Sequence and Comparative Analysis of the Rabbit α-Like Globin Gene Cluster Reveals a Rapid Mode of Evolution in a G+C-rich Region of Mammalian Genomes

Ross Hardison[1]†, Dan Krane[1]‡, David Vandenbergh[1]§, Jan-Fang Cheng[1]||
James Mansberger[1], John Taddie[1], Scott Schwartz[2], Xiaoqiu Huang[2]¶
and Webb Miller[2]

[1]Department of Molecular and Cell Biology
[2]Department of Computer Science and
Institute of Molecular Evolutionary Genetics
The Pennsylvania State University
University Park, PA 16802, U.S.A.

(Received 8 April 1991; accepted 24 July 1991)

A sequence of 10,621 base-pairs from the α-like globin gene cluster of rabbit has been determined. It includes the sequence of gene ζ1 (a pseudogene for the rabbit embryonic ζ-globin), the functional rabbit α-globin gene, and the θ1 psuedogene, along with the sequences of eight C repeats (short interspersed repeats in rabbit) and a J sequence implicated in recombination. The region is quite G+C-rich (62%) and contains two CpG islands. As expected for a very G+C-rich region, it has an abundance of open reading frames, but few of the long open reading frames are associated with the coding regions of genes. Alignments between the sequences of the rabbit and human α-like globin gene clusters reveal matches primarily in the immediate vicinity of genes and CpG islands, while the intergenic regions of these gene clusters have many fewer matches than are seen between the β-like globin gene clusters of these two species. Furthermore, the non-coding sequences in this portion of the rabbit α-like globin gene cluster are shorter than in human, indicating a strong tendency either for sequence contraction in the rabbit gene cluster or for expansion in the human gene cluster. Thus, the intergenic regions of the α-like globin gene clusters have evolved in a relatively fast mode since the mammalian radiation, but not exclusively by nucleotide substitution. Despite this rapid mode of evolution, some strong matches are found 5′ to the start sites of the human and rabbit α genes, perhaps indicating conservation of a regulatory element. The rabbit J sequence is over 1000 base-pairs long; it contains a C repeat at its 5′ end and an internal region of homology to the 3′-untranslated region of the α-globin gene. Part of the rabbit J sequence matches with sequences within the X homology block in human. Both of these regions have been implicated as hot-spots for recombination, hence the matching sequences are good candidates for such a function. All the interspersed repeats within both gene clusters are retroposon SINEs that appear to have inserted independently in the rabbit and human lineages.

Keywords: α-globin gene cluster; CpG islands; G+C-rich isochores;
DNA sequence alignments; evolutionary rates; recombination sequences;
short interspersed repeats

† Author to whom correspondence should be addressed.

‡ Present address: Department of Genetics, Washington University School of Medicine, 4566 Scott Avenue, St Louis, MO 63110-1095, U.S.A.

§ Present address: National Institute on Drug Abuse/ Addiction Research Center, P.O. Box 5180, Baltimore,

|| Present address: Human Genome Center, 459 Donner Laboratory, Lawrence Berkeley Laboratories, Berkeley, CA 94720, U.S.A.

¶ Present address: Department of Computer Science, Michigan Technological University, Houghton, MI 49931, U.S.A.

1. Introduction

Much work has been devoted to understanding the mechanisms involved in the co-ordinate, temporal and tissue-specific regulation of α and β-globin genes (for a review, see Collins & Weissman, 1984; Evans *et al.*, 1990; Orkin, 1990). Given that equal amounts of α-like and β-like globin must be synthesized to produce the hemoglobin tetramer, $\alpha_2\beta_2$, one might anticipate that this co-ordinate expression would be achieved by utilizing identical regulatory schemes. However, that is certainly not the case. The human α-globin gene is expressed permissively in a variety of cell lines after transfection (Mellon *et al.*, 1981), whereas the β-globin gene requires either a viral enhancer *in cis* or erythroid induction to be expressed after transfection (Banerji *et al.*, 1981; Humphries *et al.*, 1982; Wright *et al.*, 1984; Charnay *et al.*, 1984). This permissive expression of α-globin genes in non-erythroid cells is observed for genes from both human and rabbit, but not mouse (Cheng *et al.*, 1986; Whitelaw *et al.*, 1989). A dominant control element for the β-like globin gene cluster is the locus control region (LCR†) located 5 to 20 kb 5′ to the ε-globin gene (Grosveld *et al.*, 1987; Forrester *et al.*, 1987). A strong positive regulatory region has also been found 40 kb 5′ to the human ζ-globin gene (Higgs *et al.*, 1990); although the α-globin LCR shares some properties with that of the β-globin gene cluster, it is not clear that the regulatory features of these two LCRs are identical. Also, mammalian adult β-globin genes reach full induction at later stages of development than adult α-globin genes (Rohrbaugh & Hardison, 1983; Peschle *et al.*, 1985). Some of the critical sequences that account for these differences are within or 3′ to these genes (Charnay *et al.*, 1984; Wright *et al.*, 1984).

