropriate hybridisation stringencies human globin DNA probes detected specific genomic DNA fragments in all species examined. The hybridisation probes were either specific globin gene cDNAs or cloned fragments of human genomic DNA containing globin gene exon sequences and single copy extragenic sequences (Barrie _et_ _al_., 1981). Although some cross-hybridisation occurred with these probes, particularly in the lemur, the relative autoradiographic intensities of specific fragments permitted all major genomic DNA fragments to be assigned as either ε-, γ- or β-like.

*Figures for this Chapter follow the text.

Additional gene sequences were also found by hybridisation of primate genomic DNAs with a rabbit adult β-globin cDNA probe at lower hybridisation stringencies (Barrie _et al._, 1981). In the gorilla and yellow baboon these additional fragments were assumed to contain the Ψβ1 and Ψβ2 genes (Fritsch _et al._, 1980) as they were electrophoretically similar to the additional fragments found in man. One additional fragment was also seen in the owl monkey but none were found in the lemur, suggesting that the entire complement of β-related globin genes had already been detected within the genome of this lemur by the human globin gene probes. The possibility therefore exists that the history of the human Ψβ1 gene may be represented in these established β-globin gene clusters of the primate lineage.

While available hybridisation probes can distinguish the Ψβ1 gene in the genomic DNA of man and other primates (Barrie _et al._, 1981) these probes are not gene specific as they contain the coding regions (exons and proximal flanking regions) that are relatively well conserved between the different globin genes. Probes cross-hybridise to other β-globin sequences at the hybridisation stringencies required to detect human genomic DNA fragments presumed to contain the Ψβ1 gene. Such cross-hybridisation would make difficult the physical mapping and molecular cloning of other primate Ψβ1-related sequences. Also, a cautionary example of the problems associated with using cDNA constructs as hybridisation probes at low stringencies emerged during the course of this work as a result of the further analysis of the human "Ψβ2" gene (Shen and Smithies, 1982). Sequencing of the region thought to contain the Ψβ2 gene (by hybridisation) revealed no β-like structure. The basis for assignment of this region as a Ψβ gene appears to be several runs of poly(dT), which may have hybridised with the poly(dA) sequence of the cDNA probe used.

In order to confirm and facilitate the further analysis of the potential Ψβ1-related sequences in the other primates a gene-specific hybridisation probe was therefore required. The mutual divergence of the non-coding DNA sequences (distal flanking and intron regions) observed between different members of the human β-globin gene family suggested these regions from the human Ψβ1 gene would provide ideal gene-specific probes (see below). This chapter describes the preparation of gene-specific hybridisation probes from the human Ψβ1 gene and their use in identifying the presence of Ψβ1-related sequences in the primates and other mammals.

3.2 The detection of sequence homologies and the determination of gene orthologies by dot-matrix analysis

Computer generated dot-matrices provide an unbiased representation of the regions of homology shared between two genes over their coding and non-coding sequences. In the simplest form of dot-matrix the regions of homology are depicted as a series of dots or diagonal lines within a two-dimensional plot (Konkel *et al.*, 1979), each dot or line representing the relative position in one sequence (x-axis) for which an equivalent sequence has been found in another sequence (y-axis). The "stringency", or degree of homology required between two sequences before a match is plotted, can be varied by reducing the number of correct matches required ("hits") within the given number of consecutive nucleotides under comparison (the "window"). Also, by varying the size of the window, and the number of hits within a particular window, it is possible to reduce the level of chance background matches in a given comparison without losing the alignment that indicates sequence homology between two distantly related sequences (own results not shown, see White et al., 1984). The loss of background 'noise' may also result in the loss of useful information concerning the nucleotide composition of the two sequences, and therefore a number of different windows and stringencies are usually employed (own results not shown, see White et al., 1984).

In general, the regions which show strongest homology between two related genes are the exons and conserved 5' and, to a lesser extent, 3' non-coding regions. For example, the human β-globin genes (ε, γ, Ψβ1, δ and β) can be aligned over their coding and immediate flanking sequences but are generally unalignable by dot-matrix analysis over most of their non-coding regions (particularly over intron 2); even at low stringency dot-matrix criteria capable of detecting homologies between sequences upto 40% diverged (for example see figure 3.2 which shows the alignment of the human β-globin gene against the other members of the human globin gene family). The discrete nature of the non-coding regions of the different human β-globin genes suggests that each gene must have evolved independently for a considerable time for them to have achieved ≥40% divergence at the estimated rate of non-coding DNA evolution in the primates (see discussion).

Similarly, the non-coding regions of the human Ψβ1-globin gene are unalignable against the equivalent sequences of any other member of the human β-globin gene family, even against the γ-globin gene to which it is most closely related over the coding regions (Goodman et al., 1984). This suggests the non-coding regions of the human Ψβ1 gene are at least 40% diverged from any other member of the human β-globin gene family and therefore that this gene is potentially an ancient gene, at least as ancient as the most recent of the other globin genes (see Discussion). The absence of alignment of the human Ψβ1 gene with any other human globin gene over the non-coding regions, in particular intron 2, also suggests that hybridisation probes derived from these regions of the human Ψβ1 gene would be gene-specific and not cross-hybridise with other non-Ψβ1-related globin gene sequences in the genomes of other species (see below).

Dot-matrix criteria have become increasingly important as a means of determining gene orthologies between globin genes from different mammalian orders (see Discussion). This depends on the ability to achieve interspecies alignment between non-coding DNA sequences of orthologous globin genes that are ≤40% diverged. Homology can generally be detected over the non-coding regions of orthologous genes but not non-orthologous ones, this is particularly true over the large second intron of the β-globin genes. For example, the dot-matrix of the rabbit β1 gene x human β gene (which are orthologous) shows extensive regions of homology over non-coding DNA sequences compared with that of rabbit β4 x human β (which are non-orthologous) over the same regions (Figure 3.3). Within the mammalian orders therefore orthologous β-globin genes have, in general, not yet diverged sufficiently over non-coding regions to be undetectable by dot-matrix analysis of DNA sequence data.

Dot-matrix alignments indicating sequence homologies can also be detected over non-coding regions between certain interspecies β-globin gene comparisons of paralagous genes (own results not shown, see also Discussion). These alignments can indicate the involvement of non-coding sequences in a gene conversion event between two members of a gene family such that the the regions involved in the conversion tract are no longer as diverged as expected from the level of sequence divergence in other non-coding regions. For example, the first intron of the human δ-globin gene has been corrected to look more like the equivalent human β-globin gene region than, for example, the second intron (Figure 3.2 and Discussion).

Dot-matrices are therefore a good independent unbiased means of determining gene orthology between sequences from within and between different mammalian orders, as a means of detecting interesting sequence features and as one way of detecting potential gene conversion events. Dot-matrices have been used extensively throughout this thesis (only a few examples are shown) and by other workers as a means of interpreting the relationships between different DNA sequences. It should be remembered however that as with DNA cross-hybridisation between mammalian orders the failure to detect a region of homology between two sequences may reflect the level of mutual DNA sequence divergence between the two sequences rather than a lack of evolutionary orthology.

3.3 Isolation of human Ψβ1-specific gene probes

The detailed restriction endonuclease cleavage site map determined from the available DNA sequence of the human Ψβ1 gene (S.M.Weissman per.comm.) was examined for suitable restriction endonuclease fragments that might constitute gene-specific hybridisation probes. The restriction endonuclease fragments, containing primarily non-coding DNA sequence, were chosen from the immediate 5' end of the gene and from within the second intron (Figure 3.4(a)). In order to isolate sufficient quantities of the two probes the human Ψβ1 gene was first subcloned as part of a 7 kb EcoRI fragment from λHγG4, a human λ recombinant containing the human Ψβ1 gene (Fritsch et al., 1980), into the plasmid pAT153 (Twigg and Sherratt, 1980)

λHγG4 was digested to completion with EcoRI, the DNA recovered, and all the resultant DNA fragments ligated into the phosphatased EcoRI cleaved cloning site of pAT153. A total of 100ng of ligated material was transformed into the E.coli K12 strain HB101 and transformants selected by plasmid conferred ampicillin resistance. Transformants were screened for the presence of the 7 kb EcoRI fragment by the colony hybridisation method of Grunstein and Hogness, (1975) using a ^{32}P-labelled human γ cDNA probe (see 3.4). The presence of the 7 kb EcoRI fragment in recombinant plasmids was confirmed by agarose gel electrophoresis of small scale preparations of 'positive' recombinant plasmid DNA digested with EcoRI. A single recombinant plasmid, called pHΨβ1, was selected from which the following two hybridisation probes were isolated.

Probe 1, a 439bp Sau3A fragment from 5' of the first exon, was recovered from a 2% preparative agarose gel by method 2.10(iii) after total digestion of pHΨβ1 by the restriction endonuclease Sau3A. As well as 5' non-coding DNA sequences this probe contains the first 23 bp of the pseudogene homologues of the 5' non-translated sequence in a functional globin gene.

Probe 2, a 812bp MboII fragment containing most of intron 2, was isolated from pHΨβ1 in two stages. pHΨβ1 was digested with restriction endonuclease BglII and a ~4kb fragment containing the Ψβ1 gene was recovered from a 1.5% preparative agarose gel using method 2.10(ii). This fragment was further digested with restriction endonuclease MboII and the 812bp fragment recovered by method 2.10(iii) from a 1.5% preparative agarose gel.

The relationship of λHγG4, pHΨβ1, probe 1 and probe 2 to each other and the human β-globin gene cluster is shown diagrammatically in Figure 3.4(a).

3.4 Additional hybridisation probes used during this work

Figure 3.4(a) also shows the extent of another DNA probe isolated from the human Ψβ1 gene (probe 3). This ≈1.8 kb BglII-XbaI fragment, containing most of the human Ψβ1 gene plus some 3' flanking sequences, was used to identify M13 recombinant clones involved in the sequencing of the owl monkey Ψβ1-related gene (Chapter 4). The fragment was recovered from a 1% preparative agarose gel by method 2.10(iii) after digestion of pHΨβ1 by the two restriction endonucleases BglII and XbaI

The human γ cDNA probe was isolated as part of a 1.5 kb HpaII fragment from pHγG1, a pCR1 plasmid recombinant containing a 695-720 bp insert of a cDNA copy of the human $^{G}\gamma$-globin mRNA (Little et al., 1978)

The rabbit adult β-globin gene probe PβG was isolated as a 1.5 kb *HhaI* fragment from PβG1, a pMB9 recombinant plasmid containing a 570 bp insert of a cDNA copy of the rabbit adult β-globin mRNA (Maniatis *et al*., 1976).

3.5 Establishment of the gene-specific nature of Ψβ1 probes 1 and 2

The nature of the hybridisation pattern produced by ^{32}P-labelled probes 1 and 2 was determined by filter hybridisation to human genomic DNA digested with restriction endonucleases *Eco*R1 or *Bgl*II (Figure 3.4(b)). The filters were prepared as follows. 5μg of high molecular weight human genomic DNA was digested with *Eco*RI or *Bgl*II and after recovery the DNA was electrophoresed, in the native double stranded form, on a 0.5% agarose gel. The DNA was then acid/alkali denatured *in situ* and transferred to a nitrocellulose filter by Southern blotting.

The sizes of the principal fragments detected by the two Ψβ1 gene probes (1 and 2) are those predicted from the known restriction endonuclease map for the region of the β-globin gene cluster encompassing the human Ψβ1 pseudogene (Barrie *et al*., 1981). At the relatively low stringency employed in these and subsequent hybridisations (1xSSC, 60°C) neither probe detects other human genomic DNA fragments known to contain the functional genes of the human β-globin gene cluster. In contrast, globin coding sequence probes have been show to cross-hybridise at even higher stringencies than those employed here (Barrie *et al*., 1981). These two probes do however detect other fragments within the human genome (Figure 3.4(b), Chapter 7). The two non-coding sequence probes from the human Ψβ1 gene (probe 1 and 2) have been employed throughout this thesis as essentially unique sequence hybridisation probes for the specific detection of human Ψβ1-related sequences in other species' DNAs.

3.6 The presence of human Ψβ1-related sequences in other primate groups

The distribution of human Ψβ1-related sequences in the other primate lineages was investigated by filter hybridisation of primate genomic DNAs digested with EcoRI or BglII. 5μg samples of high molecular weight genomic DNA from at least two representative species of each of the major primate groups from prosimians to man were digested with restriction endonucleases EcoRI or BglII. After recovery the digested DNAs were electrophoresed, in the native double stranded form, on a 0.5% agarose gel. The DNA was acid/alkali denatured *in situ* then transferred to nitrocellulose by Soutern blotting. Filters were hybridised, in turn, with ^{32}P-labelled Ψβ1-specific probes 1 and 2 (Figure 3.5). The results are summarised below.

The great apes and Old World monkeys

The great ape and Old World monkey β-globin gene clusters have been shown to have a very similar gene organisation and restriction endonuclease site cleavage map to that in man (Barrie *et al.*, 1981). This apparent stability in β-globin gene cluster organisation is also reflected in the size of fragments that hybridise to Ψβ1 probes 1 and 2 (Figure 3.4). In every case except one, gorilla DNA x EcoRI, the genomic DNA fragments hybridising to the human Ψβ1 gene-specific probes were electrophoretically indistinguishable in size from those detected in man. The most probable explanation for the smaller EcoRI hybridising fragment in the gorilla examined is that this individual is homozygous for an additional EcoRI site near this gene.

These hybridisation patterns are very strong evidence for the presence of a Ψβ1-related sequence within the β-globin gene cluster of the great apes and Old World monkeys at a similar position to that of the Ψβ1 pseudogene in man.

The New World monkeys

Both of the human Ψβ1 gene-specific probes hybridise to the three primate species examined in this group (owl monkey and squirrel monkey from the family Cebidae and the red-mantled tamarin from the family Callitrichidae). The Ψβ1-related sequence is apparently present on a single genomic EcoRI or BglII fragment as both probes detect the same electrophoretically indistinguishable fragments in either digest, except for owl monkey DNA x EcoRI. The different owl monkey genomic EcoRI fragments detected by probe 1 and probe 2 suggest the presence of an intragenic EcoRI site in the Ψβ1-related sequence of this species. The presence of this intragenic EcoRI site was subsequently confirmed by DNA sequencing (see Chapter 4).

As in the great apes and Old World monkeys the hybridisation of both human Ψβ1-specific probes to genomic DNA fragments of these New World monkey species is strong evidence for the presence of a complete Ψβ1-related gene in the genome of these primates. It is not possible however to determine the position of such sequences in relation to the other globin genes in any of these primates.

The prosimians

The prosimians examined were all members of the lemuridae family. Only the 5' probe from the human Ψβ1 gene (Probe 1) hybridised to genomic DNA fragments from these lemurs suggesting that only part of a Ψβ1-related sequence is present in the genome of these primates. The consistent hybridisation to probe 1 but failure to hybridise to probe 2 in all of the lemurs examined suggests that the common ancestor of these lemurs also contained an incomplete Ψβ1-related sequence and that the loss of probe 2 related sequences therefore occurred early in lemur evolution (see chapter 5 and Discussion).

3.7 Detection of Ψβ1-related sequences in other mammals

The presence of all or part of a Ψβ1-related sequence in all the primates examined suggests that a Ψβ1-related sequence was present in the ancestral β-globin gene cluster that predated the basal primate radiation ≥70 MY ago, that is, before the prosimians diverged from the simians. The functional status of the Ψβ1-related sequence in these other primates is unknown.

As Ψβ1-related sequences apparently predate the basal primate radiation, which occurred a short evolutionary time after the mammalian radiation, the possibility exists that Ψβ1-related sequences may also have been present at the time of the mammalian radiation. The mammalian radiation is thought to have occurred ≥80 MY ago. The cross-hybridisation of human Ψβ1 non-coding DNA sequence probes, under the hybridisation conditions used, suggests an accumulated sequence divergence of <30% between the most divergent of the primate Ψβ1-related sequences and those in man. As the evolutionary period between the basal primate radiation and the mammalian radiation is so short it may therefore be possible to detect Ψβ1-related sequences in the genomes of other contemporary mammals using human Ψβ1 probes 1 and 2 (see below).

To test for the presence or absence of a human $\Psi\beta1$-related sequence~~s~~ in the genomes of other mammalian species 5μg samples of high molecular weight genomic DNA from several mammalian species were digested with the restriction endonucleases <u>Eco</u>R1 or <u>Bgl</u>II. The digested DNAs were recovered then electrophoresed, in the native double stranded form, on a 0.5% agarose gel. After acid/alkali denaturation <u>in situ</u> the DNA was transferred to nitrocellulose filters by Southern blotting. Filters were hybridised, in turn, with ^{32}P-labelled $\Psi\beta1$ probes 1 and 2, (Figure 3.6).

The success of this experiment depended on the level of mutually accumulated sequence divergence between the human non-coding DNA probes (probes 1 and 2) and their equivalent sequences, if present, in the mammalian species examined. If >30% sequence divergence has been accumulated no stable DNA hybrid~~e~~s would be expected to form and that species would appear to lack a $\Psi\beta1$-related sequence within the genome, even if present. Alternatively, due to random sequence drift towards seq~~eu~~uences contained in the two probes spurious cross-hybridisation signals may be detected by the non-coding DNA probes.

The probable presence of a genuine human $\Psi\beta1$-related sequence within the genomes of the other mammals examined was therefore only inferred from hybridisation experiments if both probes hybridised to an electrophoretically indistiguishable fragment in the same genomic DNA digest. This condition was only fulfilled in the carnivores (dog, lion) and pinniped (seal) suggesting the presence of a genuine $\Psi\beta1$-related sequence in these species (Figure 3.6). All other species (bat, blackbuck, cow, dog, mouse, rabbit, roe deer) gave inconsistent or complex hybridisation patterns (results not shown). The functional status of the $\Psi\beta1$-related sequences in the dog, lion and seal, is again unknown.

3.8 A functional human Ψβ1-related sequence in the goat

During the course of this work Goodman et al., (1984) observed that the coding regions of the recently published ε^{II}-globin gene in the goat (Shapiro et al., 1983) were more closely related to the Ψβ1 gene of man than to any of the other human β-globin genes. As the goat ε^{II} gene is thought to be functional this observation implies that in at least one mammalian species an orthologue of the contemporary human Ψβ1 pseudogene retains, or has recovered, some functional role. In order to independently confirm this observation dot-matices were performed of the goat ε^{II}-gene sequence against all the published human β-globin gene sequences. The important regions for determining gene orthologies (the non-coding regions, in particular intron 2 (see 3.2)) showed the most convincing alignment between the human Ψβ1 pseudogene sequence and the goat ε^{II} gene (Figure 3.7), confirming Goodman et als' observation that these genes are almost certainly orthologous.

This observation reinforced the original aims of this thesis (see Introduction), that is, to establish the functional status of potential Ψβ1 sequences detected in other primates in order to determine phylogenetically, and for the first time, the evolutionary history of a contemporary pseudogene.

Figure 3.1

Restriction endonuclease cleavage site maps of the primate β-globin gene clusters, adapted from Barrie *et al*. (1981).

Cleavage sites are shown for restriction endonucleases *Bam*HI (B), *Bcl*I (Bc), *Bgl*II (Bg), *Eco*RI (E), *Hind*III (H), *Kpn*I (K), *Pst*I (P), and *Xba*I (X). The restriction endonuclease site map around the human Ψβ1 gene was not determined and the relative position of the human Ψβ1 globin pseudogene is taken from Fritsch *et al*. (1980). These maps show only cleavage sites that generate β-globin DNA fragments; the direction of the gene detected relative to a mapped site is indicated by ▶, or ◀. Known polymorphic cleavage sites are marked by ±. Linkage of the owl monkey ε, γ and β globin genes was not established, although a probable ε-γ linkage was indicated by the aligned maps. The leftward owl monkey β-globin gene was assigned to δ on the basis of residual map homology with man. Primate cleavage sites indistinguishable from those in man are indicated by open circles. Sites present in a primate but definitely not in man are shown by filled circles. These comparisons are most reliable in the gorilla and baboon; possible site identities in the owl monkey and lemur are shown but are less definite due the altered arrangement and number of genes in these species.

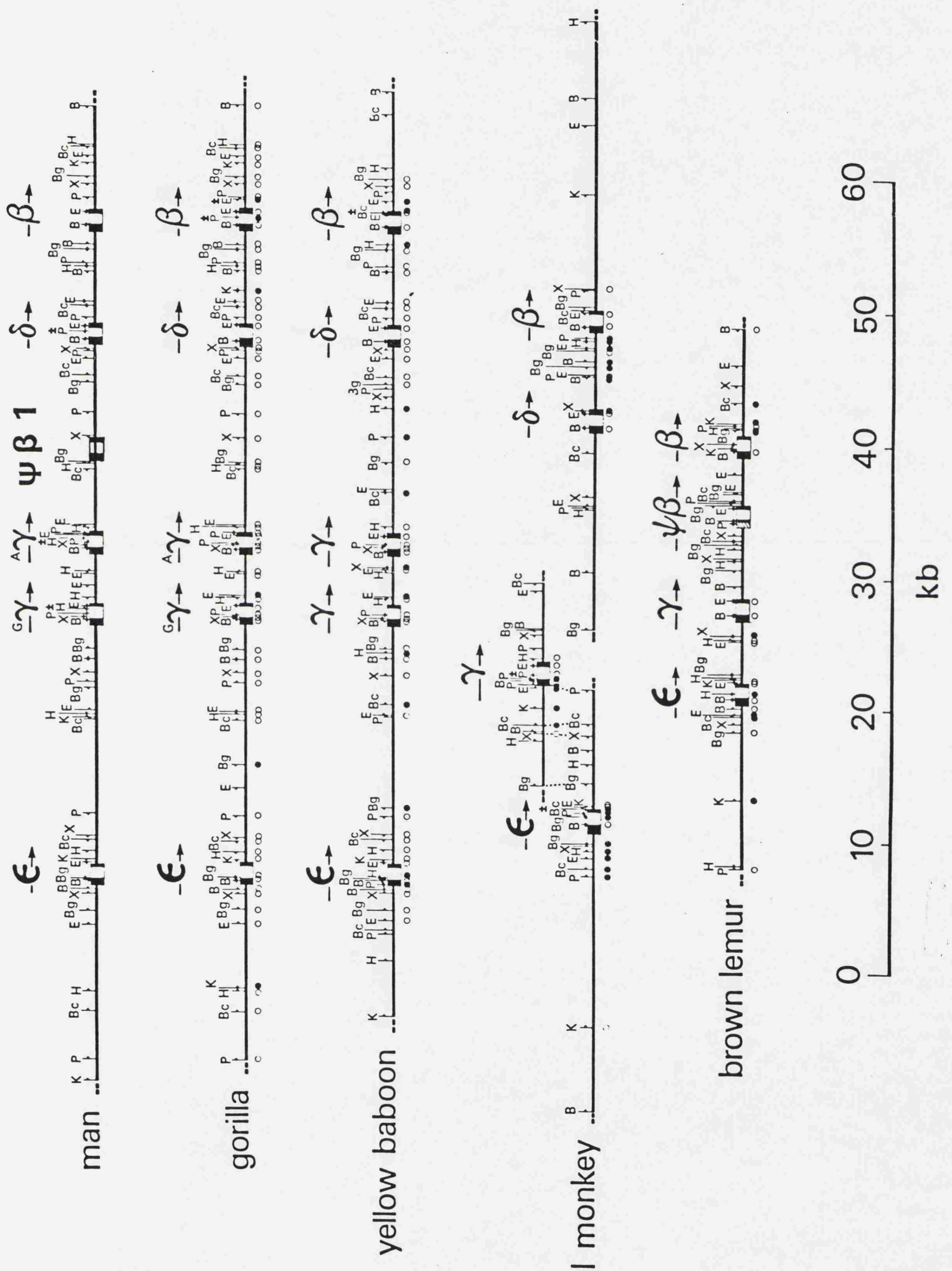

man
gorilla
yellow baboon
owl monkey
brown lemur
-ε→
Gγ→
Aγ→
ψβ1
-δ→
-β→
-γ→
-ψβ→
0
10
20
30
40
50
60
kb

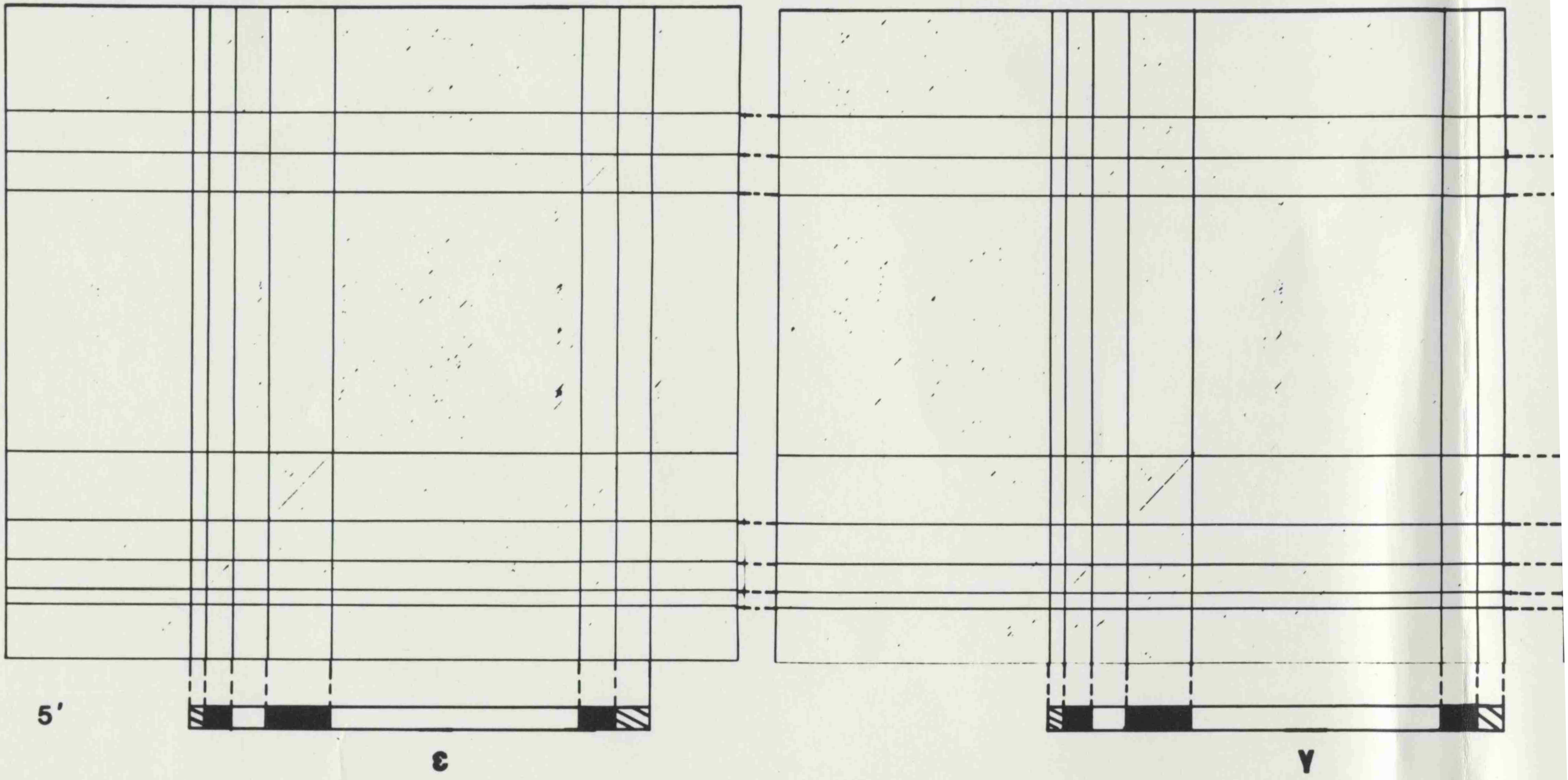
Figure 3.2

Dot-matrix comparisons of the human β-globin gene sequence with the other human β-related globin genes.

The entire available 2129 bp of the human β-globin gene (Lawn et al., 1980; Hardison, 1984) was compared with the sequences of the human ε, $^{A}\gamma$, Ψβ1 and δ globin genes taken from Baralle et al. (1980a,b); Shen et al. (1981); Chang and Slightom (1980): Weissman (pers.comm.); and Spritz et al. (1980) repectively.

Each diagonal dash represents 5 bp corresponding to the centre of a 30 bp comparison with a minimum of 17 matches between the two gene sequences, that is, a "window" of 30 bp and a minimum "hit" size of 17 bp. The main diagonal appears as a broken line at 45° across the grid. This matching criteria was established (other comparisons not shown) so as to achieve the best balance between background "noise" and genuine homology, depicted along the main diagonal. This matching criteria has been used throughout this thesis unless otherwise stated. The positions of coding sequences (filled boxes), the 5' an 3' non-translated regions of the mature mRNA (hatched boxes) and the intervening sequences (open boxes) are shown alongside the grids.

One significant feature in these comparisons is the homology detected over intron 1 and 5' flanking sequences between the human β and δ globin genes. This contrasts with the apparent absence of homology over intron 2 and 3' flanking sequences; as found in all non-coding regions of the other comparisons (see text).


5′
ε
γ

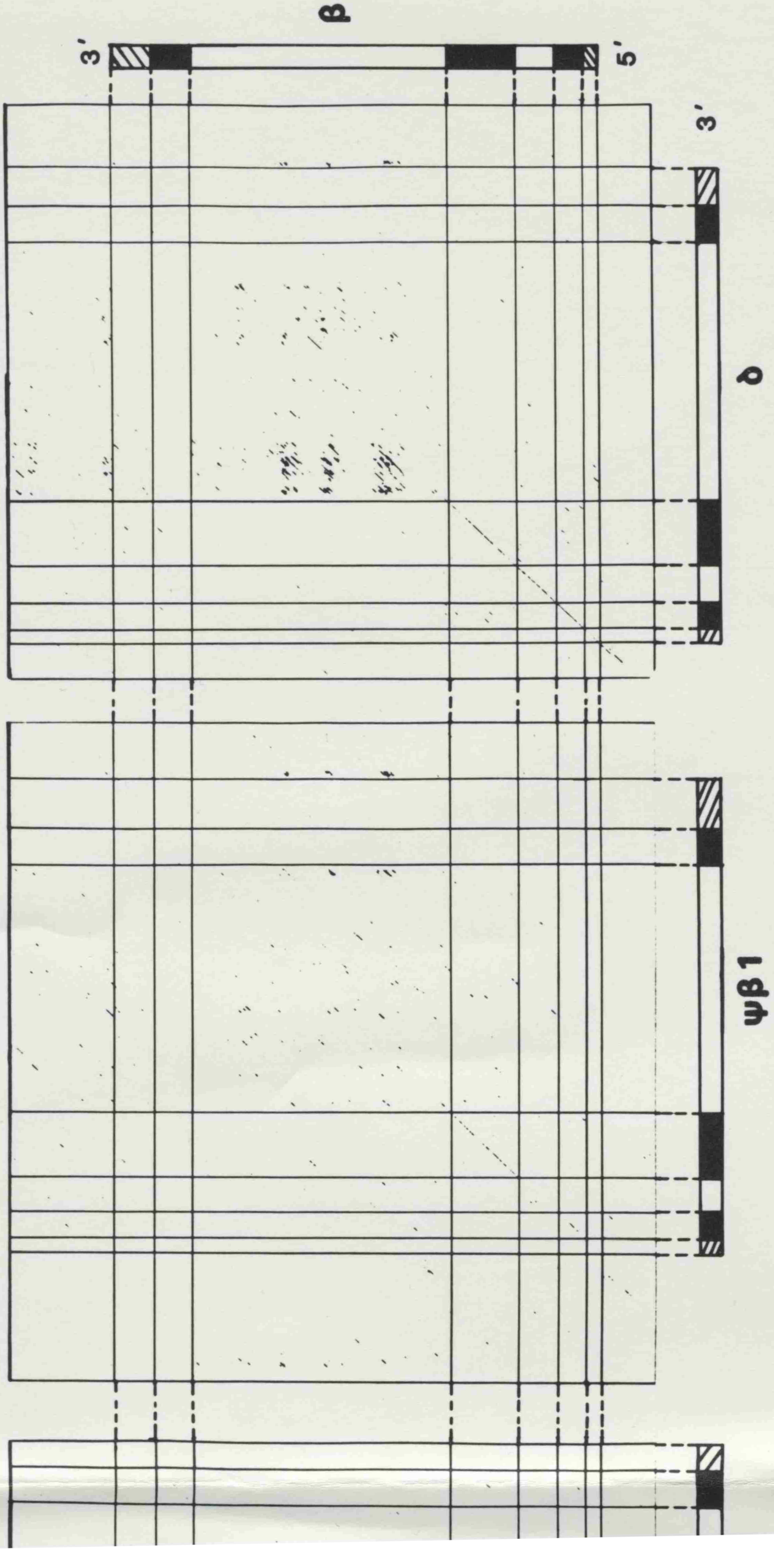

β
3'
5'
3'
δ
ψβ1

Figure 3.3

Dot-matrix comparison of the human β-globin gene with the orthologous rabbit β1 gene and non-orthologous rabbit β4 gene.

Matching criteria and sequence features depicted beside the grid corrrespond to those outlined in Figure 3.2. Rabbit sequences β1 and β4 were taken from Hardison et al. (1979) and Hardison (1984) respectively.

Orthologous sequences human β and rabbit β1, that encode the adult β-globin peptide in each species, show considerable sequence homology over all regions (exons, introns and flanking sequences). In contrast, the non-orthologous human β and rabbit β4 gene (which encodes an embryonic rabbit β-globin peptide) show no homology except over the functionally conserved coding regions.

Another feature illustrated in the orthologous comparsion of human β/ rabbit β1 sequences is the ability of this technique to detect differences due to insertion/deletion during the independent evolution of two sequences. These events shift the major diagonal by the number of bases involved but do not destroy the underlying homology. In this comparison the intron two sequences of these two genes differ by a substantial number of bases due to a major insertion/deletion event involving some ~240 bp (Hardies et al., 1984). It is possible however to distinguish even single base differences between two otherwise homologous sequences by this method, a feature which proved extremely useful when aligning two sequences.

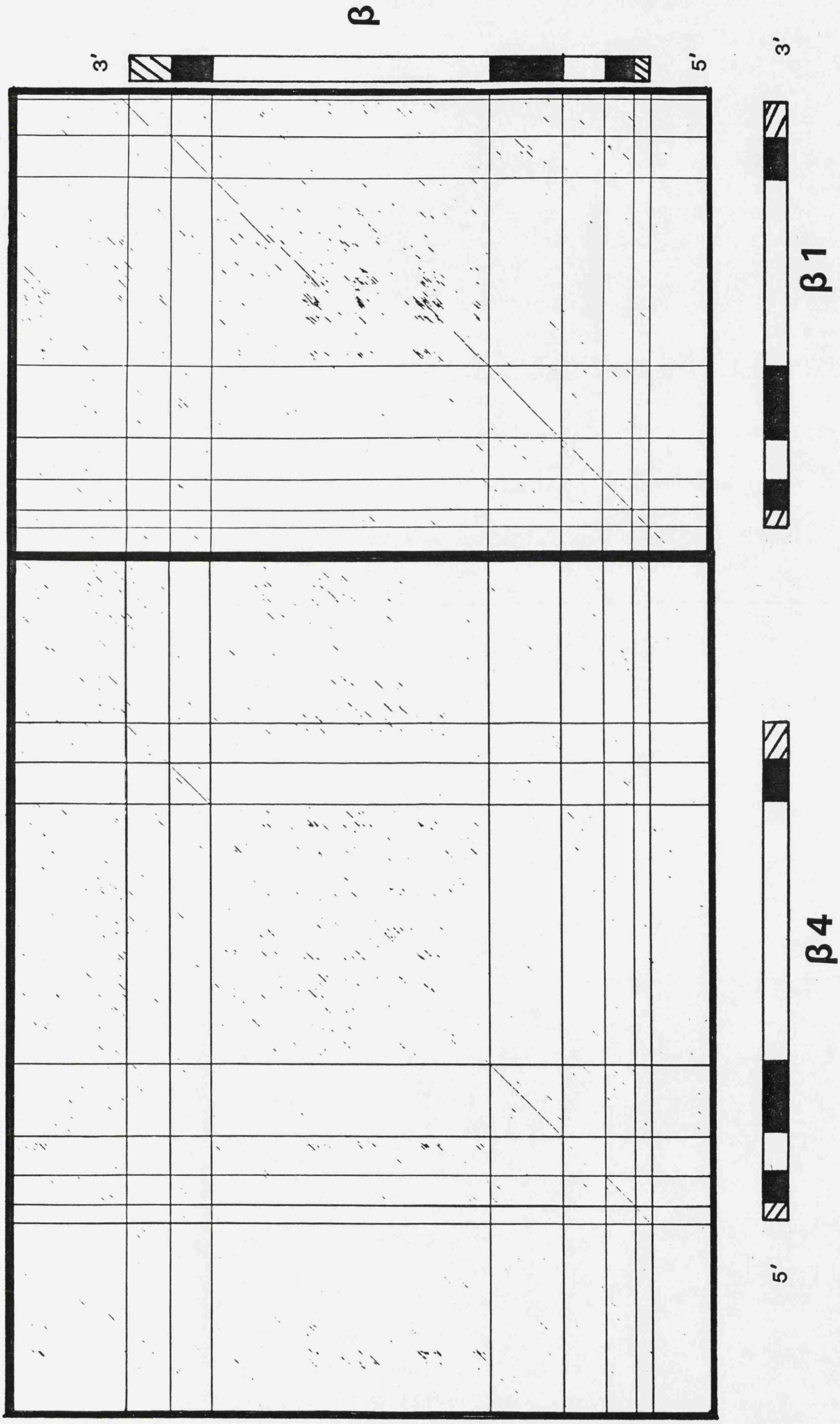

β
3'
5'
β1
3'
β4
5'

Figure 3.4

a) Isolation of human Ψβ1 DNA hybridisation probes.

The organisation of the human β-globin gene cluster is shown together with the location of the Ψβ1-containing clone λHγG4 isolated by Fritsch et al. (1980). Exons are shown by filled boxes and introns by open boxes. A 7 kb EcoRI fragment was isolated from λHγG4 DNA and cloned into pAT153 to give the subclone pHΨβ1. The detailed restriction endonuclease site map of the Ψβ1 globin pseudogene shows cleavage sites for restriction endonucleases BglII (b), EcoRI (E), MboII (M), Sau3A (S), and Xba I (X). DNA fragments containing the 5' flanking region (probe 1), intron 2 (probe 2) or essentially the entire gene (probe 3) were isolated by preparative gel electrophoresis of restriction digests of pHΨβ1 DNA (see text).

b) Hybridisation pattern observed for the pHΨβ1 derived probes against total human genomic DNA digested with EcoRI or BglII.

Samples of high molecular weight human genomic DNA (5μg) were digested with EcoRI or BglII and electrophoresed through a 0.5% horizontal agarose gel. After acid/alkaline denaturation DNA fragments were transferred to nitrocellulose filters by Southern blotting and hybridised with probe 1 (1), probe 2 (2) and the rabbit adult β-globin cDNA probe (R) labelled in vitro with ^{32}P. Hybridisations were carried out overnight in 1 x SSC (0.15M NaCl, 0.015M sodium citrate, pH 7.0) at 60°C in the presence of dextran sulphate. Autoradiographic exposures were for 5 days.

Genomic DNA fragments detected by the rabbit adult cDNA probe (R) have previously been assigned to specific human β-globin genes as labelled (Fritsch et al., 1980; Barrie et al., 1981). The open triangles correspond to the position of the additional faintly hybridising DNA fragments detected by probe 2, see Chapter 7.

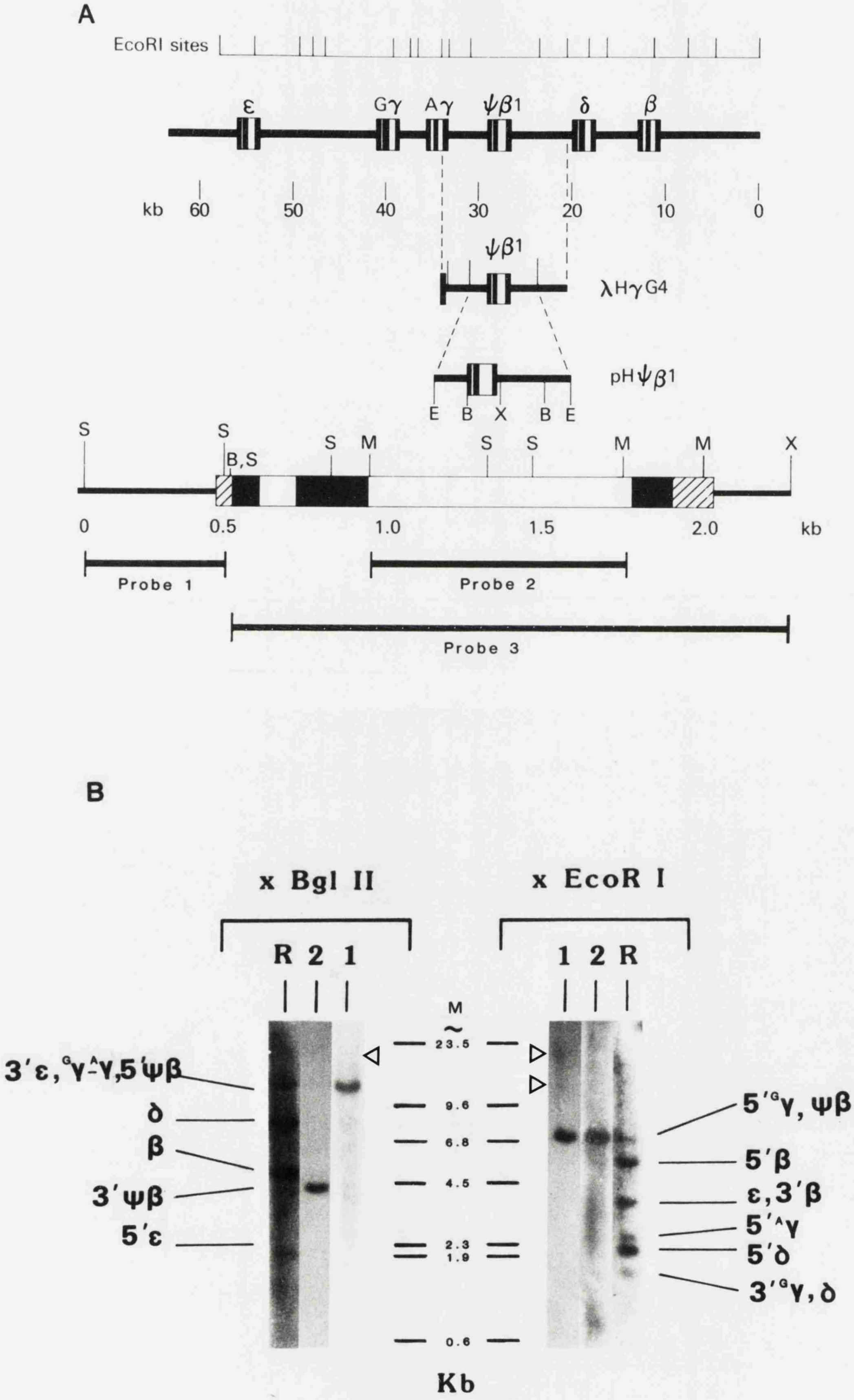

A
EcoRI sites
ε
Gγ
Aγ
ψβ1
δ
β
kb 60 50 40 30 20 10 0
ψβ1
λHγG4
pHψβ1
E B X B E
S S B,S S M S S M M X
0 0.5 1.0 1.5 2.0 kb
Probe 1
Probe 2
Probe 3
B
x Bgl II
x EcoR I
R 2 1
1 2 R
M
23.5
9.6
6.8
4.5
2.3
1.9
0.6
Kb
3′ε,ᴳγ-ᴬγ,5′ψβ
δ
β
3′ψβ
5′ε
5′ᴳγ, ψβ
5′β
ε,3′β
5′ᴬγ
5′δ
3′ᴳγ,δ

Figure 3.5

Detection of Ψβ1-related DNA fragments in human and primate genomic DNA digested with EcoRI or BglII.

Genomic DNA digests, electrophoresis, transfer to nitrocellulose filters by Southern blotting and hybridisations were as described in Figure 3.4(b). Molecular weight markers are λ x HindIII. Autoradiographic exposures were for 36 hours and one week for probe 1 and probe 2 respectively. The filled triangles indicate the position of faint or difficult to distinguish DNA fragments present more clearly on the original autoradiograph.

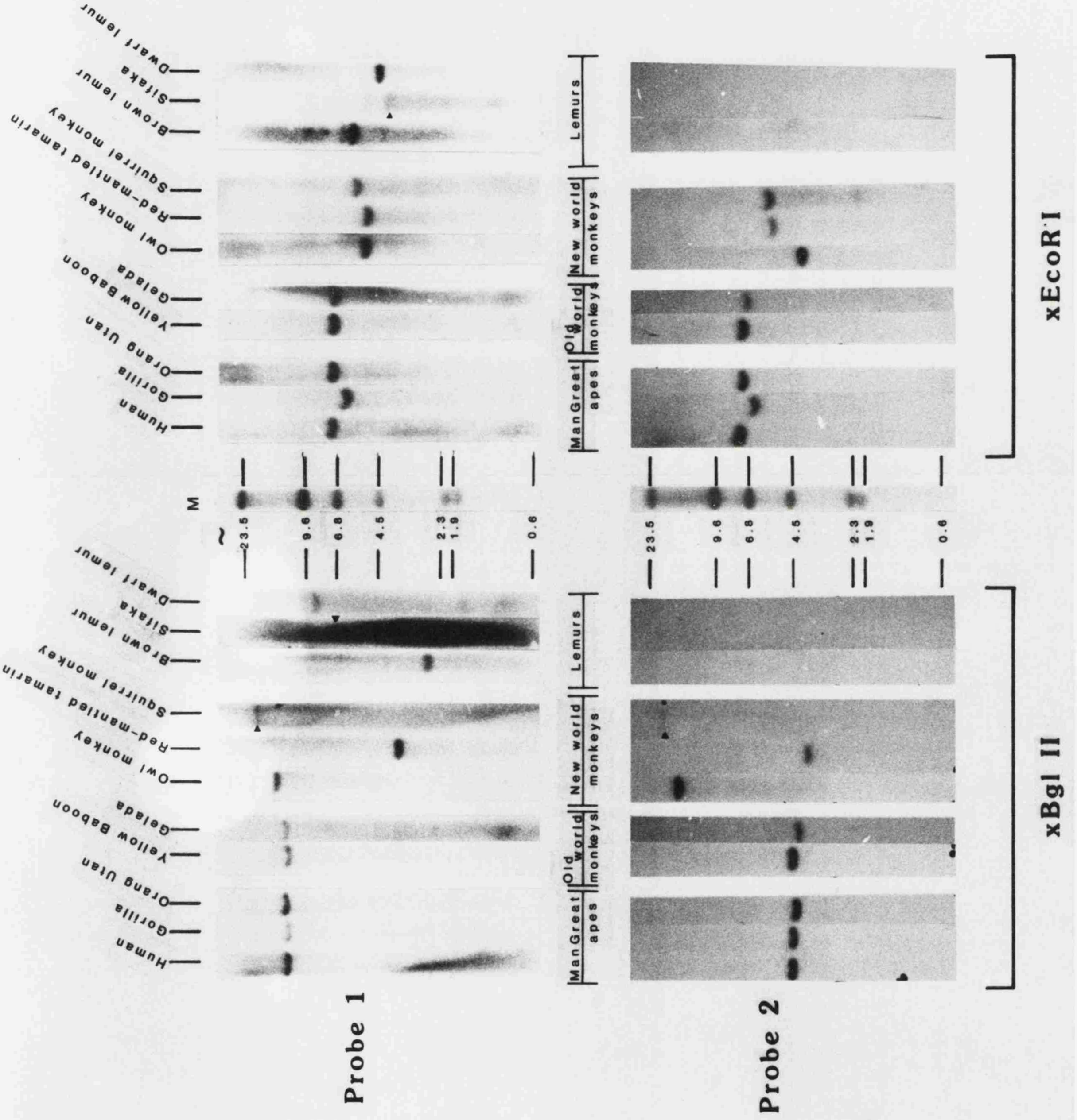

Probe 1
Probe 2
xBgl II
xEcoR I
M
23.5
9.6
6.8
4.5
2.3
1.9
0.6
Human
Gorilla
Orang Utan
Yellow Baboon
Gelada
Owl monkey
Red-mantled tamarin
Squirrel monkey
Brown lemur
Sifaka
Dwarf lemur
Man
Great apes
Old world monkeys
New world monkeys
Lemurs

Figure 3.6

Detection of Ψβ1-related sequences by pHΨβ1 derived probes in genomic DNA of lion, dog and seal digested with *Eco*RI or *Bgl*II.

Genomic DNA digests, electrophoresis, transfer to nitrocellulose filters by Southern blotting and hybridisations were as described in Figure 3.4(b). Molecular weight markers are λ x *Hin*dIII. Autoradiographic exposures were for 2 days. Open triangles indicate co-migrating DNA fragments that hybridise to probes 1 and 2 and which may therefore correspond to genuine Ψβ1-related sequences in these mammalian species.

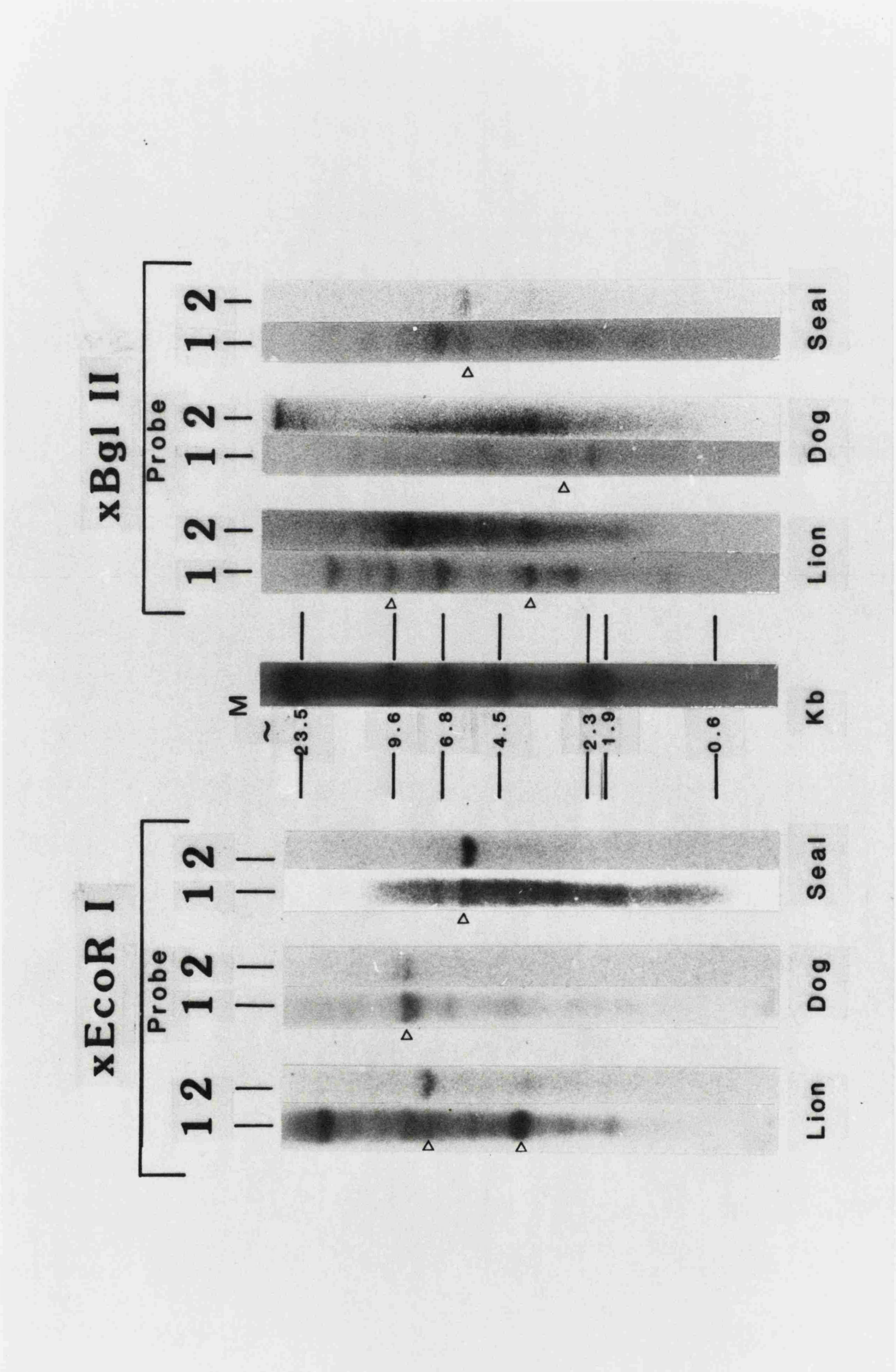
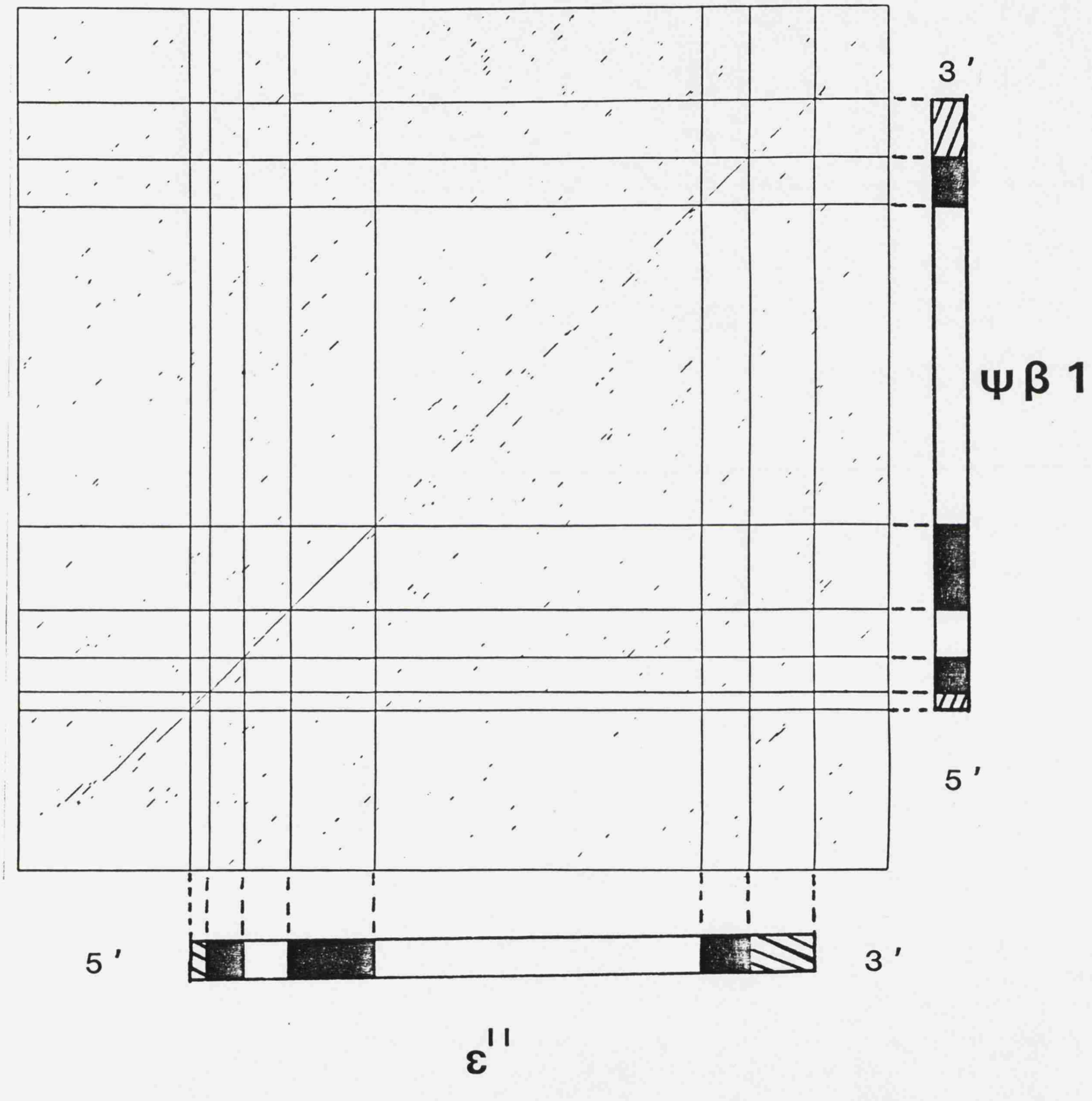

xEcoR I
Probe
1 2
1 2
1 2
Lion
Dog
Seal
M
~23.5
9.6
6.8
4.5
2.3
1.9
0.6
Kb
xBgl II
Probe
1 2
1 2
1 2
Lion
Dog
Seal

Figure 3.7

Dot-matrix comparison of the apparently functional goat ε^{II} globin gene and the human Ψβ1 globin pseudogene.

The matching criteria and sequence features depicted beside the grid correspond to those outlined in Figure 3.2. The goat ε^{II} sequence (2272 bp) is taken from Shapiro et al. (1983). The orthology of the two sequences is indicated by the homology detected over non-coding DNA sequences (see text).


3'
ψβ1
5'
5'
3'
ε''

Chapter 4

ANALYSIS OF THE OWL MONKEY β-GLOBIN GENE CLUSTER CONTAINING A Ψβ1-RELATED PSEUDOGENE

4.1 Introduction

An arrangement for the β-globin gene cluster of the owl monkey has been proposed (Barrie *et al*., 1981; Figure 4.1*), based on genomic mapping of restriction endonuclease cleavage sites both within and around structural genes and their surrounding DNA sequences using human DNA probes to discriminate between the different β-related sequences. The owl monkey β-globin gene family contains a single ε-, γ-, δ-, and β-like globin gene. While the δ and β genes were shown to be linked, linkage between the ε-γ was only provisional and there was no evidence for linkage between the γ and δ genes.

It is likely that the pattern of developmental gene expression in this globin gene cluster is similar to that in man. For example, New World monkeys, including the owl monkey, produce δ and β haemoglobins which form minor and major components of the adult haemoglobin (Boyer *et al*., 1971). Similarly, as the human ε gene and its orthologues in other mammals (mouse, rabbit and goat) are all expressed during embryonic development it is likely that the owl monkey ε-like gene is also expressed at this same stage of development. The absence of formal evidence concerning the period of expression of the single owl monkey γ-like gene means the probable foetal expression of this gene (as in man) remains

*Figures and Tables for this Chapter follow the text.

unclear, especially as the orthologues of the human γ gene in the mouse, rabbit and goat (see Hills _et_ _al_., 1984; Townes _et_ _al_., 1984) are expressed during embryonic development. The presence in one New World monkey, the marmoset, of a distinct foetal haemoglobin containing γ-chains (Huisman _et_ _al_., 1973) does however suggest that the owl monkey γ-globin may also be expressed during foetal development.

The probable presence of a complete ψβ1-related sequence in the owl monkey genome has been shown previously (see 3.6) but the location within the proposed β-globin gene cluster of this sequence was unknown. Similarly the functional status of this sequence was also unknown. It was proposed therefore to complete the characterisation of the owl monkey β-globin gene family with particular reference to the position and functional status of the ψβ1-related sequence by further genomic mapping; genomic cloning of the whole cluster and DNA sequencing.

4.2 Genomic mapping of the human ψβ1-related sequence in the owl monkey

Genomic restriction endonuclease site mapping of the owl monkey ψβ1-related sequence was performed to establish a) if this sequence lay within the proposed β-globin gene cluster and b) if so, whether its most likely position (between the γ and δ-related genes as in the higher primates) would help complete the owl monkey β-globin gene linkage map. 15μg samples of owl monkey genomic DNA were digested with single and pairwise combinations of the 8 restriction endonucleases used in previous genomic mapping studies (Barrie _et_ _al_., 1981; see Figure 4.2). The digested DNA was denatured with alkali and electrophoresed, in the single stranded form, through a 0.8% agarose gel, transferred to nitrocellulose by

Southern blotting and then filter hybridised, in turn, to the human Ψβ1 probes 1 and 2.

An example of the hybridisation pattern produced is shown in Figure 4.2(a) along with the restriction map produced around the owl monkey Ψβ1-related sequence, Figure 4.2(b). Hybridising fragments were sized by measurement of mobility relative to a set of DNA standards (λ-DNA x *Hind*III). The restriction endonuclease cleavage site map of the owl monkey Ψβ1-related sequence was orientated by the distinct nature of the fragments detected by the 5' (probe 1) and 3' (probe 2) Ψβ1 gene probes. This technique only detects the restriction cleavage site within or closest to the hybridising region detected by the specific probes.

Linkage to the γ gene

Several of the genomic restriction endonuclease cleavage sites within and around the owl monkey Ψβ1-related sequence appear to be shared in common with the genomic restriction endonuclease cleavage site map around the γ-related sequence (Barrie *et al.*, 1981), suggesting these two regions are linked. In order to confirm this gene linkage owl monkey genomic DNA was digested with two restriction endonucleases (*Bcl*I and *Eco*RI) which would be predicted to produce fragments containing both γ and Ψβ1 sequences if the genes are linked. Duplicate tracks containing 5μg of digested genomic DNA were electrophoresed, in the native double stranded form, through a 0.5% agarose gel for twice the normal distance to increase the resolution of the fragments (see 2.9). After electrophoresis the DNA was acid/alkali denatured *in situ* then transferred to nitrocellulose by Southern blotting. The duplicate halves of the filter were then hybridised to either a human γ or Ψβ1 ^{32}P-labelled probe.

Linkage between the two owl monkey globin gene sequences was confirmed (Figure 4.3(a)) by the presence of electrophoretically indistinguishable fragments detected by both the human γ and Ψβ1 (probe 1) DNA probes.

Linkage to the δ gene

The physical maps around the Ψβ1 and δ genes of the owl monkey share common restriction endonuclease cleavage sites suggesting that these regions are also linked. Linkage was confirmed (method as above) by detection of an electrophoretically indistinguishable large BglII restriction endonuclease fragment by both the human Ψβ1 gene probes and a rabbit adult β-globin DNA probe (PβG), Figure 4.3(b). PβG has previously been shown capable of detecting all the owl monkey β-like globin genes including δ; the δ gene was assigned to the large BglII fragment due to the relative position of other restriction endonuclease sites around this gene compared to those in the human β-globin gene cluster (Barrie et al., 1981).

The linkage of the owl monkey γ-Ψβ1 and Ψβ-δ sequences confirms the linkage arrangement previously proposed for this region of the β-globin gene cluster and places the Ψβ1-related sequence between the γ-δ genes as in the higher primates and man. The γ-Ψβ1-δ-β sequences in the owl monkey β-globin gene cluster are also orientated in the same direction (5'-3') and have similar intergenic distances to those observed in the higher primates and man. It is highly likely that the ε sequence is in a similar orientat~~ed~~ion though this remains provisional until linkage between ε and γ is confirmed.

4.3 Genomic cloning of the owl monkey β-globin gene cluster

The initial owl monkey genomic library was prepared by Dr P.A.Barrie by the following method. Size selected (11-20 kb) *Sau*3A partial digestion products of owl monkey genomic DNA were recovered from a preparative agarose gel using method 2.10(i). Similarly the two phage "arms" of the λ replacement vector λL47.1 (Leonen and Brammer, 1980) were isolated from the inessential central fragment, after complete digestion with the restriction endonuclease *Bam*HI, by recovery from a preparative agarose gel using method 2.10(i). Complementary *Sau*3A and *Bam*HI termini were ligated together in a reaction mix containing 4μg of size selected partials and 4μg of λL47.1 arms. 1μg of recombinant DNA was packaged *in vitro* and then infected into the *E.coli* strain ED8910. The cells were then distributed over four 9cm diameter BBL agar plates at a density of ≧200,000 p.f.u./plate.

A total of 1.25-2 x 10^6 plaques were screened for β-globin related sequences by the filter hybridisation method of Benton and Davis, (1977) using ^{32}P-labelled PβG and Ψβ1(probe 2) as DNA probes (2.15). Approximately 15 positively hybridising regions were detected of which 3 were further purified by at least 3 rounds of purification onto the *E.coli* strain ED8910 before amplification (2.16). Recombinant λ-phage DNA were prepared from liquid lysates before characterisation (2.17).

A physical restriction endonuclease site map was constructed for each recombinant by digestion with the four restriction endonucleases *Eco*R1, *Bam*H1, *Bgl*II and *Hind*III in single and double digests containing 0.5μg of recombinant phage DNA per digest. After digestion the DNAs were electrophoresed in the native double stranded form, photographed and then acid/alkali denatured before transfer to nitrocellulose by Southern blotting. DNA fragments containing sequences with homology to human globin sequences were visualised by filter hybridisation to ^{32}P-labelled probes (rabbit adult β-globin cDNA PβG and human Ψβ1 probes 1 and 2). The sizes of hybridising fragments were measured by relative mobility against known DNA standards (λ-DNA x HindIII).

The λ-recombinants could be positioned relative to the physical genomic restriction endonuclease cleavage site map by the similarity in arrangement of the restriction endonuclease cleavage sites within the insert DNA. Figure 4.4 illustrates the general approach to mapping λ-recombinants with λAT.1 as an example: a λ-recombinant containing Ψβ1, δ and part of the β-related sequences of the owl monkey. The λ-recombinants λAT.2 and λAT.3 both contain the owl monkey ε and γ-related sequences on similar sized genomic DNA fragments but in opposite orientations relative to the λL47.1 "arms". These recombinants confirm the linkage of the ε-γ sequences; the orientation of the ε gene is as suggested by the genomic mapping data (Barrie *et al*, 1981), but the intergenic distance between the two genes is much less than that between the ε and γ genes in the higher primates and man (see Discussion).

The complete linkage of ε-, γ-, Ψβ1, δ- and β-globin related sequences into a single cluster covering ~38 kb of owl monkey genomic DNA is confirmed by these λ-recombinants and the genomic mapping data presented in 4.2. The relationship of these clones to the owl monkey β-globin gene cluster is summarised in Figure 4.5. Recombinant fragment sizes were consistantly lower than those estimated by genomic mapping where the DNA was denatured before electrophoresis. This phenomenon has been observed by others (Baralle *et al.*, 1980) and may reflect the different amounts of DNA loaded under the different conditions and their mobilities.

4.4 Sequencing of the owl monkey Ψβ1-related sequence

In order to determine whether the owl monkey Ψβ1-related sequence is a pseudogene the region containing this sequence was subcloned as two restriction endonuclease fragments from λAT.1 into the plasmid pUC13 (Figure 4.5) prior to sequencing. The two recombinant plasmids (pAT.1.5 and pAT.1.~~3~~7) were each then sonicated to produce random DNA fragments which were end-repaired before ligation into the phosphatased blunt-ended *SmaI* restriction endonuclease site of the M13 vector M13mp8 (see 2.26 for standard protocols). Recombinant M13 DNA was transformed into the *E.coli* strain JM103 and after overnight growth on indicator plates "white" M13 recombinants were grown up for 6 hours then single-stranded DNA prepared (Weller *et al.*, (1984) and screened for sequences of interest by filter hybridisation using a ^{32}P-labelled Ψβ1 gene probe (probe 3, Figure 3.4). Positively hybridising M13 recombinants were sequenced by the M13-dideoxyribonuclease chain-termination method developed by Sangers' group (2.27). Sequences of individual clones were compared with the human Ψβ1 gene sequence using the programs developed by Staden (1980) on a PDP 11/44 minicomputer and melded into a complete sequence (Figure 4.6). The 2408 bp of sequence was determined from M13 recombinant clones from both strands.

4.4 The orthology and structural features of the owl monkey Ψβ1-related sequence

Dot-matrices performed between the owl monkey Ψβ1-related gene, all the human β-globin genes and the goat ε^{II} and β globin genes confirmed the orthology between this sequence, the human Ψβ1 pseudogene and the functional goat ε^{II} gene, Figure 4.7 (for clarity only dot matrices showing orthologies are presented). The 2408 bp of sequence encompassing the owl monkey Ψβ1-related gene is shown in Figure 4.6 aligned against the exons of the goat ε^{II} gene (see below). The sequence begins 405 bp 5' of the first base of the translation initiation codon and extends 577 bp 3' of the termination codon. The gene has the characteristic β-globin gene structural organisation of three exons interrupted by two introns, 120 bp and 875 bp long respectively, and is surrounded by signal sequences implicated in the transcription, mRNA maturation and translation of eukaryotic genes (though some of these appear to be abnormal, Table 4.1).

Alignment of the 5' and 3' flanking and exon regions of the owl monkey Ψβ1-related sequence against those of the functional goat ε^{II} gene allows the position of defects which could potentially silence this gene to be distinguished (Figure 4.6). Potential defects are numbered as encountered 5' to 3' and are described in detail in Table 4.1. It is not possible to determine which of these defects, if any, were responsible for the initial silencing of this gene.

4.5 Summary

The linkage arrangement of the owl monkey β-globin gene cluster has been confirmed by further genomic mapping and the isolation of a set of overlapping genomic λ-recombinants. While very similar in organisation to the β-globin gene cluster in higher primates and man in the region 3' of the γ gene this work confirms the presence of a single γ gene and a shorter intergenic ε-γ gene distance in this primate species. The gene family includes a ψβ1-related sequence at a similar position, between γ and δ globin genes, to that in higher primates and man. Sequencing of the owl monkey ψβ1 gene has shown that this gene has the archetypal β-globin gene organisation. Comparison of the owl monkey ψβ1-related sequence against its functional orthologue in the goat reveals several potential silencing defects that suggest this gene would not be expressed in the owl monkey and therefore constitutes a non-processed pseudogene orthologous to the pseudogene found in the β-globin gene cluster of man.

Figure 4.1

Restriction endonuclease cleavage site map of the owl monkey β-globin gene cluster, taken from Barrie (1982).

Cleavage sites shown are for restriction endonucleases *Bam*HI (B), *Bcl*I (Bc), *Bgl*II (Bg), *Eco*RI (E), *Hind*III (H), *Kpn*I (K), *Pst*I (P), and *Xba*I (X). The map shows only cleavage sites that generate β-globin DNA fragments; the direction of the gene detected relative to a mapped site is indicated by ▶, or ◀. Sites known to vary polymorphically are marked ±.

Linkage of the owl monkey ε, γ and β globin genes was not established although a probable ε-γ linakge is indicated by the aligned maps. The leftward β-globin gene was assigned to δ on the basis of residual map homology with man. Owl monkey cleavage sites probably identical to those in man are indicated by open circles. Sites present in the owl monkey but definitely not in man are shown by filled circles.

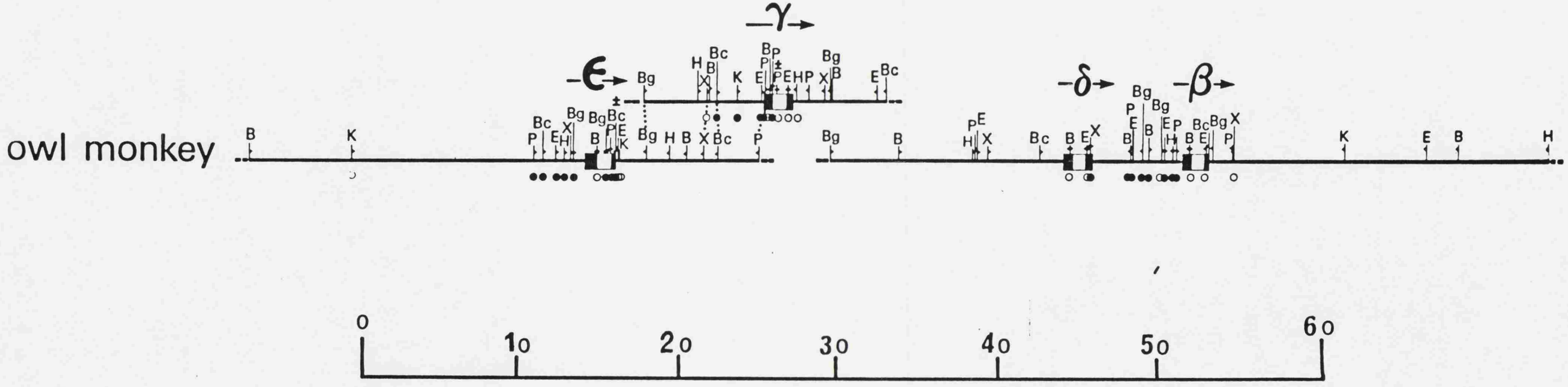

owl monkey
-ε→
-γ→
-δ→
-β→
B
K
P
Bc
E
X
H
Bg
±
0
10
20
30
40
50
60

Figure 4.2

Genomic mapping of the owl monkey Ψβ1 globin gene.

a) Example of the double digest strategy for restriction mapping of the Ψβ1 gene in owl monkey genomic DNA.

15μg per track of owl monkey genomic DNA was digested with the indicated restriction endonuclease (abbreviations as in Figure 4.1) and the DNA containing Ψβ1 globin gene sequences detected by hybridisation to ^{32}P-labelled Ψβ1 probes 1 and 2. The hybridisation pattern for probe 2 is shown. DNA digests were denatured with alkali prior to electrophoresis then transferred directly to nitrocellulose filters by Southern blotting. Hybridisations were as described in Figure 3.4. Autoradiograpghic exposures were for two days. Molecular weight markers were λ x HindIII. The filled triangles indicate faint hybridising fragments more readily apparent on the original autoradiograph.

b) Restriction endonuclease cleavage site map encompassing the owl monkey Ψβ1 globin gene.

Cleavage sites shown are the same as those mentioned in Figure 4.1. This map shows only those genomic cleavage sites that generate Ψβ1 globin gene fragments related to probe 1 and 2. The relative orientation of sites distinguished by both probes are indicated by ▶, a single probe by ◀. The hatched and filled boxes indicate the m~~ax~~inimum genomic DNA fragments detected by probe 1 and probe 2 respectively. This map is complete except for the location of the KpnI site 3' of the gene.

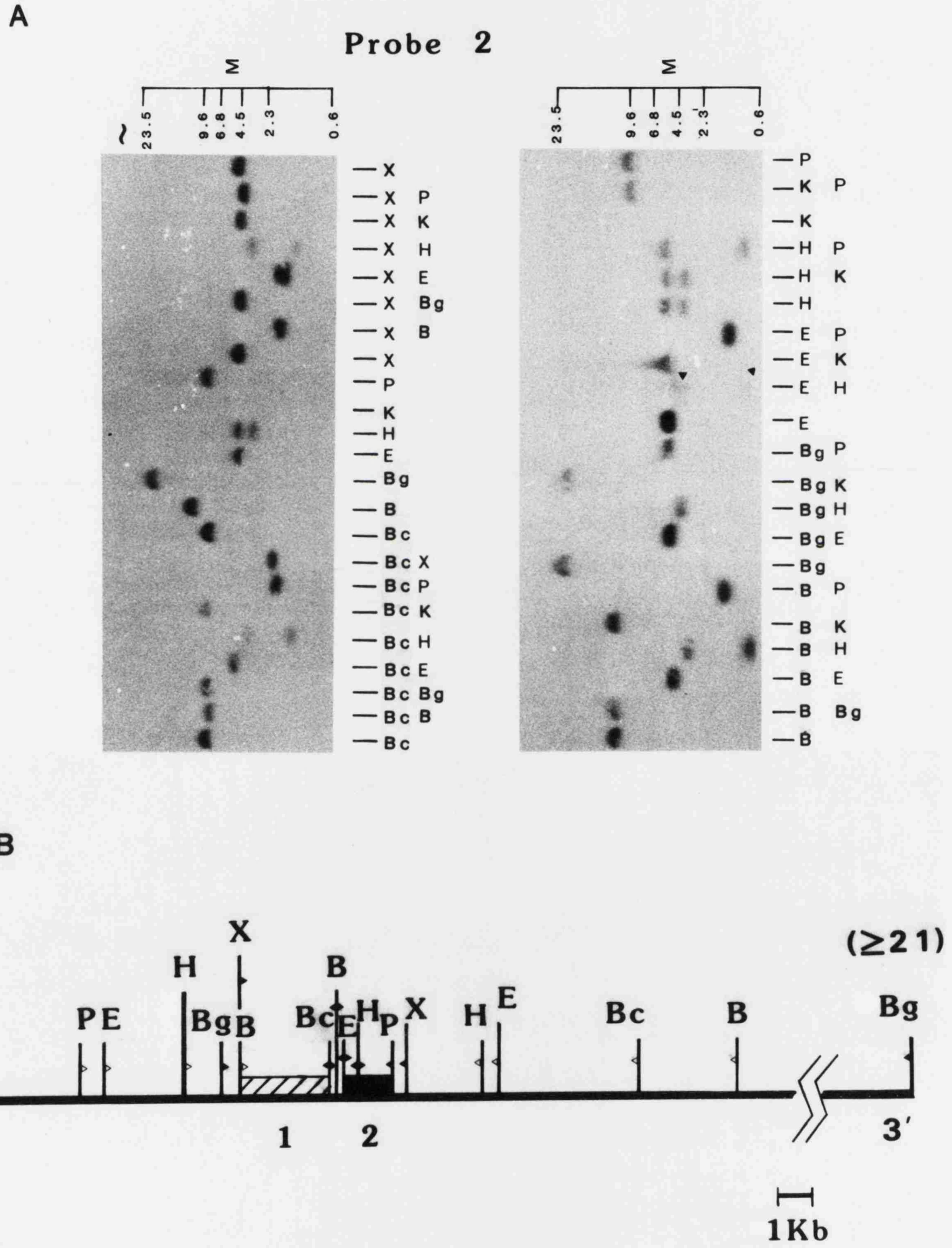

A
Probe 2
M
~
23.5
9.6
6.8
4.5
2.3
0.6
X
X P
X K
X H
X E
X Bg
X B
X
P
K
H
E
Bg
B
Bc
Bc X
Bc P
Bc K
Bc H
Bc E
Bc Bg
Bc B
Bc
M
23.5
9.6
6.8
4.5
2.3
0.6
P
K P
K
H P
H K
H
E P
E K
E H
E
Bg P
Bg K
Bg H
Bg E
Bg
B P
B K
B H
B E
B Bg
B
B
X
H
B
(≥21)
Bc
K
P
E
H
Bg
B
Bc
E
H
P
X
H
E
Bc
B
Bg
5′
1
2
3′
1Kb

Figure 4.3

Establishment of linkage of the owl monkey Ψβ1 gene to the γ- and δ-like globin genes.

The probable alignment of the restriction endonuclease cleavage site maps, established by genomic mapping, are shown for the owl monkey γ, Ψβ1 and δ/β globin genes. Abbreviations for the restriction endonuclease cleavage sites are the same as in Figure 4.1. The probable position of the genes are represented by the filled boxes (exons) and open boxes (introns). The restriction endonuclease cleavage sites used to establish linkage are indicated by the broken lines.

a) Linkage to the γ gene

Samples (10μg) of owl monkey genomic DNA were digested with EcoRI or BclI and electrophoresed on a 0.5% agarose gel. After acid/alkali denaturation in situ, DNA was transferred to nitrocellulose filters by Southern blotting and DNA fragments containing globin DNA sequences detected by hybridisation to ^{32}P-labelled Ψβ1 probe 1 (1) and a human γ- globin cDNA probe (Hγ). Hybridisations were performed overnight in 1 x SSC at 65°C in the presence of dextran sulphate. Autoradiographic exposures were for 1 week. Molecular weight markers are λ x HindIII. Open triangles indicate the position of faintly hybridising fragments present more clearly on the original autoradiograph. In both digests, the single DNA fragment detected by the Ψβ1 probe 1 is also detected by the human γ- globin cDNA thereby establishing linkage of these two sequences.

b) Linkage to the δ gene

Hybridisation filters of owl monkey genomic DNA samples digested with BglII were prepared as above. Hybridisation with ^{32}P-labelled Ψβ1 probes 1 and 2 and the rabbit adult β-globin cDNA (R) were performed as outlined in Figure 3.4. Autoradiographic exposure was for 2 days. Molecular weight markers are as above. The identity of DNA fragments detected by the rabbit cDNA probe and assigned to specific globin genes are shown (Barrie et al., 1981). The single DNA fragment detected by both Ψβ1 gene probes corresponds to that previously assigned to the δ-like globin gene of the owl monkey thereby establishing the linkage of these two sequences.

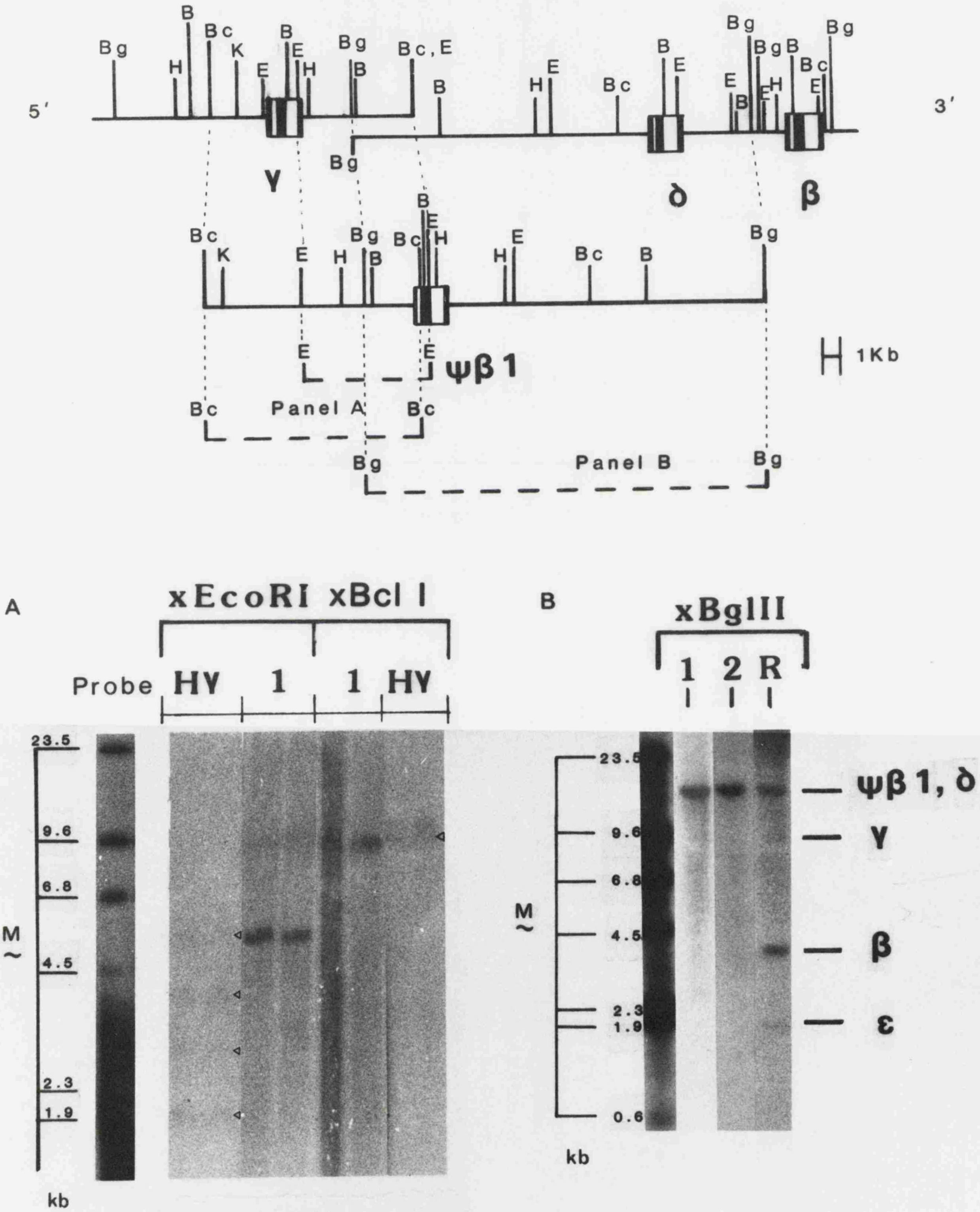

5′
3′
Bg
B
H
Bc
K
E
B
E
H
Bg
B
Bc, E
B
H
E
Bc
B
E
Bg
Bg
B
E
B
E
H
Bg
B
Bc
E
Bg
γ
δ
β
Bc
K
E
H
Bg
B
Bc
B
E
H
H
E
Bc
B
Bg
E
E
ψβ1
1Kb
Bc
Panel A
Bc
Bg
Panel B
Bg
A
xEcoRI xBcl I
Probe
Hγ
1
1
Hγ
23.5
9.6
6.8
4.5
2.3
1.9
M
~
kb
B
xBglII
1
2
R
23.5
9.6
6.8
4.5
2.3
1.9
0.6
M
~
kb
ψβ1, δ
γ
β
ε

Figure 4.4

Example of λ-recombinant characterisation

a) Characterisation of λAT.1

0.5 μg samples of λ-recombinant DNA were digested with the indicated restriction endonucleases (abbreviations as in Figure 4.1), electrophoresed on a 0.5% agarose gel and the gel photographed. DNA fragments were acid/alkali denatured *in situ* then transferred to nitrocellulose filters by Southern blotting. Fragments containing β-globin related sequences were detected by hybridisation with ^{32}P-labelled DNA probes from the human Ψβ1 gene and the rabbit adult β-globin cDNA. Hybridisations were overnight in 1 x SSC at 65°C in the absence of dextran sulphate. Autoradiographic exposure was overnight. Molecular weight markers are λ x *Hin*dIII. *Hin*dIII digestion was incomplete in several of the digests.

b) Restriction endonuclease cleavage site map of λAT.1

Restriction endonuclease cleavage site map of the recombinant λ phage λAT.1 containing a ~18.5 kb insert of owl monkey genomic DNA. Hybridisation analysis and alignment of this restriction endonuclease site map with the established genomic map of the owl monkey suggests this recombinant contains all of the owl monkey Ψβ1 and δ genes plus the 5' portion of the β globin gene. The approximate position of the exons and introns of these genes are represented by filled boxes and open boxes respectively. The horizontal bars labelled 1 and 2 indicate the m~~ax~~inimum ~~genomic~~ DNA fragments detected by probe 1 and probe 2 respectively.

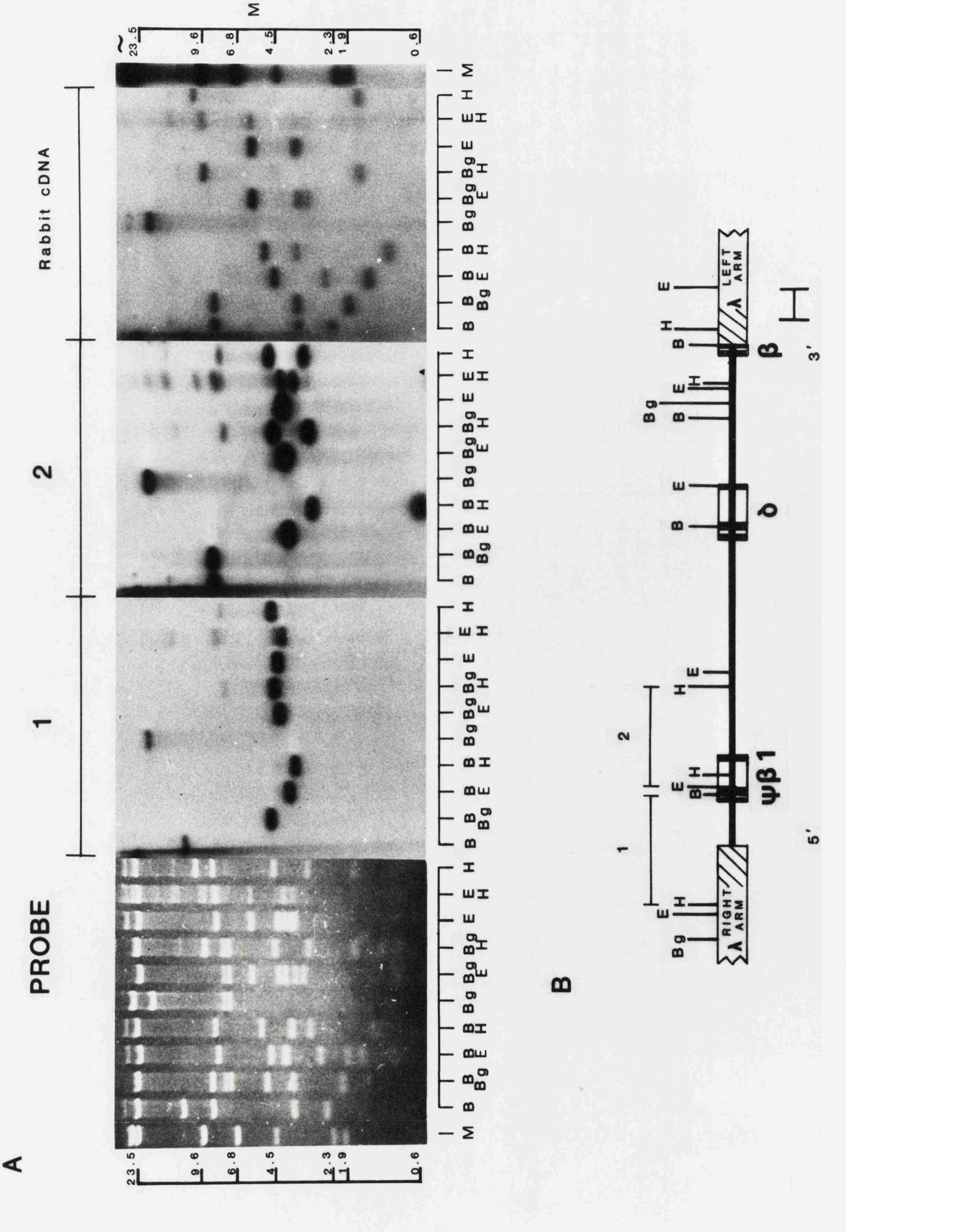

A
PROBE
1
2
Rabbit cDNA
23.5
9.6
6.8
4.5
2.3
1.9
0.6
M
M B BBg BE BH Bg BgE BgH E EH H
B
Bg E H
RIGHT ARM
λ
1
2
B E B H
H E
ψβ1
5'
B E
δ
Bg B E H
B H E
β
LEFT ARM
3'

Figure 4.5

Organisation of the β-globin gene cluster isolated from owl monkey (*Aotus trivirgatus*) genomic DNA

The relative positions of the three λ-recombinants isolated from a library of *Sau*3A partials of owl monkey DNA cloned into the bacteriophage vector λ47.1 are shown. λAT DNAs were isolated from the library by plaque hybridisation to ^{32}P-labelled Ψβ1 probe 2 and the rabbit adult β-globin cDNA (see 3.3/4). λ recombinants were characterised as outlined in Figure 4.4. Abbreviations for restriction endonuclease cleavage sites are as in Figure 4.1. Genes were located by hybridisation of λAT DNAs with rabbit adult β-globin cDNA and human Ψβ1 gene probes 1 and 2. ε, γ, δ, and β globin genes were identified by comparison of the restriction maps of λAT 1-3 with the genomic maps previously established from genomic mapping of the owl monkey β-globin genes with human ε, γ, Ψβ1 and β globin gene probes (see Figure 4.1/2).

The linkage between the γ globin gene and Ψβ1 was determined by hybridising Southern blots of *Eco*RI and *Bcl*I digests of owl monkey genomic DNA with human Ψβ1 and γ cDNA probes (see Figure 4.3). Location of *Eco*RI and *Bcl*I sites in λAT 1-3 DNA establishes that the γ-globin and Ψβ1 genes are separated by 4.7 kb of DNA, as shown.

The owl monkey Ψβ1 gene was further isolated by subcloning *Cla*I-*Hin*dIII and *Eco*RI fragments into pUC13 to give the subclones pAT.1.5 and pAT.1.7 respectively. Both plasmids were sheared, shotgun cloned in to M13mp8 and recombinant phage containing the Ψβ1 pseudogene were identified by hybridisation with the human Ψβ1 probe 3 (see Figure 3.4).

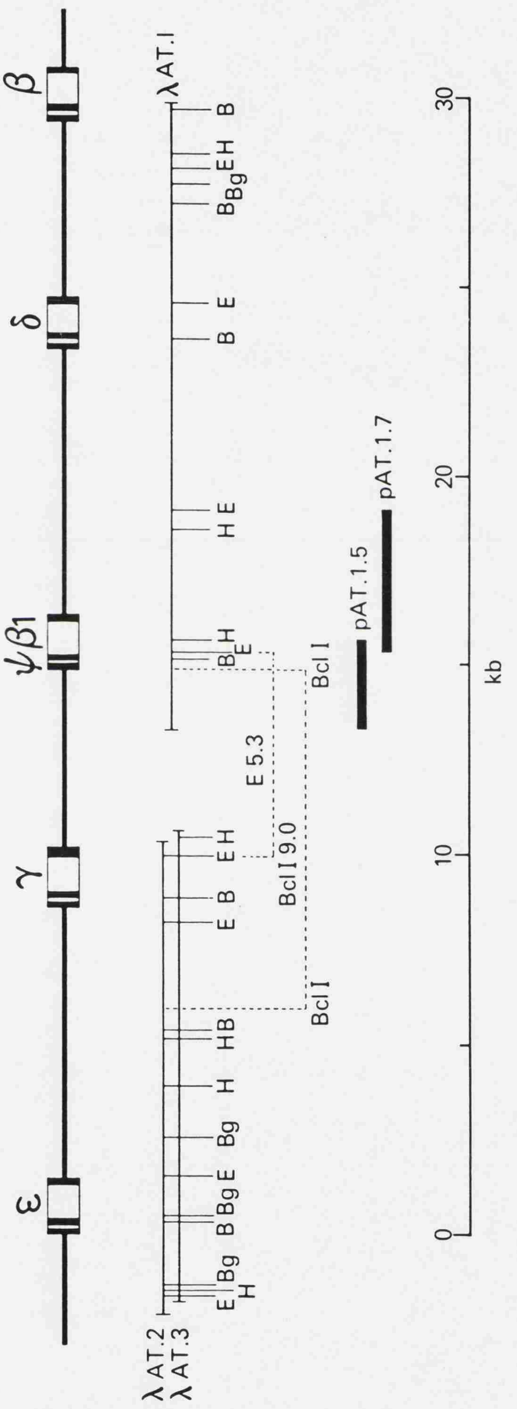

ε
γ
ψβ1
δ
β
λAT.1
λAT.2
λAT.3
E
H
Bg
B
Bg
E
Bg
H
H B
E B
E H
E 5.3
BclI 9.0
BclI
B
E
H
H E
B E
B
Bg
E H
B
BclI
pAT.1.5
pAT.1.7
0
10
20
30
kb

Figure 4.6

DNA sequence comparison of the owl monkey (Aotus) Ψβ1 pseudogene and the goat ε^{II} gene.

The sequence of the owl monkey Ψβ1 pseudogene (2408 bp) is aligned with the goat ε^{II} gene (Shapiro et al., 1983). Only differences between the two sequences are shown for the goat ε^{II} gene. A dash indicates the absence of a nucleotide in one sequence relative to the other. Sequences present in the mature mRNA are shown in uppercase letters. Homologues of sequences implicated in β-globin gene transcription, mRNA maturation and translation are indicated by bold underlined characters. The translated amino-acid sequence (numbered below the line) is that of the functional goat ε^{II} gene. Potential defects in the owl monkey Ψβ1 sequence are numbered above the sequence for reference (see also Table 4.1). In some instances the position of a microinsertion/deletion is ambiguous, within a few nucleotides, and the indicated position is therefore placed arbitarily within these limits (for example see defect 5).

```
                    10        20        30        40        50        60        70        80        90       100
AOTUS ψβ1 : tgattcttatccttttcagttctaacttactcctatttgtcagcattcaggttattaggggtcagtggtgatgaagaccttgagatataaactgtacatg
GOAT  ε"  : ----------------------------------------------------------------------------------------------------

AOTUS ψβ1 : gtggaggtagtggagaaaaatagatgggaaagaagagaagtttcaaattaagcctgaacagcaaagtccccttgagtggg-aacacagatgctatcagaa
GOAT  ε"  : ---------------------------------------------------------  t   g   at  c   ta a cc  cacc --        g

                                                      ----1---
AOTUS ψβ1 : actcgaatgtccatcttgc-aaaacttcttttgcctaaaccccacctttgg-agtcacaacccacccttgaccaatagattcatttcactgggagaggca
GOAT  ε"  :   a a              tg cc ac c-     c  gtt     cc   c   g  c   t g t           tc      t t    g a  -g

                                      ---2--
AOTUS ψβ1 : aagggc-tggggatcaa-agaggagaagtaaaaagccacaca-tgaagcagcagtgcAGACATGCTTCTGGCTCATCTGT-GATCACCAGGAAACTCCCA
GOAT  ε"  :       c     --  gc   t   g a     g    tg  g         g ca     T          C   - A G         T  G

                       -3-                                                ---4-      ---5--
AOTUS ψβ1 : GACCTGACACCGCGGTGCATTTCACTGCCGACGAAAGGGCTGCTGCCACTAGCCTATGGAGCTAGGTTGATGTG-AGAAGGCTGGAGGTGAGATTCTGGG
GOAT  ε"  :   -----    AT         T   A    G  G A        TTG    T  G  GC  A A  GA     G  GT   TC C       GC  A C
                       MetValHisPheThrThrGluGluLysAlaAlaValAlaSerLeuTrpAlaLysValAsnValGluValValGlyGlyGluSerLeuGl
                           1                          10                               20                     30

AOTUS ψβ1 : CAGgtagctactagaagc---cagagcaag-gtgcagaaaggcagaaagtgttcctg-aaagagggattagccagttgtcttacatattctgactttgca
GOAT  ε"  : a     agc g g  ca aggt  gag g a   t c      t           a a         c g tt  g t         c        ctt
            yAr

                                                                               -6-
AOTUS ψβ1 : tctgctctttgattatgattatcccacagTCTCCTGGTTGTCTACCCATGGATCCAGAAGTACTT-GAAAGTCTTGGCAATCTGGGCTCTGACTGTGTAA
GOAT  ε"  :     t   g   c      c      t  G  T   A C             C     G  T   T  C   T    T  CT AT       G C  CC
                                         gLeuLeuIleValTyrProTrpThrGlnArgPhePheAspSerPheGlyAsnLeuCysSerGluSerAlaI
                                                                  40                            50

              --7--                      -8-                                                   --9---
AOTUS ψβ1 : TAATGG-CAACCCCAAAGTCAAGGCACAT-GCAAGAAGGTGCTGATCTCCTTCAGAAAAGCTGTTATGCTCATGGATGCTCATAGTAAAGGCACCTTTGC
GOAT  ε"  :       G         G        C  CG   G           A      TG    T  CA   A  A        A TC  --- G
            LeMetGlyAsnProLysValLysAlaHisGlyArgLysValLeuAsnSerPheGlyAsnAlaIleLysHisMetAspAspLeuL---ysGlyThrPheAl
                              60                            70                         80

                                                                   10
AOTUS ψβ1 : TATACTGAGTGACCTGCACTGTGACAAGCTACACATGGACCCTGAGAACTTCCTTgagaattctaagcacactcatgctttcttccttccccttaaatat
GOAT  ε"  : AGAT  A  C  G               T G   G    T   CC       AGG t  g    gg t tg ctg      g   t   at -c gg t
            aAspLeuSerGluLeuHisCysAspLysLeuHisValAspProProAsnPheArg
                       90                            100

AOTUS ψβ1 : ttgcacgatggctacttttgaaagcaga--ggtggct-ttctcttgtgctat--gagtcagctgtgggatatgatatttcagtgttgggatagattttga
GOAT  ε"  : ca  tg tg  ----   a a t -   acac  a  caag  aa  g tga - tt t t       gg a          a  aaa c t   c   g

AOTUS ψβ1 : gagttatgtt-----ggtccaaatagcat-gcctaa--aatttggtagagtaaggactacgaatagtggaaggccacttaccatttgatagctctgaaaaac
GOAT  ε"  : t  g -    ccaga    -   -- gt  a a c  gt   ---------- ctc   t  cc      -  t    ggt   c     at        -

AOTUS ψβ1 : acatcttataaaaaat-tctgg--ccagaattgaa-atgag---tgtttgt-gg-atgagggaacaaaagttgaggtagagaaaataacatcttgccttt
GOAT  ε"  : ------- g       g gaa aa acc  cca   cc    cta   g t c cc  - a   t c  acc    c t c g  g  t

AOTUS ψβ1 : ggtcactgaaattttctgtgaaaattaat----agacacttttctgcctagtcctggaggttagaaaaaga-taacctagaacagagtaatgggaagctt
GOAT  ε"  :  a a         cc  a   g  cc   gaat  t tt c   a  --- t tt agtt   a gg g  c   a a- gg   a a  - a c a

AOTUS ψβ1 : ttaaaaa---gaagattgtt--tccctct-gaatgatgatgatatgcttttgtacacatgttacaggattttttgttatgagtgtttgcaaaaattgtgt
GOAT  ε"  :   c c  tgt tgcc  c  ctctca   ct gca        a   t      t   a g g t      cc- cgtca a  ga  tcc   caa c

AOTUS ψβ1 : gtgtg-----------tgtgataactgggaacttatcctatcccttactgttccttgaagtactattatcctactttttaaaaggatggagcctttaaaa
GOAT  ε"  :  aa  tatgtaatatca -   g  ----   ta  - a    ---  g cc-    tc  c    a  gtt      -   gta      ---------

AOTUS ψβ1 : aggtgaaacaattagtcacaatgtgctgtgg-taatgagttggcatagcaagtaagagaaggataggacacaatggaaggtgcagggctggcagtcatat
GOAT  ε"  :   cg ---    cca g g caa    ca  c tgaac  c a  g  a  a  gg        gta     g   g-   t  -      g     --c

AOTUS ψβ1 : tgaagctgatgtctagcccataatggtgtacttacctttgtgagaataagactctgagagttgctcaaactcttgtcaaaagaaaataagtgttgtatt
GOAT  ε"  :     t ca   c ag t tc        c    ----     g   c t g ca   a  c      cc     a g    g  ccggc -        c

AOTUS ψβ1 : tatttat-tgcaagtccagcttgaggcctgttattcactatgtaccatttcctttttatcttcactccctccccagTTCTTAGGCAATGTGATATTGATG
GOAT  ε"  :      gct  g   gg    c    a   tc c    t       tg c t        t-----------     C  C    A  CA          T
                                                                                        LeuLeuGlyAsnMetIleLeuIle
                                                                                                            110

                                                                              -------11-------
AOTUS ψβ1 : GTTTTGGCAGCCCACTTCAGCGAGGAGTTTACCCTACAGATACAGGCTTCCGGGCAAAGACCAATAAA------------TGCCATGGCCCACAATTATC
GOAT  ε"  :   C      A           A    A       CG     G      G TT   G  AG TG CC CGCTGTGGCTAA   TC          G  C
            ValLeuAlaThrHisPheSerLysGluPheThrProGlnMetGlnAlaAlaTrpGlnLysLeuThrAsnAlaValAlaAsnAlaLeuAlaHisLysTyrH
                                  120                           130                           140

              -12-
AOTUS ψβ1 : ACTGAGCTCCCGGTCCACTGTGTGTGTACCTA-CTGGTCCGCCATGTTTGTACCCATATCCTAAAAGCTCATCTCCTTTAGATGGAGGATGTTGGGGAGA
GOAT  ε"  :    -  T G  T G  ----------    ATG     G CTATCTGAA GC  AG G   C G   T         GA   CA    -----------G
            is***

AOTUS ψβ1 : AGAGCAATATCCTGCCTGCAGATTCAGCTCCTGCATGATAAAAATAAAATAAAGAAAT-GTGCTGTTAAAGAAATATCagtatatattttttctgtcttt
GOAT  ε"  :       GCT T    AGA --  AT A    AAT  A   C    G       -- G CA  G C AT  TG   GC ct g   t  c c t   t

AOTUS ψβ1 : atgttttat--gttgattcagccaaaaggatgcaccatttctgatggaaatgggaacactggagaatgagagtttaagaatagagaagagactttcttgcaa
GOAT  ε"  : gta      aa     n  ctat ------------------------------------------------------------------------------

AOTUS ψβ1 : tcctgaaagataagagagaacttgtgggtggatttagtggggtagctactcctgcaaaggggaggtcatctctagaataatacaatgtctttaaagaaag
GOAT  ε"  : ------------------------- a   ac  c    -     t  a  t ag tg   a t  a a   c   gatgg    g       g a g

AOTUS ψβ1 : ggaggggaatggaggtactcttgaaagatgtaagaggattgttgatagtgtgtaaagataagttaggactcaaattaaaaattctgtacatgttattatt
GOAT  ε"  :  a a ac-      t      a  g     c  a t     c a      a ct-------        a    g  -gggac     t gcc g

AOTUS ψβ1 : tgtatgaactcaggatacagctcatttggtgactgtggttcacttctacttattttaaacaacatatttttatcatttataa
GOAT  ε"  :         - gt     g    tt  c acc   t ta--------------------------------------------
```

Figure 4.7

Dot-matrix comparison of the owl monkey (Aotus) pseudogene.

The 2408 bp of the owl monkey Ψβ1 pseudogene is shown compared to the human Ψβ1 pseudogene (Chang and Slightom, 1984; Weissman per.comm.) and the goat ε^{II} gene (Shapiro *et al.*, 1983). Matching criteria and sequence features depicted beside the grid correspond to those outlined in Figure 3.2. The extensive homology that exists over the non-coding DNA sequences (5', 3' flanking and intron regions) in these comparisons strongly suggests these sequences are orthologously related. No such homology was detected in other comparisons between the owl monkey pseudogene and other human β-like globin genes (results not shown).

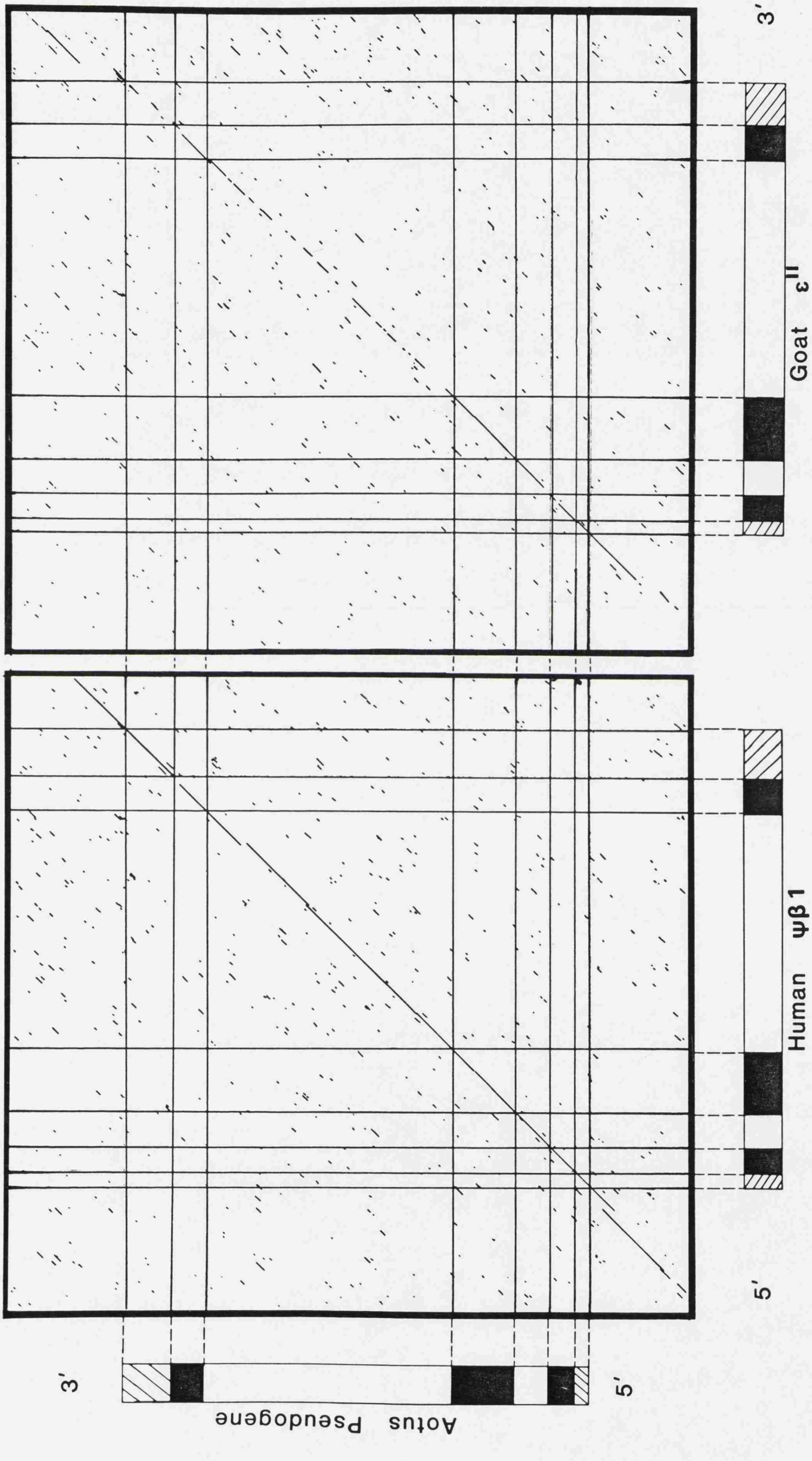

3'
Aotus Pseudogene
5'
5'
Human ψβ1
3'
Goat ε^II

Table 4.1

Analysis of potential defects in the owl monkey Ψβ1 gene.

Given the alignment of the owl monkey Ψβ1 pseudogene and the goat ε^{II} gene shown in Figure 4.6, the coding sequences and 5' and 3' flanking sequences implicated in eukaryotic gene expression were examined for potential defects. The potential consequence of each defect (numbers in the body of Table 4.1 correspond to those in Figure 4.6) was examined in isolation as the order in which the defects were accumulated during evolution is unknown, with the possible exception of the primary silencing defect (see Discussion). Confirmation that, in isolation, specific signal sequence defects (1, 2, 3) would affect correct developmental expression of this gene remains to be tested.

Table 4.1

Number	Position in sequence	Potential defect	Consequence of defect
1	'CACCCCTG'	'CACCTTTG'	Alteration in what may constitute a signal sequence specifying embryonic developmental expression (Shapiro et al.,1983).
2	'ATA'-box	ATA - GTA	Unknown effect on signal sequence implicated in the correct initiation of transcription of many eukaryotic genes.
3	Initiation codon	ATG - GCG	Absent translation initiation.
4	Codon 17	AAA - TAG	Premature translation termination.
5	Codon 20/21	Deletion (-1)	Frameshift resulting in premature translation termination at an in frame TAA codon in exon 2.
6	Codon 42	Deletion (-1)	Frameshift resulting in premature translation termination at an in frame TAA codon ~32 bp downstream.
7	Codon 55/56	Deletion (-1)	Frameshift resulting in premature translation termination at an in frame TAG codon ~74 bp downstream.
8	Codon 64	Deletion (-1)	Frameshift resulting in premature translation termination at an in frame TGA codon ~13 bp downstream.
9	Codon 83	Insertion (+3)	Frameshift resulting in aberrant amino-acid sequence till the end of exon 2, no effect on exon 3.
10	Exon 2/Intron 2	GT - GA	Absent or incorrect mRNA splicing (however see Fischer et al.,1984).
11	Codon 135 - 139	Deletion (-12)	Absence of three amino-acids, in particular Val (137) associated with heme contact (Goodman et al.,1981), may affect overall structure of polypeptide.
12	Termination codon	Insert (+1)	No probable effect other than to alter termination codon from TAG to TGA.

Chapter 5

ANALYSIS OF THE HUMAN ΨB1-RELATED SEQUENCE OF THE PROSIMIAN BROWN LEMUR

5.1 Introduction

Of the prosimians the β-globin gene cluster of a lemur, the brown lemur, is the best characterised (Figure 5.1*). Gene orthologies and a physical restriction endonuclease cleavage site map of the cluster have been deduced by filter hybridisation of brown lemur genomic DNA digests with human ε-, γ-, and β-globin gene specific probes (Barrie et al., 1981). The gene family forms a linked cluster ≈20 kb long containing a single ε-, γ- and β-like globin gene plus a peculiar hybrid gene "Ψβ" composed of a segment of sequence closely homologous (by hybridisation) to the 3' end of the human β-globin gene preceded by sequences only detected by hybridisation with the 5' end of the human ε-globin gene (Barrie et al., 1981). The origins and the functional status of this peculiar "Ψβ" gene are unclear. This overall cluster arrangement is probably representative of the lemurs as a group as this cluster arrangement is thought to be identical in another lemur species, the ruffed lemur (Barrie et al., 1981).

Little is known concerning globin gene expression in the lemurs. The adult brown lemur has been shown to produce a β haemoglobin chain (Huisman et al., 1973) which presumably results from the expression of the

*Figures and Tables for this Chapter follow the text.

3' β-globin gene of this cluster (see Chapter 6). Similarly it is assumed, though not proven, that the brown lemur expresses an embryonic haemoglobin chain from the ε-like globin gene, as is the case for other mammals. The period of developmental expression of the γ-like haemoglobin chain, at either the foetal or embryonic stage as in the higher primates or the other mammals respectively, is unclear (see Discussion).

Interestingly, the lemurs, unlike the other simians examined, do not hybridise to both of the human Ψβ1 gene probes (see 3.6). The genomic hybridisation results suggest the presence within lemurs of sequences with homology to 5' (probe 1) but not 3' (probe 2) regions of the human Ψβ1 gene. Furthermore the sizes of the hybridising fragments detected by Ψβ1 probe 1 in the brown lemur apparently correspond to those from the 5' region of the brown lemur "Ψβ" gene (see below). The possibility exists therefore that the brown lemur β-globin gene family may contain sequences related to the human Ψβ1 pseudogene. As contemporary lemurs represent the most ancient of the primate groups, having diverged from simians ≥70 MYs ago, the nature of the Ψβ1-related sequence in this gene cluster is of considerable interest and ~~as~~ has therefore been further characterised by genomic Southern blot analysis and the sequencing of λ-recombinants that were available containing the brown lemur "Ψβ" sequence.

5.2 Human Ψβ1-related sequences correspond to the 5' region of the brown lemur "Ψβ" gene

Previous hybridisation analysis, using human ε-, γ-, and β-globin gene probes had suggested that the hybrid brown lemur "Ψβ" gene was composed of 3' β-related sequences preceded by 5' ε-related sequences. The switch in homology from 3' β-like to 5' ε-like sequence was mapped by

Southern blot analysis to the region of the intergenic BamHI site within the gene (Barrie et al., 1981). As mentioned above, when compared against the size of restriction endonuclease fragments predicted from the characterised brown lemur β-globin gene cluster, the size of genomic BglII and EcoRI DNA fragments detected by Ψβ1 probe 1 suggested this probe hybridised to a fragment from the "Ψβ" region of the cluster.

In order to confirm the observation that the restriction fragment from the 5' region of the brown lemur "Ψβ" gene most likely corresponded to Ψβ1-related sequences the filters used in Figure 3.6 were washed several times in H_2O for 30 minutes at 65°C (to remove previous ^{32}P-labelled probe) then rehybridised with a ^{32}P-labelled rabbit adult β-globin gene probe (see 3.4). This rabbit adult β-globin gene probe has previously been shown to be capable of detecting all of the genomic restriction endonuclease fragments containing β-related globin genes in the brown lemur (Barrie et al., 1981).

Comparison of the relative mobility of the hybridising brown lemur fragment detected using Ψβ1 probe 1, against the mobility of hybridising fragments of previously proposed orthology, as detected by the rabbit cDNA probe (Barrie et al., 1981), shows that the fragment detected by the human Ψβ1 probe 1 is in fact electrophoretically indistinguishable from that known to contain the "Ψβ" gene (Figure 5.2). This suggests therefore that the fragment detected by the Ψβ1 probe 1 corresponds to that containing the "Ψβ" gene. Furthermore, the relative strength of the hybridisation signal obtained using the human Ψβ1 5' probe suggests that the 5' region of the hybrid "Ψβ" gene corresponds to human Ψβ1-related sequences rather than human ε-related sequences. The lack of detectable homology to the 3' human Ψβ1 probe (probe 2) in brown lemur genomic DNA is consistant with the absence of these sequences in the brown lemur "Ψβ" gene, the region 3' of the intergenic BamHI cleavage site being β-like. In order to determine the exact nature of this hybrid gene the brown lemur "Ψβ" gene has been isolated and characterised by sequencing.

5.3 Sequencing of the brown lemur "Ψβ" hybrid gene

This work was initiated by Dr P.A.Barrie and Dr A.J.Jeffreys. They isolated the complete β-globin gene cluster of the brown lemur as a set of 5 overlapping recombinants from a λL47.1 genomic library (Barrie, 1982). The "Ψβ" gene was then subcloned from the λ-recombinant λBL9 as two overlapping restriction endonuclease fragments into suitable cloning sites of the plasmid vector pAT153. Prior to sequencing a detailed restriction endonuclease cleavage map was constructed for each plasmid (pBL9.1 and pBL9.8) by partial digestion of uniquely ^{32}P end-labelled restriction endonuclease fragments covering the region of the "Ψβ" gene, after the method of Smith and Birnsteil, 1976. These steps are summarised in Figure 5.3 and my own contribution to the sequencing stage of this project are illustrated as part of Figure 5.4 which shows the sequencing stategy employed.

The general approach to sequencing the gene was as follows. ~10μg of plasmid DNA (or a specific large restriction endonuclease fragment thereof) was digested with the desired restriction endonuclease. The complete digest was recovered and fragments end-labelled with a) a suitable α-^{32}P-dNTP (fill-in reaction) or b) after first having removed the 5' terminal phosphate group with alkaline phosphatase, with γ-^{32}P-ATP (kinase reaction). After recovery, the DNA was digested with a second restriction endonuclease chosen to produce asymmetric uniquely end-labelled DNA fragments and electrophoresed on an agarose gel to resolve the products. Specific end-labelled fragments were recovered from preparative agarose gels by method 2.10(i) or 2.10(iii) and aliquots subjected to the five chemical degradation reactions (G, G+A, C, C+T, A>C) exactly as described by Maxam and Gilbert (1977, 1980).

The five reactions for each substrate were electrophoresed through 40cm 8% or 6% polyacrylamide sequencing gels at ~1500V for 5-8 hours. Three loadings were performed at 90 minute intervals; the reaction tubes being heated at 90°C for 3 minutes to denature the DNA before each loading. After electrophoresis glass plates were separated and the gel covered with aluminium foil and autoradiographed, with or without a screen, at -80°C for up to 2 weeks. All of the 5' and 3' non-translated, exonic and >95% of non-coding DNA sequences were determined on both DNA strands.

5.4 The structure and orthology of the "Ψβ" gene of the brown lemur

The 2105 bp of sequence encompassing the brown lemur "Ψβ" gene extends 495 bp 5' of the initiation codon and 268 bp 3' of the termination codon. The gene has the characteristic β-globin gene organisation of three exons and two introns, of 118 bp and 778 bp long respectively, and has associated 5' and 3' transcription and translation signal sequences implicated in eukaryotic gene expression as well as the exon-intron boundary (GT.AG) signals involved in mRNA maturation, except for the first exon-intron boundary (see Table 5.1). Alignment of the exons of the brown lemur "Ψβ" gene against those of the functional human β-globin gene show that this gene has several potential silencing defects (Figure 5.5 and Table 5.1), four of which are codon defects and the other three defects which may affect transcription, mRNA maturation and translation. This gene is unlikely to be expressed in the brown lemur and therefore constitutes a non-processed pseudogene in this species.

As mentioned, the hybridisation data suggested the brown lemur "Ψβ" gene was composed of sequence with regions of homology to several human β-globin gene probes. Comparison of the brown lemur "Ψβ" gene sequence against each of the published human ε-, γ-, Ψβ1-, δ- and β-globin gene sequences was therefore performed by dot-matrix analysis as a means of obtaining an unbiased representation of the orthology of this gene. None of the dot-matrices between the brown lemur "Ψβ" gene and the human ε-, γ-, or β-globin genes gave any clear indication of orthology over the diagnostic non-coding (flanking or intron) regions. In contrast, these regions of the "Ψβ" gene gave strong indications of alignment with the human Ψβ1 and δ gene sequences (Figure 5.6); for clarity only the dot-matrices which resulted in the strong alignments between the brown lemur "Ψβ" sequence and human Ψβ1 and δ sequences are shown in Figure 5.6.

Different regions of the two human Ψβ1 and δ genes gave good alignments against the brown lemur "Ψβ" gene sequence. Sequences 5' of a point within the 2nd exon of the brown lemur "Ψβ" gene exhibited strongest homology to 5' human Ψβ1 sequences while sequences 3' of this same region exhibited strongest homology to 3' human δ sequences (this alignment being particularly striking over intron 2), Figure 5.6.

The switch from human Ψβ1-like to human δ-like was located more precisely within the second exon by examination of the aligned sequences themselves over this region, Figure 5.7. The asterisks in Figure 5.7 indicate a position held in common between the brown lemur "Ψβ" sequence and one or other, but not both, of the human Ψβ1 or δ sequences. As can be seen the brown lemur "Ψβ" exon two sequence matches predominantly human Ψβ1-like sequences 5' to a position corresponding approximately to codon 86-87 and with human δ-like sequences 3' to this position. This bias in sequence match is particularly obvious on entering the non-coding intronic regions shown in Figure 5.7 and supports the dot-matrix analysis result shown in Figure 5.6.

5.5 Summary

The brown lemur "Ψβ" gene is a β-globin related gene with orthology to two members of the human β-globin gene cluster, the Ψβ1 and δ genes. The switch of orthology within the gene occurs in exon 2 of the sequence at, or near, the position corresponding to codon 86-87. This gene is therefore a hybrid Ψβ1-δ gene. Several potential silencing defects can be discerned in this sequence suggesting this gene is not expressed and therefore corresponds to a non-processed pseudogene.

Figure 5.1

Restriction endonuclease cleavage site map of the brown lemur β-globin gene cluster, taken from Barrie (1982).

Cleavage sites shown are for restriction endonucleases BamHI (B), BclI (Bc), BglII (Bg), EcoRI (E), HindIII (H), KpnI (K), PstI (P), and XbaI (X). The map shows only cleavage sites that generate β-globin DNA fragments; the direction of the gene detected relative to a mapped site is indicated by ▶, or ◀. Brown lemur cleavage sites probably identical to those in man are indicated by open circles. Sites present in the brown lemur but definitely not in man are shown by filled circles.

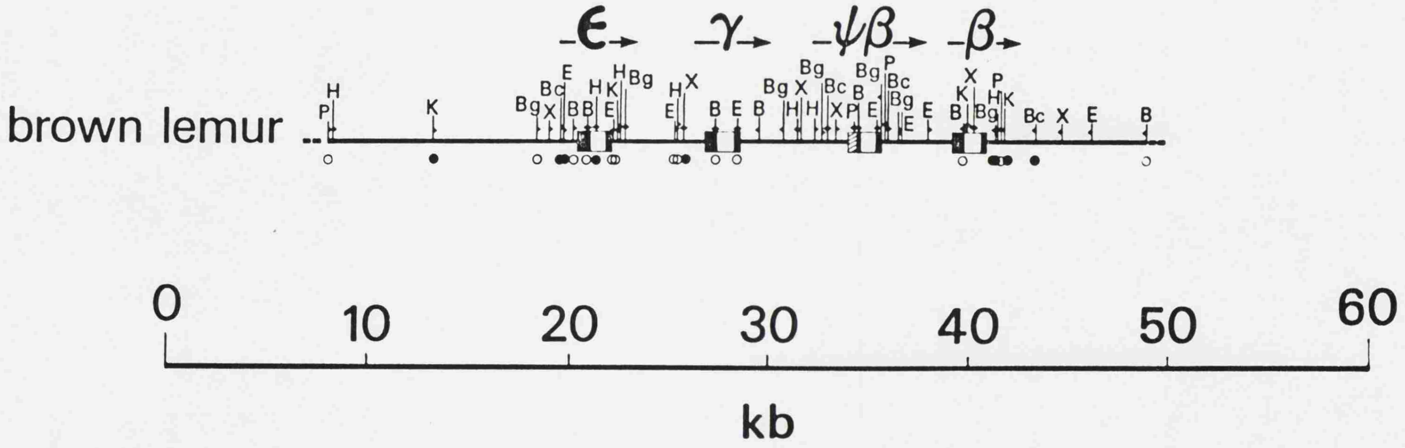

brown lemur
ε
γ
ψβ
β
H
P
K
Bg
X
Bc
E
B
Bg
kb
0
10
20
30
40
50
60

Figure 5.2

Hybridisation evidence that the brown lemur hybrid "Ψβ" contains sequences with homology to the human Ψβ1 pseudogene.