These differences in regulation may be related to the substantially different genomic contexts observed for the α and β-like globin gene clusters. After the α-like and β-like globin gene clusters moved to different chromosomes in the progenitor to birds and mammals (for a review, see Collins & Weissman, 1984; Hardison, 1991), they evolved into very different segments of the genome in some mammalian lineages. For example, the G+C-rich α-like gene cluster in humans contains several CpG-rich islands (Fischel-Ghodsian *et al.*, 1987) that are never methylated (Bird *et al.*, 1987), and this gene cluster is found in the most dense (most G+C-rich) isochore in both human and rabbit genomes (Bernardi *et al.*, 1985). Isochores are very long (probably thousands of kb) segments of homogeneous base composition that may correspond to the Giemsa light and dark bands seen in metaphase chromosomes (Bernardi, 1989). Unlike most tissue-specific genes, the α-like globin gene clusters are replicated early in S phase in both erythroid and non-erythroid cells (Calza *et al.*, 1984; Goldman *et al.*, 1984). These characteristics contrast with the A+T-rich and CpG-deficient β-like gene cluster, which has been shown to be methylated in non-erythroid tissues in human (van der Ploeg & Flavell, 1980) and rabbit (Shen & Maniatis, 1980). The β-like globin gene clusters, like the bulk of mammalian genomic DNA, are present in low-density isochores (Bernardi *et al.*, 1985), and like most tissue-specific genes, they are replicated late in S phase in non-erythroid cells but early in erythroid genes (Dhar *et al.*, 1988; Epner *et al.*, 1988).

The correlation between striking differences in genomic context and in types of regulation suggests that a detailed comparison of both these gene clusters between mammalian species would be productive in generating insights into their regulation. Indeed, analyses of the extensively sequenced human and rabbit β-like globin gene clusters (73·4 and 44·6 kb, respectively) reveal long stretches of sequence similarity that extend through and even between each of the β-like globin genes (Margot *et al.*, 1989). The sequence similarity between the mouse and human β-like globin gene clusters is less extensive (Shehee *et al.*, 1989), as is expected given the more rapid rate of evolution in rodents (Wu & Li, 1985). One notable segment of extragenic sequence, located about 6 kb 5′ to the human ε-globin gene, has been highly conserved between human, rabbit and mouse (Hardison, 1991), and it has been shown to he part of the LCR of the β-like globin gene cluster (Forrester *et al.*, 1987; Grosveld *et al.*, 1987). A similar comparison is carried out in this paper for the α-like globin gene clusters of rabbit and human.

The rabbit α-like globin gene cluster is located in a Geimsa-light band at the terminus of the long arm of chromosome 6 (Xu & Hardison, 1991). The minimal gene cluster includes one adult α-globin gene, five homologs to the embryonic ζ-globin gene, and two θ-globin pseudogenes, arranged in the order 5′-ζ0-ζ1-α-θ1-ζ2-ζ3-θ2-ζ4-3′ within a 38 kb DNA segment (Cheng *et al.*, 1986, 1987, 1988). This gene cluster probably evolved by a duplication of a large DNA block containing the ζ-ζ-α-θ gene set, followed by deletion of the α-globin gene in the 3′ duplicated ζ2-ζ3-θ2 gene set (Cheng *et al.*, 1987). The rabbit α-like globin gene cluster is highly polymorphic both for the number of duplicated gene sets as well as for restriction fragment lengths around ζ0 and ζ1 (Cheng & Hardison, 1988). A similar sequence is found at the breakpoints proposed for the recombinations involved in duplications of ζ, block duplications of ζ-ζ-α-θ, and deletion of α; this common junction sequence is called a J sequence (Cheng *et al.*, 1987). Part of the J sequence is very similar to the 3′-untranslated sequence of the α-globin gene, and this homology is likely to have been involved in the recombination that deleted the α-globin gene from the ζ2-ζ3-θ2 gene set. The deletions that occurred frequently during the propagation of λ clones carrying rabbit genomic DNA containing this

† Abbreviations used: LCR, locus control region; kb,

(Cheng *et al.*, 1987), arguing that these sequences could constitute a hot-spot for recombination. This gene cluster contains at least one active gene, the adult α-globin gene (Cheng *et al.*, 1986), and the gene ζ0 is the most likely candidate to encode the embryonic ζ-globin found in rabbit. The remaining ζ-globin genes appear to be pseudogenes (Cheng *et al.*, 1988).