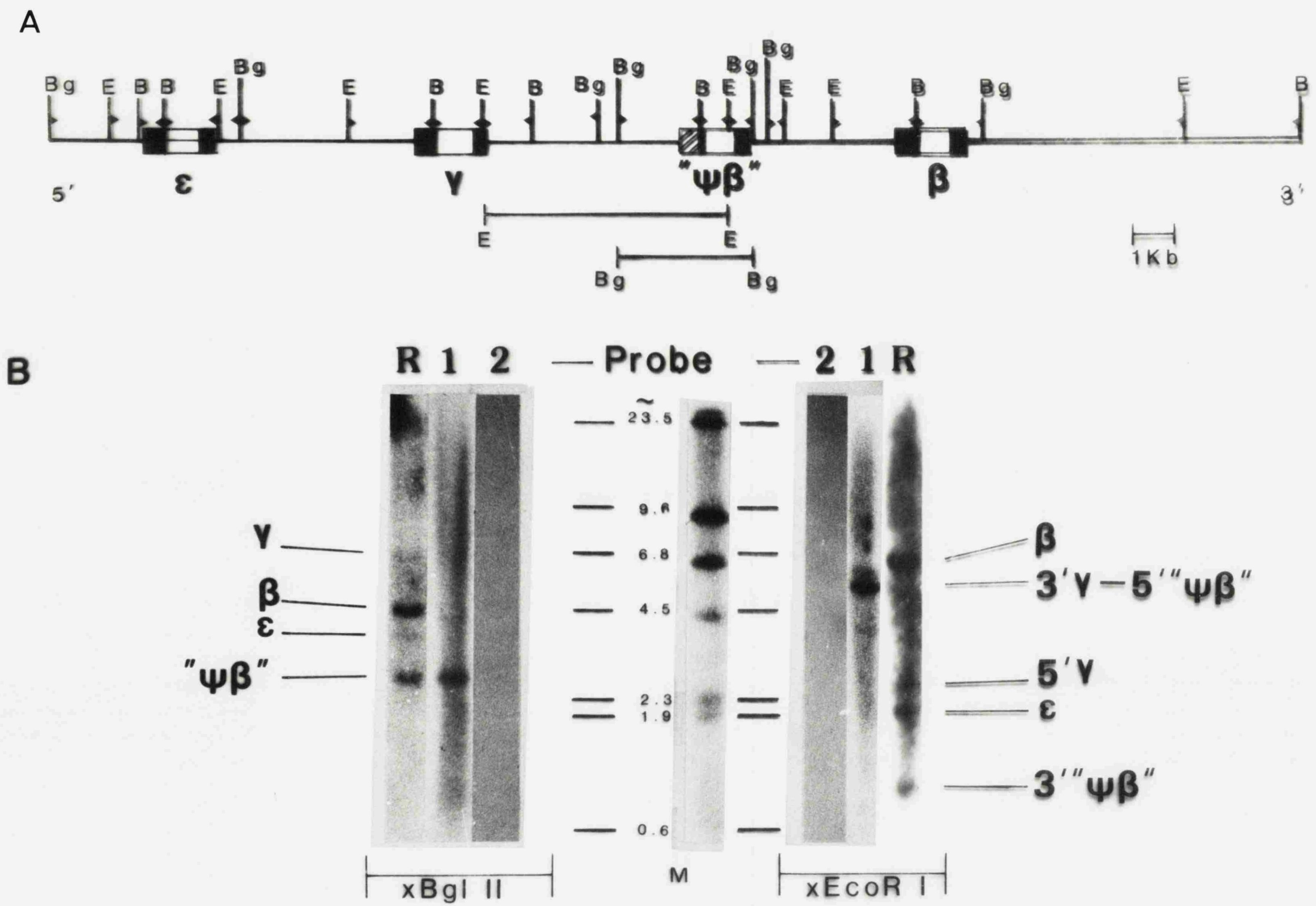
a) Restriction endonuclease cleavage site map of the brown lemur β-globin gene cluster.

A limited number of restriction endonuclease cleavage sites taken from Figure 5.1 are shown, with abbreviations and regions of the genes depicted the same as in Figure 5.1. The genomic EcoRI and BglII DNA fragments of the "Ψβ" gene thought to correspond to the fragments detected by ^{32}P-labelled human Ψβ1 probe 1 are shown below the map.

b) Hybridisation analysis of brown lemur genomic DNA fragments with homology to globin gene probes.

Preparation of Southern blot hybridisation filters and hybridisation to ^{32}P-labelled rabbit adult β-globin cDNA (R) and human Ψβ1 probes 1 (1) and 2 (2) was performed as described for Figure 3.5. Autoradiographic exposures were for 2 days (lanes 1 and 2) and 5 days (lanes R) respectively. Molecular weight markers are λ x HindIII. The previously established identity of ε, γ, Ψβ and β like genomic DNA fragments detected by the rabbit cDNA probe are shown for reference (Barrie *et al.*, 1981).

The single strongly hybridising DNA fragment detected by the human Ψβ1 probe 1 is electrophoretically indistinquishable from that previously assigned to the 5' region of the brown lemur Ψβ gene, suggesting this region of the brown lemur β-globin gene cluster is more closely related to the 5' region of the human Ψβ1 gene than the ε gene. In contrast, the 3' human Ψβ1 probe (probe 2) fails to detect any brown lemur genomic DNA fragments (even after long exposure, results not shown). Similar results were obtained for other lemur DNAs examined (see Figure 3.5) suggesting the evolutionary event that led to the apparent loss of these sequences from the brown lemur occurred in a common ancestor of the lemurs early in their evolution.



A
Bg E B B E Bg E B E B Bg Bg B E Bg Bg E E B Bg E B
5′ ε γ "ψβ" β 3′
E E
Bg Bg
1Kb
B
R 1 2 — Probe — 2 1 R
23.5 9.6 6.8 4.5 2.3 1.9 0.6
M
γ
β
ε
"ψβ"
β
3′γ–5′"ψβ"
5′γ
ε
3′"ψβ"
xBgl II
xEcoR I

Figure 5.3

Isolation of DNA segments from the β-globin gene cluster of the brown lemur.

The following work was performed by Dr P.A.Barrie and Dr A.J.Jeffreys

A brown lemur library made by ligating lemur Sau3A partials into the BamHI site of the replacement vector λL47.1 was screened for recombinants containing β-globin DNA sequences. The recombinant λBL.9 was isolated and mapped as described in Figure 4.4. Also shown is a map of the entire lemur β-globin gene cluster. Abbreviations shown for restriction endonuclease sites are as in Figure 5.1. Alignment of λBL.9 with the genomic map shows an accurate corespondence over the "Ψβ"-β globin region with the exception of the HindIII site 5' to the "Ψβ" globin gene; however, this site could only be located in genomic mapping by measurement from a distal KpnI site within the β-globin gene, and experimental errors in fragment length determination were sufficient to account for this discrepancy. Note that those sites in λBL.9 not indicated in the genomic map do not generate β-globin DNA fragments.

BglII and EcoRI digests of λBL.9 were cloned into pAT153 and recombinant plasmids containing the "Ψβ" 5' BglII fragment (pBL9.1) and the 3' EcoRI fragment (pBL9.8) were isolated. The detailed composite restriction endonuclease cleavage map of pBL9.8 and the 3' end of pBL9.1 was established by partial restriction endonuclease digestion of uniquely ^{32}P end-labelled fragments from pBL9.1 and pBL9.8, after the method of Smith and Birnstiel, 1976 (see Figure 6.2).

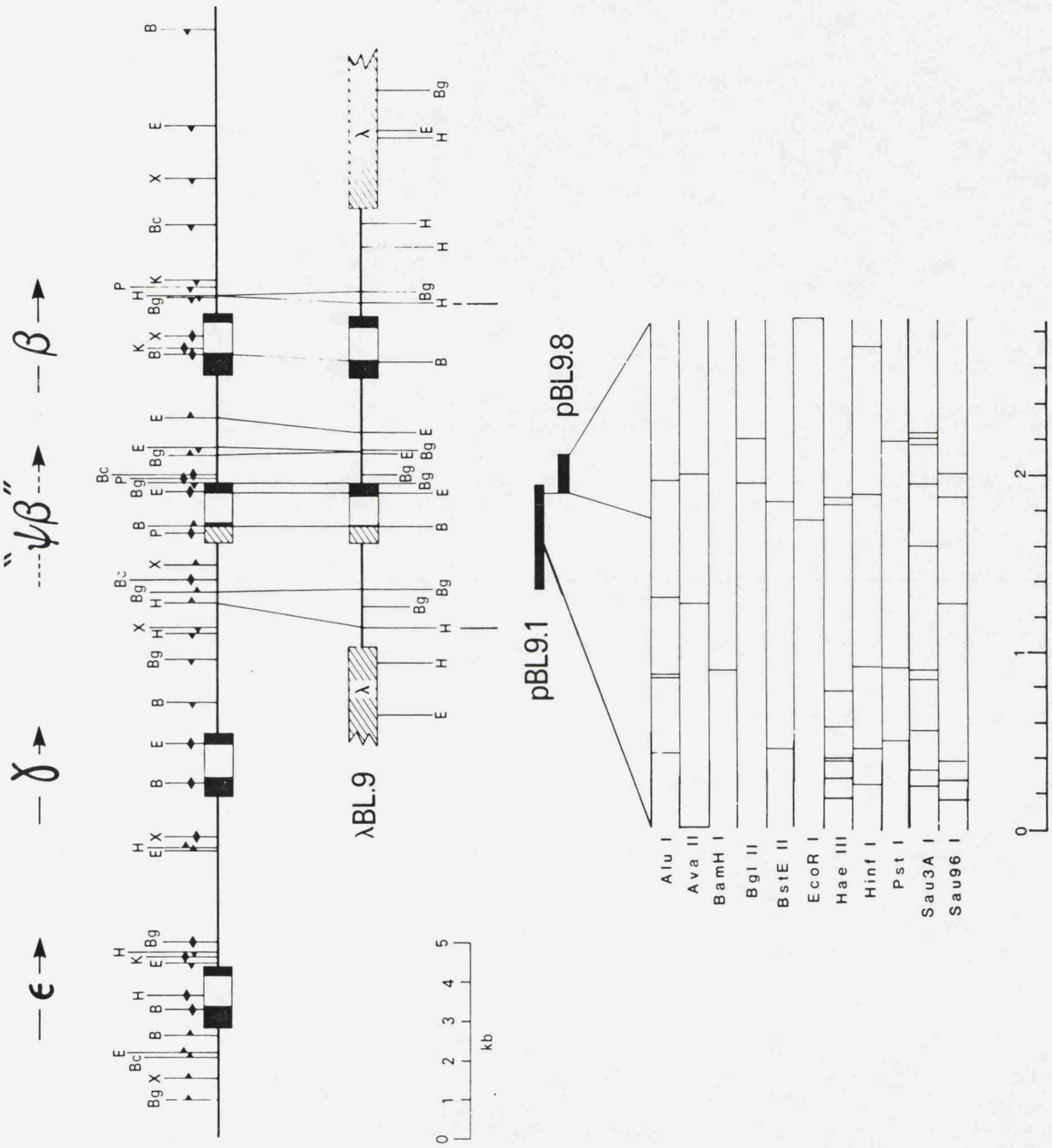

ε
γ
"ψβ"
β
λBL.9
λ
pBL9.1
pBL9.8
Alu I
Ava II
BamH I
Bgl II
BstE II
EcoR I
Hae III
Hinf I
Pst I
Sau3A I
Sau96 I
0
1
2
kb
0
1
2
3
4
5
kb
B
E
X
Bc
K
H
P
Bg
Bl

Figure 5.4

Sequencing strategy for the brown lemur "$\Psi\beta$" globin gene.

Regions homologous to the coding sequence in active β-related globin genes are shown by filled boxes, introns by open boxes and the homologues of 5' and 3' non-coding regions in the mature mRNA by hatched boxes. Restriction endonuclease cleavage sites used for end-labelling are indicated; additional sites are not shown in this map (see figure 5.3). Horizontal lines indicate the DNA fragments that were sequenced. Arrows pointing to the right refer to sequences determined from the "transcribed" strand, and to the left, from the "non-transcribed" strand. Sequences determined from pBL9.1 are shown by filled circles, and from pBL9.8 by open circles. All sequences were determined by the method of Maxam and Gilbert (1977, 1980). Sequences determined by the author are indicated by double headed arrows.

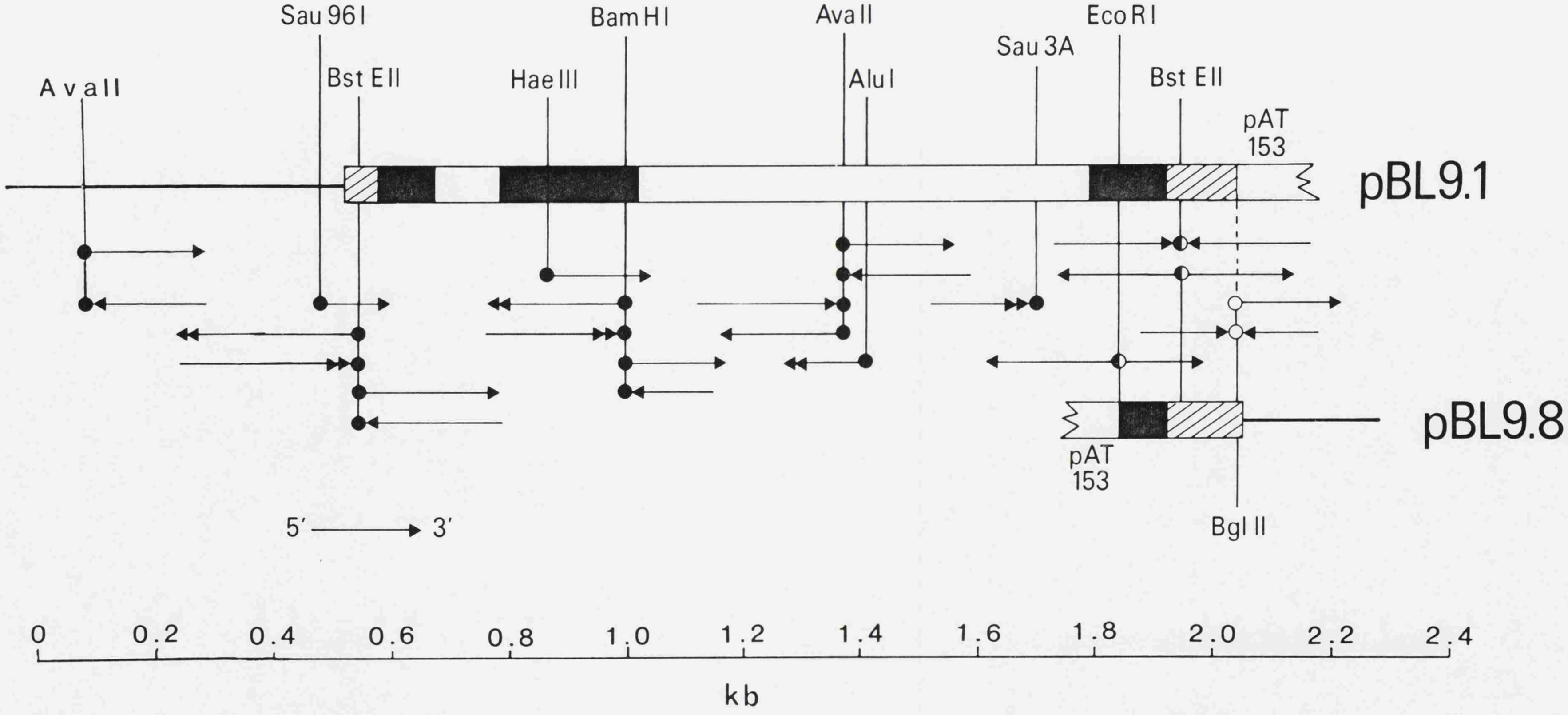

Sau 96 I
Bam H I
Ava II
Eco R I
Sau 3A
A v a II
Bst E II
Hae III
Alu I
Bst E II
pAT 153
pBL9.1
pBL9.8
pAT 153
Bgl II
5′ → 3′
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
kb

Figure 5.5

The complete nucleotide sequence of the brown lemur "Ψβ" globin gene.

The entire 2105 bp of the lemur "Ψβ" globin gene is shown. Sequences homologous to those of a mature globin mRNA are shown in uppercase. Globin consensus sequences implicated in transcription, mRNA maturation and translation are indicated by bold underlined characters. The presumptive location of the 5' and 3' non-translated portions of the mature mRNA were assigned by homology to the goat ε^{II} globin gene (Shapiro *et al.*, 1983) and the human δ-globin gene (Efstratiadis *et al.*, 1980) respectively. Protein coding regions are shown aligned with the equivalent human β-globin gene sequence (Hsa β), only differences are shown. Potential defects in the lemur "Ψβ" gene are numbered above the sequence for reference (see Table 5.1 for description of each defect).

```
                    10        20        30        40        50        60        70        80        90       100
Lfu "ψβ"  : ggacggtgcttcttccacatacagcctggctggataacagaataatacactgtcgcaactccttacccaggaggtgatggctaagagatttatatttctt
Lfu "ψβ"  : attgcttttggttctaatctattcctcataatcagcc+tcaggttatagtgggctagtggtgatggggcccttgagaaataaattgcacacttgatggtg
Lfu "ψβ"  : agggatagtgtaccaaaaaaagaggggaaagaagtgaggtttaaaattgatcctgaatggcagagtcccctgagggggccaccctgagacacaaatatcc
                                         ----1---                    --2--
Lfu "ψβ"  : atctcataaagcctcccttgcccaaactccacccttgggatcacaacccgcccttgaacaatagcctcatttcattaggagagacaaacggctgggggcc
                                                            ---3--                                        -4-
Lfu "ψβ"  : agagatgaagaataaaaggccatggagagaagcagcagtacAGGTGAGCTTCTAACTCATCCGTGGTCACCAGCAGACTCGCAGACCTGACGCTGTG.GT
Hsa   β   :                                                                                               A
                                              -5-
                                               10                                      20
Lfu "ψβ"  : G.CAT.TTC.ACT.GCA.GAG.GCA.AAG.GCT.-CT.GCG.GCT.AGC.CTG.CCA.GGC.AAC.ATG.AAT.GTG.GAG.GAG.GCT.GGA.GGC.AA
Hsa   β   :     C CTG     C T      AG     T   G C  TT A   GC      TGG       G G     C       T   A  T    T   T G
                          30-6-
Lfu "ψβ"  : G.ATC.CTG.GGC.AGgcaggcactggaggccagggtcaggagcagaaaggcagaaagtgttcctgaaagaggggatagccagttatcctatacagtatg
Hsa   β   :   GC             t
                                                            ---7---                      40
Lfu "ψβ"  : actttgcatctgttttgtgacgactgccccatagG.TTC.TTG.CTT.GTTTGTC.TAT.TCA.TGT.TCC.CAG.AAG.TTC.TTC.AGT.AAT.TTT.
Hsa   β   :                                     C G C   G G   ----    C C T   G A        G        T GAG TCC
            --8-              50                                      60                                      70
Lfu "ψβ"  : GGGG.AAT.TTG.TCC.TCT.GGC.TCT.GTG.CTA.ATG.GCC.AAC.CCC.AAG.GTC.AAG.TCT.TAT.GGC.AAG.AAG.CAA.CTG.ACC.TCT
Hsa   β   :   -  G   C       A   CCT GA   CT G T      G        T       G     G   C             A GTG   C GGT G C
                 ---9---                          80                                      90
Lfu "ψβ"  : .TTG.G-A.AAA.GAT.GTT.ACG.TGC.ACT.GAT.GAT.CTC.AAA.GGC.AAC.TTT.GCT.GAG.CTG.AGT.GAG.CTG.CAC.TGT.GAC.AAG
Hsa   β   :    T AGT G T  GC C G G T CA  CTG   C A C       G      CC       C ACA
                             100
Lfu "ψβ"  : .TTG.CAT.GTG.GAT.CCT.GAG.AAC.TGC.AGG.gtgagtctgggagatgttccgtttttccctttctctttctagttttcactctagttcttta
Hsa   β   :  C     C                      T
Lfu "ψβ"  : cctatgtgttctttctacacattcatttttactttaccatattttatcatttaacacttttcaaattttgtcaattttcttctttctacattctgtctt
Lfu "ψβ"  : ctttccttttgcacaatcttacttttattgaattttaaatttactatcctgtcatttgcctgtatctctcccatcccccatttatttttttctccaa
Lfu "ψβ"  : ccacaacccaaattatgcatatcagttctcatctgctagttctacactttgaaaaatccttctgtctcttcatatgggggtagaagatggtccaactcaa
Lfu "ψβ"  : agaggagaggcacagaatgctgttttagaagctataaatcattttaaaatgaataataattgaaattttataaattcaggaataaatgaaatgaaagaa
Lfu "ψβ"  : atggaaagtaaatatctgagggtgaaaggttaaaagttcatactggaagcagggcaccagtttttggtaagaggcagactgtcatcacactaatcaattt
Lfu "ψβ"  : atttgtatataatatatatgtacatacatatatactgtatactacttaagtatccagtattatatatgtattatgtacatatatacatacatatacttaa
Lfu "ψβ"  : cgctggtgtgaatgacatggagatcaacttgggctaggacatgggcagaaaaagaaagccaatattgatttctttgttaaccatacctatgtgtctactt
                                               110                                     120
Lfu "ψβ"  : acctcttccccacagCAC.CTG.GGC.AAC.GTG.CTG.GTG.GTT.GTG.CTG.GCT.GAA.CAG.TTT.GGC.AAG.GAA.TTC.ACC.CCA.CAG.
Hsa   β   :                 T                        C TG            C C T   C           A                  CA.
                            130                                     140
Lfu "ψβ"  : GTG.CAG.GCT.GCC.TAT.CAG.AAG.GTG.GTG.GCT.GGT.GTG.GCT.AAT.GCC.CTG.GCT.CAC.AAG.TAT.CAC.TGAGGCCTCGGACCAT
Hsa   β   :                           A                                       C                  A
Lfu "ψβ"  : TTCTTGGTGACCAGTGGAAGGCCCTATTTCCACAGATTCTCTCTTCTGTAATTGGGGAAATAGTGCTTACCTCAAGGGTATGGCATCTGCCTAATAAAGA
Lfu "ψβ"  : TCTTTCAGCTCAACTTTctgatttattttatttttgtctgggaatgtgagaaggtccctgagggtc tacagatagagagttctcatgtcttatacaaaa
Lfu "ψβ"  : ggtcaagagaaatgagaaaaggaagggagccagacacagacactaatgggtgaca
```

Figure 5.6

Dot-matrix comparison of the brown lemur "Ψβ" globin gene

The entire 2105 bp of the brown lemur "Ψβ" gene is shown compared with the human Ψβ1 pseudogene (Chang and Slightom, 1984; Weissman, per.comm.) and the human δ globin gene (Spritz _et_ _al_., 1980). Matching criteria and sequence features depicted beside the grid are the same as in Figure 3.2. The extensive sequence homology between the brown lemur gene and the two human genes apparently alters within exon 2; in particular, intron 2 and 3' flanking non-coding DNA sequences show strong homology with the human δ gene while intron 1 and 5' flanking sequences show strong homology to the human Ψβ1 pseudogene sequence. From this it can be concluded that the brown lemur "Ψβ" gene is a hybrid gene composed of sequences with orthology to the human Ψβ1 and δ globin genes, that is, a hybrid Ψβ1 -δ globin gene.

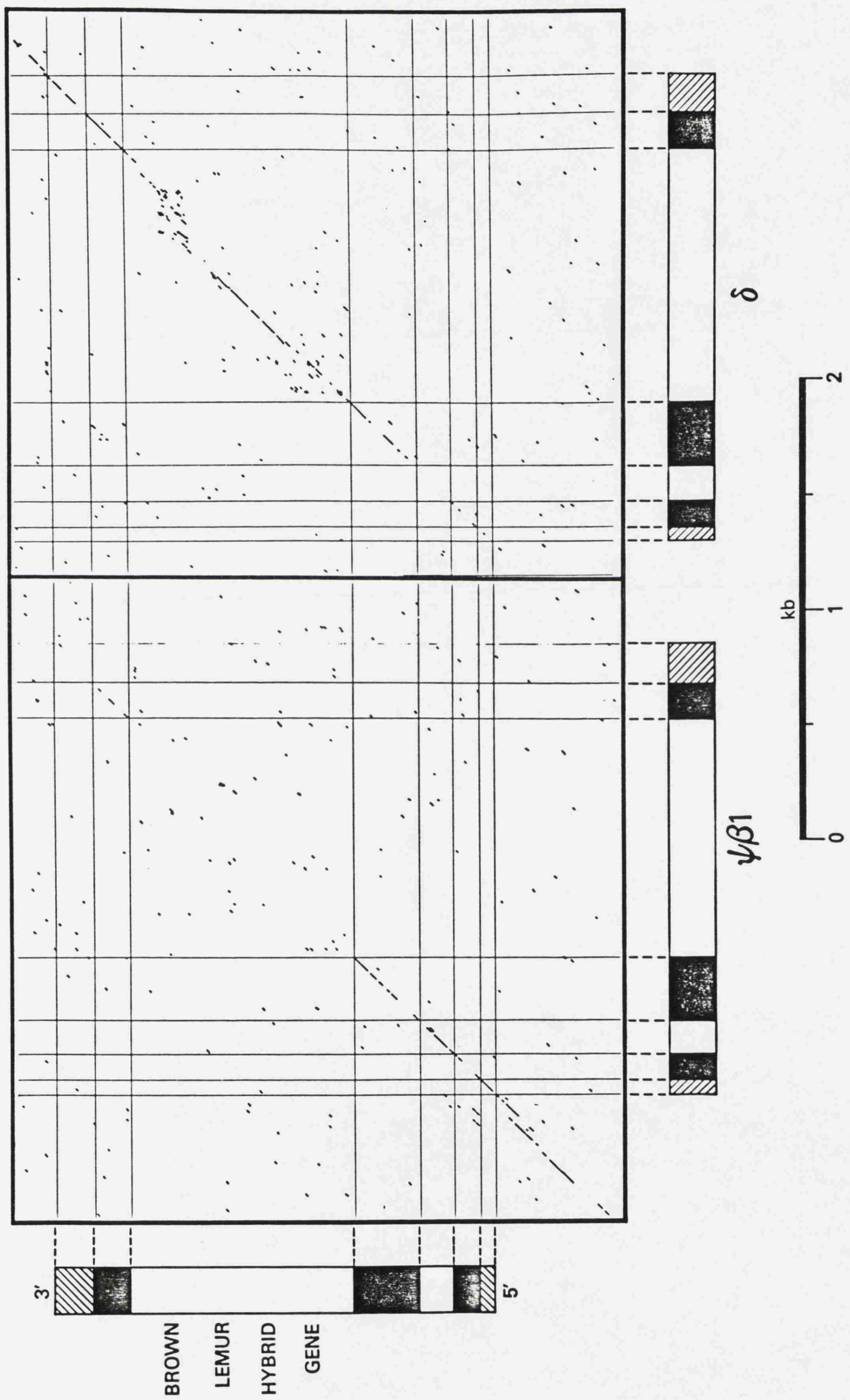

BROWN
LEMUR
HYBRID
GENE
3'
5'
ψβ1
δ
kb
0
1
2

Figure 5.7

Establishment of the probable Ψβ1-δ crossover point in the hybrid pseudogene of the brown lemur.

Exon 2 (uppercase) and flanking intron sequences are shown for the human Ψβ1 pseudogene (HsaPB1), brown lemur hybrid Ψβ1-δ pseudogene (LfuPBD) and the human δ-globin gene (HsaDEL; Spritz *et al.*, 1980). Asterisks indicate bases which are identical between lemur and one, but not both, of the human sequences. A probable Ψβ1 -δ exchange point in the lemur pseudogene sequence is indicated corresponding to codon 86/87 of the exon.

```
HsaPB1  aaaggcagaaagtgtttctgaaagagggattagcccgttgtcttacatagtctgactttgcacctgctctgtgattatgactatcccacagT.CTC.CTG.GTT.GT----C.TAC.CCA.
          *  * **** ***  *** * *  *   * **  ***  * **     * ****** *** * * *   * *    ** *     *       *       *               *
LfuPBD  aaaggcagaaagtgttcctgaaagaggggatagccagttatcctatacagtatgactttgcatctgttttgtgac---gactgccccatagG.TTC.TTG.CTT.GTTTGTC.TAT.TCA.
                                     *                                      *                        *
HsaDEL  aatggaaactgggcatgtgtagacagagaagactcttgggtttctgataggcactgactctctgtcccttgggctgttttcctaccctcagA.TTA.CTG.GTG.GT----C.TAC.CCT.

                    40                                       50                                      60
HsaPB1  TGG.ACC.TAG.AGG.TAC.TTT.GAA.AGT.TTT.GGA.-TAT.CTG.GGC.TCT.GAC.TGT.GCA.ATA.ATG.GGC.AAC.CCC.AAA.GTC.AAG.GCA.CAT.GGC.AAG.AAG.
                                    * *                          * * *         *               *       *
LfuPBD  TGT.TCC.CAG.AAG.TTC.TTC.AGT.AAT.TTT.GGG.GAAT.TTG.TCC.TCT.GGC.TCT.GTG.CTA.ATG.GCC.AAC.CCC.AAG.GTC.AAG.TCT.TAT.GGC.AAG.AAG.
                *        *                    *          **                                        *           *
HsaDEL  TGG.ACC.CAG.AGG.TTC.TTT.GAG.TCC.TTT.GGG.-GAT.CTG.TCC.TCT.CCT.GAT.GCT.GTT.ATG.GGC.AAC.CCT.AAG.GTG.AAG.GCT.CAT.GGC.AAG.AAG.

                    70                                      80                                      90
HsaPB1  GTG.CTG.ATC.TCC.TTC.GGA.AAA.GCT.GTT.ATG.CTC.ACG.GAT.GAC.CTC.AAA.GGC.ACC.TTT.GCT.ACA.CTG.AGT.GAC.CTG.CAC.TGT.AAC.AAG.CTG.
              * * * *       * * * *   * * * * *     **    * *         *       *     *
LfuPBD  CAA.CTG.ACC.TCT.TTG.G-A.AAA.GAT.GTT.ACG.TGC.ACT.GAT.GAT.CTC.AAA.GGC.AAC.TTT.GCT.GAG.CTG.AGT.GAG.CTG.CAC.TGT.GAC.AAG.TTG.
                                             *                                           **           *             *
HsaDEL  GTG.CTA.GGT.GCC.TTT.AGT.GAT.GGC.CTG.GCT.CAC.CTG.GAC.AAC.CTC.AAG.GGC.ACT.TTT.TCT.CAG.CTG.AGT.GAG.CTG.CAC.TGT.GAC.AAG.CTG.

                    100
HsaPB1  CAC.GTG.GAC.CCT.GAG.AAC.TTC.CTGgtgagtagtaagtacactcacgctttcttctttacccttagatatttgcactatgggtacttttgaaagcagaggtggctttctcttg
                                                         *          *              *     *                           **
LfuPBD  CAT.GTG.GAT.CCT.GAG.AAC.TGC.AGGtgagtctgggagatgttccgttttttccctttctctttctagtttttcactctagttcttt-acctatgtgttctttctacacattc
                  *                 **       *  **** **   *  *     *       *  **  *            * **   *****  *  ** *    ******
HsaDEL  CAC.GTG.GAT.CCT.GAG.AAC.TTC.AGGtgagtccaggagatgcttcacttttctctttttactttctaatcttacattttggttcttttacctacctgctcttctcccacattt
```

Table 5.1

Analysis of potential defects in the brown lemur "Ψβ" gene.

Given the alignment of the brown lemur "Ψβ" pseudogene and the human β-globin gene shown in Figure 5.5 the coding sequences and 5' and 3' flanking sequences implicated in eukaryotic gene expression were examined for potential defects. The potential consequence of each defect (numbers in the body of Table 5.1 correspond to those in Figure 5.5) was examined in isolation as the order in which the defects were accumulated during evolution is unknown, with the possible exception of the primary silencing defect (see Discussion). If transcribed, codon defects in the gene would result in premature translation termination prior to the end of exon 2, if not before, resulting in a truncated and almost certainly inactive peptide. Confirmation that, in isolation, specific signal sequence defects (1, 2, 3, 4) would affect correct developmental expression of this gene remains to be tested.

Table 5.1

Number	Position in sequence	Potential defect	Consequence of defect
1	'CACCCCTG'	'CACCCTTG'	Alteration in what may constitute a signal sequence specifying embryonic developmental expression (Shapiro et al.,1983).
2	'CCAAT'-box	'ACAAT'	Partial though not necessarily total reduction in transcription (see Dierks et al.,1983).
3	'CTTCTG'	'CTTCTA'	Unknown effect on ribosome binding to the mature mRNA.
4	Initiation codon	ATG - GTG	Absent or incorrect translation initiation, it is as yet unknown if GTG can initiate translation as in prokaryotes.
5	Codon 10	Deletion (-1)	Frameshift resulting in premature translation termination at an in frame TGA codon ~24 bp downstream.
6	Exon 1/Intron 1	GT - GC	Absent or incorrect mRNA splicing of mammalian globin transcripts, however see Fischer et al.,1984 for case of correct splicing of such a site from chicken α^D globin RNA in vivo.
7	Codon 34	Insertion (+4)	Frameshift resulting in premature translation termination at an in frame TAA codon ~27 bp downstream.
8	Codon 46	Insertion (+1)	Frameshift resulting in premature translation termination at an in frame TGA codon ~132 bp downstream.
9	Codon 72/73	Deletion (-1)	Frameshift resulting in premature translation termination at an in frame TGA codon ~32 bp downstream.

Chapter 6

THE FUNCTIONAL BROWN LEMUR β-GLOBIN GENE

6.1 Introduction

The brown lemur β-globin gene family has been shown by hybridisation analysis and molecular cloning to contain a single ε-, γ- and β-like globin gene plus a peculiar hybrid "Ψβ" gene within a 20 kb region of genomic DNA (see Figure 5.1). It has since been shown by sequence analysis that the hybrid "Ψβ" gene in this cluster is a non-processed pseudogene with orthology to two members of the human β-globin gene family, the Ψβ1 and δ genes (Chapter 5 and Discussion). A similar globin gene cluster organisation has been inferred for another lemur, the ruffed lemur (Barrie et al., 1981), suggesting this organisation may be common to the lemurs in general; certainly the hybrid pseudogene was apparently established early in lemur evolution as all the lemurs examined contained only part of a Ψβ1-related sequence (Chapter 3).

The probable unequal exchange that gave rise to the hybrid Ψβ1-δ pseudogene in the brown lemur is of considerable interest as the concomitant deletion has resulted in the elimination of a region of intergenic DNA that in man has been implicated in the switch from γ → β gene expression late in gestation (see Collins and Weissman, 1984). The deletion of this region in the brown lemur might therefore disrupt the normal order of developmental gene expression in this gene cluster. However, the β-like chain of the adult brown lemur has been sequenced and shows homology to the human β-haemoglobin protein rather than the ε- or γ-haemoglobin proteins (Maita et al., 1979). This suggests that the adult β-like globin gene (by hybridisation) present at the 3' end of this gene cluster is being expressed in this species. As the protein sequence of the adult brown lemur β-globin is known, the DNA sequence of the 3' β-like globin gene of the brown lemur would establish whether this gene encodes the adult β-globin in this primate.

No information is available concerning the expression of the ε and γ globin genes in this gene cluster (see Chapter5). Initial sequencing over the coding regions of the brown lemur ε and γ genes, since the completion of the work in this thesis, does however confirm the potential of these genes to encode non-adult globin polypeptides, Jeffreys et al., per.comm.

The brown lemur β-globin gene sequence would also be ideal for the comparative analysis of the mode and tempo of sequence divergence within the different regions of the primate β-globin gene (exons, introns and flanking non-coding sequences). Of particular interest in relation to this thesis is the mode and tempo of non-coding sequence evolution within the β-globin gene (flanking and intron regions) compared to that established for the primate globin pseudogene ψβ1 (see Discussion).

The brown lemur β-globin gene has therefore been sequenced in order to establish a) whether this gene codes for the adult β-globin protein of this species and b) to allow the comparative analysis of the mode and tempo of DNA sequence divergence within a functional (assumed at this point) β-globin gene of a primate that diverged early in primate evolution.

6.2 Sequencing of the brown lemur β-globin gene

The brown lemur β-globin gene was available subcloned as part of a large genomic HindIII fragment inserted into the HindIII cloning site of pAT153 to produce the recombinant plasmid pBL9.13 (the work of Dr A.J.Jeffreys). Figure 6.1* illustrates the relationship of this subclone to the genomic restriction endonuclease cleavage site map covering the β-globin gene cluster of the brown lemur and to λBL9, a λL47.1 genomic recombinant from which the subcloned HindIII fragment was derived (Barrie, 1982).

Before a Maxam-Gilbert sequencing strategy could be embarked upon a detailed restriction endonuclease cleavage site map covering the β-globin gene was required. The general method for construction of such a map is described in Chapter 2.24. A detailed restriction endonuclease site cleavage map covering the β-globin gene and surrounding DNA was produced by partial restriction endonuclease digestion of a uniquely ^{32}P end-labelled EcoRI-HindIII fragment recovered from pBL9.13, see Figure 6.2(a). The partial restriction endonuclease site cleavage patterns produced by the various enzymes are shown in Figure 6.2(b), along with the resultant cleavage map. The sequencing strategy employed for the brown lemur β-globin gene is shown in Figure 6.3. The general approach to Maxam-Gilbert sequencing is described in Chapter 5. The majority of sequence was confirmed by determination on both strands (>90%), all but 9 bp of the sequence determined on a single strand was from the 3' flanking region of the gene (see Figure 6.~~4~~3).

*Figures for this Chapter follow the text.

6.3 Structural features and orthology of the brown lemur β-globin gene

The 2215 bp DNA sequence encompassing the β-globin gene, extending 323 bp 5' of the initiation codon and 453 bp 3' of the termination codon has been aligned against the human β-globin gene sequence (see Figure 6.4). The alignment over the non-coding regions (flanking and intron sequences) was obtained with reference to the major stretches of homology detected by dot-matrix analysis between these two gene sequences (see below). This alignment confirms the functional potential of the brown lemur β-globin gene; the important transcription, translation and mRNA processing signal sequences implicated in globin gene expression being present within and around the brown lemur β-globin gene. The base positions of transcription initiation and poly(A) addition in the brown lemur β-globin sequence are inferred from their positions in the human DNA sequence which has been determined experimentally. The gene has the archetypal β-globin gene organisation, the coding sequence (three exons) interrupted by two introns of 130 bp and 868 bp long respectively.

Translation of the exon sequences of the brown lemur β-globin gene gives an amino acid sequence which is identical to that described by Maita et al. (1979) establishing that this gene specifies the adult β-globin in this primate. Where the amino acid residues determined by Maita et al. (1979) differed from that described by other workers (Bonaventura et al., 1974) the translated DNA sequence for that amino acid position within the polypeptide always favoured the sequence of Maita et al. (1979).

The brown lemur β-globin protein sequence differs at 26 amino acid residues out of 146 when compared to the human adult β-globin protein (Table 6.1). Of the 26 amino acid differences 8 occur at positions so far attributed with known functional roles (Goodman, 1981). The majority (18/26) of the changes occur at positions of unknown function, 9/18 occuring in the least well conserved exon (exon 1) which codes for only 30 out of a total of 146 amino acid residues (Table 6.1 and Discussion).

Dot-matrix analysis of the brown lemur β-globin gene sequence against the other members of the human β-globin gene cluster, and other mammalian β-globin gene sequences, demonstrated extensive regions of homology over all non-coding as well as coding regions establishing that the lemur, human and other mammalian β-globin genes are all orthologous (Figure 6.5 illustrates the orthology between the brown lemur and human β-globin gene sequences when compared by dot-matrix analysis). The brown lemur β-globin gene also showed alignment against the human δ-globin gene over the first intron (results not shown). This would be expected if this region of the contemporary human δ gene had been converted by β-globin gene sequences during simian evolution and therefore resembled β rather than δ sequences over this region (see Discussion).

6.4 Summary

A 2215 bp region of the brown lemur genome encompassing the β-globin gene has been sequenced. Comparison of the translation product of this adult β-globin gene confirms the potential of this gene to encode the adult β-globin polypeptide of the brown lemur. Dot-matrix analysis confirms previous hybridisation results (Barrie _et_ _al_., 1981) that suggested this globin gene is orthologous to the human and mammalian adult β-globin genes. The brown lemur β-globin gene has the archetypal β-globin gene structural organisation and the associated transcription, mRNA maturation and translation signal sequences implicated in mammalian β-globin gene expression.

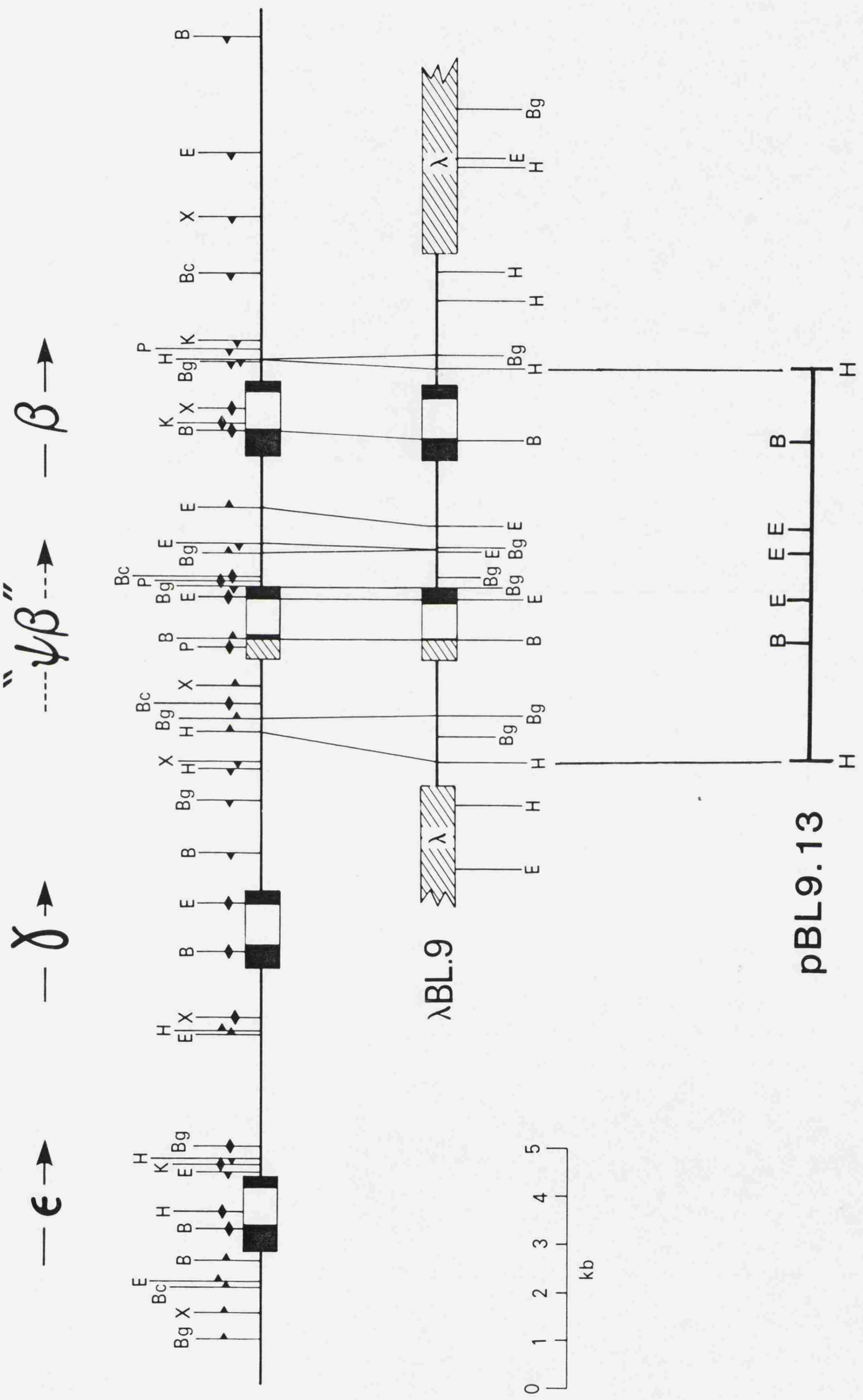
Figure 6.1

Isolation of the brown lemur β-globin gene.

The following work was conducted by Dr P.A.Barrie and Dr A.J.Jeffreys

For details concerning the restriction endonuclease cleavage site map of the brown lemur β-globin gene cluster and the isolation and characterisation of the genomic λ-recombinant λBL.9 see Figure 5.3.

The brown lemur β-globin gene was subsequently subcloned into the *Hin*dIII cloning site of the plasmid pAT153 as part of the ~8 kb *Hin*dIII fragment from λBL.9 to give the plasmid recombinant pBL9.13.



ε
γ
"ψβ"
β
λBL.9
pBL9.13
λ
Bg
X
Bc
E
B
H
K
P
0
1
2
3
4
5
kb

Figure 6.2

Restriction mapping of the brown lemur β-globin gene prior to sequencing.

a) Preparation of a ^{32}P end-labelled fragment for restriction mapping of the β-globin gene.

A limited number of restriction endonuclease cleavage site positions are shown for the plasmid recombinant pBL9.13. Abbreviations for restriction endonuclease sites and key to indicated gene sequences are as described for Figure 3.2. Plasmid sequences are depicted by the solid line. The β-globin gene was isolated as part of a ~3.2 kb *Eco*RI fragment as shown. After labelling both the 5' terminal phosphates with ^{32}P (step 1) this fragment was digested to completion with *Hin*dIII (step 2) to give two DNA fragments of ~3.17 kb and 30 bp (for details see 2.24). There was no need to separate the two fragments prior to partial mapping as the small *Eco*RI-*Hin*dIII fragment produced could not generate further restriction endonuclease fragments >30 bp in size.

b) Detailed mapping of the brown lemur β-globin gene region.

The detailed composite restriction endonuclease cleavage site map shown for the β-globin gene region of pBL9.13 was established by partial digestion of the ^{32}P end-labelled *Eco*RI-*Hin*dIII fragment after the method of Smith and Birnteil, 1976 (see 2.24). After partial digestion DNA fragments were electrophoresed on a 1% (w/v) horizontal agarose gel. The agarose gel was dried down then autoradiographed overnight to give the pattern shown. Molecular weight markers are a mixture of λ x *Hin*dIII and pBR322 x *Sau*3A. The approximate position of the sequences of the β-globin gene are shown in relation to the restriction map, regions of the gene represented as in Figure 3.2.

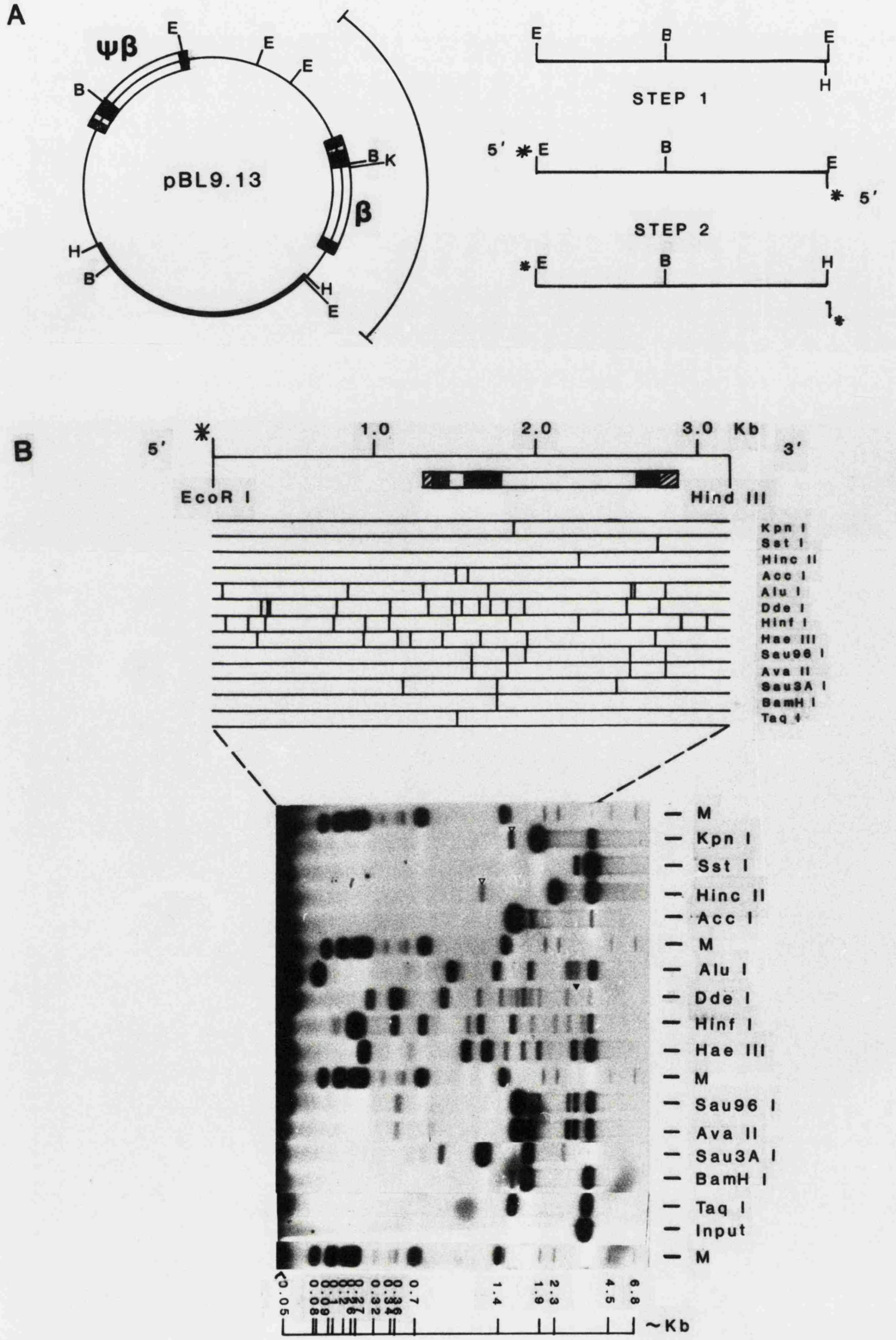

A
ψβ
E
E
E
B
B
K
pBL9.13
β
H
B
H
E
E
B
E
H
STEP 1
5'
*
E
B
E
* 5'
STEP 2
*E
B
H
]*
B
*
1.0
2.0
3.0 Kb
5'
3'
EcoR I
Hind III
Kpn I
Sst I
Hinc II
Acc I
Alu I
Dde I
Hinf I
Hae III
Sau96 I
Ava II
Sau3A I
BamH I
Taq I
M
Kpn I
Sst I
Hinc II
Acc I
M
Alu I
Dde I
Hinf I
Hae III
M
Sau96 I
Ava II
Sau3A I
BamH I
Taq I
Input
M
0.05
0.08
0.09
0.12
0.26
0.27
0.32
0.34
0.36
0.7
1.4
1.9
2.3
4.5
6.8
~Kb

Figure 6.3

Sequencing strategy for the brown lemur β globin gene.

Regions homologous to the coding sequence in active β-related globin genes are shown by filled boxes, introns and flanking non-coding DNA sequences by open boxes and the 5' and 3' non-coding regions in the mature mRNA by hatched boxes. Restriction endonuclease cleavage sites used for end-labelling are as indicated; additional sites are not shown in this map (see Figure 6.3). Horizontal lines indicate the DNA fragments that were sequenced. Arrow heads indicate whether the sequences determined were from the sense strand or from the nonsense strand. All sequences were determined by the method of Maxam and Gilbert (1977,1980). With the exception of a 9 bp region across the 5' Sau3A site the bars above the diagram indicate the regions determined on only a single strand.

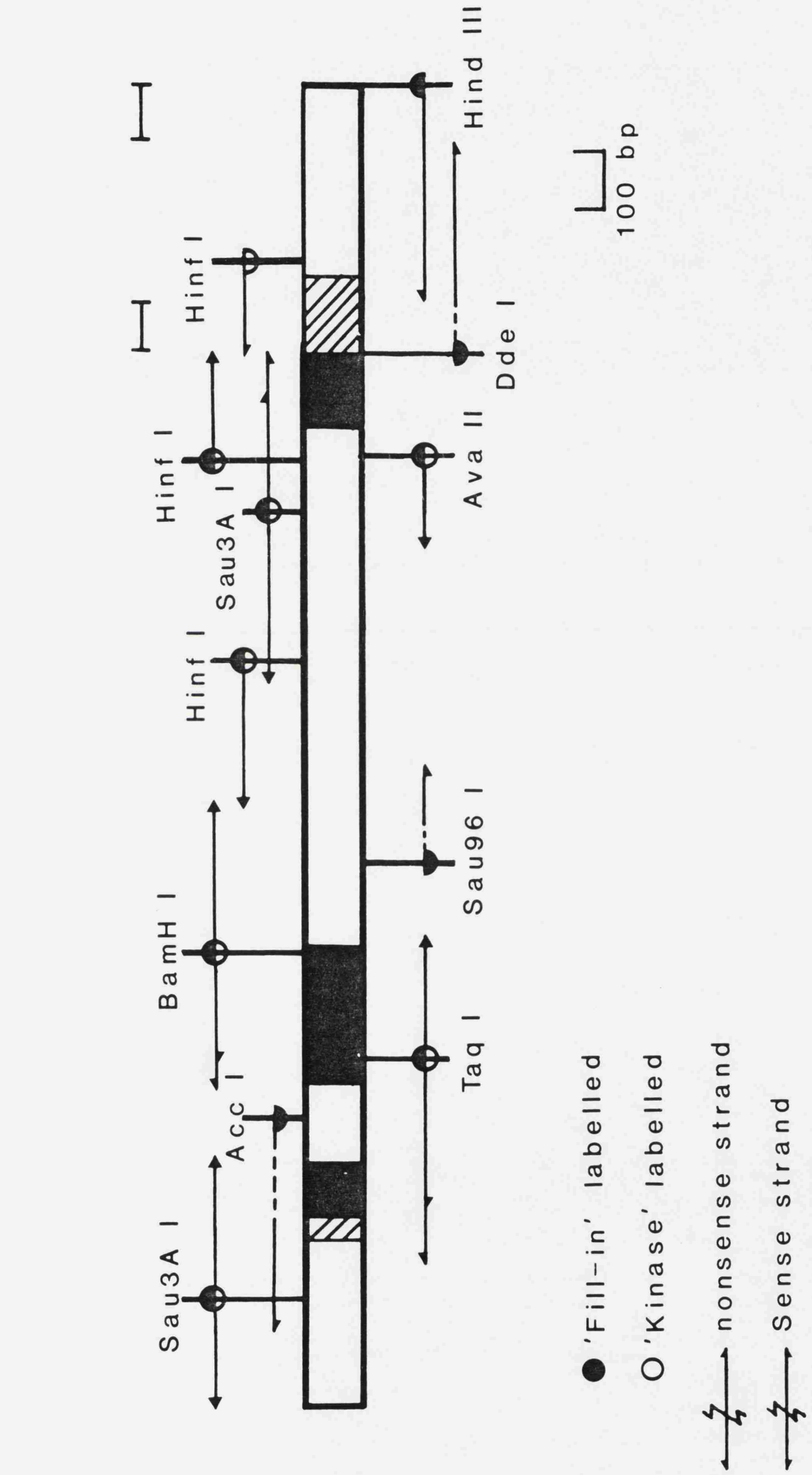

Sau3A I
Acc I
BamH I
Hinf I
Sau3A I
Hinf I
Hinf I
Taq I
Sau96 I
Ava II
Dde I
Hind III
100 bp
● 'Fill-in' labelled
O 'Kinase' labelled
nonsense strand
Sense strand

Figure 6.4

The complete nucleotide sequence of the brown lemur β-globin gene.

The entire 2215 bp of the lemur β globin gene (Lfu) is shown aligned against the human β-globin gene (Hsa), only differences between the two sequences are shown for the human sequence. A dash indicates the absence of a nucleotide in one sequence relative to the other. Sequences homologous to those of a mature globin mRNA are shown in uppercase. Globin consensus sequences implicated in adult globin transcription, mRNA maturation and translation are indicated by bold underlined characters (Hardison, 1984). The presumptive location of the 5' and 3' non-translated portions of the mature mRNA were assigned by homology to the human β-globin gene (Efstratiadis *et al.*, 1980). Amino-acid differences between the two coding sequences are indicated by the presence of the equivalent human amino-acid below the relevant codon (see also Table 6.1).

```
              10        20        30        40        50        60        70        80        90       100

Lfu : cccaaaaattgtatgagtaaacttgccaaggaggatgttttagtagcaattcatgctattcggatggaaccaggagatacatatagagggagggctgag
Hsa : --------------------------   -------------------------------------------------------------------

Lfu : ggtctgaagttagactcctaaaccagtttcagaactgccaaggacaagtactgctgccaccattcaggcctcaccctggagatccacaccctgggcttgg
Hsa :   t t     cca       g      gc     ga          g    g    t  t  c t  a          tg  g         a  g

Lfu : ccaatctgctcataggagcagggagggcaggagccagggctgggcataaaagtcagggtggggacagctactgcttACACTTGCTTCTGACACAACTGTG
Hsa :        a   cc                                              ca a c  t   t        T

                                                                         10
                                Met Thr Leu Leu Ser Ala Glu Glu Asn Ala His Val Thr Ser Leu Trp Gly Lys Va
Lfu : TTCACTAGCAAC----AGCAGACACCATG.ACT.TTG.CTG.AGT.GCT.GAG.GAG.AAT.GCT.CAT.GTC.ACC.TCT.CTG.TGG.GGC.AAG.GT
Hsa :             CTCA A            GTG CAC      C  C             G T   GCC   T   T G C
                                    Val His     Thr Pro         Lys Ser Ala         Ala

             20                                     30
      l Asp Val Glu Lys Val Gly Gly Glu Ala Leu Gly Ar
Lfu : G.GAT.GTA.GAG.AAA.GTT.GGT.GGC.GAG.GCC.TTG.GGC.AGgttggtatcggggttatagggcaggcttaaggagacaaatggaaactgagcc
Hsa :   A C   G   T G             T         C                   aa    c a a    t          c   a        g a
        Asn     Asp Glu

                                                                                   g Leu Leu Val Val Tyr
Lfu : tgtggagccagggtagactcctgggtttctgacaggtattgactctctgtctcctgggttgctttcacccccctcagG.CTG.CTG.GTC.GTC.TAC.
Hsa :        a   a a      t           t   c c           c  at    c at    c a    t

                       40                                      50                                      60
      Pro Trp Thr Gln Arg Phe Phe Glu Ser Phe Gly Asp Leu Ser Ser Pro Ser Ala Val Met Gly Asn Pro Lys Val
Lfu : CCA.TGG.ACC.CAG.AGG.TTC.TTC.GAG.TCC.TTT.GGG.GAC.CTG.TCC.TCT.CCT.TCT.GCT.GTT.ATG.GGG.AAC.CCT.AAG.GTG.
Hsa :   T                       T                   T         A       GA                C
                                                              Thr     Asp

                                           70                                      80
      Lys Ala His Gly Lys Lys Val Leu Ser Ala Phe Ser Glu Gly Leu His His Leu Asp Asn Leu Lys Gly Thr Phe
Lfu : AAG.GCC.CAT.GGC.AAG.AAG.GTG.CTG.AGT.GCC.TTT.AGT.GAA.GGT.CTG.CAT.CAC.CTG.GAC.AAC.CTC.AAG.GGC.ACC.TTT.
Hsa :       T               A       C G               T   C     GC
                                      Gly             Asp         Ala

                       90                                     100
      Ala Gln Leu Ser Glu Leu His Cys Asp Lys Leu His Val Asp Pro Gln Asn Phe Thr
Lfu : GCT.CAA.CTG.AGT.GAG.CTG.CAC.TGT.GAC.AAG.TTG.CAC.GTG.GAT.CCT.CAG.AAC.TTC.ACT.gtgagtctatgggaccttcaatgt
Hsa :   C AC                                  C                   G            GG                 c tg
          Thr                                                     Glu         Arg

Lfu : ttctcttcttctttctctctattgccaagttcatgttatgtggaggggcatgggtaccaagtccagtttagaatgggaaagaggcacctggtggcatcac
Hsa :   tct   cc   ct t     g tt          c  ag a     gaa ta agg    a                 c  a gaa  a tc     g

Lfu : tgtgggcccctaagggtta-tttagtttcttttattctctgctcacaacctttgttttctattatttatttcttgtttcctctttttttttctgtcattt
Hsa :      aagt  c   a cgt                tg   t   t   aa         t  g    a      c    t t      c tctc gcaa

Lfu : attctgttttttttatttaatttttta--------tgtaaccaaggggaaatatctagaagataacttaagtgaccaaaaaaaaaaaaaaaaaaaac
Hsa : t  t ac a  a--  c      gcc taacattgtgtat a - aa          ctg     ca      a ‘tt         -------------

Lfu : cctgaagaacttaaacagtctgcctagtatgctgctatttagatatgtatgtgta-ttgtttggatattaataatctacctactttattt-atttattt
Hsa : --------ct   c               cat a      g  -  a        gc a    c     c       c            tc

Lfu : ttatttatgtttagttgacacataatgattgtacatatttatggtgtagagtgtgatgttttgattcatatatatatatacatggtgtaatgaccaaat
Hsa :    -----------a    t       c   a             gt  aa    a       a  --  g g  c c   t--------

Lfu : caaagtaatttggcatttgtgattttttaaaaaaccctttcttctttttagtccactttttggtttttgtcttacttcttatgaacc------------
Hsa :   gg       t        a     --      tg             a at         ----   a     t       actttccctaatctctt

Lfu : -------------------------atatcatgctgctttgcatcattctaaagaatgacagtgacaatttaatttctaggttatagtaatagggaggg-
Hsa : tctttcagggcaataatgatacaatg        cc       c             a       -----        g     ag c     ca tatt

Lfu : -----------acttttttgcatgtaaatcatggctgatatggaaggtgtcatattggtagtagcagctaagatccagcagttgttccgcttttttgtttt
Hsa : tctgcatataa ta   c     a     tg aa     g aag    tt       c  a         ca       tacca   t  -    a

Lfu : atggttacagtgctaagcacagcttgagatgaggctgaaatattctgagtccaagctgggtccctctactaatcatgtccatatctctcgtctcttcccc
Hsa :       ------------------- g   a      g t                 a  c  t t g          t    c     a   tcct

                               110                                     120
           Leu Leu Gly Asn Val Leu Val Val Val Leu Ala Glu His Phe Gly Asn Ala Phe Ser Pro Ala Val Gln Ala
Lfu : acag.CTC.CTG.GGC.AAC.GTG.CTG.GTG.GTT.GTG.CTG.GCT.GAA.CAC.TTT.GGC.AAT.GCA.TTC.AGC.CCG.GCG.GTG.CAG.GCT
Hsa :                                C TG          C C T               A  A       C    A C A
                                       Cys             His             Lys Glu     Thr     Pro

           130              .                    140
       Ala Phe Gln Lys Val Val Ala Gly Val Ala Asn Ala Leu Ala His Lys Tyr His ***
Lfu : .GCC.TTT.CAG.AAG.GTG.GTG.GCT.GGT.GTG.GCC.AAT.GCT.CTG.GCT.CAC.AAG.TAC.CAC.TGA.GCTCCCTTTCCTGCTGCTCAATT
Hsa :       A        A                       T       C       C           T      A      G     T     TC
           Tyr

Lfu : CCGATTAAAGGTCCCTTTGTCCCCAGAGCCCAACTGCAAAAC-ATAGGTAATTATAAAGGACTTTGAGCATCTGGCTTCCGGCTAATAAAGAACAATTAT
Hsa : T T         T       T   TA  T      ACT    TGGG  AT     G    G C            A   T C        A    T

Lfu : TTTCACTGCaacgaaatgtgttcagtgtgtttatttctgaatctctcactcaaaagggcacatgaaaaagcaaggcatttaagaaataaagaaatgaggg
Hsa :      T     t --       t aa----            a t t   a       a tg  gg ggt  gt        a c            t a

Lfu : gttagttcagaccttgggacaataca------------------------actgtgagtctgtgtacaactaatgtgtgtggcaatagcccctgccgccg
Hsa :  c -     a         a      ctatatcttaaactccatgaaagaag     g   caac  g      caca t     c        at   t

Lfu : atgcctta-ccactcctcaaaaaacgattcaagtggaggcttgattagggggtagaattttgctat------ttttaattatttatttctttagact
Hsa :         tt  t      g    g     tt  a           t ca  t  a           gctgta     c    c     g t   - ctg

Lfu : tcctcataaatgtcttttctctctccaatcgcttgtcctgaatttcat--agcctctactc
Hsa :         g          a  ac   t t    a     c  c  tctc     tg
```

Figure 6.5

Dot-matrix comparison of the brown lemur and human β-globin genes.

The entire 2215 bp of the brown lemur β-globin gene is shown compared to the human β-globin gene (Lawn *et al.*, 1980; Hardison, 1984). Matching criteria and sequence features depicted beside the grid are the same as in Figure 3.2. The strong homology between these two sequences, in particular over the non-coding DNA sequences is indicative of the orthology of these two sequences. It is apparent that this homology may extend further than the presently available sequence data set. The large number of possible homologous alignments generated within parts of the second intron correspond to regions of sequence in both genes rich in T residues.

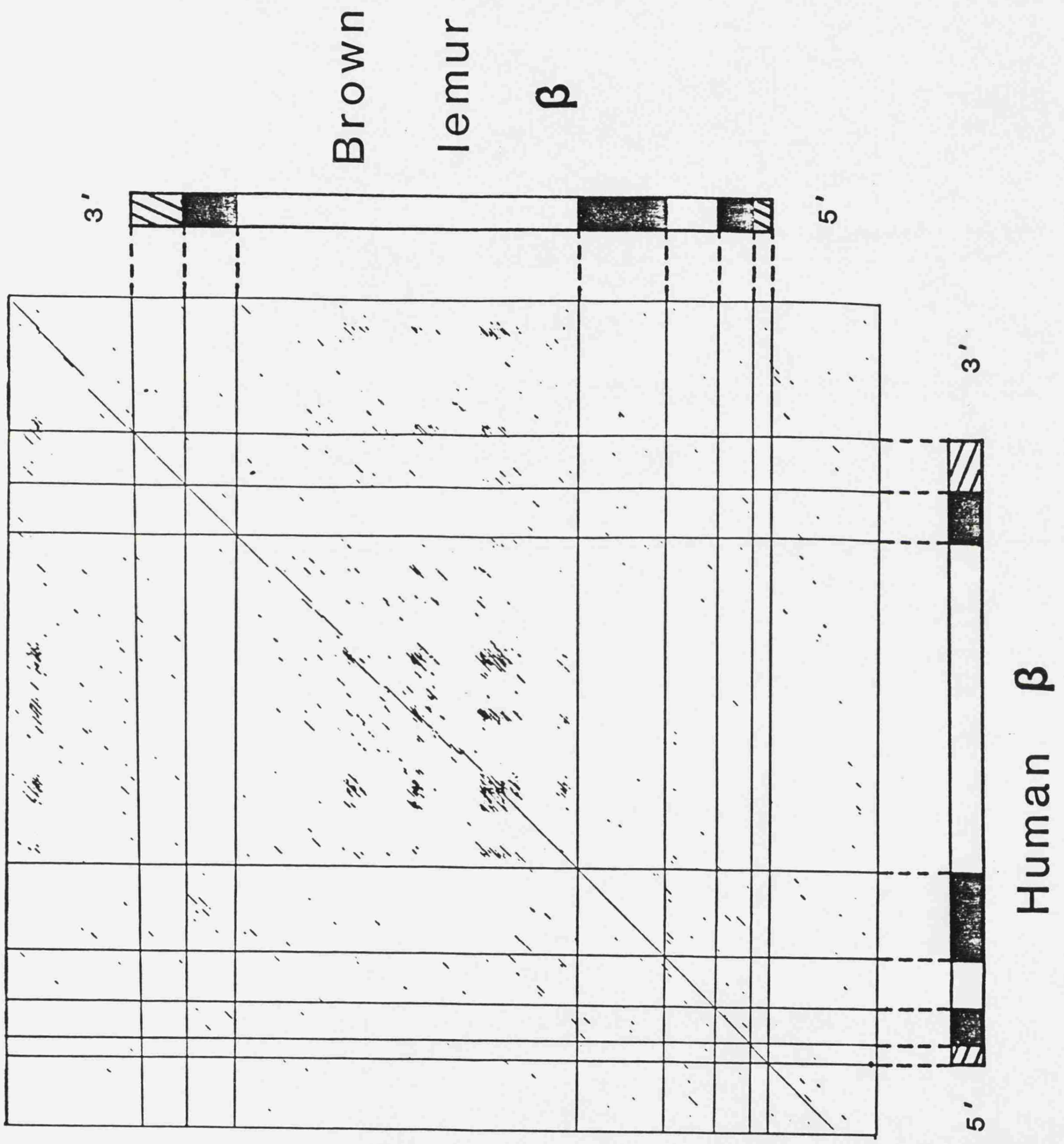

Brown lemur β
3'
5'
Human β
5'
3'

Table 6.1

Comparison of the amino-acid differences between the brown lemur and human β-globin genes.

The 26 amino-acid residues that differ between the brown lemur and human β-globin genes are shown (codon positions correspond to those in Figure 6.4). The three letter code is used to represent the amino-acid residues. Amino-acid residues are classified according to their side chain characteristics as indicated by the key at the foot of the Table. Assuming an overall charge of 0 for the human β-globin polypeptide the observed differences in the brown lemur β-globin polypeptide would result in a small (+1) change in overall charge.

[a] Where known, the functional significance of a particular amino-acid residue is shown according to the following key (taken from Goodman *et al.*, 1981);

DPG = 2,3-diphosphoglycerate binding site (regulation function)

$\alpha_1\beta_1$= non-bohr associated, $\alpha_1\beta_1$ contact sites (cooperative tetramer)

$\alpha_1\beta_2$= non-heme, $\alpha_1\beta_2$ contact sites (cooperative dimer)

INT = interior position (stabilising tertiary structure)

[b] The helical position of each altered amino-acid residue within the protein is taken from the alignments of Dickerson and Geis (1983). This alignment allows the comparison of functionally equivalent residues between globin polypeptides from different vertebrate, invertebrate and plant species. The presence of the amino-acid found in the brown lemur β-globin gene in other α-like (α), β-like (β), myoglobin (M), and invertebrate or plant (O) globin polypeptides is shown in the final column.

Table 6.1

Codon position	Helical^{b} position	Amino-acid Change			Function^{a}	Presence in other^{b} globin genes			
		Human	Brown lemur						
Exon 1						α	β	M	O
1	NA1	Val(N,B)	Thr(P,N)	0	DPG	-	-	-	-
2	NA2	His(P,L)	Leu(N,B)	-	DPG	-	-	-	-
4	A1	Thr(P,N)	Ser(P,N)	0	-	√	√	√	√
5	A2	Pro(N,B)	Ala(P,B)	0	-	√	√	-	√
8	A5	Lys(P,L)	Asn(P,N)	-	-	-	-	-	-
9	A6	Ser(P,N)	Ala(P,B)	0	-	√	√	-	√
10	A7	Ala(P,B)	His(P,L)	+	-	-	-	√	-
13	A10	Ala(P,B)	Ser(P,N)	0	-	√	√	-	√
19	A16	Asn(P,N)	Asp(P,L)	-	-	-	√	√	√
21	B3	Asp(P,L)	Glu(P,L)	0	-	√	√	-	√
22	B4	Glu(P,L)	Lys(P,L)	+	-	-	-	-	√
Exon 2									
50	D1	Thr(P,N)	Ser(P,N)	0	-	-	√	√	√
52	D3	Asp(P,L)	Ser(P,N)	+	-	-	√	-	-
69	E13	Gly(N,B)	Ser(P,N)	0	-	-	-	-	-
73	E17	Asp(P,L)	Glu(P,L)	0	-	-	√	-	√
76	E20	Ala(P,B)	His(P,L)	+	-	-	-	-	-
87	F3	Thr(P,N)	Gln(P,N)	0	-	-	√	-	√
101	G3	Glu(P,L)	Gln(P,N)	+	$\alpha_1\beta_2$	-	-	-	√
104	G6	Arg(P,L)	Thr(P,N)	-	$\alpha_1\beta_1$	-	-	-	-
Exon 3									
112	G14	Cys(P,N)	Val(N,B)	0	$\alpha_1\beta_1$	√	√	-	-
116	G18	His(P,L)	Glu(P,L)	-	$\alpha_1\beta_1$	-	-	√	√
120	GH3	Lys(P,L)	Asn(P,N)	-	$\alpha_1\beta_1$	-	√	-	-
121	GH4	Glu(P,L)	Ala(P,B)	+	-	-	-	-	-
123	H1	Thr(P,N)	Ser(P,N)	0	$\alpha_1\beta_1$	√	√	-	√
125	H3	Pro(N,B)	Ala(P,B)	0	$\alpha_1\beta_1$	√	-	-	√
130	H8	Tyr(P,L)	Phe(N,B)	+	Int	-	√	√	√

(x,y); x= polar residues (P), nonpolar residues (N)
y= neutral (N), hybrophobic (B), hydrophilic (L)

Chapter 7

IDENTIFICATION OF ADDITIONAL Ψβ1-RELATED SEQUENCES IN MAN AND OTHER PRIMATES

7.1 Introduction

The human β-globin gene cluster consists of five functional genes (ε, $^G\gamma$, $^A\gamma$, δ, and β) and a non-processed pseudogene (Ψβ1), see Figure 3.1. The pseudogene in this gene cluster is apparently the result of an ancient gene duplication event; an observation based on the failure to align the non-coding regions of the Ψβ1 gene against equivalent sequences of the other human functional genes and the fact that non-coding DNA hybridisation probes isolated from the pseudogene are capable of detecting, in essentially a gene-specific manner, Ψβ1-related sequences in other primates and mammals that diverged up to ~85 MYs ago (Chapter 3).

However, several additional genomic DNA fragments were also faintly detected by these hybridisation probes, in particular by the human Ψβ1 intron 2 probe (probe 2, Figure 3.4 and below). One possibility is that these additional hybridising fragments are the result of the random drift of non-coding DNA sequences that, by chance, have come to resemble the human Ψβ1 intron 2 probe sufficiently well to be detected at the relatively low hybridisation stringencies used. Another, more interesting possibility, is that these additional human Ψβ1-related fragments may correspond to dispersed, presumably non-processed (as they are detected by the intron 2 probe), Ψβ1 derived pseudogenes that have arisen during

*Figures for this Chapter follow the text.

primate evolution. The theoretical possibility of such sequences being present in the genome has been appreciated for some time, though few examples have been described (see Introduction).

This Chapter describes the further characterisation of these additional fragments and attempts to isolate them from a human genomic λ-recombinant library.

7.2 Detection of additional Ψβ1-related sequences in man and other primates

Figure 7.1 shows more clearly the additional genomic DNA fragments detected in man and several other primates by ^{32}P-labelled human Ψβ1 intron 2 probe (probe 2, Figure 3.4). The larger of the two additional fragments detected by the human Ψβ1 intron 2 probe (probe 2) is also found in other primates; in BglII digests of gorilla, and yellow baboon genomic DNA this larger hybridising fragment is electrophoretically indistinguishable, thought less intense, from the equivalent fragment present in man. This suggests that this fragment may correspond to a region containing homology to Ψβ1 intron 2 that has been present for some time during primate evolution.

The smaller of the two additional hybridising fragments detected by probe 2 in man is apparently absent from the genomes of the other primates examined. However, the apparent absence of an equivalent fragment in the other primates may represent the mutual sequence divergence between those sequences that may be present in the other primates and probe 2, such that they are not detected under the hybridisation conditions used.

This hybridisation to additional fragments in the primates is similar to the additional sequences detected by human Ψβ1 probes in other mammals (see Chapter 3). These additional hybridising fragments were investigated further in man (Figure 7.2) using additional probes capable of detecting human Ψβ1 related sequences. In this experiment duplicate samples of *BglII* digested human DNA were electrophoresed, in the native double stranded form, on a 0.5% agarose gel for approximately twice the normal distance before *in situ* acid/alkali denaturation and Southern transfer to nitrocellulose in order to obtain higher resolution of the larger additional fragments. Hybridisation conditions were the same as those employed previously (1xSSC, 60°C in the presence of dextran sulphate) except they were performed in the presence of human competitor DNA (50µg/ml) to reduce any spurious hybridisation signals that may have arisen from the presence of repetitive DNA sequences in any of the probes employed.

The two additional fragments detected by the human Ψβ1 intron 2 probe (Figure 7.2, track 2) do not comigrate with any of the previously assigned β-globin containing genomic DNA fragments detected by hybridisation to the rabbit adult β-globin probe (Figure 7.2, track 1). In addition, the larger of these two fragments (~18 kb) exceeds the size of any of the *BglII* restriction endonuclease fragments within the human β-globin gene cluster suggesting that this additional fragment must lie outside the characterised human β-globin gene cluster. Of the two fragments, only the larger additional fragment is also detected by the human 5' Ψβ1 probe 1 (Figure 7.2, track 3), suggesting that only this fragment contains a genuine Ψβ1-related sequence (the criterion of hybridisation of both Ψβ1 probes 1 and 2 to similar sized genomic DNA fragments was used previously, Chapter 3, to determine the probable presence of genuine Ψβ1-related sequences in other mammals). Furthermore, the larger of the additional fragments is also detected by the rabbit adult β-globin gene probe, suggesting that this fragment contains sequences with homology to β-globin coding sequence as well as 5' flanking and intron 2 Ψβ1-related sequences. Neither of these additional probes detect the smaller of the two additional fragments detected by probe 2 making this a less likely candidate for a genuine Ψβ1-related sequence.

The results of further hybridisation analysis of the additional sequences detected by the intron 2 probe from the human Ψβ1 gene therefore suggest the existence of one genuine additional Ψβ1-related sequence in the human genome. The nature of this sequence and its position relative to the rest of the human β-globin gene cluster is of great interest as it may constitute an example of a dispersed non-processed pseudogene, the first to be found in man. An attempt was therefore made to isolate and characterise additional Ψβ1-related sequences from a human genomic library.

7.3 Isolation and characterisation of human DNA regions related in sequence to intron 2 of the human Ψβ1 pseudogene

a) Isolation from a human Sau3A partial genomic recombinant library

The human genomic library, a λL47.1 genomic library consisting of ~10^6 recombinants, was produced by Miss P.Weller by the same method described in Chapter 4 for production of the owl monkey genomic library. The human genomic library was screened by the filter hybridisation method of Benton and Davis using ^{32}P-labelled human Ψβ1 probe 2. Hybridisations were performed in 3 x SSC at 60°, a low stringency designed to help distinguish additional faintly hybridising recombinants in the library. Nine positively hybridising regions of differing signal intensity were picked and purified by three rounds of rescreening before amplification and characterisation. Large scale phage DNA preparations were performed on four recombinants λSH3, 9, 7, and 2 representing high to low levels of relative hybridisation signal intensity to the human Ψβ1 probe 2 during screening.

b) Mapping of the potential human Ψβ1-related λ-recombinants

The four λ-recombinants were initially characterised by digestion of 0.5μg samples of DNA with the restriction endonuclease EcoRI and electrophorsis through a 0.5% agarose gel. Two out of four of the recombinants (λSH9 and 7) had identical fragment patterns suggesting they contained identical genomic DNA inserts (results not shown). More detailed restriction endonuclease mapping was therefore only performed on λSH3, 7, and 2 using the restriction ednonucleases EcoRI, HindIII, BglII and BamHI. 0.5μg samples of DNA were digested in all combinations of single and double digests using these four restriction enzymes. The digests were electrophoresed through 0.5% agarose gels, photographed and the DNA denatured *in situ* by acid/alkali treatment before transfer to nitrocellulose by Southern transfer. The filters were then hybridised, in turn, with the rabbit adult β-globin gene probe and the human Ψβ1 gene probes (1 and 2) to distinguish fragments with homology to these probes in each recombinant.

λSH3 was the only one of the three recombinants to give a consistent hybridisation with the rabbit adult β-globin gene probe and the human Ψβ1 probes (results not shown). Comparison of the physical restriction endonuclease map of this recombinant against the physical map of the human β-globin gene cluster established that this recombinant constituted an independent isolate of the region encompassing the Ψβ1 pseudogene from the human β-globin gene cluster.

The failure of the other recombinants (λSH7 and 2) to hybridise to any probe other than the Ψβ1 intron 2 probe (Figure 7.3) caused concern as to the relationship of the genomic inserts of λSH7 and 2 to the human Ψβ1 intron 2 related sequences previously identified (Figure 7.2). To test whether the sequences which hybridised to the human Ψβ1 probe 2 within these clones corresponded to either of the additional human genomic DNA fragments in Figure 7.2 restriction endonuclease fragments that hybridised to probe 2 were isolated from λSH2 and λSH7 (Figure 7.3) and used as DNA probes in filter hybridisations against human genomic DNA digested with EcoRI or BglII (Figure 7.4).

Neither recombinant derived probe detects the additional fragments described previously; instead each hybridised to a different specific genomic DNA fragment while also hybridising to other sequences present throughout the genome, even at the high strengency used and in the presence of human competitor DNA. This general background hybridisation (particularly in the case of the λSH2 fragment) may reflect the presence of repetitive sequences as part of the probe. These recombinants do not therefore appear to correspond to either of the additional Ψβ1 intron 2 related sequences previously detected in human genomic DNA.

The source of homology between these recombinant sequences and the human Ψβ1 intron 2 probe used to isolate them from the human genomic library was investigated further by sequencing of the smallest hybridising fragment detected by probe 2 from the genomic recombinant λSH7. This region was sequenced as it was relatively small and amenable to rapid sequence analysis.

7.4 Sequence analysis of the human Ψβ1 intron 2 related sequence of λSH7

The region of λSH7 which hybridises to human Ψβ1 probe 2 is present within a small BamHI-EcoRI double digest fragment of ~520 bp (Figure 7.3). The strategy for sequencing this fragment took advantage of the M13mp8 and mp9 cloning and sequencing vectors which have their universal cloning sites orientated in opposite directions relative to the universal primer-annealing site used in M13 sequencing. Standard M13 sequencing gels allow ≥350 bp to be read per substrate therefore by force cloning the λSH7 BamHI-EcoRI fragment in opposite orientations into M13mp8 and mp9 the ≈520 bp fragment could be provisionally sequenced by two opposing sequencing runs on opposite single strands.

The ≈520 bp fragment from λSH7 was prepared by double digestion of 16μg of λSH7 DNA with restriction endonucleases EcoRI and BamHI. The 520 bp fragment was isolated from the majority of digestion products by the DE81 paper recovery method (2.10(iii)) after electrophoresis through a 1% agarose gel. As the fragment migrated close to an EcoRI single digest product of similar size, both fragments were recovered and ≈80ng of each were ligated with 20ng of M13mp8 and mp9 DNA, also digested with BamHI and EcoRI. A total of 50ng of ligated material was used to transform the E.coli M13 host strain JM103 by the MOPS-RbCl$_2$ method.

One set of ligated material gave a very low level of transformation, similar to that of the self-ligated M13mp8 and mp9 controls, and was presumed to correspond to the recovered fragment with the two EcoRI termini and reflected the inability of this fragment to ligate with and recircularise the vector. Recombinant M13 clones were picked from the successful transformation with the 520 bp double digest fragment and single stranded M13 sequencing substrate DNA prepared. Sequencing of several of the M13mp8 and mp9 recombinants by the dideoxyribonuclease chain termination method confirmed they all corresponded to a single insert. The two orientations of the 520 bp fragment gave sequence from opposing strands which overlapped at the ends distal to the primer-annealing site. The complete provisional sequence is shown in Figure 7.5(a).

Dot-matrices of this sequence against intron 2 of the human Ψβ1 gene revealed only one small region with any homology, Figure 7.5(b). This region corresponds to a tandem repeat of the four base sequence TATG that is present in both sequences surrounded by non-homolous DNA. 10 perfect copies of this basic 4 bp repeat are present in the λSH7 sequence while 5 perfect copies are present in the human Ψβ1 intron 2 sequence. There appear to no other features in common between these two sequences.

It was concluded therefore that the human λ-recombinant λSH7 (and also possibly λSH2) was related to the human Ψβ1 intron 2 probe by this small repetitive region as no other homology was observed. In retrospect, a combination of hybridisation probes at a higher hybridisation stringency may have proved more successful for isolating additional Ψβ1-related sequences.

7.5 Summary

The additional genomic DNA fragments detected by the human Ψβ1 intron 2 probe have been characterised further by hybridisation with two other probes capable of detecting Ψβ1-related sequences in man and other primates. The larger of the two additional fragments detected by the intron 2 probe in human genomic DNA digests is also detected in the higher primates and is also, in man, detected by other probes known to hybridise to human Ψβ1 sequences suggesting that this fragment may well correspond to a genuine additional Ψβ1-related sequence within the human genome.

In contrast, the smaller of the two additional fragments detected by the intron 2 probe in human genomic DNA digests is not detected in genomic digests of other primate DNAs. Neither, in man, is this fragment detected by other probes capable of detecting Ψβ1-related sequences in man and other primates. This fragment is therefore thought less likely to correspond to a genuine additional Ψβ1-related sequence, though the mutual sequence divergence betwee~~e~~n this sequence and the probes used may preclude detection under the hybridisation conditions employed.

An attempt to isolate these additional Ψβ1-related sequences from the human genome resulted in the isolation of sequences with limited sequence homology to Ψβ1 intron 2 sequences. The homology is confined to a short repetitive sequence that may, by chance, have been generated elsewhere in the genome and may have been isolated due to the reduced hybridisation stringencies used during the screening of genomic clones.

Figure 7.1

Detection of additional Ψβ1 intron 2 related sequences in man and other primates

DNA digestion (5μg per track), gel electrophoresis and Southern blotting were performed as described in Figure 3.4. Hybridisation to ^{32}P-labelled Ψβ1 intron 2 probe (probe 2, Figure 3.4) was performed overnight in 1 x SSC at 60°C in the presence of dextran sulphate. Autoradiographic exposure was for 3 days. Molecular weight markers are λ x HindIII.

The size of the principal genomic DNA fragment detected in man and these primates (solid circle) corresponds to that predicted from the restriction endonuclease site map encompassing the Ψβ1 pseudogene region of the relevant β-globin gene cluster. The two additional DNA fragments detected in man are indicated, in order of signal intensity, by filled and open triangles respectively.

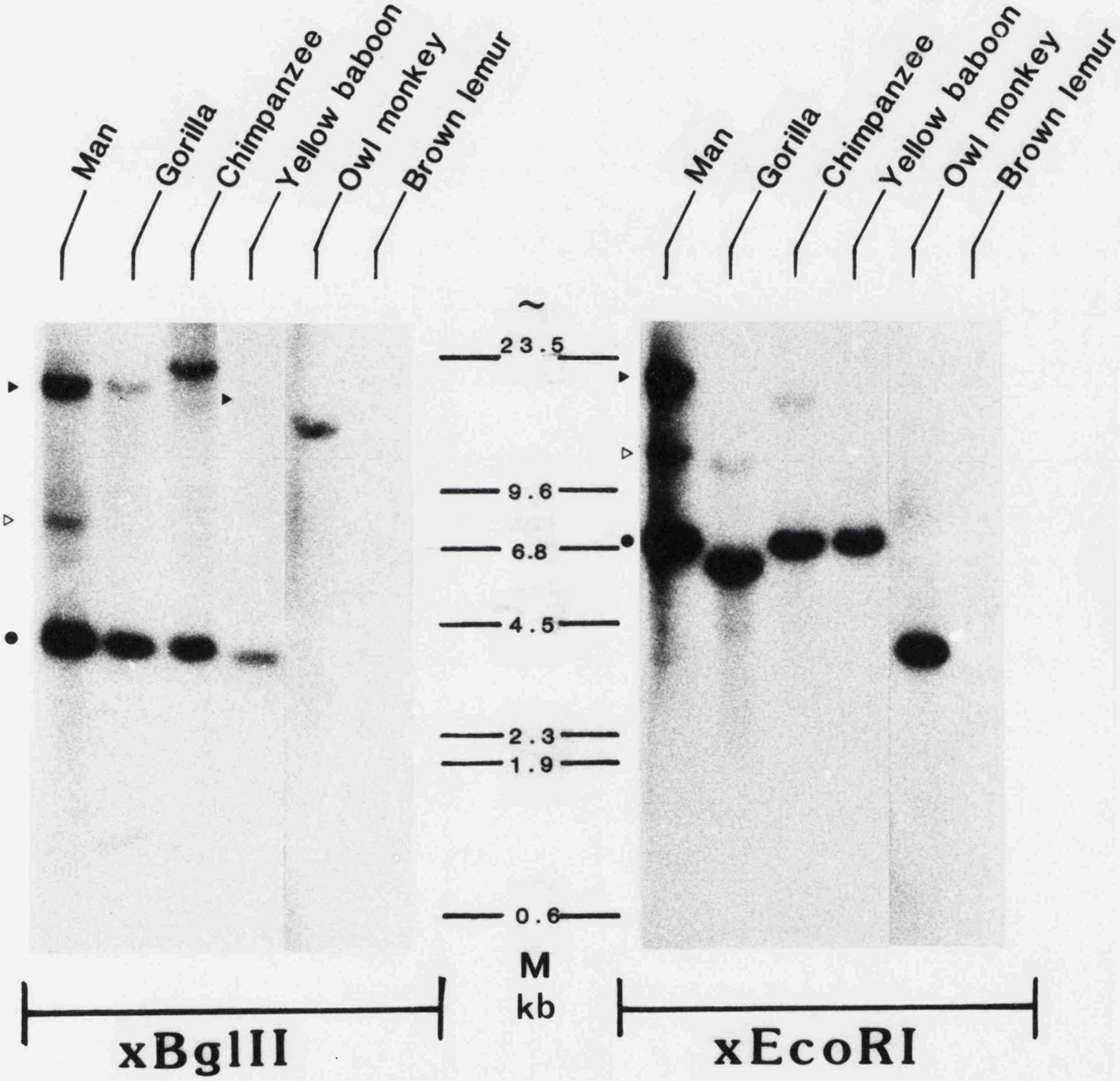

Man
Gorilla
Chimpanzee
Yellow baboon
Owl monkey
Brown lemur
Man
Gorilla
Chimpanzee
Yellow baboon
Owl monkey
Brown lemur
~
23.5
9.6
6.8
4.5
2.3
1.9
0.6
M
kb
xBglII
xEcoRI

Figure 7.2

Characterisation of additional human Ψβ1 intron 2 related sequences by hybridisation to additional probes capable of detecting globin related sequences.

DNA digestion, gel electrophoresis and Southern transfer to nitrocellulose were as described in Figure 3.4 except that the DNA was electrophoresed for ~twice the previous distance. Hybridisations were performed overnight in 1 x SSC at 60°C in the presence of dextran sulphate and human competitor DNA (50μg/ml). Tracks R, 1 and 2 correspond to hybridisations with ^{32}P-labelled rabbit adult β-globin cDNA probe and human Ψβ1 probe 1 and 2 respectively. Autoradiographic exposures were for 3 days. Molecular weight markers are λ x HindIII.

Previously characterised DNA fragments containing globin related sequences and detected by the rabbit adult β-globin cDNA probe (Barrie et al., 1980; Fritsch et al., 1980) are indicated beside tracks R. The filled triangles indicate the DNA fragments that correspond to the large, probably genuine, additional Ψβ1 intron 2 related fragment. The open triangle indicates the Ψβ1 intron 2 related fragment not detected by the two additional probes (see text).

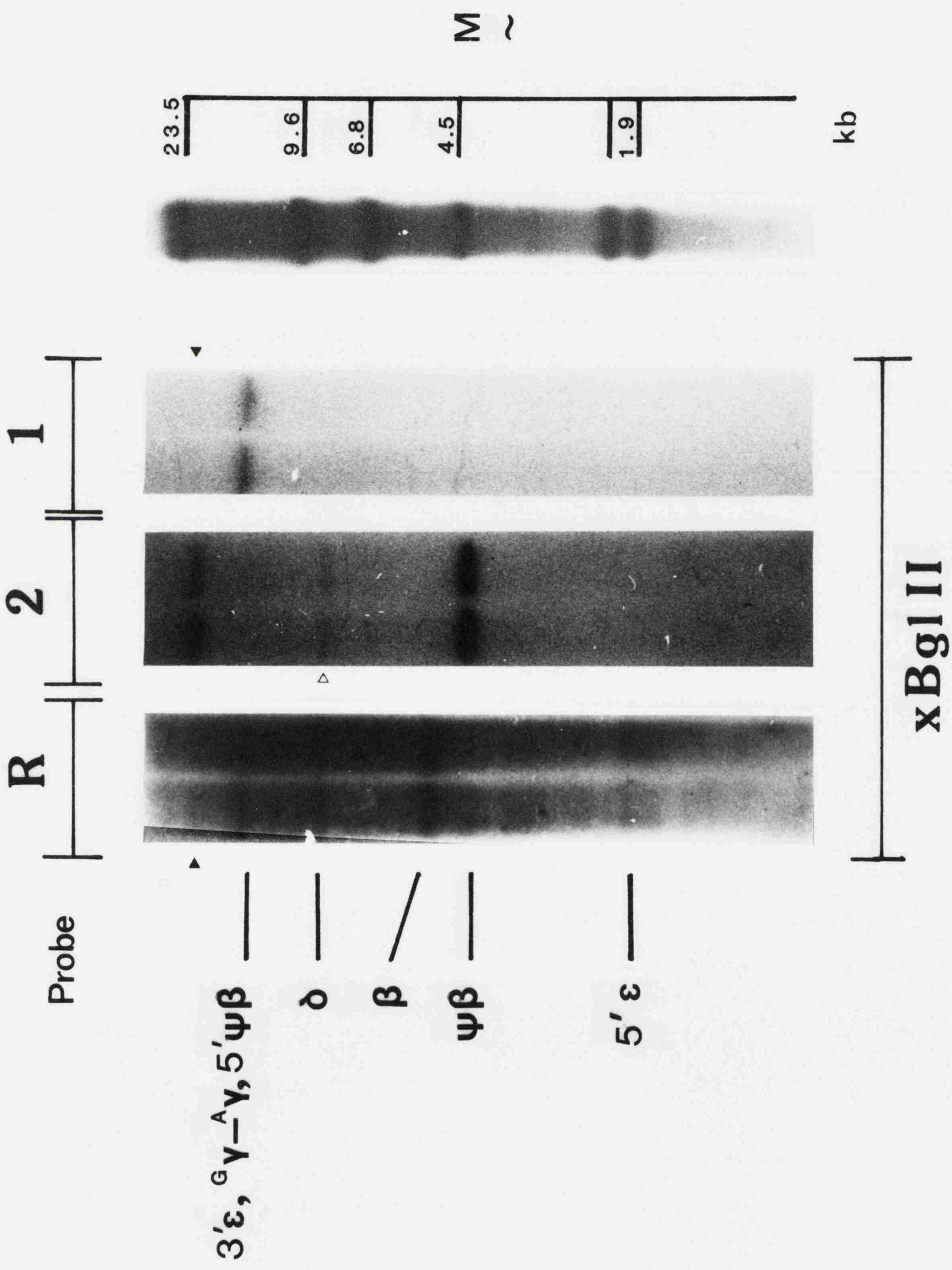

Probe
3′ε, Gγ–Aγ, 5′ψβ
δ
β
ψβ
5′ε
R
2
1
xBgl II
M
23.5
9.6
6.8
4.5
1.9
kb

Figure 7.3

Characterisation of human genomic recombinants λSH2 and λSH7.

DNA digestion, gel electrophoresis and Southern blotting were performed as described in Figure 4.4. Hybridisation against ^{32}P-labelled probes were performed overnight in 1 x SSC at 65°C in the presence of human competitor DNA (50μg/ml) and absence of dextran sulphate. Autoradiographic exposure was for 10 days. Molecular weight markers are λ x Hind III.

The λSH7 DNA was contaminated by DNA corresponding to the recombinant λSH3 (the independent isolate of the human Ψβ1 pseudogene). Hybridising DNA fragments that correspond to this clone (open triangles) are distinguishable by their strong hybridisation signal and the absence of equivalent sized DNA fragments in the accompanying agarose gel photograph.

Fragments isolated from λSH2 and λSH7 in order to determine whether these recombinants correspond to the additional fragments detected in human genomic digests (see Figure 7.5) are indicated by the open circles.

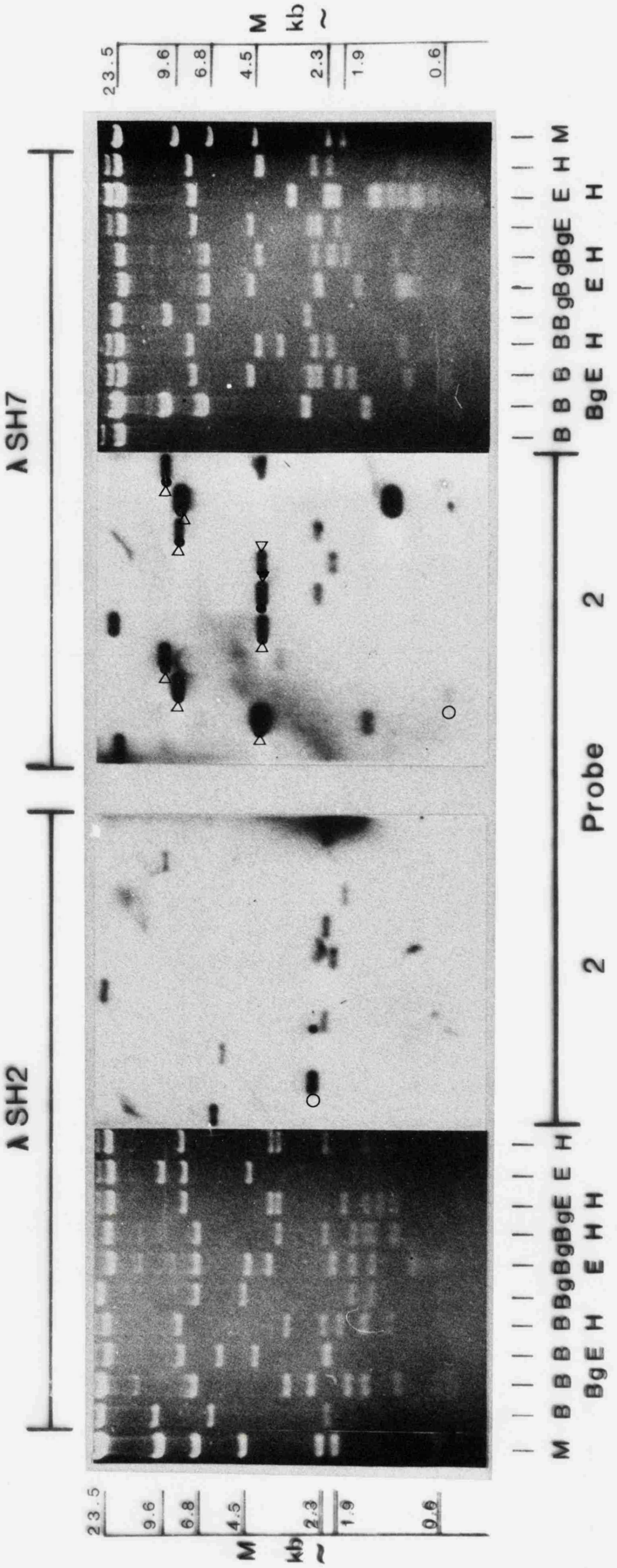

λSH7
λSH2
Probe
2
2
M
kb
~
23.5
9.6
6.8
4.5
2.3
1.9
0.6
B B B B BgBgBgE E H M
BgE H E H H
M B B B B BgBgBgE E H
BgE H E H H

Figure 7.4

Hybridisation of λSH2 and λSH7 DNA fragments to total human genomic DNA digests

DNA digests, gel electrophoresis and Southern blotting were as described for Figure 7.2. Hybridisations against ^{32}P-labelled probes were performed overnight in 1 x SSC at 65°C in the presence of dextran sulphate and human competitor DNA (50μg/ml). The BglII x probe 2 tracks are taken from Figure 7.2. Filters hybridised to probes isolated from λ-recombinants were given a stringent wash in 0.2 x SSC. Autoradiographic exposures were for 3-10 days depending on signal intensity. Molecular weight markers were λ x HindIII.

Principal fragments detected by recombinant probes are indicated by the filled triangles. As can be seen neither recombinant derived probe detects the additional fragments detected by the human Ψβ1 intron 2 probe (see text).

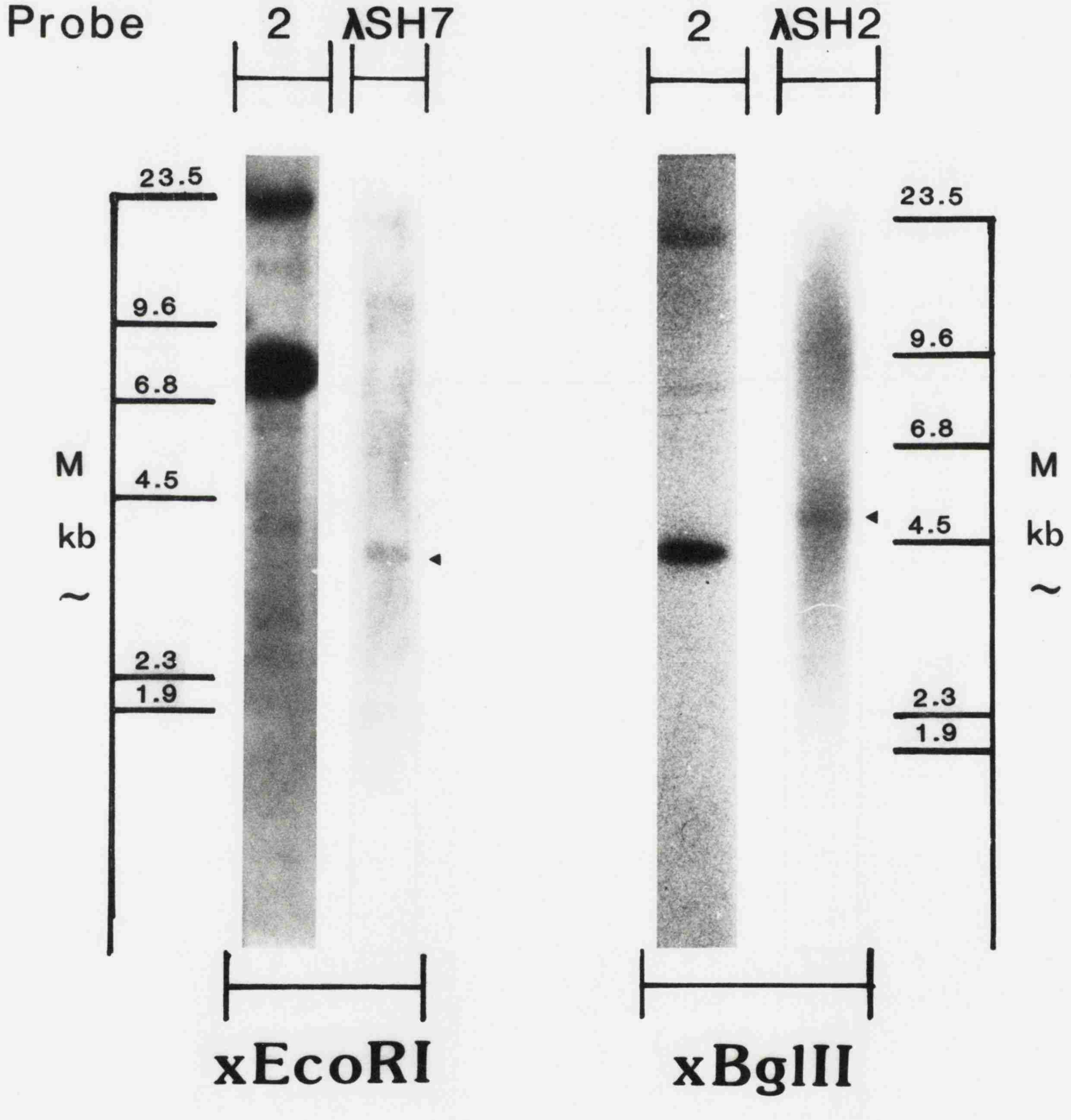

Probe
2
λSH7
2
λSH2
23.5
9.6
6.8
M
4.5
kb
~
2.3
1.9
23.5
9.6
6.8
M
4.5
kb
~
2.3
1.9
xEcoRI
xBglII

Figure 7.5

Characterisation of the ~520 bp *Eco*RI-*Bam*HI fragment fron λSH7.

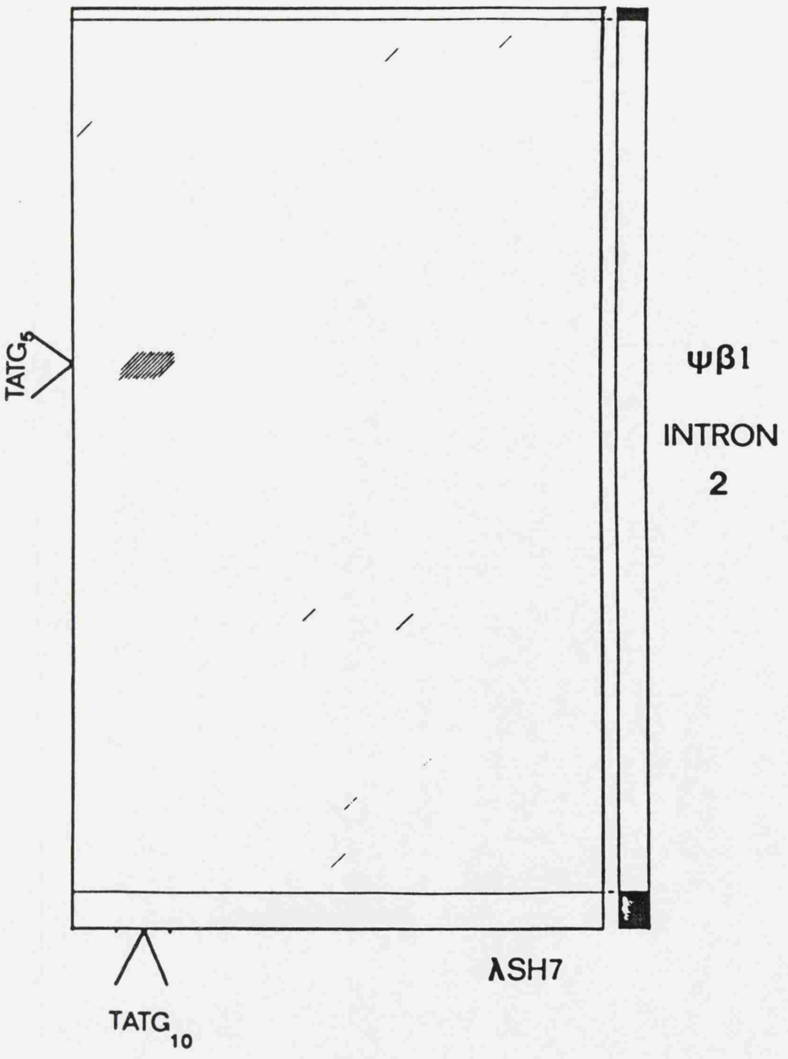
a) Sequence of the *Eco*RI-*Bam*HI fragment isolated from λSH7.

The 520 bps of sequence shown was determined after force cloning into M13mp8 and mp9 (see text). The sequence was confirmed by sequencing several identical M13 clones. The sequence, shown as the nonsense strand, was orientated 5' → 3' after dot-matrix comparison against the human Ψβ1 sequence (see below). Single strand sequence was read in the direction shown by the arrows either as presented (top arrow) or as the complimentary strand to that shown (bottom arrow).

b) Dot-matrix comparison of the λSH7 sequence with the intron 2 sequence of the human Ψβ1 pseudogene.

The 520 bp sequence determined for the λSH7 recombinant (and its compliment, results not shown) is shown aligned against intron 2 and part of exon 2 and exon 3 of the human Ψβ1 pseudogene. Only a small region of homology is detected in this analysis (matching criteria as for Figure 3.2). Altering the matching criteria did not increase the apparent homology between these two sequences but did increase the background 'spurious' matches. The sequence feature corresponding to this region of homology (TATG$_n$) is shown beside the grid. Similar dot-matrix comparisons failed to distinguish any homology between this λSH7 sequence fragment and any of the other human β-like globin genes (results not shown).

```
              10        20        30        40        50        60
λSH7  :  TCCGGTCCTCAAGCTCTGATCGGAGACGTTGTATCGTTCTGGGGTAGAGTTTATGTATGT

              70        80        90       100       110       120
λSH7  :  ATGTATGTATGTATGTATGTATGTATGTATGTACGTATTTTTCAGAAACCGTAAGTAAAG

             130       140       150       160       170       180
λSH7  :  TAGGTCGTAATTGAACGTCTTCTGTGAGAGTCTTCTGATTTCGGTTATTTAACAATTGTA

             190       200       210       220       230       240
λSH7  :  TAGATTTTACTTCTGAGAAGTATAGTCTATTTTAGTTTTAATTTATTGAGTGTTCTTTAA

             250       260       270       280       290       300
λSH7  :  TTACTTACGTTTTTTTTCCTTTAACCTTCTCTTATCTTACACCTCTTTAATAAGATTTTG

             310       320       330       340       350       360
λSH7  :  TAACGGTTAATTTGTACCATTCAAACTGAATGTACCTTCCACTAGTCGGANTTTTCTTACG

             370       380       390       400       410       420
λSH7  :  TGTCTCCGACACTTGAGAGTAAAACTTAGTCCTCTGTTCATAACTTCTTTTTATGGAGTT

             430       440       450       460       470       480
λSH7  :  TGATTTTTATCGATTCTTACAATTAGGAACCTCCTTCATATAGGTACCGACCATACTGGT

             490       500       510       520
λSH7  :  TGGTCTTTTCTTGTCTTAATACAGGGAATAAGATCNGGTT
```



Chapter 8

DISCUSSION

8.1 Introduction

During the course of this work several groups have reported the further characterisation and sequence analysis of several different mammalian β-globin gene clusters that have contributed to our understanding of many of the features of the molecular evolution of this gene family (see Collins and Weissman, 1984). However this discussion concentrates primarily on the evolutionary history of the pseudogene, Ψβ1, present in the human β-globin gene cluster which, except for the initial observation concerning the orthology of this sequence to the goat ε" gene referred to in Chapter 3, was performed independently from these other research groups. Relevent features of the molecular analysis of these other mammalian clusters are however referred to throughout and the conclusions drawn from this work are discussed in section 8.12/13 with reference to the evolution of other non-primate mammalian β-globin gene clusters.

8.2 Establishment of gene orthologies by dot-matrix criteria

Unlike the coding regions (exons) of the functional human ε-,γ-,δ-, and β-globin genes most of the flanking and non-coding regions, in particularly intron 2, are heavily diverged from one another; a feature which is also found, in general, for other mammalian intraspecific β-globin gene comparisons (Hardison, 1984; Hardies _et al._, 1984; Hill _et_ *al.*, 1984). Even very low stringency dot-matrix criteria, capable of detecting homology between globin sequences upto 40% diverged (see 3.2; White *et al.*, 1984), fail to detect significant alignment between the different functional β-globin genes of man. This substantial non-coding DNA sequence divergence suggests that discrete ε-,γ-,δ-, and β-globin genes have been in existence for a considerable time. It also implies that during primate evolution the rate of gene conversion encompassing non-coding regions of different globin genes must, in general, have been low relative to the mutation rate to have allowed the substantial divergence observed over the non-coding regions of these genes. Dot-matrix alignment over non-coding regions can however reveal regions that have been involved in gene conversion due to the resultant reduction in accmulated sequence divergence between the two contemporary DNA sequences (see Figure 3.2).

Dot-matrix alignment of non-coding DNA sequences, when comparing β-globin genes from different mammalian orders, has been employed throughout this thesis as a powerful indicator of interspecies gene orthology and has complemented other analyses based upon relative coding sequence divergence. Several other research groups have recently used dot-matrix criteria to demonstrate the presence of at least four distinct β-related globin gene sequences in other non-primate mammals and from this have inferred the presence of these sequences in a common ancestral eutherian β-globin gene cluster (see 8.12).

8.3 A complex evolutionary history for the contemporary δ-globin gene.

One example of the usefulness of dot-matrix analysis in determining gene histories is demonstrated by the elucidation of the complex evolutionary history of the δ-globin gene during primate and non-primate mammalian evolution. In man, adult haemoglobin consists primarily of HbA ($\alpha_2\beta_2$) together with a small proportion of the minor adult haemoglobin HbA_2 ($\alpha_2\delta_2$), containing the product of the δ-globin gene. Amino-acid and DNA sequence divergence of the closely related human δ and β globins initially suggested a δ/β duplication ~40 MYs ago (Dayhoff, 1972; Eftratiadis et al., 1980). This estimated divergence time accords well with the presence (detected by genomic mapping and hybridisation) of a duplicated adult β-globin gene arrangement in primate groups thought to have diverged since the postulated δ/β duplication, that is, in the simians (man, great apes and Old World and New World monkeys) but not before, that is, in the prosimians (Zimmer et al., 1980; Barrie et al., 1981). However, while both the brown and ruffed lemur (both prosimians of the lemur group, the only prosimian primates so far examined) apparently lack a duplicated β-globin gene arrangement hybridisation evidence suggested the presence of a second sequence (called Ψβ) with partial sequence homology to the human β-globin gene (Barrie et al., 1981).

Dot-matrix analysis of the brown lemur Ψβ gene (this thesis, Figure 5.6) against each of the other human β-like globin genes clearly demonstrates that this gene is orthologously related to the human δ-globin gene in the region 3' of the second exon, this homology being particularly striking over the diagnostic non-coding DNA sequences of intron 2 and the 3' flanking regions. The presence of δ-like sequences in the brown lemur suggests therefore that the δ/β gene duplication occurred not 40 MYs ago as initially estimated but some time prior to the basal primate radiation ~70 MYs ago.

Dot-matrix comparsion of the human δ and β genes and close examination of silent site and non-coding DNA sequence divergence between these two genes strongly suggest that the reason for this discrepancy is a gene conversion event between the δ and β globin genes in the lineage leading to man (Eftratiadis et al., 1980; Figure 3.2). The gene conversion event has resulted in the homogenisation of δ-globin sequences (proximal 5' flanking, exon 1, intron 1 and exon 2) such that they resemble equivalent β-globin gene sequences (on dot-matrix analysis this is particularly obvious over intron 1). In contrast, intron 2 and the 3' flanking sequences of the human δ gene show no homology to any other human β-like globin gene suggesting these regions of the δ gene have not been involved in a recent gene conversion (results not shown). The amino-acid and replacement site DNA sequence divergence of the human δ and β globins therefore establishes when the latest gene conversion occurred between these two genes rather than the initial duplication event.

The failure to detect homology between the human δ-globin gene and any of the other β-like globin genes over intron 2 and the 3' flanking regions (even at stringencies capable of detecting homology between sequences ≤40% diverged, results not shown) suggests the δ gene is the result of an ancient rather than recent duplication event. The rate of non-coding DNA sequence evolution observed in the primate Ψβ1 gene (see 8.8) gives a minimum estimate for the period of independent evolution of the δ globin gene of ≈140 MYs (this is the time it would take to accumulate ≥40% divergence between two related non-coding DNA sequences at the primate pseudogene rate). As this estimate exceeds the divergence time estimated for the mammalian radiation (≈80 MYs ago) a prediction of the dot-matrix analysis of the human δ-globin gene is that orthologues of this gene may be present in other mammalian β-globin gene families.

The arrangement of several other mammalian β-globin gene families has been established, and many of the genes sequenced (see Collins and Weissman, 1984). In both the mouse (Hill _et_ _al_., 1984; Hardies _et_ _al_., 1984) and the rabbit (Hardison, 1984) it has been shown by dot-matrix criteria and parsimony analysis of coding regions that evolutionary orthologues exist to each of the functional human genes, including δ. In the rabbit the orthologue of the human δ gene is a recently silenced pseudogene (Ψβ2) that has apparently undergone a gene conversion against the rabbit adult β1 gene in a similar manner, and over a similar region of the gene, to that in the human lineage (Lacy and Maniatis, 1980; Hardison and Margot, 1984). In the mouse pseudogenes Ψβh2 and Ψβh3 both show homology to the human δ gene and have also apparently evolved in concert with adult β-globin genes in the cluster such that the 5' regions of these genes are also β-like (Phillips _et_ _al_., 1984; Hardies _et_ _al_., 1984, however see 8.12). The presence of genes with orthology to the human δ-globin gene in other mammalian lineages strongly suggests the common mammalian ancestor contained a distinct δ-globin gene, as predicted above from the mutual sequence divergence over intron 2 of the human δ-globin gene and the other human β-like globin genes.

The silencing of the δ-globin gene in all but a few primate lineages (man, great apes and New World monkeys) suggests that whatever the functional role of the ancestral δ-globin gene, if any, expression of the contemporary δ-globin gene is no longer essential in these species. In man it has been suggested that one role of the δ polypeptide may be to prevent gelation of haemoglobin in the erythrocytes of people with sickle cell anaemia (Nagel et al., 1979). However while the function of HbA_2 ($\alpha_2\delta_2$) in normal individuals remains unclear this function is apparently no longer essential in other primates, as the δ gene is a pseudogene in the Old World monkey lineage (Kimura and Tagaki, 1983; Martin et al., 1983) and in the lemurs (this thesis). It has been suggested (Martin et al., 1983) that the low level expression of the δ gene may be the result of the conversion of the 5' region of the δ gene by the β gene early in simian evolution. Whether this represents partial "reactivation" of a previously silent δ gene or a reduction in the level of expression of an active δ gene, due to disruption of the 5' transcription signals of the δ gene, is unknown. In the absence of a contemporary δ-globin gene that has not been involved in recombinational exchanges with other members of the β-globin gene family (see below) the functional status of the ancestral δ gene is unlikely to be resolved.

The absence of a contemporary representative of the ancestral δ-gene is apparently due to the propensity of this locus, particularly the 5' regions, to be involved in recombinational exchanges with other members of the β-globin gene cluster. In rabbit, mouse and primate evolution the δ-like globin gene has undergone gene conversions with the neighbouring adult β-globin gene that has resulted in varying extents of the gene being converted to resemble the β-globin locus. In the lemurs the δ gene has apparently been involved in an unequal exchange with the neighbouring Ψβ1 pseudogene (see 8.5). The reason for the apparent recombinational activity of this gene is as yet unknown. In man at least, this susceptibility to recombinational exchange may reflect the position of the δ gene in the β-globin gene cluster, as the δ gene resides within a 9.1 kb region that is thought to contain a recombinational hotspot (see Orkin and Kazazian, 1984). Whether equivalent hotspots exist within the rabbit and mouse and whether these may have influenced the evolution of the δ globin locus in these lineages has yet to be determined.

In summary, contemporary mammalian δ-globin genes are probably descended from a common ancestral gene that arose as the result of an adult gene duplication at least 140 MYs ago. In the absence of an intact contemporary representative of this gene, due to the propensity of this locus to undergo recombinatinal exchanges with other members of the gene family, the functional status of the ancestral gene remains unclear. For example, the ancestral primate δ-globin gene that given rise to contemporary δ gene sequences could have been either a functional gene or a pseudogene. A genuine δ globin gene may be present in those unexamined prosimian and simian primates which diverged from the higher primates prior to the gene conversion involving β; a particularly strong candidate for such a study would be the tarsier which has been reported to contain a HbA_2-like haemoglobin (Beard _et al_., 1976).

8.4 Evidence for an ancient Ψβ1-like gene prior to the mammalian radiation

Computer mediated dot-matrix analysis of the DNA sequence of the human Ψβ1 globin gene fails to detect any significant homology between non-coding regions of this gene and similar regions in any of the other functional β-related genes in man. In contrast, coding sequences, which can still be clearly distinguished even though the sequence itself is not expressed, show preferential homology to the human non-adult β-globin genes, particularly the γ-globin gene (Goodman _et al._, 1984). For the reasons discussed in the case of the human δ-globin gene (p1~~28~~96), failure of dot-matrix analysis to detect non-coding sequence homology between Ψβ1 and the other functional globin genes suggests that the human Ψβ1 gene probably arose as the result of an ancient rather than recent duplication of a non-adult β-like gene. Similarly, this predicts that Ψβ1-related sequences may therefore be present in other contemporary primate and mammalian species.

The presence of Ψβ1-related sequences in all the major primate groups, and the probable existance of at least one intact Ψβ1-related sequence in carnivore and pinniped DNAs (see Chapter 3) confirms the above prediction and strongly suggests that a distinct Ψβ1-like gene, along with ε-,γ-,δ-, and β-like genes, existed within the common ancestral β-globin gene cluster that predated the eutherian radiation 80 MY ago. Failure of cross-hybridisation to provide evidence of Ψβ-related sequences in some mammalian DNAs may simply reflect excessive non-coding DNA sequence divergence between these particular lineages and man rather than the absence of sequences related to the Ψβ1 gene in other mammals.

8.5 Involvement of Ψβ1-related sequences in an unequal exchange with the δ-globin gene in lemurs

The brown lemur β-globin gene cluster contains a single ε-, γ- and β-related globin gene plus a pseudogene between the γ and β-like genes. Characterisation of the lemur "Ψβ" gene (Chapter 5) suggests that this sequence is in fact a hybrid Lepore-type Ψβ1-δ pseudogene, which has resulted from an unequal exchange between adjacent Ψβ1 and δ-like sequences during lemur evolution. The unequal exchange and fixation of the Lepore gene probably occurred in a common ancestor of the lemurs early in their evolution as the ruffed lemur (Barrie et al., 1984), dwarf lemur and sifaka (Chapter 3) all exhibit the same hybridisation pattern against human globin DNA probes capable of detecting the brown lemur hybrid gene.

In subsequent discussions the brown lemur Ψβ1-related sequence is considered to correspond to only those sequences up to the end of exon 2. The precise point of unequal exchange is obscured by subsequent sequence divergence but most likely corresponds to a region of homology shared by the genes involved. In this case the region of homology corresponds to exon 2, the unequal exchange probably occuring near codons 86-87 (see Figure 5.6). A testable consequence of the fixation of the Lepore chromosome would be the loss from the lemur genome of a large portion of intergenic DNA corresponding to the region between the Ψβ1 and δ genes in man. Direct evidence for the deletion of Ψβ1-δ intergenic DNA from the genome of contemporary lemurs is provided by the absence of hybridisation to the human Ψβ1 intron 2 probe (Chapter 3).