The α-like globin gene cluster in human is located very close to the telomere of the short arm of chromosome 16 in a segment of very G+C-rich DNA that continues as far as 2000 kb (Harris *et al.*, 1990). The cluster contains a functional ζ2 gene encoding an embryonic ζ-globin polypeptide, a non-functional ζ1 gene that is only slightly divergent from ζ2, a highly divergent ψα2 pseudogene, a moderately divergent pseudogene ψα1 that has lost its CpG island, duplicated functional adult α-globin genes α2 and α1, and a θ gene that produces transcripts, but for which no polypeptide product has been identified (for a review, see Higgs *et al.*, 1989). The genes are arranged in the order 5′-ζ2-ζ1-ψα2-ψα1-α2-α1-θ-3′. The α-globin genes were duplicated in the stem simians (Sawada & Schmid, 1986), and the 5′ α gene found in several other mammals (Schon *et al.*, 1982) is orthologous to the human ψα1 gene, based on sequence similarities in the 5′ flank (Hardison & Gelinas, 1986; Sawada & Schmid, 1986). The duplication of α genes in higher primates has left a long homology block of about 4 kb that is divided into three regions called X, Y and Z. A sequence that confers an enhanced rate of recombination in COS cells has been mapped to the first 300 bp of the X region (Hu & Shen, 1987). Almost 20 kb of continuous sequence has been determined from the human α-like globin gene cluster (see Materials and Methods), encompassing the region from ζ1 through θ. This sequence, along with that reported for rabbit in this paper, allows a comprehensive comparison of a major portion of the gene clusters in rabbit and human, including the three major members of the α-like globin gene cluster, ζ, α and θ. Parallels and differences are discussed for these interspecies comparisons of α-like and β-like globin gene clusters.

2. Materials and Methods

(a) *Determination of DNA sequence*

Clones of rabbit DNA containing the α-like globin gene cluster were isolated from a library of rabbit genomic DNA (Maniatis *et al.*, 1978; Cheng *et al.*, 1986, 1987, 1988). The recombinant phage λRαG1 containing the genes ζ1, α and θ1 was used to generate restriction fragments that were subcloned into plasmids pBR322, pUC or pBluescript, and into the phage M13. Most of the DNA sequence was determined by the dideoxynucleotide chain termination method (Sanger *et al.*, 1977) and was confirmed in some regions with the base-specific chemical degradation method (Maxam & Gilbert, 1980). In some cases, directed deletions for rapid sequence determination were constructed by using exonuclease III and mung bean nuclease (Henikoff, 1984). The strategies employed in sequence was determined from both strands. The sequence was not determined through 5 restriction sites that were the ends of fragments used to construct subclones. These sites are internal to either C repeats or J sequences. In 3 cases, *Bgl*II*, *Pst*I* and *Bam*HI* in C47 (Fig. 1), the sequences around the restriction site match with similar sequences repeated elsewhere in the gene cluster or genome, hence it is unlikely that any sequence is missing. The short segment from 4278 to 4283 was highly compressed on the gels and hence this sequence could not be determined unambiguously. The new sequences were combined with previously determined sequences (Cheng *et al.*, 1986, 1987, 1988; Krane *et al.*, 1991) to generate a composite sequence extending from the 5′ flank of ζ1 to the 3′ flank of θ1. The ζ1-α-θ1 sequence is available as GenBank accession number M35026, and the sequence of Jθ1 is available as EMBL accession number X60985.

A composite file of sequences from the human α-like globin gene cluster was assembled from data in the following sources: the 5′ flank of gene ζ1 (Willard *et al.*, 1985), gene ζ1 (Proudfoot *et al.*, 1982), pseudogene ψα2 (Hardison *et al.*, 1986), intergenic sequence between ζ1 and ψα1 (Sawada *et al.*, 1983), pseudogene ψα1 (Proudfoot & Maniatis, 1980), homology blocks containing α2 and α1 (Liebhaber *et al.*, 1980; Michelson & ●rkin, 1980, 1983; Hess *et al.*, 1983, 1984), 3′ flank of α1 (Hardison & Gelinas, 1986), intergenic sequence between α1 and θ1 (Bailey, 1990), and θ1 (Hsu *et al.*, 1988).

(b) *Analysis of the DNA sequence*

Direct and inverted repeats, open reading frames and nucleotide strings were identified with the computer program DNA Inspector IIe (Textco) running on a Macintosh computer. Plots of G+C richness and CpG and GpC dinucleotides were made from the output of the BASIC computer program "Di-nt Frequency" (Krane, 1990) scanning windows 50 bp in length.