8.6 Silencing of the Ψβ1-like gene early in primate evolution

The availability of primate sequences orthologous to the human Ψβ1 pseudogene provides, for the first time, an opportunity to phylogenetically reconstruct the evolutionary history of a contemporary pseudogene and to answer questions concerning the mode and tempo of evolutionary change that has occurred within the different lineages of a mammalian order. For example, by comparing any defects in the primate Ψβ1-related sequences is it possible to establish when and possibly how the human Ψβ1 pseudogene was initially silenced ? (see below), and if so, how has the Ψβ1 gene sequence subsequently evolved in the absence of selective constraint when compared to sequence evolution in a functional gene ? (see 8.8 and 8.9). In order to answer these questions the human Ψβ1 pseudogene and the owl monkey and brown lemur Ψβ1-related sequences were aligned and coding sequence defects and other DNA sequence differences established (Figure 8.1*).

The defects in the human, chimpanzee and gorilla Ψβ1 genes have been discussed in detail elsewhere (Chang and Slightom, 1984) but are described here for completeness. These hominoid Ψβ1-globin genes share an initiation codon change (ATG → GTA), a termination signal at codon 15, due to a single base substitution, two in phase nonsense mutations and two single base deletions, the first of which in codon 20 introduces several downstream termination signals. All defects are commom to the Ψβ1 gene of these three species suggesting the ancestral gene was itself a pseudogene and that since man and the great apes diverged ~7 MYs ago no additional

*Figures and Tables in this Chapter form part of the text.

Figure 8.1

DNA sequence comparison of human, New World monkey and prosimian Ψβ1 pseudogenes.

The sequence of the human (*Homo sapiens*) Ψβ1 pseudogene (Hsa; Jagadeeswaran *et al*,1983; Chang and Slightom, 1984) is shown aligned with the owl monkey (*Aotus trivirgatus*) Ψβ1 pseudogene (Atr); the 5' half of the brown lemur (*Lemur macaco (fulvus) mayottensis*) hybrid Ψβ1-δ globin gene (Lfu) and the goat (*Capra hircus*) ε^{II}-globin gene (Chi; Shapiro *et al.*, 1983). Only differences from the human Ψβ1 sequence are shown. A dash indicates the absence of a nucleotide in a sequence. Homologues of coding sequences are shown in uppercase letters. Codon phasing was established by comparison with the goat ε^{II}-globin gene. Sequences implicated in globin gene transcription and mRNA maturation are indicated by bold underlined characters.

Defects within the coding sequence, each of which is sufficient to silence the gene, are numbered for easy reference to Figure 8.2 and 8.3. In some instances, the position of a microinsertion/ deletion is ambiguous, within a few nucleotides, and the indicated position is therefore placed arbitarily within these limits (for example, defect 5).

```
                  10        20        30        40        50        60        70        80        90       100
Hma Vβ : taattcttatccttctcagttgtaacttatttccatttgtcagctt-caggt-attgcggc----tgctggcgaagtccttgagaaataaactgcacact
Atr Vβ :  g            t      c       c c t           a t     t                                t        t    —
Lfu Vβ :   t       tgc t tg    c   tc     c tca aa      c t     t  a t  gtcag  g  at   a                  t
                                                                     gctag  g  at gg c

Hma Vβ : ggatggtggggg-tagtgtaggaaaa-tggaggggaaggaagtaaagtttcaaattaagcctgaacagcaaagttcccctgagaaggccacctggattct
Atr Vβ : ---     a  -     g  a    - a  t     a     ag                                   c   t  -gt  ga  aca  g
Lfu Vβ : t      a   a        cc    aaa       a      g g     a     g t       tg   g   c   -   gg          ------
Chi ε° : -----------------------------------------------------------------------| t   g   at     t ga     acc --

Hma Vβ : atcagaaactcgaatgtccatcttgc-aaaacttcc-ttgcccaaaccccacccctgg-agtcacaaccccacccttgaccaatagattcatttcactgag
Atr Vβ :                           -         tt     t         tt  -                                          g
Lfu Vβ : -- t    g  a a    a         cat-   g c   c          t      t g -        g      a        cc  .     t ag
Chi ε° :         g  a a                 tg cc ac  -        gtt          c        c   t g t              tc     t t g

Hma Vβ : ggaggcaaagggctggtcaatagattcatttcactgggagaggcaaagggctgggggccag-agaggagaaataaaagccacaca-tgaagcagcaatg
Atr Vβ : a               --------------------------------------  at   a-                          . ..       g
Lfu Vβ : a   a    c      --------------------------------------      -   t  ag a     g   tgg ga           g a
Chi ε° :    a g  g       ----------------------------------------  g   c   t   g a     g   tg  g            ggca

 "cap"
   |                                                                -1-                               -2-
Hma Vβ : cAGGCATGCTTCTGGCTCATCTGT-GATCACCAGGAAACTCCCAGATCTGACACTGTAGTGCATTTCACTGCTGACAAGAAGGCTGCTGCCACCAGCCTG
Atr Vβ :   A                    -                          C     C CG                C   G A G          T    A
Lfu Vβ :     TGA       AA        C  - G        C G    G    C    G    G                 A  GGCA       -    GG T
Chi ε° :   A T               C   - A G          T  G        -----   CATGGTGCATTTTACTACCGAGGAGAAGGCTGCTGTTGCTAGTCTG
                                                                 MetValHisPheThrThrGluGluLysAlaAlaValAlaSerLeu
                                                                                             10

 -3-   -4-      -5-
Hma Vβ : TGAAGCAAGGTTAAGGT-GAGAAGGCTGGAGGTGAGATTCTGGGCAGgtaggtactggaagcc---gggacaag-gtgcagaaaggcagaaagtgtttct
Atr Vβ :   G    T     G T  -                                    c    a      ---a ag    -                        c
Lfu Vβ : CC G       CA G  T  G   G         CA    C             c   c      g   ---a  gtc  - a                       c
Chi ε° : TGGGCCAAAGTGAATGTGGAGGTGGTCGGCGGTGAGAGCCTAGCAAG  a c g    ca aggta    g g a    t c        t               c a
         TrpAlaLysValAspValGluValValGlyGlyGluSerLeuGlyAr
                   20                      30

                                                                                                    -6-
Hma Vβ : g-aaagagggattagcccgttgtcttacatagtctgactttgcacctgctctgtgattatgactatcccacagTCTCCTGGTTGT----CTACCCATGGA
Atr Vβ : -              a            t            t      t        t                          ----
Lfu Vβ : -        ga      a   a   c  t c   a           t   t t       c---      gc     t  GT  T  C    TTGT   TT    TT
Chi ε° : a           c g tta g t          c             ctt t   t       c       tc     t  GCTCTTGATCGT----CTACCCATGGA
                                                                                 gLeuLeuIleVa----lTyrProTrpT

    -7-      -8-        -9-                             -10-                   -11-
Hma Vβ : CCTAGAGGTACTTTGAAAGTTTTGGA-TATCTGGGCTCTGACTGTGCAATAATGGGCAACCCCAAAGTCAAGGCACATGGCAAGAAGGTGCTGATCTCC
Atr Vβ : T C    A       -       C    C-A                T         -                   -
Lfu Vβ :   C    A  T    CAGT A        GGA  T  TC      G  C  TGC       C          G         T TT               CAA      C  T
Chi ε° : CCCAGAGGTTCTTTGACAGTTTTGGT-AACTTATGCTCTGGGTCTGCCATAATGGGCAACCCCAAGGTCAAGGCCCACGGCAGGAAGGTGCTGAATTCC
         hrGlnArgPhePheAspSerPheGly-AsnLeuCysSerGluSerAlaIleMetGlyAsnProLysValLysAlaHisGlyArgLysValLeuAsnSer
            40                            50                            60                           70

 -12-                           -13-
Hma Vβ : TTCGGAAAAGCTGTTATGCTCACGGATGACCTCA---AAGGCACCTTTGCTACACTGAGTGACCTGCACTGTAACAAGCTGCACGTGGACCCTGAGAACTT
Atr Vβ :    A                T     CT AT GTA                T                    G     A   A
Lfu Vβ :  -         A     C TG    T     T    ---      A         GAG         G       G     T    T     T            G
Chi ε° : TTTGGAAATGCCATTAAGCACATGGATGATCTCA---AGGGCACCTTTGCAGATCTAAGCGAGCTGCACTGTGACAAGTTGCACGTGGATCCTCCGAACTT
         PheGlyAsnAlaIleLysHisMetAspAspLeuL---ysGlyThrPheAlaAspLeuSerGluLeuHisCysAspLysLeuHisValAspProProAsnPh
                                      80                            90                            100

Hma Vβ : CCTGgtgagtagtaagtacactcacgctttcttctttaccttagatatttgcactatgggtacttttgaaagcaga--ggtggct-ttctcttgtgtta
Atr Vβ :     T a  a tc    c     t          c c     a           g  c                    --       -            c
Lfu Vβ :  AG |--------------------------------------------------------------------------------------------------
Chi ε° : CAGG       tc gg   t  ctgt       g        cat -c  g t ca tg  g    ----  a a t   -   acac  a  caag   taa c g
         eArg

Hma Vβ : t--gagtcagctatgggatatgatatttcagcagtggga----ttttgagagttatgtt---gctgtaaataacata-actaa--aatttggtagagca
Atr Vβ :  --          g                 tgt     taga                      --- g cc     g   g-c    --             t
Chi ε° :  tga    - tt t t    gg a          agaaa  cttga c   gt   g -     cca ag  c  - ggt  g  c  gt    ----------

Hma Vβ : aggactatgaataatggaaggccacttaccatttgatagctctgaaaaacacatcttataaaaaat-tctgg--ccaaaatcaaa-ctgag---tgtttt
Atr Vβ :        c     g                                                     -    --    g   tg  -a     ---     g
Chi ε° :  ctc        cc g    -  t      ggt    c      at          --------- g      g gaa aa acc   c    c      cta    g

Hma Vβ : t-gg-atgagggaacagaagttgagatagagaaaataacatctttcctttggtcagcgaaattttctataaaaattaat----agtcacttttctgcata
Atr Vβ :  -  -            a          g               g            ct              g g          ----  a           c
Chi ε° :  c  cc  - a    t c   acc     gcat c g   g   t       g       a a ct        cc      gg     cc    gaat tatt c   a   ----

Hma Vβ : gtcctggaggttagaaaaaga--tcaact-gaacaaagtagtgggaagctgttaaaaa---gaggattgtt--tccctc-cgaatgatgatggtatactt
Atr Vβ :                        -- a c  a     g    a          t       ---  a          --      -t             a  g
Chi ε° : a tt t   t  ta   gg   gac a    a- gg     a  a- a c a   t   c c   tgt  t   c   c   ctct a   t t gca       aa  gt

Hma Vβ : ttgtacgcatggtacaggattctttgttatgagtgtttgggaaaattgtatgtatgtatgta-tgtatgtgatgactggggacttatcctatccattact
Atr Vβ :       a               t               ca          g   g   g  -----------    a      a              c
Chi ε° :      ta   a    g t        c-   cgtca a   ga   tcc    caa c  aa           a a ca -   ga   ----a ta   - a     ---  g

Hma Vβ : gttccttgaagtactattatcctactttttaaaaggacgaagtctctaaaaaaaaaatgaaacaatca----caatatgttgggg-tagtgagttggcat
Atr Vβ :                                                    t g   c   t ---      gg        t gtca    g  c   t  -  a
Chi ε° : -cc     tc   c      a gtt          -    gta t g  ---------------a cgg        catcag ca    c ca    c t aac   c a   g

Hma Vβ : agcaagtaagagaaggatagggcacaatgggaggtgcagggctgccagtcatattgaagctgatatctagcccataatggtg-------------------
Atr Vβ :                                     a              g                g                  tacttacctttgtgagaa
Chi ε° :   a  a gg          gta      g     -    t  -       gg      --c      t ca   gc ag t tc            c   ----         g

Hma Vβ : ----------agagttgctcaaactctggtcaaaaggatgtaagtgttatatctatttac-tgcaagtccagcttgaggccttctattcactatgtacc
Atr Vβ : taagactctg                 t          a aa        g    t     t-                      gt
Chi ε° : c t g ca    a  c        cc      ta g    g g acc gc -      g              g t   g    gg      c      a       c     t         t

Hma Vβ : attttctttttatcttcactccctccccagCTCTTAGGCAACGTGATATTGATTGTTTTGGCAACCCACTTCAGCGAGGATTTTACCCTACAGATACAG
Atr Vβ :      c     -                  T        T         G        G                  G
Chi ε° : g c       -  t-----------       CTCCTAGGAAACATGATATTGATTGTCTTGGCAACCCACTTCAGCAAGGAATTTACCCCGCAGATGCAG
                                       LeuLeuGlyAsnMetIleLeuIleValLeuAlaThrHisPheSerLysGluOheThrProGlnMetGln
                                                   110                        120

                   -14-          -15-                 -16-
Hma Vβ : GCTTCTTGGCAGTAACTAACAAATGCTGTGGTTAATGCTGTAGCCCACAAG-ACCACTGAgttccctgtcc-actatgtttgtaccta-tggtccactat
Atr Vβ :           CG     AAG  C  T   ------------    CA G          TT T          c    g     -    g   g          c       g c
Chi ε°,: GCTGCTTGGCAGAAGCTGAACAACGCTGTGGCTAATGCTCTGGCCCACAAGTACCACT-AG  g    g  t  c   c g  g           ---------- c
         AlaAlaTrpGlnLysLeuThrAsnAlaValAlaAsnAlaLeuAlaHisLysTyrHis*****
             130                           140

Hma Vβ : gtttgtacctatgtcccaaaatctcatctcctttagatgggggaggttggggagaagagcagtatcctgcctgctga--ttcagttcctgcatgataaaa
Atr Vβ :                c  a    t    g                               a   t                    a               a  --      c
Chi ε° :   aag c-   ag            g  gt          ga    ca  a   ga  aga  ----------------   tt       a  ca  at  ac    aat   a     c

Hma Vβ : atagaataaagaaat-atgctctctaagaaatatcattgtactctttttctgtctttatattttac--cctgattcagccaaaaggacgcactatttctg
Atr Vβ :    a                 -g     g  ta                g  a   tat                 g       t--gt                          t   c
Chi ε° :    ga         -- g  c    g    a    tg     gc c       t  t  c  c  t     t    g         taagt     n   ctat  --------------------

Hma Vβ : atggaaatgagaatgttggagaatgggagtttaaggacagagaagatactttcttgcaatcctgcaagaaaagagagaactcgtgggtggatttagtggg
Atr Vβ :                   g    cac           a               a t        g                    a   t          t
Chi ε° : ----------------------------------------------------------------------------------------- a    ac   c

Hma Vβ : gtagttactcctaggaaggggaaatcgt
Atr Vβ :      c          gca       gg  a
Chi ε° :         a   t    atg    a tg aaa
```

defects have been fixed within the separate lineages. The dysfunction of the other primate Ψβ1-related sequences has been confirmed by DNA sequence analysis (see Chapter 4 and 5). The owl monkey pseudogene contains 8 codon defects including a GCG initiation codon, one in phase nonsense mutation and 6 exon frameshifts. The brown lemur Ψβ1-related sequence contains 5 codon defects including a GTG initiation codon (not thought to initiate transcription in eukaryotes) and 4 exon frameshifts. The relative position of these defects is shown more clearly in Figure 8.2.

Additional potential defects are also present in the signal sequences that have been implicated in eukaryotic gene transcription, mRNA maturation and translation (Chang and Slightom, 1984; see Results). While individual codon defects would silence the gene, signal sequence defects require further _in vivo_ and _in vitro_ expression analysis to determine their ability to silence the gene. Only codon defects are discussed below.

In an attempt to distinguish when and how the primate Ψβ1 gene was silenced the defects within the coding sequence of the human, owl monkey and brown lemur Ψβ1-related sequences have been partitioned onto different branches of the primate tree by the method of maximum parsimony (Figure 8.3). The majority of codon defects are found on only one branch of the tree (that leading to man, owl monkey or brown lemur) suggesting they have accumulated recently within the individual lineages. In such cases the base composition of the other two unaffected sequences is identical to, or closely resembles, that same position in functional non-adult β-globin genes (Figure 8.1).

Figure 8.2

Comparison of defects in the Ψβ1 pseudogene sequences of man, owl monkey, and the brown lemur.

The diagram shows the relative position of coding sequence defects, numbered as in Figure 8.1, in the Ψβ1 sequences from man, owl monkey and the brown lemur. The exons, non-translated and non-coding DNA sequences are represented as solid or hatched boxes and thick lines repectively. The position of coding sequence defects, each one of which could potentially have inactivated the gene, are indicated under each sequence as shown by the key. While most defects are lineage specific common defects between two or more sequences are shown by solid vertical lines (see text).

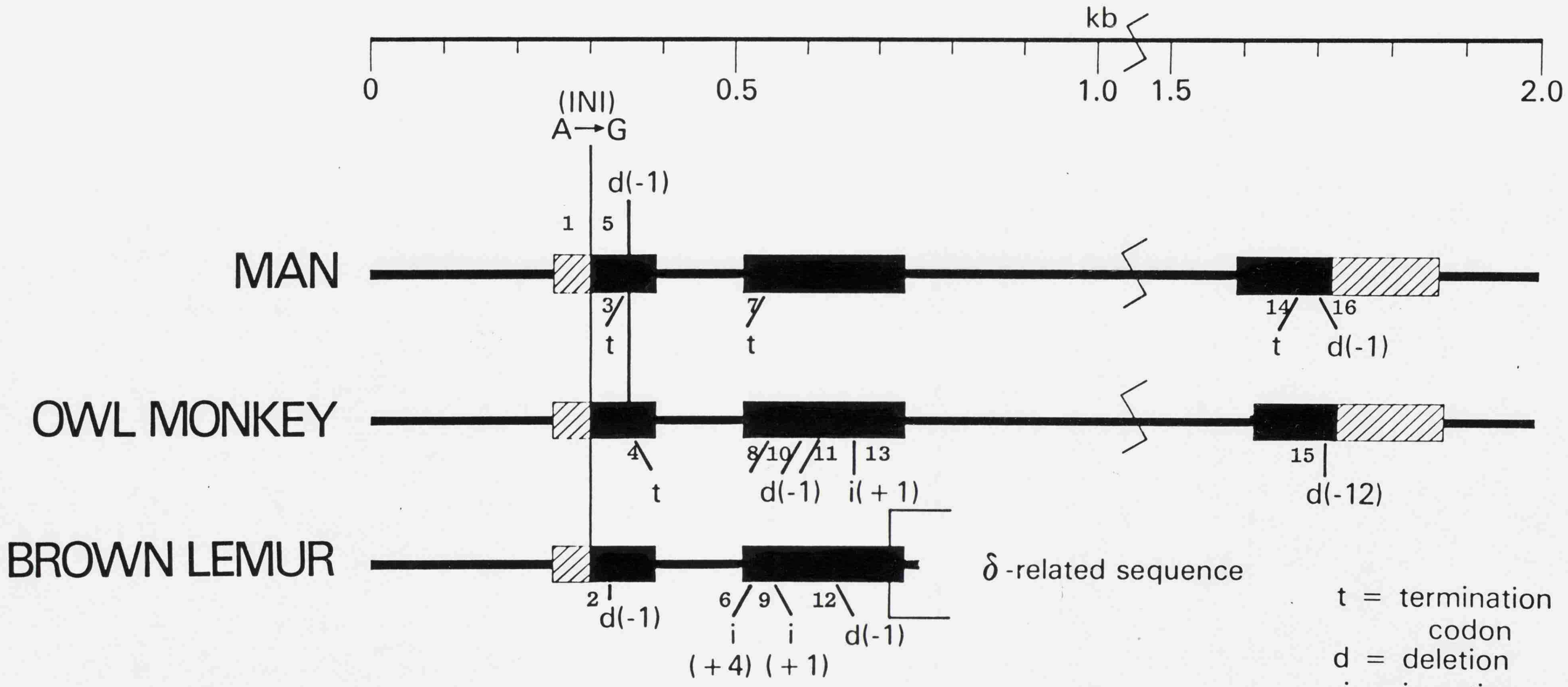

kb
0
0.5
1.0
1.5
2.0
(INI)
A→G
d(-1)
1
5
MAN
3
t
7
t
14
t
16
d(-1)
OWL MONKEY
4
t
8
10
11
d(-1)
13
i(+1)
15
d(-12)
BROWN LEMUR
2
d(-1)
6
i
(+4)
9
i
(+1)
12
d(-1)
δ-related sequence
t = termination codon
d = deletion
i = insertion

Figure 8.3

Phylogeny of the primate Ψβ1 gene.

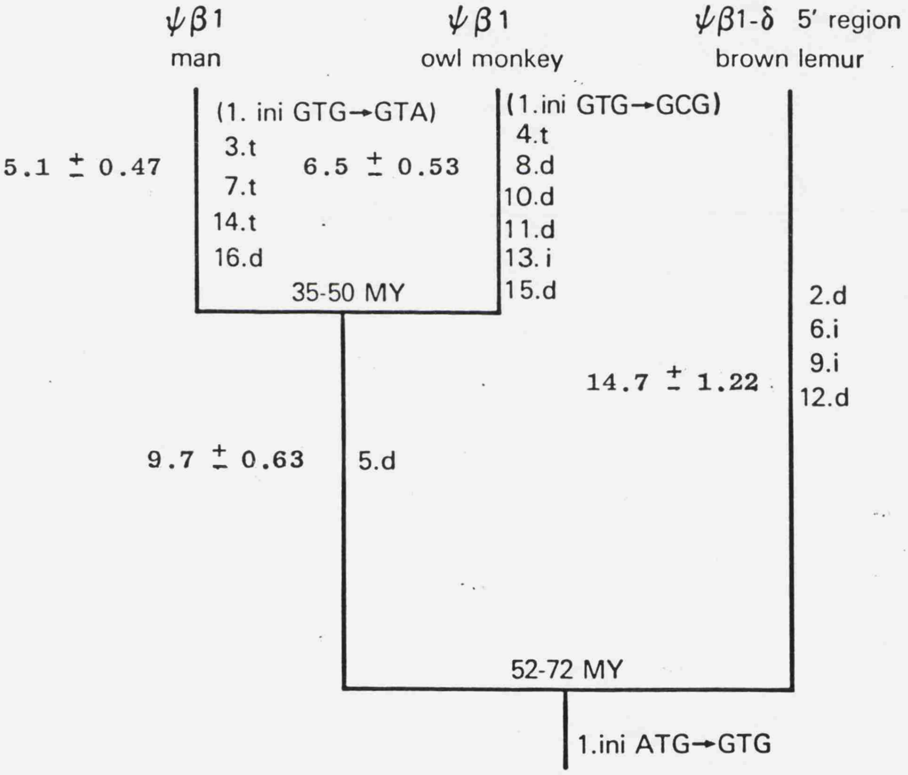
Branch lengths (substitutions per 100 bp ± S.E.) on the left of each branch were derived from a difference matrix corrected by iterative procedures for multiple substitutions with a high (71%) probability of transitions over transversions. The tree was rooted using the goat ε^{II}-globin gene as an external reference. Approximate divergence times are derived from palaeontological and protein data (Simons, 1969; Sarich and Cronin, 1977; Wilson *et al.*, 1977).

Each numbered codon defect shown in Figure 8.1 has been assigned to a branch on the basis of maximum parsimony and is identified by: ini, initiation codon defect; d, microdeletion; i, microinsertion; or t, premature termination codon. All defects in the human Ψβ1 pseudogene are shared by gorilla and chimpanzee (Chang and Slighton, 1984) indicating that the four defects on the human branch were established prior to the hominid divergence ~7 MY ago. The possible initial defect (initiation codon ATG-- GTG) is shared by all species and is placed at the root of the tree; secondary alterations in the initiation codon in man and the owl monkey are also indicated.

There is a noticeable skew in the distribution of defects towards the upper branches of the simian tree, with only one defect on the branch leading to the human/owl monkey split. However, this skew is not statistically significant. Given the branch lengths shown in the Figure and using the pattern of microinsertion/deletion and base substitution observed in intron and flanking Ψβ1 regions, coupled with computer-simulated divergence of predicted ancestral Ψβ1 sequences, it is possible to estimate the number of deletions/insertions (D) and in phase termination codons (T) which should have accumulated in the "coding sequence" of a silenced gene. The expected (observed) number of defects on each branch are:

	D	T
man/owl monkey ancestor to man	1.4 (1)	0.5 (3)
man/owl monkey ancestor to owl monkey	2.4 (5)	0.8 (1)
root to man/owl monkey ancestor	3.1 (1)	1.0 (0)
root to lemur (exon 1 and 2 only)	3.7 (4)	1.1 (0)

Thus, the skew of defects towards the top of the tree is not significant (pooling D and T defects gives χ^2[df4]=4.7 (Yate's correction), p>0.1).


ψβ1
man
ψβ1
owl monkey
ψβ1-δ 5′ region
brown lemur
(1. ini GTG→GTA)
3.t
7.t
14.t
16.d
5.1 ± 0.47
6.5 ± 0.53
(1.ini GTG→GCG)
4.t
8.d
10.d
11.d
13. i
15.d
35-50 MY
2.d
6.i
9.i
12.d
14.7 ± 1.22
9.7 ± 0.63
5.d
52-72 MY
1.ini ATG→GTG

The human and owl monkey Ψβ1 sequences do share two defects, an A → G transitional substitution at position 1 of the initiation codon and a frameshift, as the result of a single base deletion, in either codon 20 or 21 (see Figure 8.1 and 8.2). These shared defects strongly suggest the common ancestor of these two sequences was itself a pseudogene, that is, the ancestral Ψβ1-related sequence was silenced prior to the Old World monkey-New World monkey divergence some 35-50 MYs ago. Similarly the brown lemur Ψβ1-related sequence shares the initiation codon defect (A → G at position 1) that is held in common between the human and owl monkey sequences, suggesting that the common ancestor of all these species contained a Ψβ1-like gene that was itself a pseudogene, that is, the ancestral Ψβ1 gene was silenced before the prosimian-simian divergence 52-72 MYs ago. The presence of only a single defect common in all three Ψβ1-related sequences, coupled with the observation that the majority of defects are lineage specific and therefore relatively recent independent events, furthermore suggests that the ancestral Ψβ1 gene was silenced recently before the basal primate radiation; as one would have expected the accumulation of several common defects if the Ψβ1 gene had been silenced for a long period in the common ancestor of contemporary primates prior to the basal primate radiation.

To summarise, the analysis of shared defects suggests that the ancestral primate Ψβ1 sequence was initially a functional gene that was silenced, probably by an initiation codon defect, shortly before the basal primate radiation ~70 MYs ago and that the majority of contemporary pseudogene defects have subsequently accumulated independently during primate evolution. Another possibility however is that the functional gene was silenced independently during simian and prosimian evolution and that the common initiation defect has arisen, by chance, as the result of a parallel base substitution in the different lineages. Given the observed rate of pseudogene sequence evolution on the different branches of the primate lineage (see 8.8, Figure 8.3) it is possible to calculate the statistical probability of such an independent parallel change (from A → G) in each of the simian and prosimian lineages. This probability is small (p=0.009) suggesting it is unlikely that this defect evolved independently in the different primate Ψβ1 genes examined.

An additional feature of the pseudogene defects, as partitioned on to each branch of the primate tree shown in Figure 8.3, is that they have a noticeably skewed distribution towards the upper branches of the simian tree. In order to determine whether this distribution was of evolutionary significance or due to chance the expected number of codon defects was calculated for each branch given the mode and tempo of evolution observed for this pseudogene (see 8.8). Compared to the expected number of defects for a given branch length the observed skewed distribution of defects within the branches of the tree was not statistically significant (see legend Figure 8.3 for details), suggesting that the apparently large number of defects observed in the upper branches of the tree, particularly those in the owl monkey lineage, are not of evolutionary significance.

8.7 An early functional history for the Ψβ1 gene

The estimated time of silencing of the primate Ψβ1 gene (~70 MY ago) contrasts sharply with the minimum estimate for the gene duplication event which gave rise to a distinct Ψβ1 locus within the ancestral β-globin gene cluster (~140MY, see 8.3). This implies that the Ψβ1 gene may have been functional for much of its early evolutionary history and that Ψβ1-related sequences detected in other non-primate mammalian orders may still be functional, the mammalian radiation having occurred prior to the silencing of the ancestral primate Ψβ1 gene. While the functional status of Ψβ1-related sequences in the lion, dog and seal is unknown, during the course of this work Goodman *et al*. (1984) confirmed this prediction by showing that the apparently functional goat embryonic ε^{II}-globin gene is orthologous to primate Ψβ1 sequences.

The primate Ψβ1 pseudogene therefore constitutes the first case where a predicted early functional history for a contemporary pseudogene has been confirmed by the presence of a functional counterpart in another lineage. Previously described pseudogene histories for which a functional period of evolution has been suggested (see 1.4) are based on the relative level of DNA sequence divergence between the coding regions (exons) of the contemporary pseudogene and related functional genes. These comparisons, based on clock rates of DNA change, are subject to high statistical uncertainty concerning the actual time of gene duplication and silencing and may be influenced by such factors as gene conversion during the evolutionary history of the contemporary sequences. No other pseudogene history has been confirmed phylogenetically.

8.8 The mode and rate of primate Ψβ1 evolution

Assuming the common ancestor of the primate Ψβ1 gene was indeed silenced shortly before the prosimian-simian divergence 70 MYs ago, and that the pseudogene has subsequently been without any function or effect in the primate β-globin gene cluster, the contemporary pseudogene sequences should reflect the mode and tempo of neutral DNA change in this lineage. In order to test the predictions of the neutral theory the pseudogene sequences, aligned in order to maximise homologies while introducing the minimum number of gaps (Figure 8.1), have been examined in terms of their rate of nucleotide substitution and the type of mutational events that have occurred.

a) Mode of primate Ψβ1 evolution

Base substitutions within the human and owl monkey sequences have occurred in an apparently random fashion, as predicted for pseudogenes (Table 8.1(a), Kimura, 1983b). They show no preferential conservation of exon sequences, the divergence over non-coding (flanking and intron sequences) compared to exon regions is not significantly different from the overall level of divergence of the whole pseudogene (11.04 ± 0.67%). In addition, the levels of substitution at first, second and third codon positions are not significantly different (Table 8.1(b)), as would be anticipated for the codon positions of a functional gene sequence. A similar analysis covering the homologous regions of the brown lemur pseudogene and those of the human and owl monkey show a similar uniformity of divergence across all regions of the sequence, with an overall divergence of 23.41 ± 1.47% and no significant bias in substitutions at different codon positions (Table 8.2).

The apparently random distribution of altered positions within the pseudogene was examined statistically using the single runs test (Siegel, 1956). This statistical test returns a probability value that the observed distribution of base substitutions deviates significantly (either

Table 8.1

Sequence comparison of the human and owl monkey Ψβ1 sequences.

Comparisons and sequence co-ordinates are based on the sequence alignments shown in Figure 8.1. The sequences were aligned with reference to dot-matrices, that is, sequence homology was maximised with the introduction of the minimum number of gaps (microinsertions/ deletions, shown in parentheses) into either sequence. Gaps were not scored in divergence calculations. The corrected sequence divergence (hits/base) and transition frequency were calculated by an iterative procedure with a bias towards transitions over transversions of 71 %. The rate of evolution = corrected % difference/ 2 x divergence time. The divergence for these species was taken as approximately 35 MYs, see Figure 8.3.

Table 8.1

a) Regional divergence between the human and owl monkey Ψβ1 sequences

Region of gene	Co-ordinates in sequence	Number of bases compared	% Sequence divergence (uncorrected)
5' flanking	0 - 455	397 (8)	11.38 ± 1.6
Exon 1	456 - 547	91 (0)	13.18 ± 3.5
Intron 1	548 - 673	121 (0)	9.09 ± 2.6
Exon 2	674 - 904	220 (4)	8.18 ± 1.8
Intron 2	905 - 1830	840 (7)	9.88 ± 1.0
Exon 3	1831 - 1956	113 (2)	15.04 ± 3.3
3' flanking	1957 - 2327	369 (1)	13.46 ± 1.8

Overall divergence	(uncorrected)	11.04 ± 0.67	Transition frequency	69.62 ± 2.98
2146 bp	(corrected)	12.09		71.30

Rate of evolution (nuc.sub./site/year) = 1.7×10^{-9}

b) Codon analysis

Codon position	Number of base substitutions observed	expected
1	15	16
2	12	16
3	21	16
Total	48	48

χ^2 (df_2) = 2.625 $0.2<p<0.3$

c) Multiple base substitutions (>1 bp)

Consecutive number	Observed	expected
2	25	23
3	7	3
4	0	0

Single runs test $p<0.002$

d) Size distribution of insertion/deletion changes

Size (bp)	Number
1	11
3	2
4	3
5,10,12,28,38	1 each
Total	21

Table 8.2

Sequence comparison between the brown lemur Ψβ1-related region of the hybrid Ψβ1-δ pseudogene and the equivalent human and owl monkey Ψβ1 sequences.

Comparisons and sequence co-ordinates based on the alignment shown in Figure 8.1, obtained as described for Table 8.1. Corrected sequence divergences were calculated as described for Table 8.1. The rate of evolution was calculated assuming a divergence of the prosimian and simian lineages approximately 70 MYs ago.

Table 8.2

a) Regional divergence between the brown lemur Ψβ1 sequence and that of the simians

Region of gene	Co-ordinates in sequence	Number of bases compared		% Sequence divergence (uncorrected)	
		Human	owl monkey	Human	owl monkey
5' flanking	0 - 455	391 (11)	392 (8)	23.01 ± 2.12	23.98 ± 2.15
Exon 1	456 - 547	90 (2)	90 (2)	22.22 ± 4.38	23.33 ± 4.46
Intron 1	548 - 673	118 (1)	118 (1)	20.34 ± 3.70	21.19 ± 3.76
Exon 2	674 - 904	222 (3)	219 (7)	24.32 ± 2.87	25.57 ± 2.94

	Human (821)	owl monkey (819)
Overall divergence (uncorrected)	22.89 ± 1.47	23.93 ± 1.49
Transition frequency	65.96 ± 3.46	61.73 ± 3.47
Overall divergence (corrected)	28.11	29.51
Transition frequency	69.88	65.73
Rate of evolution (nuc.sub./site/year)	2×10^{-9}	2.1×10^{-9}

b) Codon analysis

vs. human sequence

Codon position	Number of base substitutions	
	observed	expected
1	24	24.3
2	21	24.3
3	28	24.3
Total	73	72.9

χ^2 (df_2) = 0.604, $0.7<p<0.8$

vs. owl monkey sequence

Codon position	observed	expected
1	26	25.3
2	23	25.3
3	27	25.3
Total	76	75.9

χ^2 (df_2) = 0.295, $p>0.9$

c) Multiple base substitutions (>1 bp)

Number	observed	expected
2	27	33
3	9	8
4	2	2
5	0	0

Single runs test $p<0.05$

Number	observed	expected
2	33	36
3	8	9
4	1	2
5	1	0

Single runs test $p<0.01$

d) Size distribution of insertion/deletion changes

Size (bp)	Number
1	12
3	1
4	2
10,38	1 each
Total	17

Size (bp)	Number
1	14
3	2
4	1
5,10	1 each
Total	19

to few or to many) from a random distribution calculated from the observed level of divergence between the two sequences. Comparison of the human and owl monkey pseudogenes using this statistical test indicated that the substitutions are not entirely random in position (p=0.002). Examination of the individual substitutions suggests this is due to a low level excess of "block" substitutions (a single mutational change involving more than one consecutive base, Table 8.1(c)) that results in an effective reduction in the observed number of "runs" compared to that expected, given the observed level of divergence between the two sequences. While the majority of changes are single randomly scattered substitutions, an excess of 3 bp "block" substitutions is noticeable when comparing exon 2 and the 3' flanking regions of the human and owl monkey Ψβ-related sequences (Figure 8.1). Similar results were obtained for comparisons between the simian and brown lemur pseudogene sequences.

The majority of base substitutions in all the Ψβ1 sequences are transitions (69% ± 3%, corrected for multiple substitutions), of which most are the result of a single independent change. In addition, microinsertions/deletions have also been fixed, apparently at random, at a rate of approximately 1 per 11 base substitutions (Table 8.1 and 8.2 (d)). Several of these microinsertion/deletion events may have resulted by strand slippage during replication resulting in local duplications and deletions (see Efstratiadis et al., 1980), for example, the 4 bp duplication in the ~~owl monkey~~ Lemur gene that gave rise to the frameshift designated defect 6 in Figure 8.1.

The uniformity in the distribution of sequence divergence within these sequences supports the conclusion that the ancestral primate Ψβ1-

related sequence was a pseudogene and that contemporary Ψβ1 sequences have evolved in the absence of any selective constraint. It also implies that no major recombinational exchanges have occurred involving these pseudogene sequences and other β-globin genes of the cluster during primate evolution, with the exception of the unequal exchange which gave rise to the hybrid Ψβ1-δ pseudogene in the lemurs (8.5).

b) Variation in the tempo of primate Ψβ1 sequence evolution

When averaged over the relevant period of evolution, the observed overall level of sequence divergence between the different primate Ψβ1 sequences (Table 8.1/8.2, corrected for multiple substitutions) gives a mean rate of evolution for these sequences of ~2 x 10^{-9} nuc.sub./ site/ year. If this reflects the true non-coding DNA sequence mutation rate or "neutral" rate in the primates it is substantially less than the previously proposed universal constant rate of mammalian non-coding DNA sequence evolution (~5 x 10^{-9} nuc.sub./ site/ year), deduced primarily from comparisons between primate-lagomorph and primate-rodent genes (Hayashida and Miyata, 1983). Similar low rates have been noted for non-coding DNA sequences in human-seal myoglobin gene comparisons, suggesting the neutral substitution rate, and therefore the mutation rate, is not constant in different mammalian lineages but may be influenced by factors such as generation time or alterations in the fidelity of DNA replication in different lineages (Weller _et_ _al_., 1984).

Within the primates there are also indications that the neutral mutation rate is variable (Figure 8.3). For rate constancy to apply in the primate lineage the time elapsed between the Old World-New World monkey divergence would have to be approximately one third that elapsed since the prosimian-simian divergence. However both the palaeontological and protein data, even given the errors associated with such methods, indicate a much shorter interval between these two divergences, suggesting that the neutral mutation rate was relatively high early in primate evolution and has subsequently decreased, particularly in the Hominoid lineage (see Goodman et al., 1984). Support for such a decrease in overall mutation rate during primate evolution is provided by the reduced rate of evolution observed between the hominoid Ψβ1 genes (1.4×10^{-9} nuc.sub./ site/ year, Chang and Slightom, 1984) and bipolar Alu-repeat sequences (1.5×10^{-9} nuc.sub./site/yr, Maeda et al., 1984). The rate has apparently slowed even further in the lineage leading to man to 1×10^{-9} nuc.sub./ site/ year (Chang and Slightom, 1984).

The formal possibility exists that the Ψβ1 sequence, or the region in which this sequence resides, has acquired a function during primate evolution and therefore has gradually evolved more slowly than expected. However, while within a region of the β-globin gene cluster repeatedly implicated in developmental regulation of the cluster no specific functional or effect has yet been ascribed to the Ψβ1 gene (see Orkin and Kazazian, 1984). Furthermore non-coding DNA sequences within the functional adult β-globin gene of the primates also evolve at a rate comparable to that found for the Ψβ1 pseudogene, suggesting these observations may be applicable to other regions of the human genome (see below).

8.9 Evolution of the functional brown lemur β-globin gene

The adult brown lemur β-like globin gene, previously characterised by hybridisation and restriction site analysis, has been isolated, sequenced and shown to have the archetypal globin gene family organisation (Chapter 6). The similarity between the amino-acid sequence encoded by the exons of this gene to the published amino-acid sequence (Maita _et_ _al_., 1979) of the brown lemur adult β-globin polypeptide is strong evidence that this gene is functional _in vivo_ and that its expression accounts for the adult β-globin polypeptide in this species.

The divergence of prosimians from simians, and therefore the human and brown lemur β-globin genes, 52-70 MY ago provides an ideal opportunity for examination of the mode and tempo of sequence evolution in the different regions of a functional gene compared to that previously established for a non-processed pseudogene (see 8.8). Particularly important to this discussion is the rate and type of change in the non-coding regions of this gene. Both should correspond to that determined comparing the primate Ψβ1 sequences if the previously observed non-coding DNA evolution is to be considered representative of such sequences in general in the primates.

Optimum alignment of the human and brown lemur adult β-globin genes (Figure 6.5) gives a corrected overall sequence divergence of 24.35 ± 0.89 % (see Table 8.3(a)), somewhat lower than that observed between the Ψβ1 sequences of these two species. However, this overall sequence divergence is composed of both non-coding DNA sequence changes and a lower level of coding and 5' flanking sequence evolution, a level that almost certainly reflects the action of purifying natural selection against base substitutions that disrupt normal adult β-globin gene expression. As mentioned previously, in relation to the rest of this discussion the

Table 8.3

Sequence comparison of the brown lemur and human β-globin genes.

Comparisons and sequence co-ordinates are based on the alignment of the brown lemur and human β-globin genes shown in Figure 6.4. The number of gaps introduced into non-coding DNA regions in order to maximise sequence homology are shown in parentheses, these regions were not scored in divergence calculations. Corrected sequence divergence in non-coding DNA regions (intron 1/2 and 3' flanking sequences) was calcluated as described for Table 8.1. The rate of evolution was calculated assuming a species divergence between brown lemur and human lineages some 70 MYs ago. Coding sequence divergences were corrected by the method of Perler *et al.* (1979).

Table 8.3 a) Regional divergence between the human and brown lemur β-globin gene sequences

Region of gene	Co-ordinates in sequence	Number of bases compared	% Sequence divergence (uncorrected)	% Sequence divergence (corrected)
5' flanking	0 - 326	226 (1)	14.16 ± 2.3	
Exon 1	327 - 418	92	26.08 ± 4.6	
Intron 1	419 - 548	130	20.77 ± 3.6	25.09
Exon 2	549 - 771	223	10.36 ± 2.0	
Intron 2	772 - 1699	790 (16)	23.79 ± 1.5	29.63
Exon 3	1700 - 1825	126	14.28 ± 3.1	
3' flanking	1826 - 2311	445 (10)	23.32 ± 2.0	28.48

Overall divergence (uncorrected)	20.47 ± 0.89	Transition frequency	62.26 ± 2.37
2032 bp (corrected)	24.35		65.56

Rate of evolution (nuc.sub./site/year) overall 1.73 x 10^{-9}, Intron 1/2, 3' flanking DNA sequences 2 x 10^{-9}.

b) Codon analysis

Codon position	Number of base substitutions observed	expected*
1	19	21.3
2	12	21.3
3	33	21.3
Total	64	63.9

χ^2 (df_2) = 10.7 p<0.01 *if random

c) Size distribution of insertion/deletion changes

Size (bps)	Number
1	8
2	5
4	3
11,12	2 each
5,6,8,9,19,21,37	1 each
Total	27

d) Coding sequence divergence (%)

Exon	1	2	3	Total
Replacement	25.7	7.1	10.7	11
Silent	58.9	26.3	33.6	34

important regions of the functional brown lemur β-globin gene are the non-coding DNA regions other than the 5' flanking region (the 5' flanking regions are not included due to the assumed influence of purifying selection on the substitution rate in this region).

As can be seen in Table 8.3(a), the corrected level of sequence divergence over intron 1, intron 2 and 3' flanking regions is very similar to the corrected overall sequence divergence observed between the pseudogene sequences of these two species (Table 8.2). Similarly, the number of microinsertion/deletion changes observed (26) corresponds almost exactly to the number predicted (25) from the pseudogene comparisons; assuming a ratio of 1 microinsertion/deletion per 11 base substitutions over intron 1/2 and the 3' flanking regions. The majority of the insertion/deletions (16/26) are ≤4 bps (Table 8.3(c)). In addition, over the transcribed regions of this gene only a single microinsertion/ deletion has occurred, a 4 bp deletion at a position 10 bp 5' of the initiation codon (Figure 6.5). This deletion is not apparently sufficient to prevent the expression of this gene.

The primary conclusion from the analysis of the brown lemur and human functional β-globin genes is therefore that the non-coding DNA sequence change within this gene during primate evolution is essentially the same as that observed for the primate Ψβ1 pseudogene. These observations are in accordance with previous rate estimates for non-coding DNA evolution, based on genomic mapping of restriction endonuclease sites in the β-globin gene cluster (Barrie et al., 1981), and encourages the belief that this mode and tempo of DNA sequence evolution may be a general feature of non-coding DNA sequences in the primate β-globin gene cluster and possibly the genome as a whole.

Base substitutions have also been examined in the polypeptide-coding regions of this gene. Nucleotide substitutions that accumulate in the coding regions of a functional gene are of two types due to redundancy in the genetic code. Replacement (non-synonymous) substitutions that lead to an amino-acid change in the peptide and silent (synonymous) substitutions that do not alter the encoded amino-acid sequence. In contrast to the essentially random distribution of base substitutions observed in the 'coding' regions of the primate pseudogene base substitutions in the functional gene have occurred in a non-random manner, the majority having occurred at position three of the codons (Table 8.3(b)). This is consistant with the mode of coding sequence evolution observed in many other functional gene comparisons (Kimura, 1983(b)).

The number of nucleotide substitutions in the coding regions of the human and brown lemur β-globin genes was estimated using the method of Perler et al. (1980). Although the assumptions involved in these particular calculations are oversimplified (Li et al., 1985), for example, that transitions and transversions are equally probable, the resulting values can be directly compared with those reported by others (for example, Efstratiadis et al., 1980). The overall replacement site divergence (11 %) and silent site divergence (34 %) between the coding regions of these sequences (Table 8.3(d)) are both somewhat higher than would be expected considering the previously established replacement site substitution rate for globins (Efstratiadis et al., 1980) and the non-coding DNA substitution rate established from primate pseudogene comparisons, a rate that should be reflected in the silent site substitution rate. Comparing the replacement and silent site divergence in the different exons of the brown lemur β-globin gene suggests this is due to a particularly high number of substitutions in both categories over exon 1 (Table 8.3(d)). Assignment of base substitutions to the human or brown lemur lineage by the maximum parsimony criterion (rabbit adult β-globin gene as outgroup) also suggests an excessive number of changes in exon 1 compared to that in the other exons.

Closer examination of exon 1 suggests this is probably due to a clustering of substitutions at the 5' end of the exon, for example, a 6 bp 'block' substitution covering codons 1 and 2 that alters the first two amino-acids from Val-His (man and other primates) to Thr-Leu (brown lemur), see Figure 6.~~5~~4. This clustering is particularly apparent over the first 13 codons. A high level of amino-acid sequence divergence has similarly been noted for other lemuroid β-globin amino-acid sequences (Hill and Buettner-Jansch, 1964; Coppenhaver *et al.*, 198~~2~~3), particularly over the first 13 residues, but not in the β-globin chains of another closely related group of prosimians, the lorises (see Maita *et al.*, 1978). Repeating the Perler analysis in the absence of these 13 codons results in a replacement site divergence of 8 % and a silent site divergence of 30 % indicating this region of divergence is contributing a large part of the overall divergence between the human and brown lemur β-globin genes, particularly replacement site divergence.

In the absence of experimental evidence concerning the functional properties of the brown lemur adult β-globin chain it is difficult to assess the *in vivo* physiological significance, if any, for the number of substitutions within the first exon of this gene. Without such information it is difficult to account for the number of changes by positive natural selection especially as, except for the first two amino-acid residues involved in 2,3-diphosphoglycerate regulation of oxygen binding affinity (see Bunn, 1980), no specific functional role has yet been attributed to the amino-acid positions which differ between the human and brown lemur β-globin genes (Table 6.1).

One recognised evolutionary event that may have led to such a clustering of substitutions is intergenic gene conversion involving one of the other β-like globin genes of the brown lemur. This is an attractive mechanism as the 'donor' sequence (that which the β-globin gene has been converted to resemble) from another β-like globin gene is likely to be compatible with the continued function of the adult β-globin gene. In order to determine whether such a clustering of base substitutions is the result of concerted evolution the first 13 codons from the brown lemur ε,γ and hybrid ψβ1-δ gene, and the human β-globin gene, were aligned against those of the brown lemur β-globin gene (Figure 8.4).

From the alignment shown in Figure 8.4 it is evident that where the brown lemur β-globin gene sequence differs from the human β-globin gene there is no preferential homology with any of the other brown lemur β-like globin genes. This suggests intergenic gene conversion by another globin locus in the cluster does not account for the observed sequence divergence over this region. However, the absence of δ-globin exon 1 sequences in the contemporary brown lemur β-globin gene cluster (or the equivalent region of a 'genuine' mammalian δ-globin gene, see 8.3) means this comparison does not include all the potential 'donors' in such a conversion. If the donor sequence was of δ-like origin this would imply a gene conversion event early in lemuroid evolution, prior to the unequal

Figure 8.4

Exon 1 alignment of the first 13 codons of the lemur β-globin gene.

The first 13 codons of the brown lemur β-globin gene (Lfu) are shown aligned with equivalent sequences from the human (Hsa) β-globin gene, the brown lemur hybrid Ψβ1-δ gene and ε-and γ-globin genes (S.Harris, Z.Wong, J.Thackeray and A.J.Jeffreys; unpublished results). Codon phasing is indicated by the punctuation. Differences in this region between the human and brown lemur β-globin genes are indicated by asterisks while differences between the brown lemur β-globin gene and other brown lemur β-globin related sequences are indicated by vertical lines. The number of base mismatches in any comparison are shown in brackets.

```
Hsa β       : GTG.CAC.CTG.ACT.CCT.GAG.GAG.AAG.TCT.GCC.GTT.ACT.GCC
              *** ***      *  *             * *   ***   *   * * *   (17)
Lfu β       : ACT.TTG.CTG.AGT.GCT.GAG.GAG.AAT.GCT.CAT.GTC.ACC.TCT
              ||| ||| | |  |    |      ||   |      |   || | | |||   (21)
Lfu Ψβ1-δ   : GTG.CAT.TTC.ACT.GCA.GAG.GCA.AAG.GCT.-CT.GCG.GCT.AGC
              ||| ||| | |  |                | | |  || | | ||| |||   (22)
Lfu γ       : GTG.CAT.TTT.ACT.GCT.GAG.GAG.AAG.TCC.ACC.ATC.CTG.AGC
              ||  ||| | |  ||   |   |   |   |      |  |    || |||   (19)
Lfu ε       : GTT.CAT.TTT.ACA.GCG.GAA.GAG.AAG.GCT.GTT.ATC.ACA.AGC
```

exchange that gave rise to the hybrid Ψβ1-δ pseudogene and concomitant loss of 5' δ sequences (see 8.5), but after the lemur-loris divergence, as the β-globin amino-acid chains of the loris group of prosimians resembles that of other primates and man rather than other lemurs.

One remaining possibility is that this region of the lemur β-globin gene has undergone conversion to resemble a moderately similar 'random' region of the lemur genome as part of the process of recombination. A model which could account for such random gene conversions has recently been proposed by Smithies and Powers (1985) based on similar observations of microconversion events between the human foetal globin genes and genes within the H-2 region of the major histcompatability complex of the mouse (Mellor _et_ _al_., 1983; Weiss _et_ _al_., 1983). They suggest that gene conversions are the consequence of a general mechanism whereby DNA strand invasion normally enables chromosomes to find their homologues during meiosis. Occassionally however, during abortive non-homologous chromosome pairing, the invading non-homologous single strand molecule might become incorporated into the invaded DNA double helix leading to a limited gene conversion. Theoretically such microconversions may therefore involve a similar but non-homologous 'donor' region from anywhere in the genome. Comparing the sequence of the first 13 codons of the human β-globin gene to that of the lemur β-globin gene (Figure 8.4) the probability of finding a similar sequence (that is one with 22 identities and 17 differences of the type observed in a total of 39 nucleotides) is ~2.3 x 10^{-5}, that is, once in ~43 kb or, given a haploid genome size of 3 x 10^{-9}, a total of ≈10^{5} 'microdonors' within the genome. The evolutionary history of the contemporary brown lemur β-globin gene sequence will remain uncertain in the absence of information from other contemporary prosimians concerning their cluster organisation and the sequence of β-like globin genes.

8.10 A fifth distinct β-globin gene in the premammalian β-globin gene cluster

The β-globin gene clusters of several non-primate mammals (rabbit, mouse, and goat) have been extensively characterised and many of the functional genes and pseudogenes sequenced (see Collins and Weissman, 1984). Intraspecies gene orthologies have been established that suggest the pre-eutherian radiation β-globin gene cluster was composed of single distinct proto ε-, γ-, δ- and β-globin genes. The simplest interpretation of contemporary mammalian β-globin gene cluster arrangements would therefore be that they reflect lineage dependent evolution of this common ancestral cluster during the past ~80 MYs.

Several lines of evidence have been put forward in this thesis that together strongly suggest the contemporary human Ψβ1 gene also has a long and distinct evolutionary history. This evidence suggests the Ψβ1 locus arose before the mammalian radiation ~80 MYs ago and that contemporary primate and non-primate mammalian β-globin gene clusters have therefore evolved from a common ancestral cluster consisting of at least a proto ε-, γ-, Ψβ1-, δ- and β-like globin gene.

In order to acknowledge the long and distinct evolutionary history of the Ψβ1 gene, and avoid confusion concerning its origin in relation to the other human β-globin genes, it is proposed that the human Ψβ1 gene be renamed Ψη (η following on from the present distinct human globin genes α to ζ), with contemporary orthologues being subsequently referred to as the Ψη pseudogene of simians; the hybrid Ψη-δ pseudogene of lemurs; the η-globin gene in the goat and η-related sequences in the lion, dog and seal. The minimum pre-eutherian cluster would therefore consist of proto ε-, γ-, η-, δ- and β-globin genes.

8.11 Ψη and the evolution of the primate β-globin gene cluster

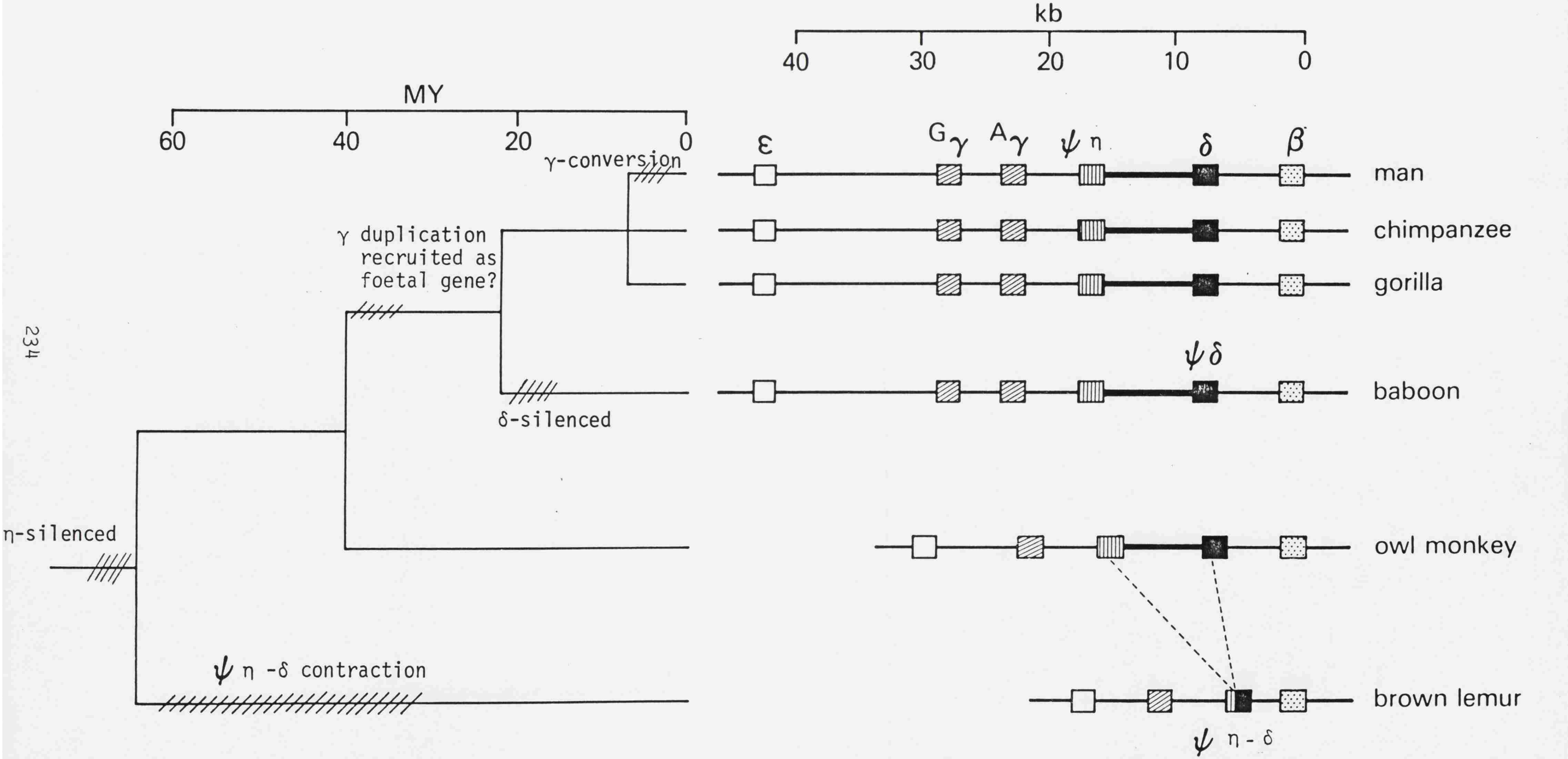
The analysis of Ψη related sequences in man, gorilla, chimpanzee, baboon, owl monkey and brown lemur (Barrie et al., 1981; Chang and Slightom, 1984; this thesis) contributes further to our understanding of the evolution of the primate β-globin gene cluster (Figure 8.5). The simplest interpretation of contemporary primate β-globin gene cluster arrangements is that they have evolved from the proposed pre-eutherian β-globin gene cluster containing single proto ε-, γ-, η-, δ- and β-like genes; this arrangement having been retained in at least one contemporary primate (the owl monkey).

Since the basal primate radiation (~70 MYs ago) the primate β-globin gene family has evolved by mechanisms of gene duplication, gene conversion, unequal exchange and transposon insertion. However it is apparent from contemporary gene cluster arrangements that these processes, or the fixation of the products of these processes, have not occurred at a constant rate throughout primate evolution. For example, the similarity in organisation of the β-globin gene clusters of the Old World monkeys, great apes and man suggest an absence of those processes (or fixation of the products of such processes) which would result in large changes in DNA content, including the region around the Ψη gene. This may reflect either an intrinsically low rate for these processes in the higher primates or

Figure 8.5

Evolution of the β-globin gene cluster in the primates.

The organisation of the β-globin gene clusters are taken from: man, Fritsch *et al*. (1980); gorilla and yellow baboon, Barrie *et al*. (1981); Chimpanzee, Barrie (1982); owl monkey, this thesis, and the brown lemur, Barrie *et al*. (1981). Where expressed, all genes are transcribed from left to right. The Ψη in Old World species was located by genomic mapping using human Ψη probes (see Figure 3.~~4~~5). The divergence times shown are taken from palaeontological and protein data and are approximate (see Figure 8.2). Major events during β-globin gene cluster evolution in the primates are indicated beside the relevent branch of the phylogeny (see text).


kb
40
30
20
10
0
MY
60
40
20
0
ε
Gγ
Aγ
ψη
δ
β
γ-conversion
man
chimpanzee
gorilla
γ duplication recruited as foetal gene?
ψδ
δ-silenced
baboon
η-silenced
owl monkey
ψ η -δ contraction
brown lemur
ψ η - δ

the effect of negative (purifying) selection against variants of what may constitute an optimal gene cluster arrangement.

During primate evolution there have been two alterations in β-globin gene number. One of these was a duplication of the non-adult γ gene resulting in a cluster expansion in the higher primates. The presence of a single γ gene in the owl monkey and duplicate γ genes in Old World monkeys suggests this gene duplication event occurred after the New World monkey-Old World monkey divergence but prior to the Old World monkey divergence from the common ancestor of great apes and man, that is, 20-40 MYs ago. The phylogenetic analysis contrasts sharp~~e~~ly with the estimate of divergence of the human γ genes based on amino-acid and DNA coding sequence divergence. This discrep~~e~~ancy is another example of the influence gene conversion can have on the apparent evolution of two contemporary sequences (see also 8.3). However, comparis~~i~~ons between those non-coding DNA sequences in the 5kb human γ duplication unit not involved in the 1.5 kb conversion tract show these sequences to be approximately 14% diverged (Shen et al., 1981). This sequence divergence exceeds that between the human and owl monkey $\Psi\eta$ sequences (~11%) suggesting that either the γ duplication occurred very soon after the New World-Old World monkey divergence or that the owl monkey initially had a duplicated γ gene arrangment but has eliminated one gene during subsequent evolution. The presence of chromosomes with three or single γ genes in contemporary human populations (Trent et al., 1981; Sukumaran et al., 1983) suggests the fixation of a chromosome with a single γ gene during owl monkey evolution is a feasable evolutionary mechanism that could have led to the contemporary β-globin gene cluster in this species.

The second alteration in gene number in primate β-globin gene cluster evolution involved an unequal exchange between adjacent Ψη and δ like genes in a common ancestor of the lemurs (see 8.5), resulting in a cluster contraction in this lineage. Interestingly this cluster contraction in the lemurs has resulted in the loss of intergenic DNA sequences equivalent to those repeatedly implicated in the developmental switch from foetal to adult β-globin gene expression in man (see Collins and Weissman, 1984). The absence of this region in the brown lemur does not apparantly disrupt the expression of the 3' adult β-globin gene (see 8.7) and again highlights the uncertainty over the exact role this region plays in developmental β-globin gene expression in man (see Orkin and Kazazian, 1984). No information is available however concerning the expression of the 5' ε- and γ-like genes in this cluster which may have been affected by this rearrangment. It is assumed that the ε gene encodes an embryonic globin as in man and the other mammalian species. The developmental expression of the γ-like locus is however unclear. In man and other primates (though not apparently the lemurs, Coppenhaver et al., 1983) the γ gene is expressed primarily during foetal development (see also below); however, in the non-primate mammals the γ-related genes are expressed during embryonic development (Czelusniak et al., 1982; Hill et al., 1984; Hardison, 1984).

The observed similarity of restriction endonuclease sites between different primates suggests that the intergenic non-coding DNA sequences in the β-globin gene cluster are evolving at a rate comparable with that found for the primate Ψη gene (Barrie et al., 1981). This stablility apparently encompasses gross rearrangements as well as base substitution as, for example, with the exception of the lemur, the 3' end of the cluster encompassing the γ - β gene region is very similar in all primates (Figure 8.5). In theory, such non-functional non-coding DNA sequences would be expected to evolve freely in terms of base composition and alterations in intergenic DNA length by insertion/deletion and transposon movement. While small alterations in intergenic DNA distance would not be apparent from genomic mapping, the presence/absence of transposons should be detectable. For example, the human β-globin gene cluster contains several members of the Alu family of retroposons (see Rogers, 1984), transposable elements with the ability to move within the genome and therefore with the potential to cause gross alterations in intergenic DNA length during primate evolution. This has apparently occurred infrequently, or if it has occurred such events have not been fixed in primate evolution, as the spacing of primate globin genes has, in general, remained unaltered.

The increased leng~~ht~~th of the ε - γ intergene distance from 7 kb in the owl monkey to 13.3 kb in the hominoids and Old World monkeys (Figure 8.5) may be an example of such an event, due to the insertion of a 6.4 kb Kpn repetitive element known to be present in this region in man (Forget *et al.*, 1981). It has been suggested (see Collins and Weissman, 1984) that the insertion of this Kpn transposon between the ε and $^{G}\gamma$ may have in fact caused, or facilitated, the shift in regulated expression of the γ genes from embryonic to foetal development during primate evolution. While an attractive proposal, the presence a foetal globin in a New World primate (the marmoset) implies that foetal expression of a γ gene can occur in the apparent absence of the 5' Kpn element (assuming the arrangement of the β-globin gene cluster in this species is the same as in the closely related owl monkey, Figure 8.5). However, only further careful examination of the ε - γ region of the New World monkeys will establish whether a portion of the Kpn element may be present and affecting the developmental expression of the γ gene in this primate group.

In general, such large alterations in intergenic distance have not occurred. The reason for this intergenic DNA stability is unclear but may reflect constraints imposed on the β-globin gene cluster in terms of the overall arrangement and spacing of functional genes such that the whole cluster evolves as a single unit (Barrie _et_ _al_., 1981). A more general phylogenetic analysis of specific retroposons within contemporary primate β-globin gene clusters is required to determine the effects, if any, that these sequences may have had on the evolution of this gene cluster and possibly the genome in general.

8.12 Mammalian β-globin gene orthologies and the evolution of contemporary mammalian β-globin gene clusters

The linkage arrangement of the β-globin gene cluster has been determined for various other non-primate mammals, notably the rabbit, mouse, and goat (Figure 8.6). The orthology of the functional and non-functional members of these clusters have also been determined relative to the human β-globin gene family using dot-matrix criteria, overall sequence homology and maximum parsimony analysis of coding sequences. As a result, various groups have proposed that these contemporary cluster arrangements reflect adaptive evolution from a common ancestral cluster composed of a proto ε-, γ-, δ- and β-like sequences.

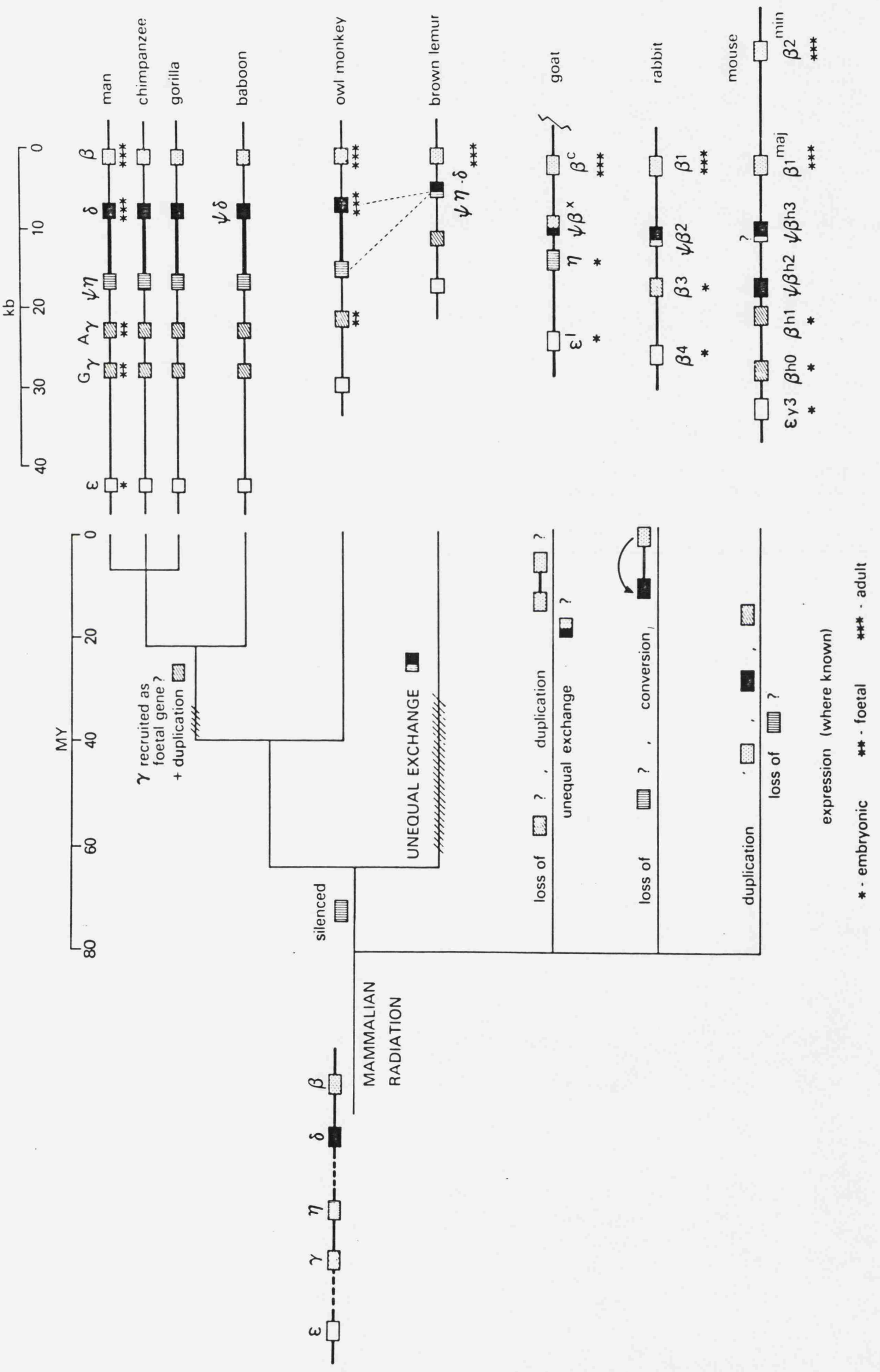
Figure 8.6

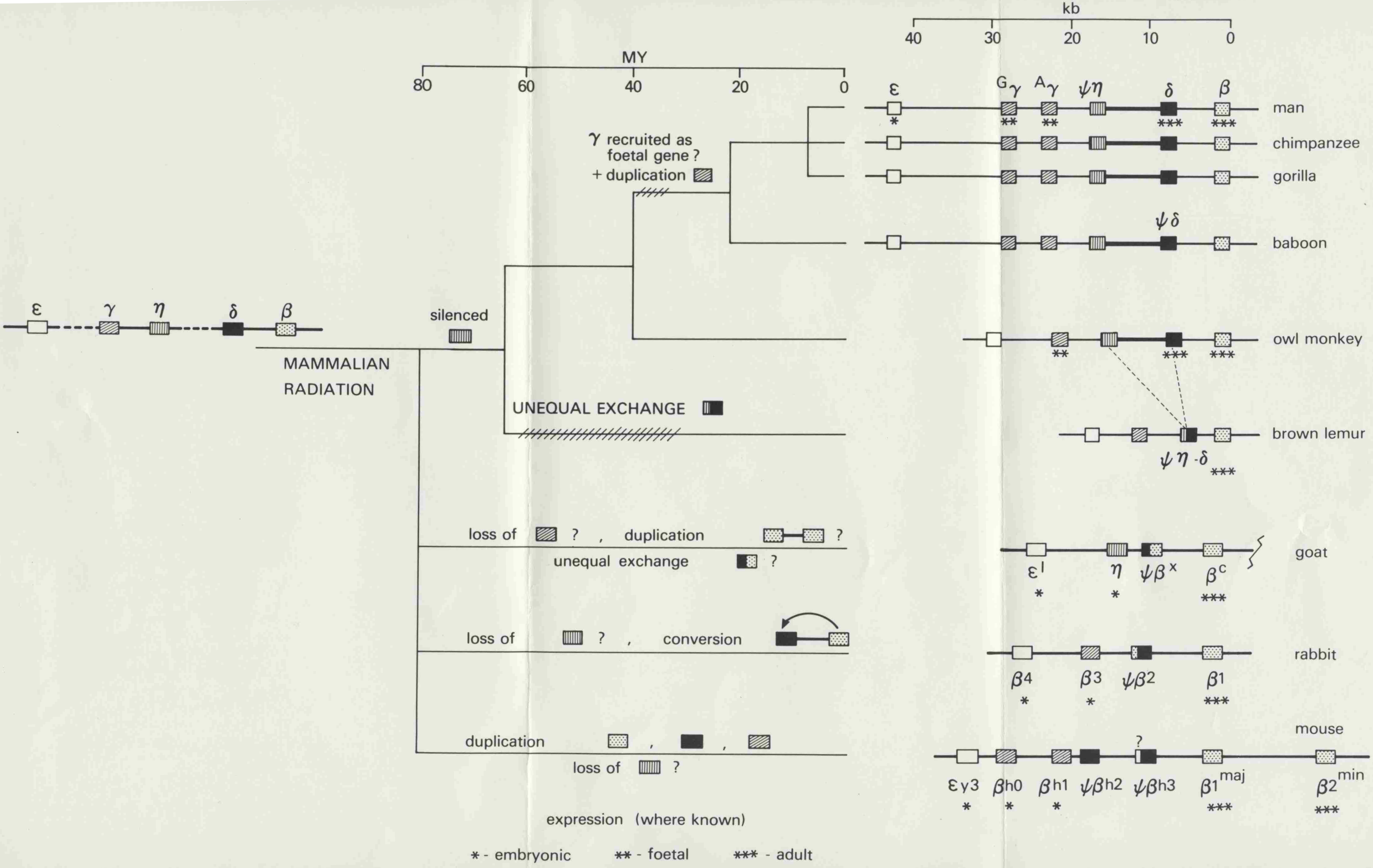
Mammalian β-globin gene cluster evolution.

The contemporary β-globin gene cluster arrangements are shown for various primate and mammalian species. Mammalian β-globin gene cluster arrangements are taken from: goat, Townes et al. (1984); rabbit, Hardison (1984); and mouse, Hill et al. (1984); Hardies et al. (1984). Gene orthologies, indicated with reference to the β-globin gene family in man, were established by comparative coding sequence homology and dot-matrix criteria. Dates are taken from the palaeontological and protein data referred to in Figure 8.4 and are approximate.

Where known, the period of developmental expression of the genes are shown according to the key; in man, great apes and the baboon this is essentially the same. Expression of a foetal globin in the owl monkey is assumed given that another New World monkey (the marmoset) has a foetal globin. For clarity, only the left (5') cluster of the triplicated four gene unit of the goat is shown; the β^{c}-globin gene in this four gene cluster is expressed during juvenile development.

Possible evolutionary events that have determined contemporary gene cluster arrangements are shown for each lineage (see Figure 8.5 for details in the primate lineage). In the primate lineage the history of the β-globin gene cluster is clearer due to the establishment of cluster arrangements from a number of primate groups whose divergence from the lineage leading to man encompasses the ~70 MY's of primate evolution. In the non-primate mammalian lineages the postulated events shown are only one possible scenario between the ancient and contemporary gene cluster arrangment.


kb
0
10
20
30
40
ε
Gγ
Aγ
ψη
δ
β
man
chimpanzee
gorilla
baboon
ψδ
owl monkey
brown lemur
ψη-δ
goat
ε'
η
ψβx
βc
rabbit
β4
β3
ψβ2
β1
mouse
εy3
βh0
βh1
ψβh2
ψβh3
β1maj
β2min
MY
0
20
40
60
80
γ recruited as foetal gene ? + duplication
UNEQUAL EXCHANGE
silenced
MAMMALIAN RADIATION
ε
γ
η
δ
β
loss of ? , duplication ?
unequal exchange ?
loss of ? , conversion
duplication , , , loss of ?
expression (where known)
* - embryonic
** - foetal
*** - adult


MY
80
60
40
20
0
kb
40
30
20
10
0
ε
γ
η
δ
β
MAMMALIAN
RADIATION
silenced
γ recruited as
foetal gene ?
+ duplication
UNEQUAL EXCHANGE
loss of ? , duplication ?
unequal exchange ?
loss of ? , conversion
duplication , ,
loss of ?
expression (where known)
* - embryonic
** - foetal
*** - adult
ε
$^G\gamma$
$^A\gamma$
ψη
δ
β
man
chimpanzee
gorilla
ψδ
baboon
owl monkey
brown lemur
ψη-δ
goat
ε^I
η
$\psi\beta^x$
β^c
rabbit
β4
β3
ψβ2
β1
mouse
εy3
βh0
βh1
ψβh2
ψβh3
$\beta1^{maj}$
$\beta2^{min}$
?

This comparative study of the human η pseudogene suggests that an additional distinct genetic locus was present in this ancient cluster, a η-like gene, in a minimum cluster composed of proto ε-, Υ-, η-, δ- and β-like sequences.

It is possible to reconstruct a minimal evolutionary history for each contemporary mammalian β-globin gene cluster from this common ancestral arrangement of five β-like globin genes (Figure 8.6). However, in the absence of a phylogenetic analysis of these mammalian orders, such as that conducted in the primates, other combinations of gene duplication, gene conversion and unequal exchange, other than those shown in Figure 8.6, could be postulated to account for the contemporary cluster arrangements.

Several features are common amongst all these different clusters. All the genes have the archetypal β-globin gene structure of three exons split by two similar sized introns. The genes are orientated in the same transcriptional orientation 5' → 3' and all conform to the basic arrangement 5'- non-adult gene(s) - pseudogene(s) - adult gene(s) -3', which also corresponds approximately to the order in which they are developmentally expressed. Each cluster contains at least one pseudogene between the 5' non-adult and 3' adult genes. In primates the pseudogene is a descendent of an ancestral η gene whereas in non-primate mammals the pseudogene(s) are decendents of an ancestral δ gene.

The functional status of the ancestral proto δ gene is unclear (see also 8.3). In all but a few primate groups, including man, the δ-like gene is silent. This genetic locus also appears to have been a focal point within the cluster for genetic recombination and, in man at least, resides close to an apparent recombination "hotspot" (see Orkin and Kazazian, 1984). Within the primates the lemur δ-gene has been involved in an unequal exchange with adjacent Ψη sequences. During early simian evolution the δ-globin gene has also apparently been involved in a gene conversion against β sequences which may have "reactivated" a Ψδ sequence, or reduced the expression of a functional δ sequence (Martin et al., 1983). In the other non-primate mammals δ sequences also appear to have been involved in either gene conversion or unequal exchange events involving other genes within the relevent cluster (Hardison, 1984; Hardies et al., 1984). The reason for the active or passive involvement of all these δ-loci in recombinational events is however unknown.

The different mammalian β-globin gene clusters have apparently evolved from a common ancestral gene arrangement and intergenic DNA framework (non-coding DNA such as introns and intergenic DNA sequences). During their independent evolutionary histories however the different mammalian species have expanded or contracted the number of genes and overall size of their β-globin gene cluster such that not all the contemporary arrangements retain the full complement of distinct genes thought to have been present in the common ancestor. For example (see Figure 8.6), the rabbit cluster consists of four genes with orthology to human ε, γ, δ and β globin genes but lacks a η-related gene. Similarly the mouse appears to lack a gene orthologous to η while the goat lacks a gene orthologous to γ, the foetal globin expressed in this species having apparently been recruited from an adult β-like gene (Schon et al., 1981).

Given that the contemporary β-globin gene clusters were derived from a common ancestral gene arrangement embedded in a unique non-coding DNA sequence framework they should retain, albeit diverged, a common non-coding DNA sequence framework in which the gene duplications and deletions have occurred. For example, the human γ duplication apparently involved a 5 kb region of the cluster surrounded by a repeat sequence that may have facilitated the duplication (Shen *et al* ., 1981). On sequencing the intergenic DNA of several of the mammalian β-globin gene clusters it may therefore be possible to distinguish other such duplication units and also the position of gene deletions. For example, in a dot-matrix type analysis, a gene deletion would appear as a discontinuity in the homology between orthologous non-coding DNA sequence frameworks at the position formally occupied, in one lineage, by the common ancestral gene. Such analysis may help distinguish more clearly the evolutionary orthology of some mammalian β-globin genes that are difficult to determine conclusively at present due to the influence of gene conversion and relative sequence divergence (see below).

The mouse β-globin gene cluster has apparently lost sequences with orthology to the ancestral proto η gene. There are however two non-adult β-like globin genes in the mouse cluster (βh0 and βh1), present between the embryonic and adult genes, the ancestry of which is unclear (see Hill *et al*., 1984). The mouse βh0 and βh1 genes are closely homologous over their coding sequences, having apparently evolved in concert during mouse evolution. However, the intergenic non-coding DNA sequence divergence between these two genes (35% over intron 2) suggests they have evolved independently since an ancient rather than recent duplication event, possibly preceeding the eutherian radiation. Their next strongest coding sequence homology is to γ-like globin genes, that is, the orthologous human γ and rabbit β3 genes, thought to have evolved from a common ancestral proto γ gene. This orthology is not reflected however in the ability to align their non-coding DNA sequences with equivalent sequences from the human γ, rabbit β3 or any other β-like gene (Hill et al., 1984). Similarly, neither of these genes show significant alignment over non-coding sequences with the human Ψη gene (own results not shown).

While orthologous sequences in man and the mouse generally show a higher level of sequence divergence than between man and the rabbit, gene orthologies can still be established from non-coding DNA sequence alignments (see Hardies et al., 1984). Failure to align non-coding DNA sequences of βh0 and βh1 against other β-like globin genes may therefore either reflect a relatively high level of sequence divergence in these genes compared to other mammalian globin genes or it may relect their origin from a distinct globin locus no longer present in other mammalian lineages. The concept of a non-coding DNA sequence framework may help resolve the ancestry of these mouse β-globin genes but awaits the complete sequencing of mammalian β-globin gene clusters and the development of even better alignment algorithms before it can be tested.

8.13 Non-processed pseudogenes as a component of multigene family evolution

It is clear that while the pseudogenes within contemporary mammalian β-globin gene clusters have no apparent function, they may contribute to cluster evolution due to their resemblance to functional sequences and may form another source of genetic variation within these clusters. These sequences may themselves engage in unequal crossover, gene conversion, gene rearrangement and gene duplication resulting in the generation of additional pseudogene copies or alterations in cluster arrangement. Many of the features of globin pseudogene evolution also apply to non-processed pseudogenes found in other multigene families. For example, the adjacent $\Psi\eta$ and δ genes of the lemur have been involved in an unequal exchange involving homologous exon 2 sequences that gave rise to a major cluster contraction in this lineage. Similarly, unequal exchange is thought to have resulted in the non-processed pseudogene found in the Drosophila cuticle protein gene cluster (Snyder *et al.*, 1982) and the Igγ3 heavy chain locus in man (Takashashi *et al.*, 1982). Also, an intragenic rearrangement most probably gave rise to the $\Psi C_{\varepsilon}1$ gene in man (Battey *et al.*, 1982; Hi~~sajima~~roshi *et al.*, 1983).

Non-processed pseudogenes can also apparently be involved in gene conversion with other members of the gene family, for example, the δ-locus in primate and non-primate mammals (rabbit and mouse) has been particularly susceptable to conversion by the adjacent adult β-globin gene (this thesis; Hardison, 1984; Hardies *et al.*, 1984). The functional primate δ gene may in fact constitute a "reactivated" pseudogene as a result of such a gene conversion by the functional β-globin gene early in simian evolution (Martin *et al.*, 1983). In the human α-globin gene cluster the $\Psi\zeta$ gene is also thought to have been involved in a recent conversion event with the adjacent functional ζ-globin gene (Proudfoot *et al.*, 1982).

While there are at present no examples of the independent duplication of a non-processed pseudogene, additional pseudogene copies can apparently be generated as part of a larger duplication unit. Examples include the pseudogene within the triplicated four gene β-globin cluster of the goat (Townes _et al._, 1984) and that repeated ~24,000x as part of the tandem repeat unit of the _Xenopus laevis_ oocyte 5S RNA cluster (Jacq _et al._, 1977). An additional means of non-processed pseudogene duplication, associated with transposition, is apparently via an RNA intermediate (see 1.4(c)). The RNA in this case retains the intron and flanking DNA sequences normally associated with the gene (Leder _et al._, 1981; Maxson _et al._, 1983). The precise mechanism(s) responsible for this phenomenon are as yet unclear (see Vanin, 1984), as is the question of whether the transposed gene is non-functional before, after or as a result of the duplication/transposition event.

8.14 Additional Ψη-related sequences in the human genome

At least one additional Ψη-related sequence has been detected in man and gorilla using human intron 2 DNA probes and several additional hybridising fragments were also found in the lion, dog and seal. One of these additional human sequences may constitute a genuine Ψη-related sequence as the hybridising genomic DNA fragment was also detected using another Ψη probe (probe 1) and a rabbit adult β-globin cDNA probe. In man this additional sequence is apparently dispersed from the β-globin gene cluster as the DNA fragment detected does not correspond to DNA fragments within the cluster. This sequence may therefore correspond to the rare class of pseudogene, which includes the mouse Ψα4 pseudogene, known as dispersed non-processed pseudogenes. Attempts to isolate these additional sequences from a human λ-library failed. One isolated recombinant contained homology to a short low copy number tandem repeat within the large intron of the human Ψβ1 gene which is apparently present elsewhere in the human genome.

These low copy number repeat containing fragments, which can be faintly detected in genomic DNA digests, do not appear to correspond to the additional putative dispersed Ψη-related sequence. The exact nature of the "genuine" additional Ψη-related sequence therefore remains unclear. Such a sequence is potentially very interesting as a means of establishing if the rate of non-coding DNA sequence evolution outside the functional "domain" that may correspond to the primate β-globin gene cluster is the same, or different, to that observed in other regions of the human genome. Phylogenetic analysis of additional Ψβ1-related sequences at different genomic locations would help determine whether this was in fact the case.

8.15 Summary

This study of contemporary Ψβ1 pseudogene sequences in the primates has shown that while without function this gene has been a stable component of the β-globin gene cluster during the evolution of this mammalian order. The pseudogene was probably functional early in primate evolution and was silenced recently before the basal primate radiation 60-70+ MYs ago. After silencing, the gene has evolved randomly in terms of base substitution and microinsertions/deletions at a mean rate thought to be representative of such sequences throughtout the primates, but which is less than that expected according to the current estimate of the neutral rate for such sequences. While apparently non-functional, the presence of the pseudogene has influenced the evolution of the β-globin gene cluster during primate evolution as a result of residual homology over coding regions. This is illustrated by the cluster contraction in the lemurs involving ΨB1 and δ sequences. Analysis of this pseudogene in the primates completes the characterisation of the archetype β-globin related sequences within this gene family. Evidence for an early functional history for ΨB1-related sequences prior to the eutherian radiation is supported by the detection of ΨB1-related sequences in other mammalian orders, in one species of which the ΨB1 orthologue is apparently expressed. Finally, the history of the ΨB1-like sequences in the primates and various mammalian orders suggests a previously unidentified ancient and discrete genetic locus in the β-globin gene cluster that has been termed η. The simplest interpretation of the evolution of contemporary primate and other mammalian β-globin gene clusters is that they descended from a common minimal ancestral cluster composed of proto ε-, γ-, η-, δ- and β-like sequences.

The generality of the conclusions drawn from this work concerning pseudogene longevity and sequence evolution after silencing await the phylogenetic analysis of other pseudogene sequences. It is apparant however that pseudogenes may constitute another source of genetic variation on which the process of natural selection can act in the evolution of both eukaryotic multigene gene families and the genome in general.

Bibliography

Aaij,C. and Borst,P. (1972).

The gel electrophoresis of DNA. Biochim.Biophys.Acta.,269,192-200

Allan,M. and Paul,J. (1984).

Transcription in vivo of an Alu family member upstream from the human ε-globin gene. Nuc.Acids.Res.,12,1193-1200.

Andrews,P. and Cronin,J.E. (1982).

The relationships of Sivapithecus and Ramapithecus and the evolution of the orangutan. Nature 297,541-546.

Arnheim,N., Krystal,M., Schmickel,R., Wilson,G., Ryder,O. and Zimmer,E. (1980).

Molecular evidence for genetic exchanges among ribosomal genes on non-homologous chromosomes in man and apes. Proc.Natl.Acad.Sci.USA.,17,7323-7327.

Arnheim,N. (1983).

Concerted evolution of multigene families. in Evolution of Genes and Proteins. Sinauer Associates Inc. USA.

Artymink,P.J., Blake,C.C.F. and Sippel,A.E. (1981).

Genes pieced together-exons delineate homologous structures of diverged lysozymes. Nature 290,287-288.

Bankier,A.T and Barrell,B.G. (1983).

in "Techniques in the life sciences",Vol.B5. Elsevier,Ireland.

Baralle,F.E., Shoulders,C.C. and Proudfoot,N.J. (1980).

The primary structure of the human ε-globin gene. Cell 21,621-626.

Baralle,F.E., Shoulders,C.C., Goodbourn,S., Jeffreys,A. and Proudfoot,N.J. (1980b).
The 5' flanking region of the human ε-globin gene.
Nucl.Acids.Res.,8,4393-4404.

Barrie,P.A. (1982).
A study on the β-globin gene cluster in man and the primates. Ph.D. Thesis, Leicester University, England.

Barrie,P.A., Jeffreys,A.J. and Scott,A.F. (1981).
Evolution of the β-globin gene cluster in man and the primates.
J.Mol.Biol.,149,319-336.

Battey,J., Max,E.E., McBride,W.O., Swan,D. and Leder,P. (1982).
A processed human immunoglobulin ε gene has moved to chromosome 9.
Proc.Natl.Acad.Sci.USA.,79,5956-5960.

Beard,J.M., Barnicot,N.A and Hewett-Emmett,D. (1976).
α and β chains of the major haemoglobin and a note on the minor component of Tarsius. Nature 259,338-340.

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

Berstein,L.B., Mount,S.M. and Weiner,A.M. (1983).
Pseudogenes for human small nuclear RNA U3 appear to arise by intergration of self-primed reverse transcripts of the RNA into new chromosomal sites. Cell 32,461-672.

Birnboim,H.C. and Daly,J. (1979).
A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucl.Acids Res.,7,1513-1523.

Blanchetot,A., Wilson,V., Wood,D. and Jeffreys,A.J. (1983).

The seal myoglobin gene: an unusually long globin gene. Nature 301,732-734.

Blackburn,E.H. and Szostak,J.W. (1984).

THe molecular structure of centromeres and telomeres. Ann.Rev.Biochem.,53,163-194.

Blake,C. (1983).

Exons-present from the beginning. Nature 306,535-537.

Blake,C.C.F. (1978).

Do genes-in-pieces imply proteins-in-pieces ?. Nature 273,267.

Blattner,F.R., Williams,B.G., Blechl,A.E., Denniston-Thompson,K., Faber,H.E., Furlong,L., Grunwald,D.J., Kiefer,D.O., Moore,D.D., Schumm,J.W., Sheldonm E.L. and Smithies,O. (1977).

Charon phages: safer derivatives of bacteriophage lambda for DNA cloning. Science,194,161-169.

Bonaventura,C., Sullivan,B. and Bonaventura,J. (1974).

Effects of pH and anions on functional properties of haemoglobin from lemur fulvus fulvus. J.Biol.Chem.,249,3768-3775.

Borst,P. and Grivell,L.A. (1981).

One gene's intron is another gene's exon. Nature 289,439-440.

Bouche,J.P. (1981).

The effect of spermidine on endonuclease inhibition by agarose contaminates. Anal.Biochem.,115,42-45.

Brown,A.L. (1984).

On the origin of the Alu family of repeated sequences. Nature 312,106.

Brown,W.M., Prager,E.M., Wang,A. and Wilson,A.C. (1982).
Mitochondrial DNA sequences of primates: tempo and mode of evolution. J.Mol.Evol.,18,225-239.

Brownell,E., Krystal.M., and Arnheim,N. (1983).
Structure and evolution of human and african ape rDNA pseudogenes. Mol.Biol.Evol.,1,29-37.

Bunn,F.H. (1980).
Regulation of haemoglobin function in mammals. Amer.Zool.,20,199-211.

Busch,H., Reddy,R., Rothblum,L and Choi,Y.C. (1982).
SnRNAs, snRNPs and RNA processing. Ann.Rev.Biochem.,51,617-654.

Cartmill,M. (1982).
Island primates. Science 217,1132-1133.

Chang,L.-Y.E. and Slightom,J.L. (1984).
Isolation and sequence analysis of the β-type globin pseudogene from human, gorilla and chimpanzee. J.Mol.Biol.,180,767-784.

Chu,F.K., Maley,G.F. Maley,F. and Belfort,M. (1984).
Intervening sequence in the thymiylate synthetase gene of bacteriophage T4. Proc.Natl.Acad.Sci.USA 81,3049-3053.

Cochet,M., Gaanon,F., Hen,R., Maroteaux,L., Perrin,F. and Chambon,P. (1979).
Organisation and sequence studies of the 17-piece chicken conalbumin gene. Nature 282,567-574.

Cohen,J.B. and Givol,D. (1983).
Conservation and divergence of immunoglobulin heavy chain variable region pseudogenes. The EMBO Journal 2,1795-1800.

Cohen,S.N., Chang,A.C.Y. and Hsu,L. (1972).

Nonchromosomal antibiotic resistance in bacteria: genetic transformation of Escherichia coli by R-factor DNA. Proc.Natl.Acad.Sci.U.S.A.,69,2110-2114.

Collins,F.S. and Weissman,S.M. (1984).

The molecular genetics of human haemoglobin. Prog.Nucl.Acids.Res.,31,312-462.

Coppenhaver.D.H., Dixon,J.D. and Duffy,L.K. (1983).

Prosimian haemoglobins I. The primary structure of the β-globin chain of Lemur catta. Haemoglobin 7,1-14.

Cowen,N.J. and Dudley,L. (1983).

Tubulin isotypes and the multigene gene families. Int.Rev.Cytol.,83,147-173.

Crabtree,G.R. and Kent,J. (1982).

Organisation of the rat γ-fibrinogen gene: alternative mRNA splicing patterns produce the γA and γB (γ') chains of fibrinogen. Cell 31,159-166.

Curtis,S.E. and Clegg,M.T. (1984).

Molecular evolution of chloroplast DNA sequences. Mol.Biol.Evol.,1,291-301.

Curtiss,W.C. and Vournakis,J.N. (1984).

Quantitation of base substitutions in eukaryotic 5S RNA: selection for the maintenance of RNA secondary structure. J.Mol.Biol.,20,351-361.

Czelusniak,J., Goodman,M., Hewett-Emmett,D., Weiss,M.L., Venta,P.J. and Tashian,R.E. (1982).

Phylogenetic origins and adaptive evolution of avian and mammalian haemolglobin genes which are expressed differentially during ontogeny. Nature 298,297-300.

Daniels,D.L., De Wet,J.R. and Blattner,F.R. (1980).

New map of bacteriophage lambda DNA. J.Virol.,33,390-400.

Daniel,S.B., Strusbaugh,L.D., Ehrman,L. and Armstrong,R. (1984).

Sequences homologous to P elements occur in Drosophila paulistorum. Proc.Natl.Acad.Sci.USA 81,6794-6797.

Darnell,J.E. (1978).

Implications of RNA·RNA splicing in evolution of eukaryote cells. Science 202,1257-1260.

Dayhoff,M.O., Hunt,L.T., McLaughlin,P.J. and Jones,D.D. (1972).

Gene duplications in evolution. In Atlas of Protein Sequence And Structure. Vol 5. M.O.Dayhoff,Ed., Natn.Biomed.Res.Found., Washington D.C. USA. pp. 17-30.

Dayhoff,M.O. ed. (1972).

Atlas of Protein Sequence and Structure. Vol 5., Natl.Biomed.Res.Found., Silver Spring, Md., USA.

Dayhoff,M.O. ed. (1976).

Atlas of protein sequence and structure. Vol 5, supplement 2. Natl.Biomed.Res.Found., Washington, D.C., USA.

Deininger,P.L. (1983).

Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. Anal.Biochem.,129,216-223.

Dickerson,R.E. (1971).

The structure of cytochrome c and rates of molecular evolution. J.Mol.Evol.,1,26-45.

Dickerson,R.E. and Geis,I. (1983).

in Hemoglobin: structure, function, evolution and pathology. Benjamin/Cummings Publishing Co., Menlo Park. U.S.A.

Di Nocera,P.P. and Dawid,I.B. (1983).

Interdigitated arrangement of two oligo(A)-terminated DNA sequences in Drosophila. Nuc.Acids.Res.,11,5475-5482.

Doolittle,W.F. (1978).

Genes in pieces: were they ever together ?. Nature 272,581-582.

Dover,G. (1982).

Molecular drive: a cohesive mode of species evolution. Nature 299,111-117.

Dretzen,G., Bellard,M., Sassone-Corsi,P. and Chambon,P. (1981).

A reliable method for the recovery of DNA fragments from agarose and acrylamide gels. Anal.Biochem.,112,295-298.

Duckworth,M.L., Gait,M.J., Goelet,P., Hong,G.F., Singh,M. and Titmas,R.C. (1981).

Rapid synthesis of oligodeoxyribnucleotides VI. Efficient,mechanised synthesis of heptadecadeoxyribonucleotides by an improved solid phase phosphotriester route. Nucl.Acids Res.,9,1691-1706.

Eaglen,R.M. (1984).

Parallelism, parsimony, and the phylogeny of the lemuridae. Int.J.Primatology 4,249-273.

Early,P. Rogers,J., Davis,M., Calame,K., Bond,M., Wall,R. and Hood,K. (1980).

Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing. Cell 20,313-319.

Eaton,W.A. (1980).

The relationship between coding sequences and function in haemoglobin. Nature 284,183-185.

Efstratiadis,A., Posakony,J.W., Maniatis,T., Lawn,R.M., O'Connell,C., Spritz,R.A., DeRiel,J.K., Forget,B.G., Weissman,S.M., Slightom,J.L., Blechl,A.E., Smithies,O., Baralle,F.E., Shoulders,C.C. and Proudfoot,N.J. (1980).

The structure and evolution of the human β-globin gene family. Cell 21,653-668.

Eiferman,F.A. Young,P.R., Scott,R.W. and Tilghman,S.M. (1981).

Intragenic amplification and divergence in the mouse α-fetoprotein gene. Nature 294,713-718.

Engel,J.D. and Dodgson,J.B. (1980).

Analysis of the closely linked adult α-globin genes in recombinant DNAs. Proc.Natl.Acad.Sci.USA 77,2596-2600.

Engel,J.D. Rusling,D.J., McCune,K.C. and Dodgson,J.B. (1983).

Unusual structure of the chicken embryonic α-globin gene π'. Proc.Natl.Acad.Sci.USA 80,1392-1396.

Engels,W.R. (1983).

The P family of transposable elements in Drosophila. Ann.Rev.Genetics 17,315-344.

Falconer,D.S. (1981).

Introduction to quantitative genetics. 2nd Edition. Longman.

Farris,J.S. (1970).

Methods for computing Wagner trees. Syst.Zool.,19,83-92.

Farris,J.S. (1972).

Estimating phylogenetic trees from distance matrices.

Amer.Nat.,106,645-668.

Firtel,R.A. (1981).

Multigene families encoding actin and tubulin. Cell 24,6-7.

Fischer,D.H., Dodgson,J.B., Hughes,S and Engel,J.D. (1984)

An unusual 5' splice sequence is efficiently utilised *in vivo*.

Proc.Natl.Acad.Sci.U.S.A.,81,2733-2737.

Fitch,W.M. and Margoliasch,E. (1967).

Construction of phylogenetic trees. Science 155,279-284.

Fitch,W.M. (1973).

Aspects of molecular evolution. Ann.Rev.Genetics 7,343-380.

Flanagan,J.G. and Rabbits,T.H. (1982).

Arrangement of human immunolglobulin heavy chain constant region genes implies evolutionary duplication of a seqment containing γ, ε and α genes. Nature 300,709-713.

Forget,B.G., Tuan,D., Biro,P.A., Jagadeeswaran,P. and Weissman,S.M.(1981).

Structural features of the DNA flanking the human non-alpha globin genes: implications in the control of fetal hemoglobin switching.

Trans.Assoc.Amer.Physic.,94,204-210.

Freytag,S.O., Bock,H.-G.O., Beaudet,A.L. and O'Brien,W.E. (1984).

Molecular structures of human arginosuccinate synthetase pseudogenes. J.Biol.Chem.,259,3160-3166.

Fritsch,E.F., Lawn,R.M. and Maniatis,T. (1980).

Molecular cloning and characterisation of the human β-like globin gene cluster. Cell 19,959-972.

Gideon,R., Ram,D., Glazer,L., Zakut,R. and Givol,D. (1983).

Evolutionary aspects of immunoglobulin heavy chain variable region (V_H) gene subgroups. Proc.Natl.Acad.Sci.USA 80,855-859.

Gilbert,W. (1978).

Why genes in pieces ?. Nature 271,501.

Gillespie,D. (1977).

Newly evolved repeated DNA sequences in primates. Science 196,889-891.

Gingerich,P.D. and Schoeninger,M. (1977).

The fossil record and primate phylogeny. J.Hum.Evol.,6,483-505.

Glover,D.M. (1981).

The rDNA of Drosophila melanogaster. Cell 26,297-298.

Go,M. (1981).

Correlation of DNA exonic regions with protein structural units in haemoglobin. Nature 291,90-91.

Goeddel,D.V., Leung,D.W., Dull,T.J., Gross,M., Lawn,R.M., McCandliss,R., Seeburg,P.H., Ullrich,A., Yelverton,E. and Gray,P.W. (1981)

The structure of eight distinct cloned human leukocyte interferon cDNAs. Nature 290,20-26.

Gojobori,T. (1983).

Codon substitution in evolution and the "saturation" of synonymous changes. Genetics 105,1011-1027.

Gojobori,T and Nei,M. (1984).

Concerted evolution of the immunoglobulin V_H gene family. Mol.Biol.Evol.,1,195-212.

Goodman,H.M. (1980).

Repair of overlapping DNA termini. Meth.Enz.,65,63-64.

Goodman,M. Moore,G.W. Barnabas,J. and Matsuda,G. (1974).

The phylogeny of the human globin genes investigated by the maximum parsimony method. J.Mol.Evol.,3,1-48.

Goodman,M. Moore,G.W. and Matsuda,G. (1975).

Darwinian evolution in the geneology of haemoglobin. Nature 253,603-608.

Goodman,M. (1981).

Decoding the pattern of protein evolution. Prog.Biophys.Mol.Biol.,38,105-164.

Goodman,M. (1981).

Globin evolution was apparently very rapid in early vertebrates: a reasonable case against the rate-constancy hypothesis. J.Mol.Evol.,17,114-120.

Goodman,M., Koop,B.F., Czelusniak,J., Weiss,M.L. and Slightom, J.L.(1984).

The ζ-globin gene: its long evolutionary history in the β-globin gene family of mammals. J.Mol.Biol.,180,803-823.

Grantham,R., Gantier,C., Gouy,M., Jacobzone,M. and Mercier,R. (1981).

Codon catalogue usage is a genome strategy modified for gene expressivity. Nucl.Acids.Res.,9,r43-r74.

Grunstein,M. and Hogness,D.S. (1975).

Colony hybridisation: a method for the isolation of cloned DNAs that contain a specific gene. Proc.Natl.Acad.Sci.U.S.A.,72,3961-3965.

Hall,B.G., Yokoyama,S. and Calhoun,D.H. (1983).
Role of cryptic genes in microbial evolution.
Mol.Biol.Evol.,1,109-124.

Hansen,T.H., Spinella,D.G., Lee,D.R. and Screffer.,D.C. (1984).
The immunogenetics of the mouse major histocompatability gene complex. Ann.Rev.Genet.,18,99-129.

Hardies,S.C., Edgel,M.H. and Hutchison III,C.A. (1984).
Evolution of the mammalian β-globin gene cluster.
J.Biol.Chem.,259,3748-3756.

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

Hardison,R.C. (1983).
The nucleotide sequence of the rabbit embryonic globin gene β4.
J.Mol.Chem.,258,8739-8744.

Hardison,R.C. (1984).
Comparison of the β-like globin gene families of rabbits and humans indicates that the gene cluster 5'-ε-γ-δ-β-3' predates the mammalian radiation. Mol.Biol.Evol.,1,390-410.

Hardison,R.C. and Margot,J.B. (1984).
Rabbit globin pseudogene Ψβ2 is a hybrid of δ- and β-globin gene sequences. Mol.Biol.Evol.,1,302-316.

Harris,H. (1966).
Enzyme polymorphism in man. Proc.Roy.Soc. London,Ser.B.,164,298-310

Harris,H. (1976).
Molecular evolution: the neutralist-selectionist controversy.
Fed.Proc.,35,2079-82.

Hess,J.F., Fox,M., Scmid,C. and Shen,C.-K.J. (1983).

Molecular evolution of the human adult α-globin-like gene region: insertion and deletion of Alu family repeats and non-Alu sequences. Proc.Natl.Acad.Sci.USA 80,5970-5974.

Hill,A., Hardies,S.C., Phillips,S.J., Davis,M.G., HutchisonIII,C.A. and Edgell,M.H. (1984).

Two mouse early embryonic β-globin gene sequences. J.Biol.Chem.,259,3739-3747.

Hill,R.L. and Buettner-Jansch,J. (1964).

Evolution of haemoglobin. Fed.Proc.,23,1236-1242.

Hiroshi,N. and Kornberg,R.D. (1983).

Genes and pseudogenes for mouse U1 and U2 small nuclear RNA species. J.Biol.Chem.,258,8151-8155.

Hiroshi,H., Nishida,Y., Nakai,S., Takahashi,N., Ueda,S. and Honjo,T. (1983).

Structure of the human immunoglobulin $C_{\varepsilon}2$ gene, a truncated pseudogene: Implications for its evolutionary origin. Proc.Natl.Acad.Sci.USA 80,2995-2999.

Hollis,G.F., Heiter,P.A., McBride,W.O., Swan,D. and Leder,P. (1982).

Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. Nature 296,321-325.

Hood,L., Campbell,J.H. and Elgin,S.C.R. (1975).

The organisation, expression and evolution of antibody genes and other multigene families. Ann.Rev.Genet.,9,305-353.

Hood,L., Kronenberg,M. and Hunkapillaer,T. (1985).

T cell antigen receptors and the immunoglobulin supergene family. Cell 40,25-229.

Hosbach,H.A., Wyler,T. and Weber,R. (1983).

The Xenopus laevis globin gene family: chromosomal arrangement and gene structure. Cell 32,45-53.

Ikemura,T. (1985).

Codon usage and tRNA content in unicellular and multicellular organisms. Mol.Biol.Evol.,2,13-34.

Jacq,C., Miller,J.R. and Brownlee,G.G. (1977).

A pseudogene structure in 5S DNA of Xenopus laevis. Cell 12,109-120.

Jagadeeswaran,P., Pan,P., Forget,B.G. and Weissman,S.M. (1982).

Sequences of human repetitive DNA, non-α-globin genes, and major histocompatibility locus genes. Cold Spring Harbor Symp.Quant.Biol.,47,1079-1086.

Jeffreys,A.J., Wilson,V., Wood,D., Simons,J.P., Kay,R.M. and Williams,J.G. (1980).

Linkage of adult α-and β-globin genes in X.laevis and gene duplication by tetraploidization. Cell,21,555-564.

Jeffreys,A.J. and Harris,S. (1982).

Processes of gene duplication. Nature 296,9-10.

Jeffreys,A.J., Harris,S., Barrie,P.A., Wood,D., Blanchetot,A. and Adams,S.M. (1983).

Evolution of gene families : the globin genes. in Evolution from Molecules to Men (ed. D.S. Bendall), pp.175-95. Cambridge University Press.

Jeffreys,A.J. Wilson,V., Blanchetot,A., Weller,P., van Kessel,A.G., Spurr,N., Solomon,E. and Goodfellow,P. (1984).

The human myoglobin gene: a third dispersed globin gene locus in the human genome. Nucl.Acids Res.,12,3235-3243.

Jensen,E.O., Paluden,K., Hyldig-Nielson,J.J., Jorgensen,P. and Marcker,K.A. (1981).

Structure of a chromosomal leghaemoglobin gene from soybean. Nature 291,677-679.

Kaine,B.P., Gupta,R. and Woese,C.R. (1983).

Putative introns in tRNA genes of prokaryotes. Proc.Natl.Acad.Sci.USA.,80,3309-3312.

Karin,M. and Richards,R.I. (1982).

Human metallothionein genes - primary structure of the metallotionein II gene and a related processed gene. Nature 299,797-802.

Karin,H., Westin,G. and Pettersson,U. (1982).

A pseudogene for human U4 with a remarkable structure. The EMBO Journal 1,737-740.

Kavanagh,M. (1983).

A complete guide to monkeys, apes and other primates. Oregon Press Ltd. London.

Kidwell,M.G. (1983).

Evolution of hybrid dysgenesis determinants in Drosophila melanogaster. Proc.Natl.Acad.Sci.USA.,80,1655-1659.

Kimura,M. (1983a).

The neutral theory of molecular evolution. in Evolution of genes and proteins (eds.) Nei,M and Keohn,R.K. Sinauer associates Inc, USA.

Kimura,M. (1983b).

The neutral theory of molecular evolution. Cambridge university press.

King,G.R. and Piatigorsky,J. (1983).

Alternative RNA splicing of the murine αA-crystallin gene: Protein -coding information within an intron. Cell 32,707-712.

King,L.K. and Jukes,J.H. (1969).

Non-Darwinian evolution. Science 164,788-798.

Kirby,K.S. (1957).

A new method for the isolation of deoxyribonucleic acids: evidence on the nature of bonds between deoxyribonucleic acid and protein. Biochem.J.,66,495-504.

Konkel,D.A., Maizel,J.V. and Leder,P. (1979).

The evolution and sequence comparison of two recently diverged mouse chromosomal β-globin genes. Cell 18,865-873.

Kushner,S.R. (1978).

An improved method for transformation of Escherichia coli with COL E1 derived plasmids. In Genetic engineering. Boyer,H.W. and Nicosia,S. (eds). North Holland Biomedical Prees, Amsterdam, pp17-23

Lacy,E. and Maniatis,T. (1980).

The nucleotide sequence of a rabbit β-globin pseudogene. Cell 21,545-553.

Langley,C.H. and Fitch,W.M. (1974).

An examination of the constancy of the rate of molecular evolution. J.Mol.Evol.,3,161-177.

Lawn,R.M., Efstratiadis,A., O'Connell,C and Maniatis,T. (1980).

The nucleatide sequence of the human β-globin gene. Cell 21,647-651.

Lee,J.S., Brown,G.G., and Verma,D.P.S. (1983).

Chromosomal arrangement of leghaemoglobin genes in soybean. Nuc.Acids.Res.,11,5541-5553.

Lee,M.G-S., Lewis,S.A., Wilde,D.C. and Cowan,N.J. (1983).

Evolutionary history of a multigene family: An expressed human β-tubulin gene and three processed pseudogenes. Cell 33,477-488.

Leder,A., Swan,D., Ruddle,F., D'Eustachio.,P. and Leder,P. (1980).

Dispersion of α-like globin genes of the mouse to three different chromosomes. Nature 293,196-200.

Leigh-Brown,A.J. and Ish-Horowicz,D. (1981).

Evolution of the 87A and 87C heat-shock loci in Drosophila. Nature 290,677-682.

Lemischka,I. and Sharp,P.A. (1982).

The sequences of an expressed rat α-tubulin gene and a pseudogene with an inserted repetitive element. Nature 300,330-335.

Lewin,R. (1983).

Promiscuous DNA leaps all barriers. Science 218,478-479.

Lewontin,R.C. (1974).

The genetic basis of evolutionary change. Columbia University Press, New York.

Lewontin,R.C and Hubby,J.L. (1966).

A molecular approach to the study of genic heterozygosity in natural populations,II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudobscura. Genetics 54,595-609.

Li,W-H., Gojobori,T. and Nei,M. (1981).

Pseudogenes as a paradigm of neutral evolution. Nature 292,237-39.

Li,W.-H. (1980).

Rate of gene silencing at duplicate loci: A theoretical study and interpretation of data from tetraploid fish. Genetics 95,237-258.

Li,W-H. (1984).
Retention of cryptic genes in microbial populations. Mol.Biol.Evol.,1,213-219.

Li,W-H., Wu,C.I. and Luo,C.C. (1985).
A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol.Biol.Evol.,2,150-174.

Liebhaber,S.A., Goossens,M. and Kan,Y.W. (1981).
Homology and concerted evolution at the α1 and α2 loci of human α-globin. Nature 290,26-29.

Little,P.R.F. (1982).
Globin pseudogenes. Cell 28,683-684.

Loenen,A.M. and Brammer,W.J. (1980).
A bacteriophage lambda vector for cloning large DNA fragments made with several restriction enzymes. Gene,20,249-259.

Lomedico,P., Rosenthal,N., Efstratiadis,A., Gilbert,W., Klodoner,R. and Tizard,R. (1979).
The sturucture and evolution of the two non-allelic rat preproinsulin genes. Cell 18,545-558.

Long,E.O. and Dawid,I.B. (1980).
Repeated genes in eukaryotes. Ann.Rev.Biochem.,49,727-764.

Maeda,N., Yang,F., Barnett,D.R., Bowman,B.H. and Smithies,O. (1984).
Duplication within the haptoglobin Hp^2 gene. Nature 309,131-135.

Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982).
Molecular Cloning (a laboratory manual). Cold Spring Harbor Laboratory, New York, U.S.A.

Maita,T., Goodman,M and Matsuda,G. (1978).

Amino-acid sequences of the α and β chains of adult haemoglobin of the slender loris, Loris tardigradis. J.Biochem.,84,377-383.

Martin,S.L., Vincent,K.A. and Wilson,A.C. (1983).

Rise and fall of the delta globin gene. J.Mol.Biol.,164,513-528.

Masters,J.N., Yang,J.K., Cellini,A. and Attardi,G. (1983).

A human dihydrofolate reductase pseudogene and its relationship to the multiple forms of specific messenger RNA. J.Mol.Biol.,167,23-26.

Max,E.E., Battey,J., Ney,R., Kirsch,I.R. and Leder,P. (1982).

Duplication and deletion in the human immunoglobulin ε gene. Cell 29,691-699.

Maxson,R., Cohn,R., Kedes,L. and Mohun,T. (1983).

Expression and organisation of histone genes.

Ann.Rev.Genet.,17,239-277.

McGrath,J.P., Capon,D.J., Smith,D.H., Chen,E.Y., Seeburg,P.H., Goeddel,D.V. and Levinson,A.D. (1983).

Structure and organisation of the human Ki-ras proto-oncogene and a related processed pseudogene. Nature 304,501-506.

Mellor,A.L., Weiss,E.H., Ramachandran,K. and Flavell,R.A. (1983).

A potential donor gene for the bm1 gene conversion event in the C57BL mouse. Nature 306,792-794.

Messing,J. (1983).

New M13 vectors for cloning. Meth.Enz.,101,20-89.

Messing,J., Crea,R. and Seeburg,P.H. (1981).

A system for shotgun DNA sequencing. Nucl.Acids Res.,9,309-321.

Messing,J. and Vieira,J. (1982).

A new pair of M13 vectors for selecting either strand of double-digest restriction endonuclease fragments. Gene,19,269-276.

Mill,P.J. (1972).

Respiration in the invertebrates. MacMillan Press Ltd. London.

Miyata,T. and Yasunaga,T. (1981).

Rapidly evolving mouse α-globin-related pseudogene and its evolutionary history. Proc.Natl.Acad.Sci.USA.,78,450-453.

Monstein,H-J., Hammarstrom,K., Westin,G., Zabielski,J., Philison,L. and Pettersson,U. (1983).

Loci for human U1 RNA: structural and evolutionary implications. J.Mol.Biol.,167,245-258.

Moos,M. and Gallwitz,D. (1982).

Structure of a human β-actin-related pseudogene which lacks intervening sequences. Nucl.Acids Res.,10,7843-7849.

Moos,M. and Gallwitz,D. (1983).

Structure of two human β-actin-related processed genes one of which is located next to a simple repetitive sequence. The EMBO Journal 2,727-762.

Nagel,R.L., Bookchin,R.M., Johnson,J., Labie,D., Wajeman,H., Isaac-Sodeye,W.A., Honig,G.R., Sciliro,G., Crookston,J.H. amd Matsutomo,K. (1979).

Structural bases of the inhibitory effects of hemoglobin F and hemoglobin A_2 on the polymerisation of hemoglobin S. Proc.Natl.Acad.Sci.USA.,76,670-672.

Nishioka,Y., Leder,A. and Leder,P. (1980).

Unusual α-globin-like gene that has cleanly lost both globin intervening sequences. Proc.Natl.Acad.Sci.USA.,77,2806-2809.

Ng,R., Domdey,H., Larson,G., Rossi,J.J. and Abelson,J. (1985).

A test for intron function in the yeast actin gene. Nature 314,183-184.

Ohno,S. (1970).

Evolution by gene duplication. Springer-Verlag, Heidelberg.

Orkin,S.H. and Kazazian,Jr.H.H. (1984).

The mutation and polymorphism of the human β-globin gene and its surrounding DNA. Ann.Rev.Biochem.,18,131-171.

Papayannopuolou,T., Tatsis,B., Kurachi,S., Nakamoto,B. and Stramatoyannopoulos,G. (1984).

A haemolglobin switching activity modulates hereditary persistance of fetal haemoglobin. Nature 309,71-73.

Phillips,S.J., Hardies,S.C., Jahn,C.L., Edgell,M.H. and Hutchison III,C.A. (1984).

The complete nucleotide sequence of a β-globin-like structure, βH2, from the [Hbb]^d mouse BALB/c. J.Biol.Chem..,259,7947-7954.

Proudfoot,N.J. (1980).

Pseudogenes. Nature 286,840-41.

Proudfoot,N.J. and Maniatis,T. (1980).

The structure of a human α-globin pseudogene and its relationship to α-globin gene duplication. Cell 21,537-544.

Proudfoot,N.J., Gil,A. and Maniatis,T. (1982).

The structure of the human zeta-globin gene and a closely linked, nearly identical pseudogene. Cell 31,553-564.

Proudfoot,N.J., Rytherford,T.R. and Partington,G.A. (1984).
Transcriptional analysis of human zeta globin genes. The EMBO Journal 3,1533-1540.

Radding,C.M. (1978).
Genetic recombination: strand transfer and mismatch repair. Ann.Rev.Biochem.,47,847-880.

Radding,C.M. (1982).
Homologous pairing and strand exchange in genetic recombination. Ann.Rev.Biochem.,51,405-437.

Rogers,J.H. (1984)
The origin and evolution of retroposons. Int.Rev.Cytol.Suppl.,17, in press.

Rozek,E.E. and Davidson,N. (1983).
Drosophila has one myosin heavy-chain gene with three developmentally regulated transcripts. Cell 32,23-34.

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

Sanger,F., Coulson,A.R., Hong,C.F., Hill,D.F. and Petersen,G.B.
Nucleotide sequence of bacteriophage λ DNA. J.Mol.Biol.,162,729-773.

Sarich,V.M. and Wilson,A.C. (1967).
Immunological time scale for hominid evolution. Science 158,1200-1203.

Scarpulla,R.C. and Wu,R. (1983).

Nonallelic members of the cytochrome c multigene family of the rat may arise through different messenger RNAs. Cell 32,473-482.

Schimenti,J.C. and Duncan,C.H. (1984).

Ruminant globin gene structures suggest an evolutionary role for Alu-type repeats. Nucl.Acids Res. 12,1641-1655.

Schon,E.A., Cleary,M.L., HAynes,J.R. and Lingrel,J.B. (1981).

Structure and evolution of goat γ-, β^{C}- and β^{A}-globin genes: three developmentally regulated genes contain inserted elements. Cell 27,359-369.

Schon,E.A., Wernke,S.M. and Lingrel,J.B. (1982).

Gene conversion of the functional goat α-globin gene preserves only minimal flanking sequences. J.Biol.Chem.,257,6825-6835.

Scott,A.F., Heath,P., Trusko,S., Boyer,S.H., Prass,W., Goodman,M., Czelusniak,J., Chang,L.-Y.E. and Slightom,J.L. (1984).

The sequence of the gorilla fetal globin genes: evidence for multiple gene conversions in human evolution. Mol.Biol.Evol.,1,371-389.

Selsing,E., Miller,J., Wilson,R. and Storb,V. (1982).

Evolution of mouse immunoglobulin γ genes. Proc.Natl.Acad.Sci.USA.,79,4681-4685.

Shapiro,S.G., Schon,E.A., Townes,T.M. and Lingrel,J.B. (1983).

Sequence and linkage of the goat ε^{I} and ε^{II} β-globin genes. J.Mol.Biol.,169,31-52.

Sharp,P.A. (1979).

Summary: molecular biology of viral oncogenes. Cold Spring Harbor Symp.Quant.Biol.,44,1305-1309.

Sharp,P.A. (1983).

Conversion of RNA to DNA in mammals: Alu-like elements and pseudogenes. Nature 301,471-472.

Shen,S., Slightom,J.L. and Smithies,O. (1981)

A history of the human fetal globin gene duplication. Cell 26,191-203.

Shen,S. and Slightom,O. (1982).

Human globin $\Psi\beta2$ is not a globin-related sequence. Nucl.Acids Res.,10,7809-7818.

Sibley,C.G. and Ahlquist,J.E. (1984).

The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridisation. J.Mol.Evol.,20,2-15.

Singer,M.F. (1982).

SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell 28,433.

Slightom,J.L., Blechl,A.E. and Smithies,O. (1980)

Human foetal $^{G}\gamma$ and $^{A}\gamma$ globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21,627-638.

Smith,H.O. (1980)

Recovery of DNA from gels. Meth.Enz.,65,371-380.

Smith,H.O. and Birnstiel,M.L. (1976).

A simple method for DNA restriction site mapping. Nucl.Acids Res.,3, 2387-2398.

Smithies,O. and Powers,P.A. (1985).

Gene conversions and their relationship to homologous chromosome pairing. Phil.Trans.R.Soc.Lond.B., in press.

Sneath,P.H.A. (1966).

Relationships between chemical structure and biological activity in peptides J.Theo.Biol.,12,157-195.

Sneath,P.H.A. and Sokal,R.R. (1973).

Numerical Taxonomy. Freeman. San Francisco. USA.

Snyder,M., Hunkapillar,M., Yuen,D., Silvert,D., Fristrom,J. and Davidson,N. (1982).

Cuticle protein genes of Drosophila: structure, organisation and evolution of four clustered genes. Cell 29,1027-1040.

Spritz,R.A., Deriel,J.K., Forget,B.G. and Weissman,S.M. (1980).

Complete nucleotide sequence of the human δ-globin gene. Cell 21,639-646.

Stein,J.P., Catheral,J.F., Kristo,P., Means,A.R. and O'Malley,B.W. (1980).

Ovomucoid intervening sequences specify functional domains and generate protein polymorphism. Cell 21,681-687.

Stein,J.P., Munjaal,R.P., Lagace,L., Lai,E.C., O'Malley,B.W. and Means,A.R. (1983).

Tissue-specific expression of a chicken calmodulin pseudogene lacking intervening sequences. Proc.Natl.Acad.Sci.USA.,80,6485-6489.

Sukumaran,P.K., Nakatsuji,T., Gardiner,M.B., Reese,A.L., Gilman,J.G. and Huisman,T.H.J. (1983).

Gamma thalassemia resulting from the deletion of a γ-globin gene. NUcl.Acids.Res.,11,4635-4643.

Sutcliffe,J.G., Milner,R.J., Gottesfeld,J.M. and Lerner,R.A. (1983).

Identifier sequences are transcribed specifically in brain. Nature 308,237-241.

Syvanen,M. (1984).

The evolutionary implications of mobile genetic elements. Ann.Rev.Genet.,18,271-293.

Takashashi,N., Ueda,S., Obata,M., Nikaido,T., Nakai,S. and Honjo,T.(1982).

Structure of human Immunoglobulin gamma genes: implications for evolution of a multigene family. Cell 29,671-679.

Takmun,J.W., Schwarzbauer,J.E. and Haynes,R.O. (1984).

A single rat fibronectin gene generates tree different mRNAs by alternative splicing of a complex exon. Proc.Natl.Acad.Sci.USA.,81,5140-5144.

Tajima,F. and Nei,M. (1984).

Estimation of evolutionary distance between nucleotide sequences. Mol.Biol.Evol.,1,269-285.

Tonegawa,S. (1983).

Somatic generation of antibody diversity. Nature 302,575-581.

Townes,T.M, Shapiro,S.G., Wernke,S.M. and Lingrel,J.B. (1984).

Duplication of a four-gene set during the evolution of the goat β-globin locus produced genes now expressed differentially in development. J.Biol.Chem.,259,1896-1900.

Trent,R.J., Bowden,D.K., Old,J.M., Wainscoat,J.S., Clegg,J.B. and Weatherall,D.J. (1981).

A novel rearrangement of the human β-like globin gene cluster. Nucl.Acids.Res.,9,6723-6733.

Twigg,A.J. and Sherratt,D. (1980).

Trans-complementable copy-number mutants of plasmid ColE1. Nature, 238,216-218.

Ueda,S., Sumiko,N., Yasuyoshi,N., Hiroshi,H. and Tasuku,H. (1982).

Long terminal repeat-like elements flank a human immunoglobulin ε pseudogene that lacks introns. The EMBO Journal 1,1539-1544.

Ullrich,A., Gray,A., Goeddel,D.V. and Dull,T.J. (1982).

Nucleotide sequence of a portion of human chromsome 9 containing a leukocyte interferon gene cluster. J.Mol.Biol.,156,467-486.

Ullu,E. and Tschudi,C. (1984).

Alu sequences are processed 7SL RNA genes. Nature 312,171-172.

Ullu,E. and Weiner,A.M. (1984).

Human genes and pseudogenes for the 7SL RNA component of the signal recognition particle. The EMBO Journal 3,3303-3310.

Van Arsdell,S.W., Denison,R.A., Bernstein,L.B., Weiner,A.M., Manser,T. and Gesteland,R.F. (1981)

Direct repeats flank three small nuclear RNA pseudogenes in the human genome. Cell,26,11-17.

Van Ooyen,A., Van Den Berg,J., Mantei,N. and Weissman,C. (1979)

Comparison of total sequence of a cloned rabbit β-globin gene and its flanking regions with a homologous mouse sequence. Science,206,337-344.

Vanin,E.F., Goldberg,G.I., Tucker,P.W. and Smithies,O. (1980).

A mouse α-globin-related pseudogene lacking intervening sequences. Nature 286,222-226.

Vanin,E.F. (1984).

Processed pseudogenes: Characteristics and evolution. Biochim.Biophys.Acta.,782,231-241.

Walter,P. and Blobel,G. (1982).

Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. Nature 299,691-698.

Weiss,E., Godden,L., Zakut,R., Mellor,A., Fahrner,K., Kvist,S. and Flavell,R.A. (1983).

The DNA sequence of the H-2K^b gene: evidence for gene conversion as a mechanism for the generation of polymorphism in histocompatibity antigens. EMBO Journal,2,453-462.

Weller,P., Jeffreys,A.J., Wilson,V. and Blanchtot,A. (1984).

Organisation of the human myoglobin gene. EMBO Journal,3,439-446.

Westin,G., Zabielski,J., Hammarstrom,K., Monstein,H-J., Bark,C. and Petterson,U. (1984).

Clustered genes for human U2 RNA.

Proc.Natl.Acad.Sci.USA.,81,3811-3815.

White,T.C., Hardies,S.C., Hutchinson III,C.A and Edgell,M.H. (1984)

The diagonal-traverse homology search algorithm for locating similarities between two sequences. Nucl.Acids Res.,12,751-766.

Wieringa,B., Hofer,E. and Weissman,C. (1984).

A minimal intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron. Cell 37,915-925.

Wilde,C., Crowther,C.E. and Cowan,N.J. (1982)

Diverse mechanisms in the generation of human β-tubulin pseudogenes. Science 217,549-552.

Wills,C. (1978).

Rank-order selection is capable of maintaining all genetic polymorphisms. Genetics 89,403-417.

Wilson,A.C., Carlson,S.S. and White,T.J. (1977)

Biochemical evolution. Ann.Rev.Biochem.,46,573-639.

Wilson,G.N., Knoller,M., Szura,L.L. and Schmickel,R.D. (1984).

Individual and evolutionary variation of primate ribosomal DNA transcription initiation regions. Mol.Biol.Evol.,1,221-237.

Wozney,J., Hanahan,D., Tate,V., Boedtker,H. and Doty,P. (1981).

Structure of the pro α2(I) collagen gene. Nature 294,129-135.

Yang,R.C.-A., Lis,J. and Wu,R. (1979).

Elution of DNA from agarose gels after electrophoresis. Meth.Enz.,68,176-182.

Yunis,J.J. and Prakash,O. (1982).

The origin of man: a chromosomal pictorial legacy. Science 215,1525-1529.

J. Mol. Biol. (1984) **180**, 785–801

The Primate ψβ1 Gene

An Ancient β-Globin Pseudogene

STEPHEN HARRIS, PAUL A. BARRIE, MARK L. WEISS† AND ALEC J. JEFFREYS

Department of Genetics, University of Leicester
Leicester LE1 7RH, U.K.

(Received 21 June 1984)

The human β-globin gene cluster contains five functional genes plus a single pseudogene termed ψβ1. Hybridization and comparative sequence analysis show that this pseudogene is not the product of a recent gene duplication, but is ancient and has been maintained in all major primate groups ranging from prosimians to anthropoids, at the same position as in man, between γ- and δ-globin genes. In the lemur, a prosimian, the central exons of the ψβ1 and δ-globin genes have undergone an unequal exchange, which has resulted in a contraction of the β-globin gene cluster and the formation of a Lepore-type ψβ1-δ globin pseudogene. Comparisons of defects shared by prosimian, New World monkey and human ψβ1 sequences suggest that the ancestral primate gene was probably a pseudogene with an abnormal initiation codon but few if any additional defects, and that most contemporary pseudogene defects were accumulated relatively recently by slow neutral drift. We suggest that ψβ1 arose early in primate evolution by silencing of a pre-existing discrete functional gene, and show that ψβ1-related sequences are also present in other mammalian orders. In view of the antiquity of ψβ1-related sequences, we propose that this gene be renamed the η-globin gene.

1. Introduction

Pseudogenes are a common feature of many multigene families in higher eukaryotes. Two basic classes of pseudogenes have been described so far. Non-processed pseudogenes possess the DNA organization of related functional genes and have arisen either by duplication and silencing of a functional gene, or by duplication of pre-existing non-processed pseudogenes (Lacy & Maniatis, 1980; Proudfoot, 1980; Proudfoot & Maniatis, 1980; Cleary *et al.*, 1981; Snyder *et al.*, 1982; Little, 1982; Proudfoot *et al.*, 1982). They frequently contain multiple defects, and in such cases it is not known which one, if any, of the contemporary defects was responsible for the initial gene silencing. Non-processed pseudogenes are generally found within parent gene clusters, although an example of a

† Permanent address: Department of Anthropology, Wayne State University, Detroit, Mich. 48202, U.S.A.

0022–2836/84/360785–17 $03.00/0

dispersed non-processed pseudogene is known (Leder *et al.*, 1981). In contrast, processed pseudogenes apparently arise by insertion of reverse transcripts of spliced RNA into germ line DNA and differ from non-processed pseudogenes in generally being dispersed, lacking introns, possessing a DNA relic of a poly(A) tail and being flanked by direct repeats characteristic of transposable elements (see Sharp, 1983; Rogers, 1984).

Pseudogenes have no known function and are therefore likely to be useful models for neutral evolution. Early analysis of the divergence between the "coding regions" of non-processed globin pseudogenes and their functional relatives suggested that gene duplication preceded the silencing event, and that pseudogene sequences were therefore functional during their early history (Lacy & Maniatis, 1980; Proudfoot & Maniatis, 1980). However, more detailed analysis has shown that the standard errors associated with these estimates of duplication and silencing times are high (Li *et al.*, 1981). With the exception of the recently silenced δ-globin gene in Old World monkeys (Kimura & Takagi, 1983; Martin *et al.*, 1983), there is therefore no clear evidence yet for an early functional history of a contemporary pseudogene after gene duplication. All comparisons of pseudogenes with related functional genes suggest that after silencing, pseudogenes evolve rapidly, apparently at a rate greater than, or equal to, the currently accepted minimal neutral rate of 5×10^{-9} nucleotide substitutions/site per year derived from an analysis of synonymous codon substitutions in functional mammalian genes (Li *et al.*, 1981; Miyata & Hayashida, 1981; Miyata & Yasunaga, 1981; Hayashida & Miyata, 1983).

The human β-globin gene cluster contains, in addition to five functional globin genes, a single non-processed β-globin pseudogene, termed ψβ1, which lies between the foetal $^{A}\gamma$-globin gene and the minor adult δ-globin gene and is detectable by cross-hybridization with human β- and γ-globin DNA (Fritsch *et al.*, 1980; Shen & Smithies, 1982). Sequence analysis has shown that this pseudogene has the conventional three exon–two intron structure of globin genes and contains multiple defects (Jagadeeswaran *et al.*, 1983; Chang & Slightom, 1984). In this paper we show by phylogenetic analysis that the ψβ1 pseudogene in man is ancient, has been maintained probably as a pseudogene in most or all major primate groups, and seems to have arisen early in primate evolution by silencing of a pre-existing functional globin gene, which we term the η-globin gene.

2. Materials and Methods

(a) *Preparation of DNA*

DNA was prepared, using the procedure of Jeffreys (1979), from human blood and from liver taken from a yellow baboon (*Papio cynocephalus*), owl monkey (*Aotus trivirgatus*), squirrel monkey (*Saimiri sciureus*), red-mantled tamarin (*Saguinus illigeri*), brown lemur (*Lemur macaco (fulvus) mayottensis*), ruffed lemur (*Lemur variegatus*), lion (*Panthera leo*), dog (*Canis familiaris*), C57 mouse (*Mus musculus*), rabbit (*Oryctolagus cuniculus*), blackbuck (*Antilope cervicapra*) and flying fox (*Pteropus lastal*). Other DNAs were prepared from grey seal (*Halichoerus grypus*) muscle, from cow (*Bos taurus*) thymus, and from the whole carcass of a dwarf lemur (*Cheirogaleus major*). Western lowland gorilla (*Gorilla gorilla gorilla*) and orang-utan (*Pongo pygmaeus*) blood DNA samples were generously provided by Dr A. F. Scott (Johns Hopkins University School of Medicine, Baltimore, U.S.A.).

(b) *Genomic hybridization analysis*

Samples (8 μg) of DNA were restricted under the manufacturer's recommended conditions and electrophoresed through a 0·5% (w/v) agarose gel. DNA was denatured *in situ* and transferred to a Sartorius nitrocellulose filter (Southern, 1980). DNA probes were labelled *in vitro* with ^{32}P by the method of Weller *et al.* (1984) and hybridized to Southern blots in 1 × SSC (SSC is 0·15 M-NaCl, 15 mM-sodium citrate, pH 7·0) at 60°C in the presence of dextran sulphate (Jeffreys *et al.*, 1980), plus 50 μg sheared single-stranded human DNA/ml (sheared in 0·3 M-NaOH, 20 mM-EDTA at 100°C for 20 min) to suppress hybridization to repetitive sequences.

(c) *Isolation of hybridization probes*

The phage recombinant λHγG4 containing the human ψβ1 pseudogene and generously provided by Dr T. Maniatis was grown as described by Jeffreys *et al.* (1982) and phage DNA prepared. λHγG4 DNA was cleaved with *Eco*RI, cloned into pAT153 (Twigg & Sherratt, 1980) and recombinants containing the ψβ1 gene were isolated. Probes from the ψβ1 gene were purified from suitable restriction endonuclease digests by preparative gel electrophoresis onto DE81 paper (Dretzen *et al.*, 1981).

Probes containing rabbit adult β-globin complementary DNA sequences or human $^{G}\gamma$-globin cDNA were isolated as described by Barrie *et al.* (1981).

(d) *Isolation and sequence analysis of the β-globin gene cluster from the owl monkey*

An owl monkey genomic DNA library was constructed and screened for β-globin DNA sequences as described by Jeffreys *et al.* (1982). The owl monkey ψβ1 sequence was subcloned into pUC13 (Messing, 1983). Recombinants were sheared by sonication (Deininger, 1983), end-repaired using the Klenow fragment of DNA polymerase I and fragments 500 to 1000 bp† long recovered by agarose gel electrophoresis onto DE81 paper. Fragments were blunt-end ligated into the *Sma*I site of M13mp8 RF DNA and transfected into *Escherichia coli* JM103 (Messing & Vieira, 1982). White plaques were screened for ψβ1 sequences, and phage DNA from positive plaques was prepared as described by Weller *et al.* (1984). DNA sequences were determined by the dideoxynucleotide chain-termination method of Sanger *et al.* (1977) as modified by Biggin *et al.* (1983), using the 17-mer primer (Duckworth *et al.*, 1981) generously provided by Dr P. Meacock (Biocentre, Leicester) and [α-^{35}S]dATP. Sequencing data were assembled with the aid of a Digital PDP 11/44 computer using programs developed by Staden (1980).

3. Results

(a) *Isolation of gene-specific hybridization probes from the human ψβ1 globin pseudogene*

The human ψβ1 sequence isolated by Fritsch *et al.* (1980) was subcloned into pAT153 and two potential unique sequence DNA fragments were isolated. Probe 1 contained the 5′ flanking region of ψβ1, and probe 2 included most of intron 2 (Fig. 1(a)). Hybridization of these probes to *Eco*RI digests of human DNA under low stringency conditions (1 × SSC, 60°C) detected a single major 7 kb hybridizing fragment, as predicted from the map of the human β-globin gene cluster (Fig. 1). In addition, probe 2 also detected two or three additional faintly hybridizing components in human DNA digested with *Eco*RI (Fig. 1(b)) or *Bgl*II (not shown).

† Abbreviations used: bp, base-pairs; kb, 10^3 bases or base-pairs.

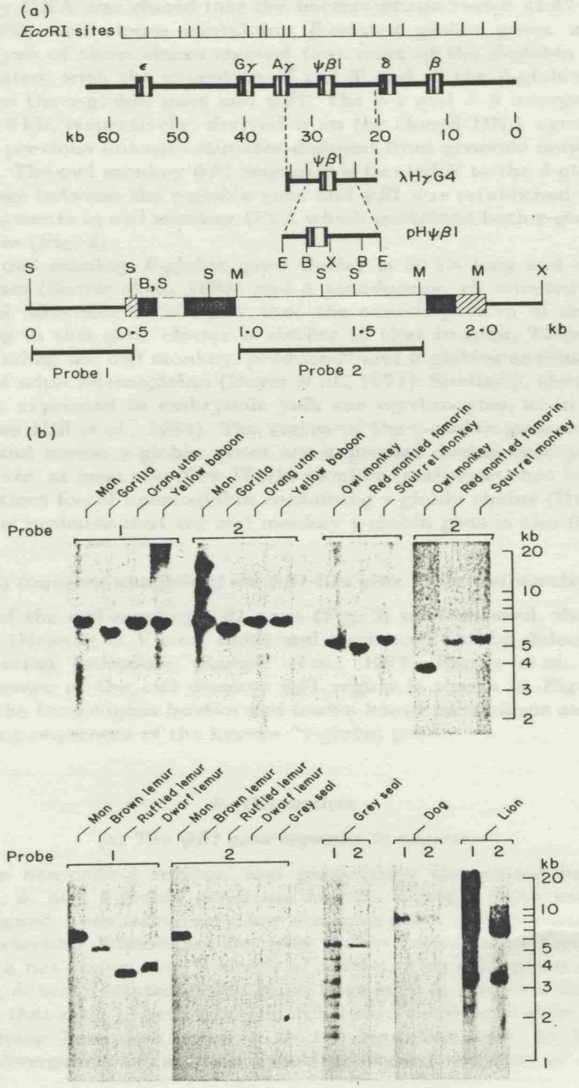


FIG. 1.

(b) *Detection of ψβ1-like DNA sequences in primate, seal and carnivore DNAs*

Both ψβ1 probes detected a single prominent ψβ1-related sequence in *Eco*RI digests of a range of great ape, Old World monkey and New World monkey DNAs (Fig. 1(b)). The orang-utan and yellow baboon ψβ1 fragments were indistinguishable in size from that of man, suggesting that the organization of this region of the β-globin gene cluster is similar in all three species. In *Eco*RI digests of two of the New World monkey DNAs (owl monkey, squirrel monkey), probe 1 and probe 2 detected different ψβ1 DNA fragments; in the owl monkey, this difference results from an *Eco*RI cleavage site at the 5′ end of ψβ1 intron 2 which generates two ψβ1 DNA fragments (see below and Fig. 2).

Human ψβ1 probe 1 also detected a single major hybridizing fragment in various lemur DNAs (Fig. 1(b)); in the brown lemur, this fragment corresponded to the 5′ region of a previously characterized hybrid δ-globin pseudogene (Jeffreys *et al.*, 1982). In contrast, probe 2 containing the second intron of ψβ1 failed to detect any homologous sequences in any of the lemur DNAs tested, indicating that this sequence may have been eliminated from the lemur genome.

This Southern blot analysis was extended to a range of non-primate mammalian DNAs. Probe 2 failed to detect clearly any consistent cross-hybridizing sequences in DNA from mouse, rabbit, bat, cow or blackbuck. In contrast, clear hybridization of probe 2 to one or more components was seen with pinniped and carnivore DNAs (grey seal, dog and lion) (Fig. 1(b)). In each case, at least one of these fragments also hybridized with human ψβ1 probe 1, which suggests that these species each contain at least one authentic ψβ1-related gene sequence.

(c) *Isolation of the β-globin gene cluster from the owl monkey*

Since sequence information is only available for mutually diverged hominoid and prosimian ψβ1-like sequences, we chose to characterize the β-globin gene cluster and ψβ1 gene from a New World monkey (owl monkey or night monkey, *Aotus trivirgatus*). Earlier work on the genomic mapping of owl monkey β-related globin genes demonstrated single ε-, γ-, δ- and β-globin genes, and established linkage between δ and β and, provisionally, between ε and γ (Barrie *et al.*, 1981).

FIG. 1. Detection of ψβ1-globin related sequences in primate, seal and carnivore DNAs. (a) Isolation of human ψβ1 DNA hybridization probes. The organization of the human β-globin gene cluster is shown together with the location of the ψβ1-containing clone λHγG4 isolated by Fritsch *et al.* (1980). Exons are shown by filled boxes and introns by open boxes. A 7 kb *Eco*RI fragment was isolated from λHγG4 DNA and cloned into pAT153 (Twigg & Sherratt, 1980) to give the subclone pHψβ1. A detailed restriction map of the ψβ1 globin pseudogene shows cleavage sites for restriction endonucleases *Bgl*II (B), *Eco*RI (E), *Mbo*II (M), *Sau*3A (S) and *Xba*I (X). DNA fragments containing the 5′ flanking region (probe 1) or intron 2 (probe 2) of the ψβ1 pseudogene were isolated by preparative gel electrophoresis of restriction endonuclease digests of pHψβ1 DNA. (b) Samples (8 μg) of primate, seal and carnivore DNAs were digested with *Eco*RI and electrophoresed through a 0·5% agarose gel. DNA fragments were transferred to nitrocellulose by the method of Southern (1980) and hybridized with probe 1 or probe 2 DNA labelled *in vitro* with ^{32}P. Hybridizations were carried out in 1 × SSC at 60°C in the presence of dextran sulphate plus 50 μg human competitor DNA/ml (see Materials and Methods). Autoradiographic exposures were for 2 days, except for the lower right panel (6 days).

Owl monkey DNA was cloned into the bacteriophage vector λL47·1 (Loenen & Brammar, 1980) and clones containing β-related globin genes were purified (Fig. 2). Analysis of these clones showed that most of the β-globin gene cluster had been isolated, with the exception of the 3′ end of the β-globin gene and a region between the γ-globin gene and ψβ1. The ε–γ and δ–β intergene distances (7·0 kb and 4·6 kb, respectively) derived from the cloned DNA agree reasonably well with our previous linkage estimates deduced from genomic mapping (~9 kb and ~5·6 kb). The owl monkey ψβ1 sequence is located 5′ to the δ-globin gene, as in man. Linkage between the γ-globin gene and ψβ1 was established by analysing restriction fragments in owl monkey DNA which contained both γ-globin and ψβ1 DNA sequences (Fig. 2).

The entire owl monkey β-globin gene cluster is 30 kb long and includes four functional genes (Barrie *et al.*, 1981) and a pseudogene, all oriented in the same transcriptional direction. It is likely that the overall pattern of developmental gene switching in this gene cluster is similar to that in man. Thus New World monkeys, including the owl monkey, produce δ- and β-globins as minor and major components of adult haemoglobin (Boyer *et al.*, 1971). Similarly, the ε-globin gene is presumably expressed in embryonic yolk sac erythrocytes, as in man, rabbit and mouse (see Hill *et al.*, 1984). The status of the γ-globin gene is less certain, since rabbit and mouse γ-globin genes are expressed during embryogenesis (see below). However, at least one New World monkey (marmoset) has been shown to produce a distinct foetal haemoglobin containing γ-globin chains (Huisman *et al.*, 1973), and it is probable that the owl monkey γ-globin gene is also foetal.

(d) *Sequence analysis of the ψβ1-like gene of the owl monkey*

Subclones of the owl monkey ψβ1 gene (Fig. 2) were sheared, shotgun cloned into M13mp8 (Messing & Vieira, 1982) and sequenced by the dideoxynucleotide chain termination procedure (Sanger *et al.*, 1977; Biggin *et al.*, 1983). The complete sequence of the owl monkey ψβ1 region is shown in Figure 3, and is aligned with the homologous human and brown lemur pseudogene sequences, and with the coding sequences of the human $^{A}\gamma$-globin gene.

4. Discussion

(a) *The ψβ1 gene sequence is ancient*

Most of the non-coding regions, and particularly the second intron, of the human ε-, γ-, δ- and β-globin genes are heavily diverged from each other and cannot be aligned, even using very low stringency dot matrix matching criteria capable of detecting homologies between globin intron sequences up to 40% diverged (data not shown, see White *et al.*, 1984). This divergence suggests that discrete ε-, γ-, δ- and β-related globin genes have been in existence for a long time, and therefore that rates of gene conversion between duplicated globin genes must, in general, have been low relative to the mutation rate to have allowed substantial divergence of at least the non-coding regions of these genes.

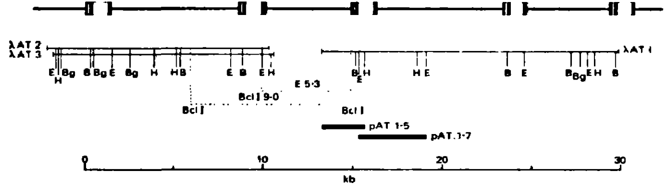


Fig. 2. Organization of the β-globin gene cluster isolated from owl monkey (*Aotus trivirgatus*) DNA. A library of *Sau*3A partials of owl monkey DNA cloned into the bacteriophage vector λL47·1 (Loenen & Brammar, 1980; see Materials and Methods) was screened by plaque hybridization with human ψβ1 probe 2 (intron 2, see Fig. 1(a)) and with rabbit β-globin cDNA isolated from the plasmid PβG1 (Maniatis *et al.*, 1976). Three hybridizing plaques (λAT.1 to 3) were detected, one of which hybridized to ψβ1 probe 2. λAT DNAs were purified and mapped with restriction endonucleases *Bam*HI (B), *Bgl*II (Bg), *Eco*RI (E) and *Hin*dIII (H). Genes were located by Southern blot hybridization of λAT DNAs with rabbit β-globin cDNA and human ψβ1 probes 1 and 2. ε-, γ-, δ- and β-globin genes were identified by comparison of the restriction maps of λAT.1 to 3 DNAs with maps of owl monkey β-related globin genes previously determined by Southern blot hybridization of owl monkey genomic DNA with human ε-, γ- and β-globin gene probes (Barrie *et al.*, 1981). The linkage between the γ-globin gene and ψβ1 was determined by hybridizing Southern blots of *Eco*RI and *Bcl*I digests of owl monkey genomic DNA with human γ-globin cDNA (Little *et al.*, 1978) and human ψβ1 probe 1 DNA (Fig. 1). In both cases, the single DNA fragment detected by the ψβ1 probe was also detected by γ-globin cDNA. Location of *Eco*RI and *Bcl*I sites in λAT.1 to 3 DNA established that the γ-globin and ψβ1 genes are separated by 4·7 kb of DNA, as shown. Other *Bcl*I sites in λAT.1 to 3 are omitted. The owl monkey ψβ1 pseudogene was further isolated by subcloning *Cla*I-*Hin*dIII (*Cla*I cleaving 198 bp from the *Bam*HI site within the right arm of λL47·1 DNA) and *Eco*RI fragments into pUC13 (Messing, 1983) to give the subclones pAT.1·5 and pAT.1·7, respectively. Both plasmids were sheared, shotgun cloned into M13mp8 (Messing & Vieira, 1982) and recombinant phage containing the ψβ1 pseudogene were identified by hybridization with a 1·8 kb *Bgl*II-*Xba*I fragment containing the human ψβ1 sequence isolated from pHψβ1 (Fig. 1(a)). Relevant M13 recombinants were sequenced by the dideoxynucleotide chain termination procedure (Sanger *et al.*, 1977; Biggin *et al.*, 1983). The sequence of the owl monkey ψβ1 pseudogene (Fig. 3) was fully determined on both DNA strands.

Interspecific similarities of β-globin gene non-coding regions, and particularly intron 2, provide a powerful indicator of gene orthologies (Jeffreys *et al.*, 1982) and have been used to demonstrate distinct ε-, γ-, δ- and β-related globin genes in non-primate mammals (Hardies *et al.*, 1984; Hardison, 1984).

Similarly, intron 2 of the human ψβ1 pseudogene cannot be aligned with the corresponding region of any of the functional genes in the human β-globin gene cluster; this region in ψβ1 is therefore at least 40% diverged from all functional human globin genes. In contrast, the exon regions show preferential homology with ε- and γ-globin genes (data not shown, see Jeffreys *et al.*, 1982; Goodman *et al.*, 1984). It thus appears that the ψβ1 sequence is not the product of a recent gene duplication in primates, but arose by duplication of a non-adult globin gene at least 140 million years ago. This is the minimal time required to achieve >40% divergence of ψβ1 intron sequences from corresponding non-adult globin gene sequences at the rate of evolution of 2×10^{-9} substitutions/site per year, with a strong bias towards transitions, as derived from primate ψβ1 sequence comparisons (see sections (d) and (e), below).

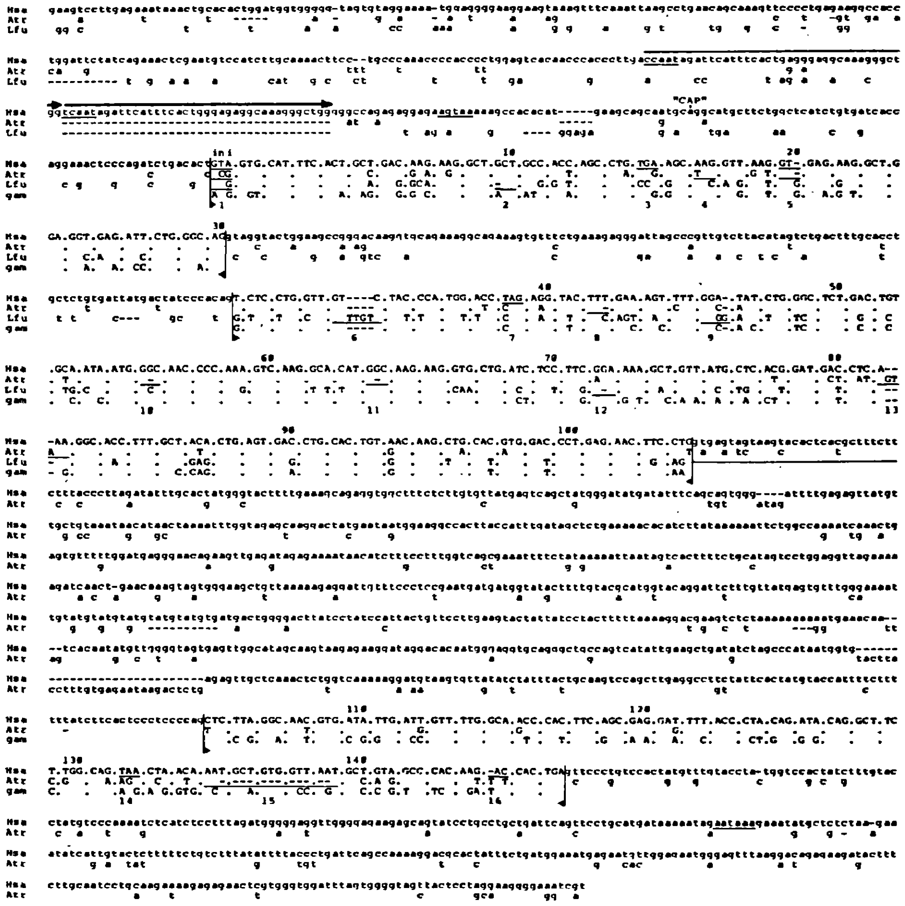


FIG. 3. DNA sequence comparison of human, New World monkey and prosimian $\psi\beta 1$ pseudogenes. The sequence of the human (*Homo sapiens*) $\psi\beta 1$ pseudogene (Hsa) taken from Chang & Slightom (1984) is aligned with owl monkey (*Aotus trivirgatus*) $\psi\beta 1$ (Atr) and with the 5′ half of the brown lemur (*Lemur macaco (fulvus) mayottensis*) hybrid $\psi\beta 1$-δ globin gene (Lfu) (Jeffreys *et al.*, 1982). Only differences from the human $\psi\beta 1$ sequence are shown. A dash indicates the absence of a nucleotide in a sequence. Homologues of coding sequences are shown in uppercase letters, and codon phasing was established by comparison with the human $^{A}\gamma$-globin gene coding region (gam) (Slightom *et al.*, 1980) and the human ε-globin gene (Baralle *et al.*, 1980; not shown). The non-coding regions of the human $^{A}\gamma$-globin gene could not be aligned with the $\psi\beta 1$ sequences, and are omitted. Probable homologues of the cap site, C-C-A-A-T and T-A-T-A promoter elements, and the A-A-T-A-A-A polyadenylation sequence are underlined. Codons within the $\psi\beta 1$ sequences which are defective (nonsense and frameshift mutations) are underlined and numbered for reference. In some instances, the position of a microdeletion/insertion is ambiguous, within a few nucleotides, and the indicated position is therefore placed arbitrarily within these limits (see for example defect 8). Sequence hyphens have been omitted for clarity.

This conclusion is supported by cross-hybridization of human $\psi\beta 1$ non-coding regions with other mammalian DNAs (Fig. 1). The presence of a $\psi\beta 1$-related sequence in all major primate groups, which had a basal radiation about 60 million years ago (Simons, 1969; Sarich & Cronin, 1977; Wilson *et al.*, 1977), and the probable existence of at least one intact $\psi\beta 1$-like sequence in seal and carnivores, suggest that the $\psi\beta 1$ gene existed along with distinct ε-, γ-, δ- and β-globin genes prior to the eutherian radiation 80 million years ago. Failure to detect $\psi\beta 1$-related sequences by hybridization to other mammalian DNAs might simply be due to excessive DNA divergence in these lineages, rather than to absence of $\psi\beta 1$ in most mammals, although none of the β-related globin genes or pseudogenes characterized in the rabbit and mouse β-globin gene clusters shows clear orthology to the primate $\psi\beta 1$ sequence (Goodman *et al.*, 1984).

(b) *The $\psi\beta 1$ sequence has undergone an unequal exchange with the δ-globin gene in lemurs*

The brown lemur β-globin gene cluster contains single ε-, γ- and β-related globin genes plus a pseudogene between the γ- and β-globin genes (Barrie *et al.*, 1981). Hybridization and sequence analysis of the pseudogene show it to be a Lepore-type of gene which contains the 5′ end of a $\psi\beta 1$-globin gene, previously identified as an ε- or γ-related sequence (Jeffreys *et al.*, 1982), fused to the 3′ end of a δ-globin gene (Jeffreys *et al.*, 1982; Jagadeeswaran *et al.*, 1983). Comparison of the $\psi\beta 1$–δ hybrid pseudogene with human $\psi\beta 1$ and δ-globin sequences shows that the unequal exchange which generated the hybrid gene occurred in exon 2 (Fig. 4). The precise position is partially obscured by subsequent divergence between lemur and human sequences, and by the homogenization of the 5′ regions of δ- and β-globin genes at some stage during simian evolution (Jeffreys *et al.*, 1982; Martin *et al.*, 1983). Nevertheless, the lemur pseudogene sequence does appear to switch from $\psi\beta 1$-like to δ-like at a defined position towards the 3′ end of exon 2, at codons 86 and 87 (Fig. 4). In further discussions, the brown lemur $\psi\beta 1$ sequence is taken to be the region extending 5′ from codon 86.

(c) *The ancestral primate $\psi\beta 1$-related sequence was probably a pseudogene*

The defects in the human and brown lemur $\psi\beta 1$-related pseudogenes have been described elsewhere (Jeffreys *et al.*, 1982; Jagadeeswaran *et al.*, 1983; Chang & Slightom, 1984), and are shown in Figure 3. The owl monkey $\psi\beta 1$ sequence is also a pseudogene, and contains eight definite defects including a GCG initiation codon, one nonsense mutation and six exon frameshifts (Fig. 3). Each of these defects alone would be sufficient to render the sequence a pseudogene.

Most of the $\psi\beta 1$ pseudogene defects are found only in one of the three species (man, owl monkey, lemur); in all such cases, the corresponding position in the other two species' pseudogenes closely resembles, or is identical to, functional ε- and γ-globin gene sequences. These species-specific defects are therefore likely to

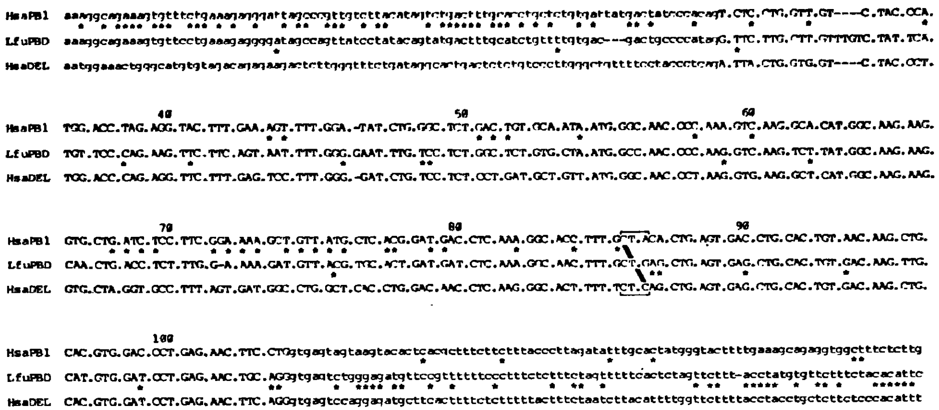

HsaPB1 aaaggcagaaagtgtttctgaaagagggattagcccgttgtcttacatagtctgactttgcacctgctctgtgattatgactatcccacagT.CTC.CTG.GTT.GT---C.TAC.CCA.
LfuPBD aaaggcagaaagtgttcctgaaagaggggatagccagttatcctatacagtatgactttgcatctgttttgtgac---gactgccccatagG.TTC.TTG.CTT.GTTTGTC.TAT.TCA.
HsaDEL aatggaaactgggcatgtgtagacagagaagactcttgggtttctgataggcactgactctctgtcccttgggctgttttcctaccctcagA.TTA.CTG.GTG.GT---C.TAC.CCT.
40
50
60
HsaPB1 TGG.ACC.TAG.AGG.TAC.TTT.GAA.AGT.TTT.GGA.-TAT.CTG.GGC.TCT.GAC.TGT.GCA.ATA.ATG.GGC.AAC.CCC.AAA.GTC.AAG.GCA.CAT.GGC.AAG.AAG.
LfuPBD TGT.TCC.CAG.AAG.TTC.TTC.AGT.AAT.TTT.GGG.GAAT.TTG.TCC.TCT.GGC.TCT.GTG.CTA.ATG.GCC.AAC.CCC.AAG.GTC.AAG.TCT.TAT.GGC.AAG.AAG.
HsaDEL TGG.ACC.CAG.AGG.TTC.TTT.GAG.TCC.TTT.GGG.-GAT.CTG.TCC.TCT.CCT.GAT.GCT.GTT.ATG.GGC.AAC.CCT.AAG.GTG.AAG.GCT.CAT.GGC.AAG.AAG.
70
80
90
HsaPB1 GTG.CTG.ATC.TCC.TTC.GGA.AAA.GCT.GTT.ATG.CTC.ACG.GAT.GAC.CTC.AAA.GGC.ACC.TTT.GCT.ACA.CTG.AGT.GAC.CTG.CAC.TGT.AAC.AAG.CTG.
LfuPBD CAA.CTG.ACC.TCT.TTG.G-A.AAA.GAT.GTT.ACG.TGC.ACT.GAT.GAT.CTC.AAA.GGC.AAC.TTT.GCT.GAG.CTG.AGT.GAG.CTG.CAC.TGT.GAC.AAG.TTG.
HsaDEL GTG.CTA.GGT.GCC.TTT.AGT.GAT.GGC.CTG.GCT.CAC.CTG.GAC.AAC.CTC.AAG.GGC.ACT.TTT.TCT.CAG.CTG.AGT.GAG.CTG.CAC.TGT.GAC.AAG.CTG.
100
HsaPB1 CAC.GTG.GAC.CCT.GAG.AAC.TTC.CTGgtgagtagtaagtacactcacgctttcttctttacccttagatatttgcactatgggtacttttgaaagcagaggtggctttctcttg
LfuPBD CAT.GTG.GAT.CCT.GAG.AAC.TGC.AGGgtgagtctgggagatgttccgtttttcccttttctctttctagttttcactctagttcttt-acctatgtgttctttctacacattc
HsaDEL CAC.GTG.GAT.CCT.GAG.AAC.TTC.AGGgtgagtccaggagatgcttcacttttctcttttactttctaatcttacatttggttcttttacctacctgctcttctcccacattt

FIG. 4. The $\psi\beta1$-δ crossover point in the hybrid pseudogene of the brown lemur. Exon 2 (upper case) and flanking intron sequences (lower case) are shown for the human $\psi\beta1$ pseudogene (HsaPB1), brown lemur hybrid $\psi\beta1$-δ pseudogene (LfuPBD; Jeffreys *et al.*, 1982), and the human δ-globin gene (HsaDEL; Spritz *et al.*, 1980). Asterisks indicate bases which are identical between lemur and one, but not both, of the human sequences. A probable $\psi\beta1$-δ exchange point in the lemur pseudogene sequence is indicated. The divergence between human and lemur $\psi\beta1$ sequences upstream of this exchange point (23·0 ± 1·5%, Fig. 3) is identical to the divergence between human and lemur δ-globin sequences downstream of this point (22·5 ± 1·3% over intron 2, exon 3 and 3′ flanking DNA, except for a diverged simple sequence element in intron 2; see Jeffreys *et al.* (1982)).

have arisen recently in evolution. Two defects (an A → G transition in the initiation codon and a frameshift at codon 20) are shared by man and the owl monkey, establishing that the $\psi\beta1$ sequence in the common ancestor of these two species was a pseudogene. The A → G defect in the initiation codon is also seen in the brown lemur $\psi\beta1$ sequence. From the known divergence of pseudogene sequences (Fig. 5, see below), it is unlikely that the prosimian–simian ancestor had an ATG initiation codon which gained the same A → G defect independently in the prosimian and simian lineages ($p = 0{\cdot}009$). Thus, the ancestral primate $\psi\beta1$ sequence was probably a pseudogene, with a defective initiation codon but few if any additional defects.

The phylogenetic distribution of $\psi\beta1$ defects is summarized in Figure 5, and indicates that most defects in modern primate $\psi\beta1$ genes have accumulated recently, and well after the initial silencing event, which may have been an alteration of the initiation codon from ATG to GTG. As far as is known, GTG is not used as an initiation codon in eukaryotes. The lack of major defects in the ancestral primate $\psi\beta1$ gene, other than the initiation codon, suggests that the $\psi\beta1$ gene was silenced only shortly before the prosimian/simian divergence, perhaps about 70 million years ago and well after the appearance of a discrete $\psi\beta1$-like sequence at least 140 million years ago. This implies that $\psi\beta1$ was a functional gene over much of its early history and may still be functional in other mammalian orders. This prediction has been confirmed by Goodman *et al.* (1984), who show that the apparently functional goat ε^{II}-globin gene (Shapiro *et al.*, 1983) is orthologous to primate $\psi\beta1$ sequences.

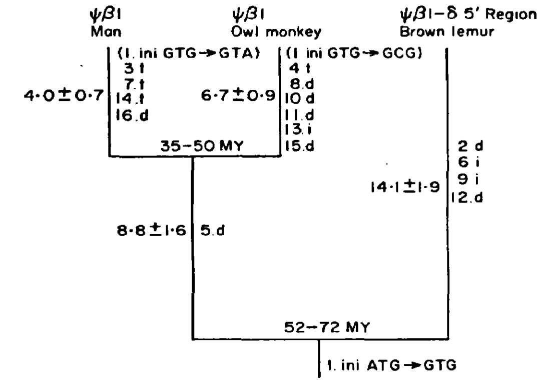


FIG. 5. Phylogeny of the primate $\psi\beta1$ gene. Branch lengths (substitutions per 100 bp ± S.E.) are given on the left of each branch, and were derived from a difference matrix corrected by iterative procedures for multiple substitutions with a high (71%) probability of transitions over transversions. The tree was approximately rooted using human ε- and γ-globin coding sequences as an external reference. Approximate divergence times are derived from palaeontological and protein data (Simons, 1969; Sarich & Cronin, 1977; Wilson *et al.*, 1977). Each numbered defect shown in Fig. 3 is assigned to a branch on the basis of maximum parsimony, and is identified by: ini, initiation codon defect; d, microdeletion; i, microinsertion; or t, premature termination codon. All defects in the human $\psi\beta1$ pseudogene are shared by gorilla and chimpanzee (Chang & Slightom, 1984), which indicates that the four defects on the human branch were established prior to the hominoid divergence approximately 7 million years (MY) ago. The possible initial defect (initiation codon *ATG* → *GTG*) is shared by all species and is placed at the root of the tree. Secondary alterations of the initiation codon in man and owl monkey are also indicated. There is a noticeable skew in the distribution of defects towards the upper branches of the simian tree, with only one defect on the branch leading to the human/owl monkey split. However, this skew is not statistically significant. We have used the patterns of microdeletion/insertion and base substitution in intron and flanking $\psi\beta1$ regions, coupled with computer-simulated divergence of predicted ancestral $\psi\beta1$ sequences according to the branch lengths in the Figure, to predict the number of deletions/insertions (D) and in-phase termination codons (T) which should have accumulated in the "coding sequence" of a silenced gene. The expected (observed) number of defects on each branch are: man/owl monkey ancester → man, $D = 1{\cdot}4(1)$, $T = 0{\cdot}5(3)$; man/owl monkey ancestor → owl monkey, $D = 2{\cdot}4(5)$, $T = 0{\cdot}8(1)$; root → man/owl monkey ancestor, $D = 3{\cdot}1(1)$, $T = 1{\cdot}0(0)$; root → lemur (exon 1 and 2 only), $D = 3{\cdot}7(4)$, $T = 1{\cdot}1(0)$. Thus, the skew of defects towards the top of the tree is not significant (pooling D and T defects gives $\chi^2[4] = 4{\cdot}7$ (Yates' correction), $p > 0{\cdot}1$).

It is not known whether any of the primate $\psi\beta1$ pseudogene sequences so far analysed are transcribed *in vivo*. The brown lemur and owl monkey genes both have recognizable remnants of the C-C-A-A-T and T-A-T-A boxes thought to be components of the RNA polymerase II promoter. Interestingly, the C-C-A-A-T box has been duplicated recently in the human $\psi\beta1$ gene, as part of a 38 bp tandem duplication in the 5′ flanking region of this gene (Fig. 3).

(d) *Modes of ψβ1 sequence evolution*

The human, owl monkey and brown lemur ψβ1 genes have diverged from what was probably a common ancestral pseudogene. Assuming that this pseudogene has been without function or effect in the β-globin gene cluster, the pattern of divergence of ψβ1 should reflect the rate and mode of neutral DNA change in this lineage.

The human and owl monkey ψβ1 sequences have diverged as predicted for pseudogenes, and do not show conservation of exon sequences; the divergence of each region (exons, introns, flanking regions) is not significantly different from the overall level of divergence of the whole pseudogene (10·7±0·7%). In addition, the levels of substitution at first, second and third codon positions do not significantly deviate from each other ($p > 0{\cdot}1$). Repeating this analysis with the brown lemur ψβ1 region again shows uniformity of divergence of all regions against man and owl monkey ψβ1 sequences. This uniformity supports our conclusion that the ancestral primate ψβ1 sequence was a pseudogene, and suggests that none of the primate ψβ1 sequences has been involved subsequently in any major recombinational exchange or gene conversion with other β-related globin genes, with the exception of the ψβ1–δ unequal exchange in lemurs.

Detailed analysis of the distribution of changed sites along the human and owl monkey pseudogenes using the single runs test (Siegel, 1956) shows that substitutions are in fact not entirely random in position ($p = 0{\cdot}002$). Instead, there is some substitution clustering, apparently due to a low level of "block" substitutions (a single event leading to the substitution of two or more consecutive bases) in addition to randomly scattered ($p > 0{\cdot}1$) single substitutions. An excess of 3 bp "block" substitutions is particularly noticeable in a comparison of the 3′ flanking region of human and owl monkey ψβ1 sequences (Fig. 3). Similar results were found in comparison of simian and lemur ψβ1 sequences.

Most single substitutions in the ψβ1 sequence are transitions (71%±3%, corrected for multiple substitutions). In addition, there is a relatively high level of microdeletions/insertions, which have been fixed at the approximate rate of one per 12 single substititions. Several of these have resulted from local duplications or deletions of short tandem repeats (see Efstratiadis *et al.*, 1980). Finally, as shown above, there is evidence for a low level of "block" substitutions in all pseudogene comparisons, although most substitutions (>95%) arise from single independent hits.

(e) *Primate ψβ1 sequences have evolved slowly*

The mean rate of evolution of primate ψβ1 sequences is 2×10^{-9} substitutions/site per year (Fig. 5). This probable neutral rate is substantially less than the supposedly constant rate of mammalian non-coding DNA evolution (5×10^{-9} substitutions/site per year) deduced primarily by comparing primate–rodent and primate–lagomorph genes (Hayashida & Miyata, 1983). A similar low rate has also been noted in human–seal myoglobin gene comparisons, and strongly suggests that the neutral substitution, and therefore mutation, rate is not constant in different mammalian orders, but may be influenced by generation time or by lineage-specific changes in the fidelity of DNA replication (Weller *et al.*, 1984).

There are also indications that the neutral rate is not only low but also variable within the primates. For rate constancy in this mammalian order, the age of the Old World–New World monkey divergence would have to be only one-third that of the prosimian–simian split (Fig. 5). In contrast, both palaeontological and protein data suggest that the interval between the two divergences was relatively short (Fig. 5). It is therefore possible that the neutral rate was relatively high early in primate evolution and subsequently decreased, particularly in the catarrhine lineage (see also Goodman *et al.*, 1984). This deceleration is also supported by the low neutral rate ($1{\cdot}4 \times 10^{-9}$ substitutions/site per year) derived from comparisons of great ape and human ψβ1 sequences (Chang & Slightom, 1984). Further pseudogene sequencing could provide a powerful test for localized rate fluctuations which would generate trees with significantly asymmetric branch lengths.

(f) *ψβ1 and the evolution of the primate β-globin gene cluster*

Identification of primate ψβ1 sequences consolidates our understanding of the evolution of the β-globin gene cluster in man, gorilla, chimpanzee, baboon, owl monkey and lemur (see Barrie *et al.*, 1981; Barrie, 1982; Jeffreys *et al.*, 1982). Figure 6 shows that the organization of the entire human β-globin gene cluster was established over 20 million years ago, prior to the emergence of the hominoids. In addition, the organization of the 3′ end of the cluster, from the γ- to the β-globin gene, is very similar in Old World and New World monkeys. The increased length of the ε–γ intergene region from 7·0 kb in the owl monkey to 13·3 kb in man, great apes and baboon might have been due to insertion of a 6·4 kb *Kpn* repetitive element known to be present in this region in man (Forget *et al.*, 1981). Clearly, major rearrangements (duplications, large deletions, acquisition and loss of transposable elements) are seldom fixed in this region of the genome. It is not known whether these events are intrinsically rare in higher primates, or whether most are deleterious to the function of the β-globin gene cluster and are eliminated by selection (see also Barrie *et al.*, 1981).

Only two changes in gene number have been identified so far in primate evolution. One of these is the γ-globin gene duplication which probably occurred early in the evolution of catarrhines (Fig. 6) and which has been discussed elsewhere (Barrie *et al.*, 1981; Shen *et al.*, 1981). The second event was a contraction of the β-globin gene cluster by unequal exchange between a ψβ1 sequence and the neighbouring δ-globin gene in an ancestor of the brown lemur. A hybrid ψβ1–δ pseudogene also exists in the ruffed lemur (Jeffreys *et al.*, 1982) and probably also in the dwarf lemur (Fig. 1), and suggests that the unequal exchange occurred early in the evolution of the lemurs. The exchange was by homologous recombination between the conserved central exons which specify the haem-binding sector of globin, and has probably resulted in the complete elimination

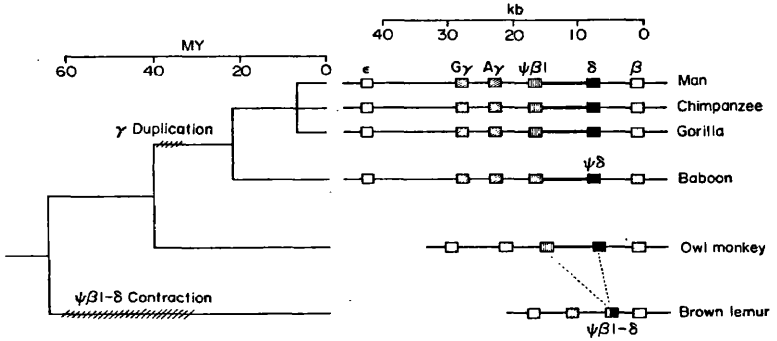


FIG. 6. Evolution of the β-globin gene cluster in primates. The organizations of β-globin gene clusters are taken from: man, Fritsch *et al.* (1980); gorilla and yellow baboon, Barrie *et al.* (1981); chimpanzee, Barrie (1982); owl monkey, present paper, and the brown lemur, Barrie *et al.* (1981), Barrie (1982) and Jeffreys *et al.* (1982). All genes are transcribed from left to right. The ψβ1 sequence in Old World species was located by genomic mapping using human ψβ1 probes (data not shown, see Fig. 1(a) and Chang & Slightom, 1984). The δ-globin gene is a pseudogene in Old World monkeys (Kimura & Takagi, 1983; Martin *et al.*, 1983) and has undergone homogenization with the β-globin gene over the 5′ region of the gene, possibly in an ancestor of Old World and New World monkeys (not shown, see Efstratiadis *et al.*, 1980; Jeffreys *et al.*, 1982; Martin *et al.*, 1983). Similarly, the human $^{G}\gamma$- and $^{A}\gamma$-globin genes have undergone multiple localized gene conversions in recent evolution (Slightom *et al.*, 1980; Shen *et al.*, 1981; Scott *et al.*, 1984). The remainder of the 5 kb γ-globin duplication units in man show 14% sequence divergence (Shen *et al.*, 1981), compared with the 10·7% divergence between human and owl monkey ψβ1 pseudogenes. This suggests either that the γ-globin gene duplication was fixed very shortly after the divergence of Old World and New World monkeys, as shown, or that the γ-globin gene duplication arose prior to this divergence and was subsequently eliminated by unequal exchange in New World monkeys. The divergence times shown are taken from palaeontological and protein data (see Fig. 5) and are approximate. MY, million years.

from lemurs of DNA sequences found between ψβ1 and δ-globin genes in higher primates (Fig. 6). The absence of ψβ1 intron 2 sequences from all lemurs tested (Fig. 1) provides direct evidence for this deletion.

There are suggestions that the region between the ψβ1 pseudogene and the δ-globin gene in man is involved in the switch from γ-globin to β-globin production late in gestation (see Collins & Weissman, 1984). In the brown lemur, loss of this region has not inhibited the expression of the downstream β-like globin gene: sequence analysis of this gene shows that it specifies adult lemur β-globin (S. Harris, P. A. Barrie & A. J. Jeffreys, unpublished results). However, neonate lemur blood contains only adult haemoglobin (Coppenhaver *et al.*, 1983), and it is not yet known whether the brown lemur γ-related globin gene is functional, nor whether it is expressed during foetal development. In non-primate mammals (rabbit, mouse), γ-related globin genes are expressed instead during early embryogenesis (Czelusniak *et al.*, 1982; Hill *et al.*, 1984; Hardison, 1984), and it is possible that the lemur γ-globin gene is also embryonic.

The phylogeny in Figure 6 strongly suggests that the organization of the ancestral primate β-globin gene cluster was ε-γ-ψβ1-δ-β and that this organization has been maintained in at least one present-day New World monkey.

The existence of discrete ε-, γ-, ψβ1-, δ- and β-related gene sequences in other mammalian orders (see above) indicates that this organization of the β-globin gene cluster may have predated the eutherian radiation 80 million years ago.

(g) *The possible existence of additional ψβ1-related genes*

Two to three additional sequences have been detected by the human ψβ1 intron 2 probe in DNA from man and gorilla (Fig. 1). Similarly, multiple hybridizing fragments containing ψβ1-like sequences have been found in seal and carnivore DNAs. In man, these sequences are apparently dispersed, since none of the additional restriction fragments containing ψβ1-related sequences corresponds to DNA fragments from the β-globin gene cluster. It seems possible that the ψβ1 gene has duplicated on more than one occasion and we are now isolating and sequencing these human ψβ1-related sequences in order to analyse the origins of these putative pseudogene derivatives.

(h) *Concluding remarks*

Analysis of the ψβ1 pseudogene in primates shows that, while it may be without present function, it has been a stable component of the β-globin gene cluster during the evolution of an entire mammalian order, and has probably existed as a pseudogene since the initial primate radiation. Furthermore, the persistence of this pseudogene has influenced the evolution of a gene family, as illustrated by the major ψβ1–δ contraction in lemurs. It is not known whether this contraction silenced the early prosimian δ-globin gene, or whether instead the δ-globin locus was also a pseudogene at the moment of this exchange (see Martin *et al.*, 1983; Hardison & Margot, 1984). In addition, it seems likely that ψβ1-like sequences were functional genes early in mammalian evolution, before the eutherian radiation, and may still be functional in mammalian orders other than primates. In view of the long and discrete history of ψβ1-like sequences, we propose that this gene be renamed η (following on from currently named α–ζ globin genes of man), with contemporary orthologues being the ψη pseudogene in simians, the hybrid ψη–δ pseudogene in lemurs, the η-globin gene in the goat and η-related genes of unknown functional status in seal and carnivores.

Phylogenetic analysis of other pseudogenes will be required to show whether the antiquity of the ψη-globin gene is unusual, or whether most pseudogenes have had long histories and therefore give important information about the early evolution of gene families.

We are grateful to Victoria Wilson for excellent technical assistance, to Robert Semeonoff and John Brookfield for much helpful advice on statistical analysis, and to Zoanne Nugent (Oxford) and Sue Jeffreys for help with dot matrix analyses. Our thanks also to Christine Hawkey (London Zoo) for generous donations of primate tissue samples. A.J.J. is a Lister Institute Research Fellow, and this work was supported by a grant to A.J.J. from the Medical Research Council and by grants to M.W. from the National Science Foundation (BNS 82-3077) and the Wenner Gren Foundation.

REFERENCES

Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980). *Cell,* **21**, 621–626.
Barrie, P. A. (1982). Ph.D. thesis, Leicester University.
Barrie, P. A., Jeffreys, A. J. & Scott, A. F. (1981). *J. Mol. Biol.* **149**, 319–336.
Biggin, M. D., Gibson, T. J. & Hong, H. F. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 3963–3965.
Boyer, S. H., Crosby, E. F., Noyes, A. N., Fuller, G. F., Leslie, S. E., Donaldson, L. J., Vrablik, G. R., Schaefer, E. W. & Thurmon, T. F. (1971). *Biochem. Genet.* **5**, 405–448.
Chang, L.-Y. E. & Slightom, J. L. (1984). *J. Mol. Biol.* **180**, 767–784.
Cleary, M. L., Schon, E. A. & Lingrel, J. B. (1981). *Cell,* **26**, 181–190.
Collins, F. S. & Weissman, S. M. (1984). *Progr. Nucl. Acids Res. Mol. Biol.* In the press.
Coppenhaver, D. H., Dixon, J. D. & Duffy, L. K. (1983). *Hemoglobin,* **7**, 1–14.
Czelusniak, J., Goodman, M., Hewett-Emmett, D., Weiss, M. L., Venta, P. J. & Tashian, R. E. (1982). *Nature (London),* **298**, 297–300.
Deininger, P. (1983). *Anal. Biochem.* **129**, 216–223.
Dretzen, G., Bellard, M., Sassone-Corri, P. & Chambon, P. (1981). *Anal. Biochem.* **112**, 295–298.
Duckworth, M. L., Gait, M. J., Goelet, P., Hong, G. F., Singh, M. & Titmas, R. C. (1981). *Nucl. Acids Res.* **9**, 1691–1706.
Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980). *Cell,* **21**, 653–668.
Forget, B. G., Tuan, D., Biro, P. A., Jagadeeswaran, P. & Weissman, S. M. (1981). *Trans. Assoc. Amer. Physic.* **94**, 204–210.
Fritsch, E. F., Lawn, R. M. & Maniatis, T. (1980). *Cell,* **19**, 959–972.
Goodman, M., Koop, B. F., Czelusniak, J., Weiss, M. & Slightom, J. L. (1984). *J. Mol. Biol.* **180**, 803–823.
Hardies, S. C., Edgell, M. H. & Hutchison, C. A. (1984). *J. Biol. Chem.* **259**, 3748–3756.
Hardison, R. C. (1984). *Mol. Biol. Evol.* **1**, 390–410.
Hardison, R. C. & Margot, J. B. (1984). *Mol. Biol. Evol.* **1**, 302–316.
Hayashida, H. & Miyata, T. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 2671–2675.
Hill, A., Hardies, S. C., Phillips, S. J., Davis, M. G., Hutchison, C. A. & Edgell, M. H. (1984). *J. Biol. Chem.* **259**, 3739–3747.
Huisman, T. H. J., Schroeder, W. A., Keeling, M. E., Gengozian, N., Miller, A., Brodie, A. R., Shelton, J. R., Shelton, J. B. & Apell, G. (1973). *Biochem. Genet.* **10**, 309–318.
Jagadeeswaran, P., Pan, J., Forget, B. G. & Weissman, S. M. (1983). *Cold Spring Harbor Symp. Quant. Biol.* **47**, 1079–1086.
Jeffreys, A. J. (1979). *Cell,* **18**, 1–10.
Jeffreys, A. J., Wilson, V., Wood, D., Simons, J. P., Kay, R. M. & Williams, J. G. (1980). *Cell,* **21**, 555–564.
Jeffreys, A. J., Barrie, P. A., Harris, S., Fawcett, D. H., Nugent, Z. J. & Boyd, A. C. (1982). *J. Mol. Biol.* **156**, 487–503.
Kimura, A. & Takagi, Y. (1983). *Nucl. Acids Res.* **11**, 2541–2550.
Lacy, E. & Maniatis, T. (1980). *Cell,* **21**, 545–553.
Leder, A., Swan, D., Ruddle, F., D'Eustachio, P. & Leder, P. (1981). *Nature (London),* **293**, 196–200.
Li, W. H., Gojobori, T. & Nei, M. (1981). *Nature (London),* **292**, 237–239.
Little, P. F. R. (1982). *Cell,* **28**, 683–684.
Little, P. F. R., Curtiss, P., Coutelle, C., Van Den Berg, J., Dalgleish, R., Malcolm, S., Courtney, M., Westaway, D. & Williamson, R. (1978). *Nature (London),* **273**, 640–643.
Loenen, W. A. M. & Brammar, W. J. (1980). *Gene,* **20**, 249–259.
Maniatis, T., Kee, S. G., Efstratiadis, A. & Kafatos, F. C. (1976). *Cell,* **8**, 163–182.
Martin, S. L., Vincent, K. A. & Wilson, A. C. (1983). *J. Mol. Biol.* **164**, 513–528.
Messing, J. (1983). *Methods Enzymol.* **101**, 20–89.

Messing, J. & Vieira, J. (1982). *Gene,* **19**, 269–276.
Miyata, T. & Hayashida, H. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 5739–5743.
Miyata, T. & Yasunaga, T. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 450–453.
Proudfoot, N. J. (1980). *Nature (London),* **286**, 840–841.
Proudfoot, N. J. & Maniatis, T. (1980). *Cell,* **21**, 537–544.
Proudfoot, N. J., Gil, A. & Maniatis, T. (1982). *Cell,* **31**, 553–563.
Rogers, J. H. (1984). *Int. Rev. Cytol. suppl.* **17**, in the press.
Sanger, F., Nicklen, S. & Coulson, A. R. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 5463–5467.
Sarich, V. M. & Cronin, J. E. (1977). *Nature (London),* **269**, 354.
Scott, A. F., Heath, P., Trusko, S., Boyer, S. H., Prass, W., Goodman, M., Czelusniak, J., Chang, L.-Y. E. & Slightom, J. L. (1984). *Mol. Biol. Evol.* **1**, 371–389.
Shapiro, S. G., Schon, E. A., Townes, T. M. & Lingrel, J. B. (1983). *J. Mol. Biol.* **169**, 31–52.
Sharp, P. A. (1983). *Nature (London),* **301**, 471–472.
Shen, S. & Smithies, O. (1982). *Nucl. Acids Res.* **10**, 7809–7818.
Shen, S., Slightom, J. L. & Smithies, O. (1981). *Cell,* **26**, 191–203.
Siegel, S. (1956). *Nonparametric Statistics for the Behavioural Sciences,* McGraw-Hill Kogakusha, Tokyo.
Simons, E. L. (1969). *Ann. N.Y. Acad. Sci.* **167(1)**, 319–331.
Slightom, J. L., Blechl, A. E. & Smithies, O. (1980). *Cell,* **21**, 627–638.
Snyder, M., Hunkapiller, D., Yuen, D. S., Fristrom, J. & Davidson, N. (1982). *Cell,* **29**, 1027–1040.
Southern, E. M. (1980). *Methods Enzymol.* **68**, 152–176.
Spritz, R. A., DeRiel, J. K., Forget, B. G. & Weissman, S. M. (1980). *Cell,* **21**, 639–646.
Staden, R. (1980). *Nucl. Acids Res.* **8**, 3673–3694.
Twigg, A. J. & Sherratt, D. (1980). *Nature (London),* **283**, 216–218.
Weller, P., Jeffreys, A. J., Wilson, V. & Blanchetot, A. (1984). *EMBO J.* **3**, 439–446.
White, C. T., Hardies, S. C., Hutchison, C. A. & Edgell, M. H. (1984). *Nucl. Acids Res.* **12**, 751–766.
Wilson, A. C., Carlson, S. S. & White, T. J. (1977). *Annu. Rev. Biochem.* **46**, 573–639.

Edited by P. Chambon

J. Mol. Biol. (1982) **156**, 487–503

Isolation and Sequence Analysis of a Hybrid δ-globin Pseudogene from the Brown Lemur

A. J. Jeffreys, P. A. Barrie, S. Harris
D. H. Fawcett, Z. J. Nugent

Genetics Department, University of Leicester
University Road, Leicester, LE1 7RH, England

AND

A. C. Boyd

Biochemistry Department, University of Leicester
University Road, Leicester, LE1 7RH, England

(Received 14 December 1981)

The β-globin gene cluster of the brown lemur, a prosimian, is very short and contains a single ϵ-, γ- and β-globin gene, with an additional β-related gene sequence between the γ- and β-globin genes. Brown lemur DNA was cloned into the bacteriophage vector λL47.1 and a recombinant was isolated which contained an 11×10^3 base insert including the β-globin gene and the additional putative β-globin pseudogene. The nucleotide sequence of this β-related gene was completely determined. A complete gene sequence was found, containing four frameshift mutations sufficient to establish its pseudogene status. The gene was interrupted by two intervening sequences with sizes and locations typical of mammalian β-related globin genes. The pseudogene sequence was compared in detail with human ϵ-, γ-, δ- and β-globin genes. The beginning of the pseudogene, from the 5′ flanking region to the second exon, was homologous to the corresponding regions of the human ϵ- and γ-globin genes. In contrast, the second intron, third exon and 3′ flanking region showed a remarkably close homology to the δ-globin, but not β-globin, gene of man. This suggests that the δ-globin gene is not the product of a recent gene duplication, but instead is present in most or all primates. This gene has been silenced on at least two separate occasions in primate evolution (in lemurs and in old world monkeys). In addition, the 5′ end of the lemur $\psi\delta$ gene appears to have exchanged sequences with an ϵ- or γ-globin gene, and an analogous exchange with the β-globin gene seems to have occurred recently in the human δ-globin gene. The evolution and function of the δ-globin gene are discussed.

1. Introduction

Human globins are specified by two unlinked clusters of globin genes (see Efstratiadis *et al.*, 1980). The β-globin gene cluster contains five active globin genes arranged in the order 5′-$\epsilon{}^{G}\gamma{}^{A}\gamma\delta\beta$-3′. The ϵ-globin gene is expressed during early embryogenesis, the foetal ${}^{G}\gamma$- and ${}^{A}\gamma$-globin genes during later foetal development,

0022–2836/82/110487–17 $03.00/0

and the minor (δ) and major (β) adult genes from late gestation onwards. The β-globin gene cluster is about 60×10^3 bases long and contains two additional pseudogene sequences (Fritsch *et al.*, 1980). Only 5% of the DNA in the cluster specifies globin messenger RNAs; the remaining DNA, consisting of intergenic regions plus intervening sequences, is a complex mixture of single copy and repetitive DNA sequences (Coggins *et al.*, 1980; Fritsch *et al.*, 1980). The role, if any, of the intergenic regions is not known, though there is suggestive evidence for a control sequence between the $^A\gamma$- and δ-globin genes that regulates the $\gamma \rightarrow \beta$ switch at birth (Fritsch *et al.*, 1979; Bernards & Flavell, 1980).

The organization of the δ- and β-globin genes has been compared in man, great apes and old world monkeys by Southern blot analysis of total genomic DNA (Martin *et al.*, 1980; Zimmer *et al.*, 1980; Jeffreys & Barrie, 1981). All species examined contained both δ- and β-globin genes. However, old world monkeys do not produce δ-globin (Boyer *et al.*, 1969,1971), and it seems that this gene has recently been silenced in these primates (Martin *et al.*, 1980). Barrie *et al.* (1981) have extended this comparative analysis to include the entire β-globin gene cluster. The organization of this cluster was indistinguishable in man, a great ape and an old world monkey, and the strong evolutionary conservation of both gene organization and intergenic DNA sequences suggested that the bulk of sequences in the cluster have been under substantial selective constraint.

In contrast, a new world monkey contains a single γ-globin gene, which suggests that the human γ-globin gene duplication arose about 20 to 40 million years ago (Barrie *et al.*, 1981), and that the close homology between the $^G\gamma$- and $^A\gamma$-globin genes in man has been maintained by inter-locus recombination or gene conversion (Jeffreys, 1979; Slightom *et al.*, 1980). The most compact β-globin gene cluster has been found in the brown lemur, a prosimian (Barrie *et al.*, 1981). The cluster is only 20 kb† long and contains single ϵ-, γ- and β-related genes. The γ- and β-globin genes are separated by a curious gene consisting of a segment closely homologous (by hybridization) to the 3′ end of the human β-globin gene, preceded by sequences only detectable by hybridization with the 5′ end of the human ϵ-globin gene.

We now describe the cloning and complete sequence analysis of this hybrid gene segment in the brown lemur, and show that this region is a pseudogene consisting of the 5′ region of an ϵ- or γ-like gene fused to a sequence that shows remarkable sequence homology to the 3′ end of the human δ-globin gene.

2. Materials and Methods

(a) *Materials*

The preparation of DNA from a brown lemur *(Lemur macaco (fulvus) mayottensis)* is described by Barrie *et al.* (1981). Restriction endonuclease *Sau*3A was prepared by the method of Sussenbach *et al.* (1976). All other restriction endonucleases were purchased from Bethesda Research Labs. T4 DNA ligase, avian myeloblastosis virus reverse transcriptase and rabbit globin mRNA were generously provided by Dr R. Wilson (Leicester), Dr J. W. Beard (Life Science Inc., U.S.A.) and Professor C. Weissmann (Zürich), respectively. Polynucleotide kinase was purchased from P.L. Biochemicals Inc., [α-^{32}P]dCTP and [γ-^{32}P]ATP (both 2000 to 3000 Ci/mmol) were obtained from Amersham.

† Abbreviations used: kb, 10^3 base-pairs; cDNA, complementary DNA.

(b) *Preparation and screening of a brown lemur-λ bacteriophage library*

The *Bam*HI replacement vector λL47.1 (Loenen & Brammar, 1980) was grown on *Escherichia coli* C600 by the method of Blattner *et al.* (1977), and phage DNA isolated as described by Loenen & Brammar (1980). λL47.1 DNA was restricted with endonuclease *Bam*HI, electrophoresed in a 0·4% (w/v) agarose gel, and the left and right λ arms recovered by electroelution onto a dialysis membrane (Yang *et al.*, 1979). Agarose impurities were removed by extraction with phenol and precipitation with ethanol, and the arms annealed at 0·5 mg DNA/ml in 10 mM-MgCl$_2$, 100 mM-Tris · HCl (pH 7·5) for 1 h at 42°C.

Brown lemur DNA was cleaved partially with endonuclease *Sau*3A and electrophoresed through a preparative 0·4% agarose gel. Partial digest fragments 11 to 20 kb long were excised and recovered by electroelution (see above). Equimolar amounts of lemur partials and annealed λL47.1 arms were ligated at 70 μg DNA/ml in 66 mM-Tris · HCl (pH 7·5), 6·6 mM-MgCl$_2$, 10 mM-dithiothreitol, 0·4 mM-ATP, 250 units T4 ligase/ml at 4°C overnight.

In vitro packaging extracts from *E. coli* BHB2688 and BHB2690 (Hohn, 1979) were prepared and used as described by Enquist & Sternberg (1979). A total of 2 μg ligated lemur-λ47.1 DNA was packaged in a total volume of 200 μl, diluted on ice with 1 ml phage buffer (6 mM-Tris · HCl (pH 7·2), 10 mM-MgSO$_4$, 0·005% (w/v) gelatin) and shaken with 2 drops of CHCl$_3$. Portions (250 μl) of packaged DNA were mixed with 250 μl of an overnight culture of *E. coli* WL87 (803 *supE*, *supF*, *hsdR*$_k^-$, *hsdM*$_k^+$, *tonA*, *trpR*$^-$, *metB*) grown in Luria broth supplemented with 10 mM-MgCl$_2$, 0·2% (w/v) maltose. After 15 min absorption at room temperature, bacteria were plated in soft agar supplemented with 10 mM-MgCl$_2$, 0·2% maltose on Luria agar in 9 cm Petri dishes. Plates were incubated overnight at 37°C. Approximately 3×10^5 recombinant plaques were obtained per plate (per 0·4 μg ligated DNA).

Plaques were lifted onto 8·8 cm diameter Sartorius nitrocellulose filters (0·45 μm pore size) by the method of Benton & Davis (1977) and hybridized with ^{32}P-labelled rabbit adult β-globin complementary DNA (in a 1·5 kb *Hha*I DNA fragment isolated from the recombinant plasmid PβG1; Maniatis *et al.*, 1976). Labelling of DNA and filter hybridizations were performed as described by Jeffreys *et al.* (1980). Filters were given a post-hybridization wash at 65°C in 1 × SSC (SSC is 0·15 M-NaCl, 15 mM-sodium citrate, pH 7·0) and autoradiographed overnight using an intensifier screen; 1 to 3 positively hybridizing plaques were detected/9 cm plate.

The region of the lawn containing a positive plaque was excised, replated on *E. coli* ED8910 (803, *supE*, *supF*, *recB*21, *recC*22, *hsdS*; Loenen & Brammar, 1980) and rescreened. Individual positive plaques were replated on *E. coli* ED8910 on three 22 cm × 22 cm plates to give confluent lysis. Phage were eluted into 30 ml Luria broth plus 10 mM-MgCl$_2$ at 37°C for 2 h, clarified by centrifugation at 16,000 *g* for 10 min at 4°C, and phage pelleted by centrifugation at 165,000 *g* for 45 min at 4°C. Phage were resuspended in phage buffer, cleared again by centrifugation at 16,000 *g* for 5 min, and repelleted. Phage DNA was prepared as described above, including extraction with 2-methoxyethanol and potassium phosphate to remove agar impurities (Jeffreys *et al.*, 1980). A total of 50 μg recombinant phage DNA was recovered.

(c) *Subcloning ψδ-globin gene fragments*

pAT153 DNA (Twigg & Sherratt, 1980) was linearized by cleavage with endonuclease *Bam*HI or *Eco*RI, dephosphorylated with calf intestinal phosphatase and ligated to recombinant phage λBL.9 DNA cleaved with *Bgl*II or *Eco*RI. Ligated DNA was transformed into *E. coli* HB101 using the method of Fantoni *et al.* (1979) and transformants selected on Luria agar plus 20 μg thymine/ml and 25 μg sodium ampicillin/ml. Transformants containing β-globin DNA sequences were identified by colony hybridization (Grunstein & Hogness, 1975) with ^{32}P-labelled rabbit globin complementary DNA prepared by reverse transcription of rabbit adult globin mRNA in the presence of [α-^{32}P]dCTP (Ghosh *et al.*, 1980). Plasmid DNA was prepared by the method of Birnboim & Doly (1979), followed by banding in CsCl and ethidium bromide. Traces of RNA were removed by subsequent centrifugation through a 10% to 40% sucrose gradient.

(d) *DNA sequencing*

Restriction endonuclease digests of recombinant plasmids were labelled at the 5′ terminus using polynucleotide kinase and [γ-^{32}P]ATP (Maxam & Gilbert, 1980), or at the 3′ terminus by fill-in labelling using reverse transcriptase and [α-^{32}P]dCTP (Goodman, 1980). Sequencing substrates were isolated, after cleavage with a second restriction endonuclease, by the method of Smith (1980). All sequencing was performed by the procedure of Maxam & Gilbert (1980), using five chemical modification reactions (G, G+A, T+C, C, A > C). Sequences were determined on 40 cm 8% (w/v) polyacrylamide gels 0·35 mm thick.

(e) *Containment*

Cloning of lemur DNA into λL47.1 was carried out under category 1 containment, and subcloning into pAT153 at category 0, in accordance with Genetic Manipulations Advisory Group guidelines.

3. Results

(a) *Preparation and screening of a brown lemur genomic library*

DNA was prepared from the same brown lemur that had been used to establish a map of the β-globin gene cluster by Southern blot analysis of genomic DNA (Barrie *et al.*, 1981). Lemur DNA was partially digested with endonuclease *Sau*3A and partial digest fragments 11 to 20 kb long were isolated. Partials were ligated onto arms purified from the *Bam*HI replacement vector λL47.1 (Loenen & Brammar, 1980) cleaved with *Bam*HI. Recombinants were packaged *in vitro* and plated without amplification onto *E. coli* WL87 (rec^+) at high plaque density (~3×10^5 plaques/9 cm Petri dish).

Recombinant plaques containing β-globin DNA sequences were identified by plaque hybridization (Benton & Davis, 1977) with ^{32}P-labelled cloned adult rabbit β-globin cDNA (Maniatis *et al.*, 1976) at low stringency. We have shown that this probe is capable of detecting all of the β-related globin genes in the brown lemur (Barrie *et al.*, 1981). From 2 μg of *in vitro* packaged DNA, ~$1{\cdot}8\times10^6$ recombinant plaques were obtained, of which 15 gave positive hybridization with rabbit β-globin cDNA.

One strongly hybridizing plaque was replated at low plaque density (on *E. coli* ED8910 (*recBC*) to minimize the risk of rearrangement of the insert) and rescreened. An isolated positive plaque was amplified on *E. coli* ED8910 and phage DNA prepared. A restriction endonuclease cleavage map of this recombinant, termed λBL.9, was established by cleavage with endonucleases *Bam*HI, *Bgl*II, *Eco*RI and *Hin*dIII (Fig. 1). Comparison of λBL.9 with the genomic map of the brown lemur β-globin gene cluster (Barrie *et al.*, 1981) showed that λBL.9 had an 11 kb insert containing both the "$\psi\beta$" and β-globin genes, separated by 2·6 kb of DNA and not 3·5 kb as estimated by Southern blot analysis of total lemur DNA (Barrie *et al.*, 1981). This discrepancy, plus slight discrepancies in the alignment of restriction endonuclease cleavage sites between λBL.9 and the genomic map, can be fully

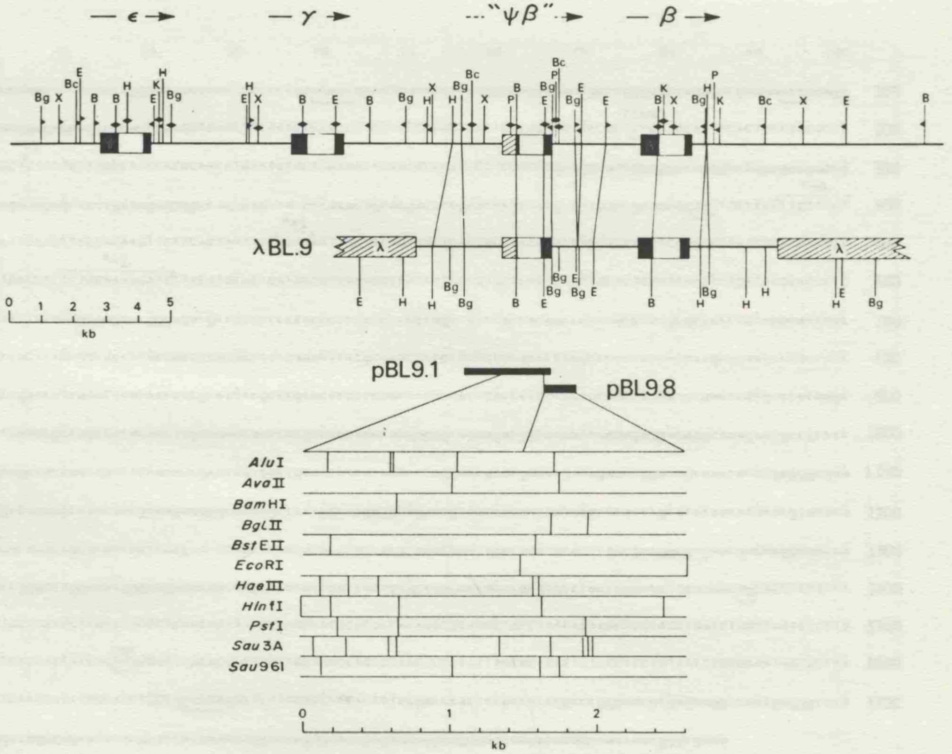


FIG. 1. Isolation of DNA segments from the β-globin gene cluster of the brown lemur.

A brown lemur library made by ligating lemur *Sau*3A partials into the *Bam*HI replacement phage vector λL47.1 was screened for recombinants containing β-globin DNA sequences. The recombinant λBL.9 was isolated and mapped by cleavage with restriction endonucleases *Bam*HI (B), *Bgl*II (Bg), *Eco*RI (E) and *Hin*dIII (H). Also shown is a map of the entire lemur cluster deduced by Southern blotting analysis of total genomic DNA, using human globin DNA probes to detect lemur globin genes (Barrie *et al.*, 1981). This genomic map also shows cleavage sites for endonucleases *Bcl*I (Bc), *Kpn*I (K), *Pst*I (P) and *Xba*I (X), and only shows sites that generate globin DNA fragments; the direction of a gene detected relative to a mapped site is indicated by ▸ or ◂. Alignment of λBL.9 with the genomic map shows an accurate correspondence over the "$\psi\beta$"–β-globin gene region. The only slight discrepancy was in the position of a *Hin*dIII site 5′ to the "$\psi\beta$"-globin gene; however, this site could only be located in the genomic map by measurement from a distal *Kpn*I site within the β-globin gene, and experimental errors in fragment length determination were sufficient to account for the *Hin*dIII site discrepancy. Note that those sites in λBL.9 not indicated in the genomic map do not generate β-globin DNA fragments.

*Bgl*II and *Eco*RI digests of λBL.9.1 were cloned into pAT153 and recombinant plasmids containing the "$\psi\beta$" 5′ *Bgl*II fragment (pBL9.1) and the 3′ *Eco*RI fragment (pBL9.8) were isolated. A detailed composite restriction endonuclease cleavage map of pBL9.8 and the 3′ end of pBL9.1 is shown.

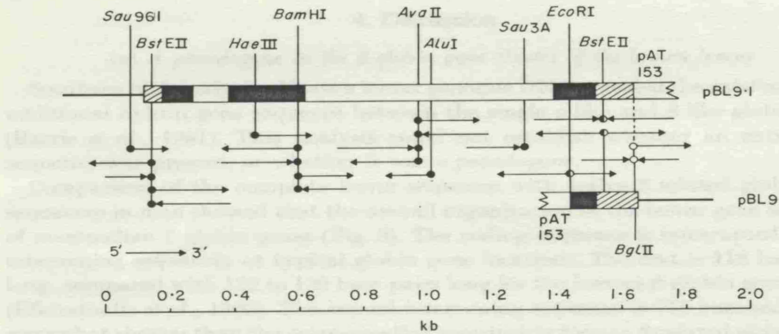


FIG. 2. Sequencing strategy for the lemur $\psi\delta$-globin gene.

Regions homologous to coding sequence in active β-related globin genes are shown by filled boxes and the homologues of 5' and 3' non-coding regions in the mature mRNA by hatched boxes. Restriction endonuclease cleavage sites used for end-labelling are indicated; additional sites are not shown in this map (see Fig. 1). Horizontal lines indicate the DNA segments that were sequenced. Arrows pointing to the right refer to sequences determined from the "transcribed" strand, and to the left, from the "non-transcribed" strand. Sequences determined from pBL9.1 are shown by filled circles, and from pBL9.8 by open circles. All sequences were determined by the method of Maxam & Gilbert (1977,1980).

accounted for by errors in fragment length determination inherent in genomic mapping (Baralle *et al.*, 1980*a*). We conclude that the insert in λBL.9 is derived from lemur DNA without any obvious signs of rearrangement. The locations of the "$\psi\beta$" and β-globin genes in λBL.9 were confirmed by Southern blot analysis of restricted λBL.9 DNA and hybridization with rabbit β-globin cDNA to detect β-globin DNA fragments. Only those fragments predicted from the genomic map hybridized with β-globin cDNA (data not shown).

(b) *DNA sequence analysis of the "$\psi\beta$" globin gene*

A 2·7 kb *Bgl*II fragment containing the "$\psi\beta$" sequence plus 5' flanking regions, and an overlapping 1·1 kb *Eco*RI fragment containing the 3' end of this gene, were subcloned into the plasmid pAT153 (Twigg & Sherratt, 1980) linearized with *Bam*HI or *Eco*RI. Two recombinant plasmids, termed pBL9.1 and pBL9.8, respectively, were exhaustively mapped for restriction endonuclease cleavage sites (Fig. 1).

The strategy adopted for sequencing the gene is shown in Figure 2. All exons and flanking sequences, and 95% of intervening sequences, were determined on both DNA strands using the method of Maxam & Gilbert (1977,1980), and all sequences were overlapped. The complete 1786 base-pair sequence, extending from 134 base-pairs before the start of the gene to 138 base-pairs beyond the homologue of the poly(A) addition site in active β-globin genes, is shown in Figure 3.

```
        10        20        30        40        50        60        70        80        90       100

gcccaaactccacccttgggatcacaacccgcccttgaacaatagcctcatttcattaggagagacaaacggctgggggccagagatgaagaataaaagg   100
                                                                           (ini)
ccatggagagaagcagcagtacAGGTGAGCTTCTAACTCATCCGTGGTCACCAGCAGACTCGCAGACCTGACGCTGTGGTGCATTTCACTGCAGAGGCAA   200
    *-1
AGGCTCTGCGGCTAGCCTGCCAGGCAACATGAATGTGGAGGAGGCTGGAGGCAAGATCCTGGGCAGgcaggcactggaggccagggtcaggagcagaaag   300
                                                                                              *+4
gcagaaagtgttcctgaaagaggggatagccagttatcctatacagtatgactttgcatctgttttgtgacgactgccccatagGTTCTTGCTTGTTTGT   400
                                    *+1
CTATTCATGTTCCCAGAAGTTCTTCAGTAATTTTGGGGAATTTGTCCTCTGGCTCTGTGCTAATGGCCAACCCCAAGGTCAAGTCTTATGGCAAGAAGCA   500
            *-1
ACTGACCTCTTTGGAAAAGATGTTACGTGCACTGATGATCTCAAAGGCAACTTTGCTGAGCTGAGTGAGCTGCACTGTGACAAGTTGCATGTGGATCCTG   600
AGAACTGCAGGgtgagtctgggagatgttccgtttttccctttctctttctagttttcactctagttctttacctatgtgttctttctacacattcat   700
ttttactttaccatattttatcatttaacacttttcaaattttgtcaatttcttctttctacattctgtcttctttcctttgcacaatcttactttt   800
tattgaattttaaatttactatcctgtcatttgcctgtatctctcccatcccccatttatttttttctccaaccacaacccaaattatgcatatcagt   900
tctcatctgctagttctacactttgaaaaatccttctgtctcttcatatgggggtagaagatggtccaactcaaagaggagaggcacagaatgctgtttt   1000
agaagctataaatcattttaaaatgaataataattgaaattttataaattcaggaataaatgaaatgaaagaaatggaaagtaaatatctgagggtgaa   1100
aggttaaagttcatactggaagcagggcaccagttttggtcagaggcagactgtcatcacactaatcaatttatttgtatataatatatatgtacata   1200
catatatactgtatactacttaagtatccagtattatatatgtattatgtacatatatacatacatatacttaacgctggtgtgaatgacatggagatca   1300
acttgggctaggacatgggcagaaaagaaagccaatattgatttctttgttaaccatacctatgtgtctacttacctcttccccacagCACCTGGGCAA   1400
CGTGCTGGTGGTTGTGCTGGCTGAACAGTTTGGCAAGGAATTCACCCCACAGGTGCAGGCTGCCTATCAGAAGGTGGTGGCTGGTGTGGCTAATGCCCTG   1500
                ter
GCTCACAAGTATCACTGAGGCCTCGGACCATTTCTTGGTGACCAGTGGAAGGCCCTATTTCCACAGATTCTCTCTTCTGTAATTGGGGAAATAGTGCTTA   1600
CCTCAAGGGTATGGCATCTGCCTAATAAAGATCTTTCAGCTCAACTTTctgatttatttatttttgtctgggaatgtgagaaggtccctgagggtcta   1700
cagatagagagttctcatgtcttatacaaaaggtcaagagaaatgagaaaaggaagggagccagacacagacactaatgggtgaca
```

FIG. 3. The complete nucleotide sequence of the hybrid $\psi\delta$-globin gene in the brown lemur.

Sequences homologous to exons in functional globin genes are shown in upper case, and additional flanking sequences and intervening sequences shown in lower case. The homologous protein-coding regions were readily detected by comparison with human β-related globin gene sequences (see Fig. 4). The presumptive location of regions homologous to the 5' non-translated region of mature mRNA was determined by detailed comparison with the sequence of the human $^{A}\gamma$-globin gene (Slightom *et al.*, 1980) and of the 3' non-translated region by homology with corresponding sequences in the human β- and δ-globin genes (Lawn *et al.*, 1980; Spritz *et al.*, 1980). Relevant sequences are underlined; these include the ATA box and the A-C-A-A-T equivalent of the C-C-A-A-T box in the 5' flanking region, the abnormal (GTG) initiation codon (ini), intron/exon junction sequences that conform to the G·T-A·G rule of Breathnach *et al.* (1978), the termination codon (ter), and the A-A-T-A-A-A sequence in the 3' non-translated region (Proudfoot & Brownlee, 1976). The approximate locations of frameshift mutations that prevent the gene from coding for globin are shown by an asterisk; these include single base insertions (+1) and deletions (−1), and a 4 base-pair tandem repeat (+4). The hyphens have been omitted for clarity.

4. Discussion

(a) *A pseudogene in the β-globin gene cluster of the brown lemur*

Southern blot analysis of brown lemur genomic DNA revealed the existence of an additional hybrid gene sequence between the single γ-like and β-like globin genes (Barrie *et al.*, 1981). This analysis could not establish whether an entire gene sequence was present, or whether it was a pseudogene.

Comparison of the complete lemur sequence with active β-related globin gene sequences in man showed that the overall organization of the lemur gene is typical of mammalian β-globin genes (Fig. 3). The coding sequence is interrupted by two intervening sequences at typical globin gene locations. The first is 118 base-pairs long, compared with 122 to 130 base-pairs long for the human β-globin gene family (Efstratiadis *et al.*, 1980). The second intervening sequence is 778 base-pairs long, somewhat shorter than the corresponding sequence in human β-related globin genes (850 to 904 base-pairs). With the exception of the junction between exon 1 and intron 1, all intron/exon junctions follow the G·T-A·G rule of Breathnach *et al.* (1978) (see Fig. 3). Normal length homologues of 5′ and 3′ non-coding regions in mature globin mRNA could be discerned (Figs 2 and 3), including in the 3′ non-coding region the presence of an A-A-T-A-A-A sequence that commonly occurs in polyadenylated RNAs (Proudfoot & Brownlee, 1976). The homologue of putative promoter elements in the 5′ flanking region could also be detected; in particular, the $\overset{*}{\mathrm{A}}$TA box is present 30 base-pairs (asterisk) before the homologue of the mRNA capping site (cf. 26 to 34 base-pairs in numerous other eukaryotic genes; Efstratiadis *et al.*, 1980). The sequence $\overset{*}{\mathrm{A}}$-C-A-A-T 83 base-pairs before the capping site is similar in sequence and position to the canonical C-C-A-A-T box typical of many animal genes (Efstratiadis *et al.*, 1980).

This β-related gene sequence in the brown lemur contains a number of features which establish that the sequence is a pseudogene that cannot code for globin (see Fig. 3). The initiation codon is altered from ATG to GTG, although this would not necessarily prevent correct initiation of translation. The first exon contains a single base deletion, 27 base-pairs from the initiation codon, that brings into phase a TGA termination codon 25 base-pairs further on. The second exon contains three frameshift mutations, including a 4 base-pair insertion apparently resulting from a direct duplication of a T-T-G-T tetranucleotide. The third exon shows no irregularities. The only other obviously atypical sequence is the G·T → G·C alteration at the junction of exon 1 and intron 1. Whilst this lemur pseudogene cannot code for globin, the 5′ flanking region shows no obviously unusual sequences that would prevent transcription of this gene.

The abnormalities seen in this lemur pseudogene are typical of those found in various other globin pseudogenes, such as the rabbit ψβ2 globin gene (Lacy & Maniatis, 1980), the human ψα1 globin gene (Proudfoot & Maniatis, 1980) and the goat $\psi\beta^{\times}$ gene (Cleary *et al.*, 1980).

(b) *Comparison of the brown lemur pseudogene with human β-related globin genes*

Complete nucleotide sequences have been published for the human β-globin

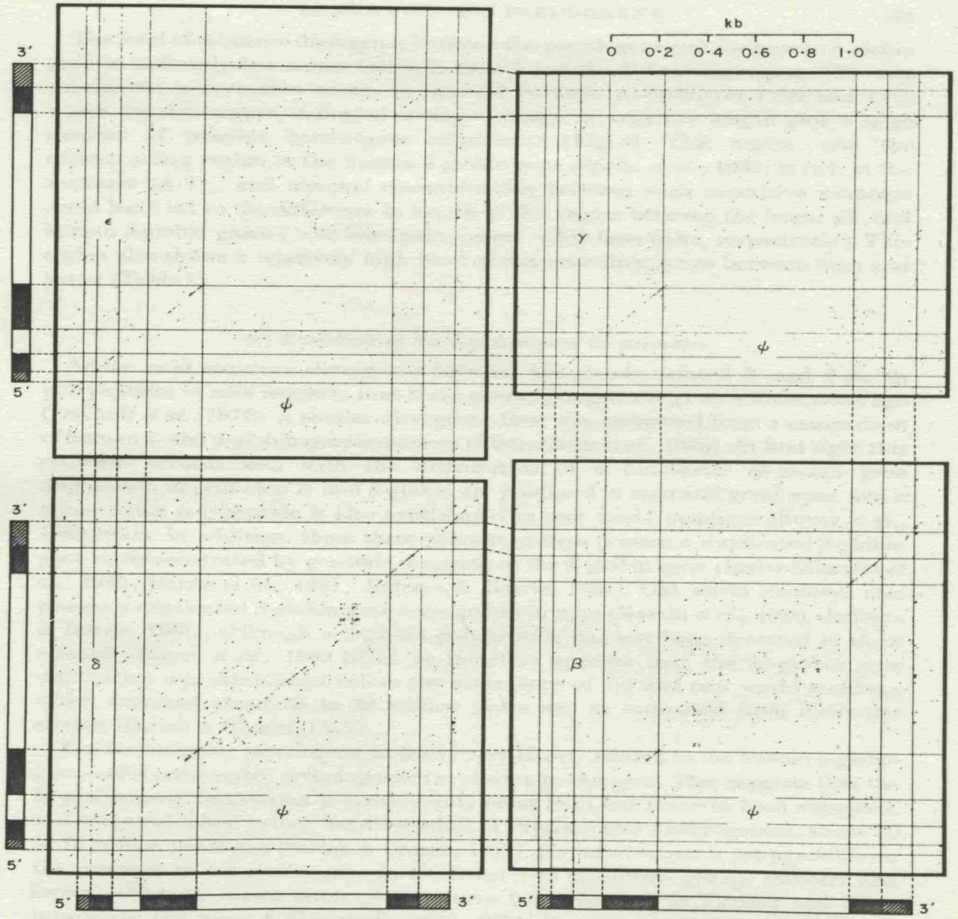


FIG. 4. Dot matrix comparisons of the lemur pseudogene with human β-related globin genes.

The entire coding sequence, flanking sequences and intervening sequences of the lemur pseudogene were compared with corresponding sequences of the human ε-, $^{A}\gamma$-, δ- and β-globin genes, taken from Baralle *et al.* (1980*b*), Slightom *et al.* (1980), Spritz *et al.* (1980) and Lawn *et al.* (1980), respectively, using the dot matrix method of Konkel *et al.* (1979). Each short slanting stroke represents the centre of a 6 base-pair identity between 2 genes. Homologies appear as lines at 45° across the grid. The positions of coding sequences (filled boxes), the 5′ and 3′ non-translated regions of mature mRNA (hatched boxes) and intervening sequences (open boxes) are shown alongside the grids. Reduction of the length of a matching sequence to 5 base-pairs did not noticeably improve the length of homologous regions detected at this scale, but substantially increased the background noise plus the time taken to compute and plot the matrix (data not shown).

(Lawn *et al.*, 1980), $^{G}\gamma$- and $^{A}\gamma$-globin (Slightom *et al.*, 1980), δ-globin (Spritz *et al.*, 1980) and ε-globin genes (Baralle *et al.*, 1980*b*). In order to compare these sequences reasonably objectively with the lemur pseudogene sequence, we used the dot matrix comparison method of Konkel *et al.* (1979). Figure 4 shows the homology matrices determined for this pseudogene compared with human β-, $^{A}\gamma$-, δ- and ε-globin genes. The pseudogene showed clear homologies with the ε- and $^{A}\gamma$-globin genes, particularly over the 5′ flanking region, exon 1 and exon 2. Homology with the human β-globin gene was instead most pronounced in the coding region of exon 3. In remarkable contrast, there were substantial regions of close homology with the human δ-globin gene, commencing within exon 2 (at a region highly conserved in all globin genes) and extending, with one major interruption, throughout intron 2, exon 3 and the 3′ flanking region.

Sequence divergences (corrected for multiple substitutions at a single site) are given in Table 1. These data and Figure 4 show that the pseudogene is most closely related to the human ε- and γ-globin genes probably up to some point within exon 2, past which it assumes close homology with the human δ-globin but not β-globin gene. This hybrid gene structure is in agreement with previous genomic analyses that showed that the 5′ region of this gene could be detected by human ε-globin DNA but not by β-globin cDNA (Barrie *et al.*, 1981). However, the sequence comparison cannot rule out the possibility that the 5′ region of the pseudogene is γ-related rather than ε-like.

TABLE 1

Sequence divergence between the lemur ψδ-globin gene and human β-related globin genes

			% Sequence divergence *versus*		
ψδ region		Co-ordinates	ε	$^{A}\gamma$	δ
5′ Flanking		1–122	43 (2)	55 (2)	99 (2)
Exon 1		123–266	38 (1)	43 (1)	45 (3)
Intron 1		267–384	82 (3)	82 (1)	>100
Exon 2		385–611	35 (3)	29 (3)	38 (3)
Intron 2	A	612–1179	>100	>100	30 (8)
	B	1180–1270	>100	>100	74 (6)
	C	1271–1389	>100	>100	27 (1)
Exon 3		1390–1648	38 (3)	53 (2)	21 (0)
3′ Flanking		1649–1786	>100	—	25 (3)

The ψδ sequence was aligned in turn with human ε-, $^{A}\gamma$- and δ-globin gene sequences, in accordance with the regions of homology detected in the dot matrix analysis (Fig. 4). High resolution dot matrices scoring 2 base-pair matches were used to align relatively diverged regions (data not shown). Percent sequence divergences were calculated for exons, introns and flanking sequences (co-ordinates taken from Fig. 2), ignoring tracts that had been inserted or deleted from the ψδ sequence plus the single nucleotides immediately flanking these tracts that were used to define the positions of each microdeletion/insertion. The minimal number of deletions/insertions required for sequence alignment is given in parentheses. Percent sequence divergences were corrected for multiple substitutions at single sites (see Jeffreys, 1981). Values >100% indicate highly diverged regions that could not be aligned unequivocably. The 3′ flanking sequence of the human $^{A}\gamma$-globin gene is not available. Intron 2 is divided into 3 regions; the central B region (co-ordinates 1180 to 1270) contains the (A-T)$_n$-rich segments (Fig. 3) and substantially differs in length between the lemur ψβ- and human δ-globin genes (Fig. 4).

The level of sequence divergence between the pseudogene and the human δ-globin gene is uniformly low across intron 2, exon 3 and the 3′ flanking region. The only substantial interruption occurs in intron 2 between co-ordinates 1180 and 1270 where the dot matrix indicated a major change in sequence length plus a large number of possible homologous alignments (Fig. 4). This region, and the corresponding region in the human δ-globin gene (Spritz *et al.*, 1980) is rich in the sequence (A-T)$_n$, and unequal recombination between such repetitive elements could have led to the difference in length of this region between the lemur ψδ- and human δ-globin genes (~90 base-pairs *versus* ~200 base-pairs, respectively). This region also shows a relatively high level of sequence divergence between man and lemur (Table 1).

(c) *Evolution of the δ-globin gene in primates*

Amino acid sequence divergence between the closely related δ- and β-globin polypeptides of man suggests that these genes diverged about 40 million years ago (Dayhoff *et al.*, 1972). A similar divergence time was estimated from a comparison of human δ- and β-globin gene sequences (Efstratiadis *et al.*, 1980). At first sight this estimate accords well with the distribution of a functional δβ-globin gene duplication in primates. δ- and β-globin are produced in man and great apes, and a minor δ-like polypeptide is also synthesized in new world monkeys (Boyer *et al.*, 1969,1971). In addition, these three primate groups possess a duplicated β-globin gene as demonstrated by genomic mapping of the β-globin gene cluster (Zimmer *et al.*, 1980; Barrie *et al.*, 1981; Jeffreys & Barrie, 1981). Old world monkeys also possess a duplicated β-globin gene arranged as in man (Martin *et al.*, 1980; Jeffreys & Barrie, 1981), although a δ-globin polypeptide has not been detected in these animals (Boyer *et al.*, 1969,1971). It therefore appears that the δβ-globin gene duplication was established before the divergence of old and new world monkeys which occurred about 35 to 38 million years ago as estimated from molecular studies (Sarich & Cronin, 1977).

The brown lemur pseudogene is clearly very closely related to the human δ-globin gene and is presumably orthologous in evolution to this gene. This suggests that the δβ-globin gene duplication is considerably older than has hitherto been supposed, and was established before the divergence of Prosimii and Anthropoidea, about 70 to 75 million years ago (Sarich & Cronin, 1977). No information is yet available on the presence of a δ-globin gene in the other two prosimian groups (tarsiers and lorises) although minor adult globins have been reported in *Tarsius* and other prosimians (Barnicot & Hewett-Emmett, 1974; Beard *et al.*, 1976).

The primate δ-globin gene has had a long and complex evolutionary history (Fig. 6). It has been silenced on at least two separate occasions: in the lineage leading to old world monkeys that possess a δ-globin gene but do not produce δ-globin, and in the lineage leading to the lemur. In addition, the 5′ region of the δ-globin gene appears to have acquired sequences from other β-related genes on more than one occasion. In man, the δ- and β-globin genes are closely homologous over exon 1, intron 1 and exon 2 (Fig. 5; see Efstratiadis *et al.*, 1980) yet show little similarity in intron 2 and the 3′ flanking region which are nevertheless closely

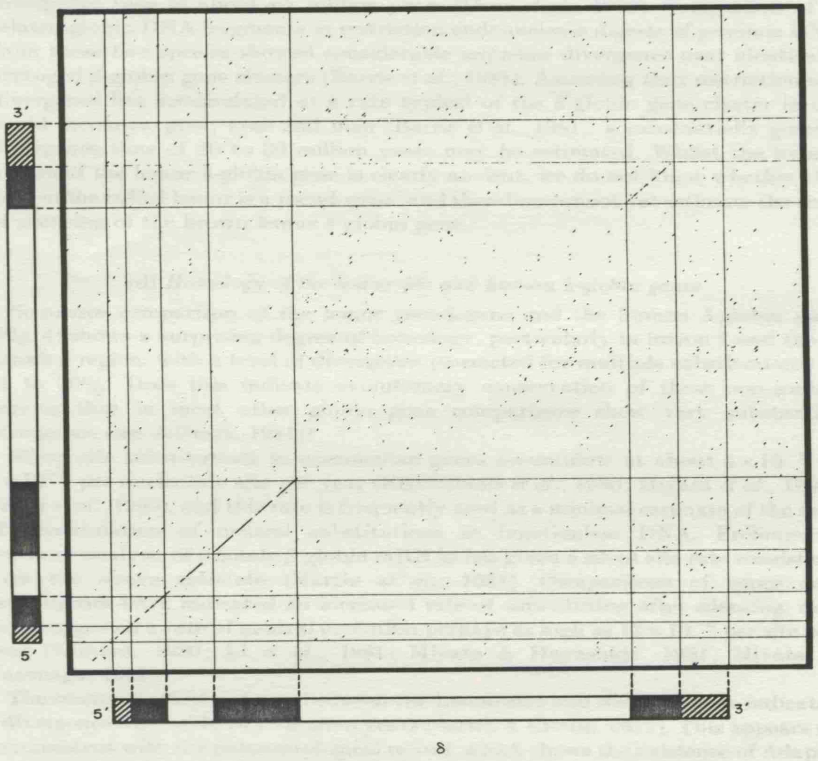


FIG. 5. Dot matrix comparison of the human δ- and β-globin genes. The comparison was performed as described in the legend to Fig. 4.

related in sequence between the human δ and lemur ψδ genes. This strongly suggests that this 5′ region of the δ-globin gene has undergone gene conversion with the β-globin gene recently in the lineage leading to man. The silent site divergence of human δ- and β-globin genes in exons 1 and 2, corrected for multiple substitutions, is 16% (Efstratiadis *et al.*, 1980). Assuming that silent site substitutions accumulate at about 5×10^{-9} per site per year (see below), this suggests that the last round of conversion of the δ-globin gene by β occurred about 16 million years ago (Fig. 6).

An analogous event appears to have occurred in the lemur ψδ globin gene, involving a similar region of the gene plus a segment of either an ϵ- or γ-globin gene. If the 5′ region of the ψδ gene is ϵ-like, then the most plausible mechanism would involve gene conversion following mispairing of ϵ- and δ-globin genes at meiosis (Barrie *et al.*, 1981). If instead the lemur pseudogene is a γδ hybrid, then another fascinating possibility exists, namely that this gene appeared as a result of unequal crossing over between γ- and δ-globin genes, analogous to the unequal recombination in man that generates the fused $^{A}\gamma$-β-globin polypeptide of Hb Kenya (see Weatherall & Clegg, 1976). If so, then the organization of the lemur cluster would originally have been 5′-ϵγγδβ-3′, an organization with a duplicated γ-globin gene very similar to that seen in man (see Efstratiadis *et al.*, 1980). The unequal crossing over would then have contracted the cluster from 5′-ϵγγδβ-3′ to 5′-ϵγ(γδ)β-3′, with the concomitant deletion of all sequences between the second γ-globin gene and the δ-globin gene. Such a deletion would be of considerable interest in view of evidence for an element in this $^{A}\gamma$–δ region in man that is involved in mediating the γ→β switch towards the end of gestation (Fritsch *et al.*, 1979; Bernards & Flavell, 1980). It is worth noting that the β-globin polypeptide of the adult brown lemur is homologous to human β-globin rather than γ-globin (Maita *et al.*, 1979) and that if a contraction of the lemur cluster has occurred, it has apparently not prevented expression of the β-globin gene at the 3′ end of the cluster (Fig. 1).

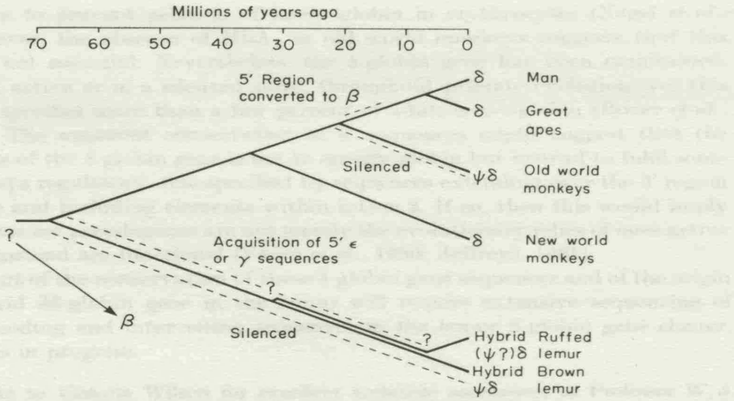


FIG. 6. Evolutionary history of the primate δ-globin gene.

Divergence times estimated from molecular studies are taken from Sarich & Cronin (1977). The divergence time of the brown and ruffed lemurs is uncertain (see Discussion). The approximate timing of major events so far discovered in the evolution of the δ-globin gene are indicated by broken lines. The age of the βδ-globin gene duplication is completely unknown, though it presumably predates the divergence of the Lemuridae and Anthropoidea.

The acquisition of γ- or ϵ-globin sequences at the 5′ end of the lemur δ-globin gene probably occurred at least several million years ago. Analysis of total genomic DNA suggested that this hybrid gene is present not only in the brown lemur (*Lemur macaco (fulvus) mayottensis*) but also in the ruffed lemur (*Lemur variegatus*) (Barrie *et al.*, 1981). The immunological distance between these two species gives a divergence time of about six million years (Dene *et al.*, 1976). Comparison of β-related globin DNA fragments in restriction endonuclease digests of genomic DNA from these two species showed considerable sequence divergence over identically arranged β-globin gene clusters (Barrie *et al.*, 1981). Assuming that restriction site divergence has accumulated at a rate typical of the β-globin gene cluster in old world monkeys, great apes and man (Barrie *et al.*, 1981), a substantially greater divergence time of 20 to 30 million years may be estimated. Whilst the hybrid nature of the lemur δ-globin gene is clearly ancient, we do not know whether this gene in the ruffed lemur is a pseudogene, and therefore cannot yet estimate the time of silencing of the brown lemur δ-globin gene.

(d) *Homology of the lemur $\psi\delta$- and human δ-globin genes*

Sequence comparison of the lemur pseudogene and the human δ-globin gene (Fig. 4) shows a surprising degree of homology, particularly in intron 2 and the 3′ flanking region, with a level of divergence (corrected for multiple substitutions) of 21 to 30%. Does this indicate evolutionary conservation of these non-coding regions that in most other globin gene comparisons show very substantial divergence (see Jeffreys, 1981)?

Silent site substitutions in mammalian genes accumulate at about 4×10^{-9} to 7×10^{-9} per nucleotide site per year (Efstratiadis *et al.*, 1980; Miyata *et al.*, 1980; Perler *et al.*, 1980), and this rate is frequently used as a minimal estimate of the rate of accumulation of neutral substitutions in functionless DNA. Preliminary sequence analysis of primate β-globin mRNAs has given a silent site rate consistent with the above estimate (Martin *et al.*, 1981). Comparisons of genes and pseudogenes have indicated an increased rate of substitution after silencing, and have suggested a rate of neutral evolution perhaps as high as 13×10^{-9} per site per year (Kimura, 1980; Li *et al.*, 1981; Miyata & Hayashida, 1981; Miyata & Yasunaga, 1981).

The immunological distance between the Lemuridae and Anthropoidea indicates a divergence time of 70 to 75 million years (Sarich & Cronin, 1977). This appears to be consistent with the palaeontological record, which shows the existence of Adapid lemuroids in the Eocene about 40 to 58 million years ago, although the relation between Adapids and contemporary Lemuridae and Anthropoidea is disputed (Gingerich & Schoeninger, 1977; Szalay & Delson, 1979).

Taking the time of common ancestry of man and lemurs as 72 million years ago, and a rate of neutral evolution of 4×10^{-9} to 13×10^{-9} per site per year, gives a predicted corrected divergence of functionless DNA of 60 to 190%, substantially greater than the divergence of 25 to 30% over intron 2 and the 3′ flanking sequence of the lemur $\psi\delta$- and human δ-globin genes (Table 1). This suggests that these non-coding regions of the δ-globin gene have been under selective constraint for much of their evolutionary history. However, these neutral rates may not necessarily be applicable to primate evolution, and there is evidence that the overall rate of nuclear DNA evolution might have decreased in lemurs (Bonner *et al.*, 1980).

If these 3′ δ-globin gene sequences have been conserved, then it raises additional problems about the role of the δ-globin gene. It has been suggested that HbA_2

($\alpha_2\delta_2$) serves to prevent gelation of haemoglobin in erythrocytes (Nagel *et al.*, 1979); however, the absence of HbA_2 in old world monkeys suggests that this function is not essential. Nevertheless, the δ-globin gene has been maintained, either in an active or in a silenced state, throughout primate evolution, yet this gene never specifies more than a few percent of adult non-α-globin (Boyer *et al.*, 1969,1971). The apparent conservation of 3′ sequences might suggest that the primary role of the δ-globin gene is not to specify globin but instead to fulfil some other, perhaps regulatory, role specified by sequences extending over the 3′ region of this gene and including elements within intron 2. If so, then this would imply that at least some pseudogenes are not merely the evolutionary relics of once active genes, but instead are functional (Vanin *et al.*, 1980; Jeffreys, 1981).

Assessment of the conservation of these δ-globin gene sequences and of the origin of the hybrid $\psi\delta$-globin gene in the lemur will require extensive sequencing of additional coding and intervening sequences in the lemur β-globin gene cluster. This work is in progress.

Our thanks to Victoria Wilson for excellent technical assistance, to Professor W. J. Brammar and Wilhelmina Loenen for advice on λ cloning, and to Peter Meacock for invaluable assistance with DNA sequencing. This work was supported by a grant from the Medical Research Council.

REFERENCES

Baralle, F. E., Shoulders, C. C., Goodbourn, S., Jeffreys, A. & Proudfoot, N. J. (1980*a*). *Nucl. Acids Res.* **8**, 4393–4404.

Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980*b*). *Cell*, **21**, 621–626.

Barnicot, N. A. & Hewett-Emmett, D. (1974). In *Prosimian Biology* (Martin, R. D., Doyle, G. A. & Walker, A. C., eds), pp. 891–902, Duckworth, London.

Barrie, P. A., Jeffreys, A. J. & Scott, A. F. (1981). *J. Mol. Biol.* **149**, 319–336.

Beard, J. M., Barnicot, N. A. & Hewett-Emmett, D. (1976). *Nature (London)*, **259**, 338–340.

Benton, W. D. & Davis, R. W. (1977). *Science*, **196**, 180–182.

Bernards, R. & Flavell, R. A. (1980). *Nucl. Acids Res.* **8**, 1521–1534.

Birnboim, H. C. & Doly, J. (1979). *Nucl. Acids Res.* **7**, 1513–1523.

Blattner, F. R., Williams, B. G., Blechl, A. E., Denniston-Thompson, K., Faber, H. E., Furlong, L.-A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schumm, J. W., Sheldon, E. L. & Smithies, O. (1977). *Science*, **194**, 161–169.

Bonner, T. I., Heinemann, R. & Todaro, G. J. (1980). *Nature (London)*, **286**, 420–423.

Boyer, S. H., Crosby, E. F., Thurmon, T. F., Noyes, A. N., Fuller, G. F., Leslie, S. E., Shepard, M. K. & Herndon, C. N. (1969). *Science*, **166**, 1428–1431.

Boyer, S. H., Crosby, E. F., Noyes, A. N., Fuller, G. F., Leslie, S. E., Donaldson, L. J., Vrablik, G. R., Schaefer, E. W. & Thurmon, T. F. (1971). *Biochem. Genet.* **5**, 405–448.

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 4853–4857.

Cleary, M. L., Haynes, J. R., Schon, E. A. & Lingrel, J. B. (1980). *Nucl. Acids Res.* **8**, 4791–4802.

Coggins, L. W., Grindlay, G. J., Vass, J. K., Slater, A. A., Montague, P., Stinson, M. A. & Paul, J. (1980). *Nucl. Acids Res.* **8**, 3319–3333.

Dayhoff, M. O., Hunt, L. T., McLaughlin, P. J. & Jones, D. D. (1972). In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, pp. 17–30, National Biomedical Research Foundation, Washington, D.C.

Dene, H. T., Goodman, M. & Prychodko, W. (1976). In *Molecular Anthropology, Genes and Proteins in the Evolutionary Ascent of the Primates* (Goodman, M. & Tashian, R. E., eds), pp. 171–195, Plenum, New York.

Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980). *Cell*, **21**, 653–668.

Enquist, L. & Sternberg, N. (1979). *Methods Enzymol.* **68**, 281–298.

Fantoni, A., Bozzoni, I., Ullu, E. & Farace, M. G. (1979). *Nucl. Acids Res.* **6**, 3505–3517.

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

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

Ghosh, P. K., Reddy, V. B., Piatak, M., Lebowitz, P. & Weissman, S. M. (1980). *Methods Enzymol.* **65**, 580–595.

Gingerich, P. D. & Schoeninger, M. (1977). *J. Human Evol.* **6**, 483–505.

Goodman, H. M. (1980). *Methods Enzymol.* **65**, 63–64.

Grunstein, M. & Hogness, D. S. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 3961–3965.

Hohn, B. (1979). *Methods Enzymol.* **68**, 299–309.

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

Jeffreys, A. J. (1981). In *Genetic Engineering* (Williamson, R., ed.), vol. 2, pp. 1–48, Academic Press, London and New York.

Jeffreys, A. J. & Barrie, P. A. (1981). *Phil. Trans. Roy. Soc. ser. B*, **292**, 133–142.

Jeffreys, A. J., Wilson, V., Wood, D., Simons, J. P., Kay, R. M. & Williams, J. G. (1980). *Cell*, **21**, 555–564.

Kimura, M. (1980). *J. Mol. Evol.* **16**, 111–120.

Konkel, D. A., Maizel, J. V. & Leder, P. (1979). *Cell*, **18**, 865–873.

Lacy, E. & Maniatis, T. (1980). *Cell*, **21**, 545–553.

Lawn, R. M., Efstratiadis, A., O'Connell, C. & Maniatis, T. (1980). *Cell*, **21**, 647–651.

Li, W.-H., Gojobori, T. & Nei, M. (1981). *Nature (London)*, **292**, 237–239.

Loenen, W. A. M. & Brammar, W. J. (1980). *Gene*, **20**, 249–259.

Maita, T., Setoguchi, M., Matsuda, G. & Goodman, M. (1979) *J. Biochem.* **85**, 755–764.

Maniatis, T., Kee, S. G., Efstratiadis, A. & Kafatos, F. C. (1976). *Cell*, **8**, 163–182.

Martin, S. L., Zimmer, E. A., Kan, Y. W. & Wilson,, A. C. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 3563–3566.

Martin, S. L., Zimmer, E. A., Davidson, W. S., Wilson, A. C. & Kan, Y. W. (1981). *Cell*, **25**, 737–741.

Maxam, A. M. & Gilbert, W. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 560–564.

Maxam, A. M. & Gilbert, W. (1980). *Methods Enzymol.* **65**, 499–560.

Miyata, T. & Hayashida, H. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 5739–5743.

Miyata, T. & Yasunaga, T. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 450–453.

Miyata, T., Yasunaga, T. & Nishida, T. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 7328–7332.

Nagel, R. L., Bookchin, R. M., Johnson, J., Labie, D., Wajcman, H., Isaac-Sodeye, W. A., Honig, G. R., Schiliro, G., Crookston, J. H. & Matsutomo, K. (1979). *Proc. Nat. Acad. Sci., U.S.A.* **76**, 670–672.

Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980). *Cell*, **20**, 555–566.

Proudfoot, N. J. & Brownlee, G. G. (1976). *Nature (London)*, **263**, 211–214.

Proudfoot, N. J. & Maniatis, T. (1980). *Cell*, **21**, 537–545.

Sarich, V. M. & Cronin, J. E. (1977). *Nature (London)*, **269**, 354.

Slightom, J. L., Blechl, A. E. & Smithies, O. (1980). *Cell*, **21**, 627–638.

Smith, H. O. (1980). *Methods Enzymol.* **65**, 371–380.

Spritz, R. A., DeRiel, J. K., Forget, B. G. & Weissman, S. M. (1980). *Cell*, **21**, 639–646.

Sussenbach, J. S., Monfoort, C. H., Schiphof, R. & Stobberingh, E. E. (1976). *Nucl. Acids Res.* **3**, 3193–3202.

Szalay, F.S. & Delson, E. (1979). *Evolutionary History of the Primates*, Academic Press, New York.

Twigg, A. J. & Sherratt, D. (1980). *Nature (London)*, **283**, 216–218.

Vanin, E. F., Goldberg, G. I., Tucker, P. W. & Smithies, O. (1980). *Nature (London)*, **286**, 222–226.

Weatherall, D. J. & Clegg, J. B. (1976). *Annu. Rev. Genet.* **10**, 157–178.

Yang, R. C.-A., Lis, T. & Wu, R. (1979). *Methods Enzymol.* **68**, 176–182.

Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W. & Wilson, A. C. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 2158–2162.

Edited by P. Chambon

Biggin,M.D., Gibson,T.J. and Hong,H.F. (1983).
Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc.Natl.Acad.Sci.USA.,80,3963-3965.

Boyer,S.H., Crosby,E.F., Noyes,A.N., Fuller,G.F., Leslie,S.E., Donaldson,L.J., Vrablik,G.R., Schaefer,E.W. and Thurmon,T.F. (1971).
Primate haemoglobins: some sequences and some proposals concerning the character of evolution and mutation. Biochem. Genet.,5,405-448.

Coppenhaver,D.H., Dixon,J.D. and Duffy,L.K. (1983).
Prosimian haemoglobins I. The primary structure of the β-globin chain of *Lemur catta*. Hemoglobin 7,1-14.

Hayashida,H. and Miyata,T. (1983).
Unusual evolutionary conservation and frequent DNA seqment exchange in class I genes of the major histocompatibility complex. Proc.Natl.Acad.Sci.USA.,80,2671-2675.

Huisman,T.H.J., Schroeder,W.A., Keeling,M.E., Gengozian,N., Miller,A., Brodie,A.R., Shelton,J.R., Shelton,J.B. and Apell,G. (1973).
Search for non-allelic structural genes for γ-chains of foetal haemoglobin in some primates. Biochem. Genet.,10,309-318.

Kimura,M. (1968).
Evolutionary rate at the molecular level. Nature 217,624-626.

Kimura,A. and Tagaki,Y. (1983).
A frameshift addition causes silencing of the δ-globin gene in an old world monkey, an anubis (*Papio doguera*). Nucl.Acids Res.,11,2541-2550.

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

Maniatis,T., Kee,S.G., Efstratiadis,A. and Kafatos,F.C. (1976).
Amplification and characterisation of a β-globin gene synthesised *in vitro*. Cell *8*,163-182.

Maita,T., Setoguchi,M., Matsuda,G. and Goodman,M. (1979).
Amino acid sequences of the α and β chains of adult haemoglobin of the brown lemur, *Lemur fulvus fulvus*. Japanese J. Biochem.,*85*,755-764.

Maxam,A.M. and Gilbert,W. (1977).
A new method for DNA sequencing. Proc.Natl.Acid.Sci.USA.,*74*,560-564.

Maxam,A.M. and Gilbert,W. (1980).
Sequencing end-labelled DNA with base-specific chemical cleavages. Meth.Enz.,*65*,499-560.

Perler,F., Efstratiadis,A., Lomedico,P., Gilbert,W., Kolodner,R. and Dodgson,J. (1980).
The evolution of genes: the chicken preproinsulin gene. Cell *20*,555-566.

Sarich,V.M. and Cronin,J.E. (1977).
Generation length and rates of hominoid molecular evolution. Nature *269*,354.

Schaffner,W., Gross,K., Telford,J. and Birnstiel,M. (1976).
Molecular analysis of the histone gene cluster of *Psammechinus miliaris*:II. The arrangement of the five coding and spacer sequences. Cell *8*,471-478.

Siegel,S. (1956).
Nonparametric statistics for the behavioural sciences. McGraw-Hill, Kogakusha, Tokyo.

Simons,E.L. (1969).

The origin and radiation of the primates. Ann. N.Y. Acad.Sci., 167(1),319-331.

Southern,E.M. (1975).

Detection of specific sequences among DNA fragments separated by gel elctrophoresis. J.Mol.Biol.,98,503-517.

Zimmer,E.A., Martin,S.L., Beverley,S.M., Kan.Y.W. and Wilson,A.C. (1980).

Rapid duplication and lose of genes coding for the α chains of haemoglobin. Proc.Natl.Acad.Sci.USA.,77,2158-2162.

Abstract Stephen Harris

A COMPARATIVE STUDY OF β-GLOBIN PSEUDOGENES IN MAN AND THE PRIMATES

The human β-globin gene family, situated on chromosome 11, consists of five functional genes (ε, $^{G}\gamma$, $^{A}\gamma$, δ and β) and a non-processed pseudogene, ψβ1. This study of contemporary ψβ1 pseudogene sequences has shown that this gene has been a stable component of the β-globin gene cluster during the evolution of the primate, and other, mammalian orders. The gene was apparently functional early in primate evolution and probably silenced recently before the basal primate radiation ~70 million years ago. After silencing, the primate ψβ1 pseudogene has evolved randomly in terms of base substitution and microinsertion/deletion, at a mean rate thought to be representative of non-functional non-coding DNA sequences throughout the primates. These conclusions are supported by the mode and tempo of non-coding DNA sequence evolution observed within the functional brown lemur β-globin gene. However, the tempo of primate ψβ1 gene evolution conflicts with views concerning the universal constant rate of neutral evolution, the rate of non-coding DNA evolution having apparently slowed within the different lineages of this mammalian order. The consequences for primate β-globin gene cluster evolution of the presence of a non-processed pseudogene are discussed.

The distinct nature of the ψβ1 gene in the human β-globin gene cluster, the history of the ψβ1 gene in the primates and the presence of sequences related to ψβ1 in various other mammalian orders suggests an additional ancient genetic locus was present in the ancestral β-globin gene cluster prior to the mammalian radiation, a locus renamed η. The simplest interpretation of the evolution of contemporary mammalian β-globin gene clusters, that is, that they resulted from a common minimal ancestral cluster composed of proto ε-, γ-, η-, δ- and β-like sequences, is discussed.

While the generality of the conclusions drawn from this work concerning pseudogene longevity and sequence evolution after silencing await the phylogenetic analysis of other pseudogene sequences, it is apparent that pseudogenes may constitute another potential source of genetic variation on which the processes of natural selection can act in the evolution of both eukaryotic multigene families and the genome in general.