Local alignments of the 2 sequences were generated using the program SIM (Huang *et al.*, 1990), run on a Sun4 workstation. SIM generates alignments between very long DNA sequences while using computer space efficiently, and it produces alignments that are optimized to parameters set by the user. All alignments discussed in this paper were obtained where matches count 1, mismatches count −1, the gap-open penalty is 4·0, and the gap-extension penalty is 0·4 per nucleotide. With the single exception of Fig. 7, the number of local alignments to be used was determined by using theoretical results (Karlin & Altschul, 1990) on the expected number of gap-free alignments. Specifically, we used only those alignments whose score exceeds a threshold τ, defined so that the probability is 0·8 that random sequences matching the given sequences in length and in nucleotide composition have a gap-free alignment scoring at least τ. Informally speaking, τ is a threshold where we expect that 2 random sequences of the given length and composition would exhibit 1 or several gap-free alignments, i.e. a dot-plot of random sequences at these criteria would contain a few specks. The large number of local alignments generated by SIM were organized and viewed by a graphical user interface called LAV (local alignment viewer; Schwartz *et al.*, 1991). Figs 6, 7 and 8 were drawn directly from the SIM alignments and from hand-generated files giving positions of sequence features (such as exons, introns and repeated sequences) using the program LAD (local align-

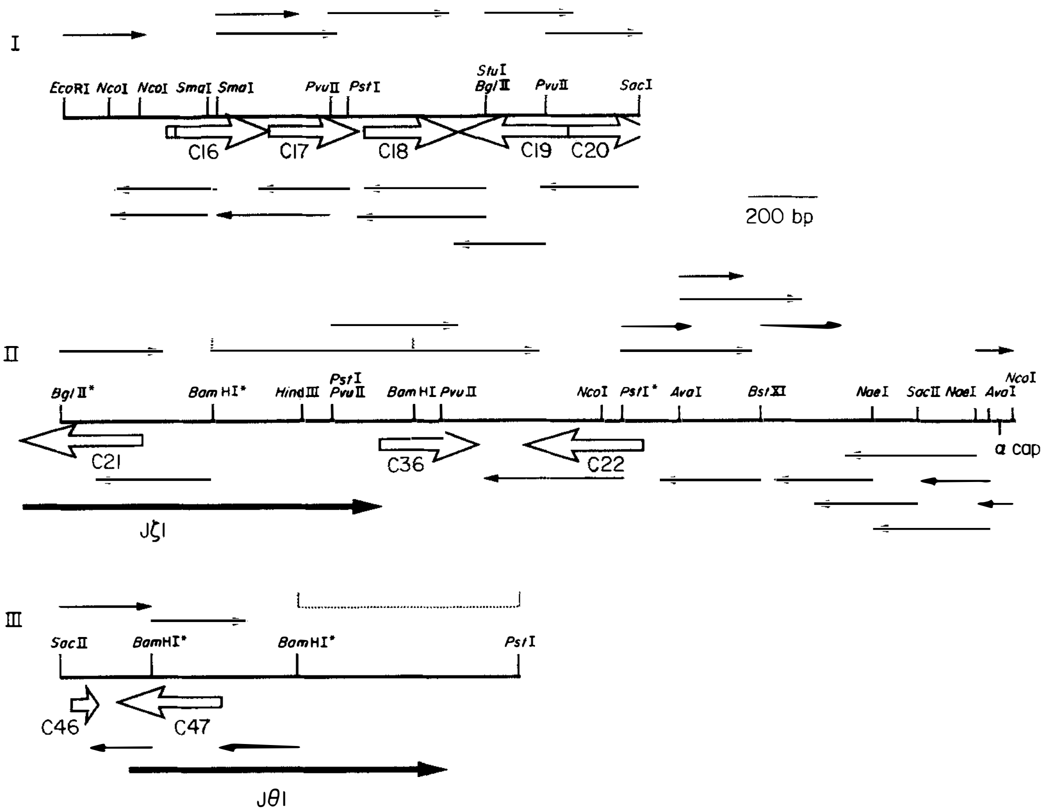

Figure 1. Strategies used to determine the DNA sequence of regions I, II and III. Arrows above the line indicate the extent of individual readings of the top strand, and arrows below the line correspond to readings of the lower strand. The dotted lines drawn above parts of regions II and III cover the sequence previously determined by Cheng *et al.* (1987). Open arrows indicate the positions of C repeats within the gene cluster; they point in the direction of their poly(A) tracts. Asterisks mark internal restriction sites through which the sequence has not been determined.

To obtain a quantitative estimate of the fraction of the α-like or β-like globin gene clusters that match between rabbit and human (see Results, section (h)), local alignments were optimally chained together to make "meta-alignments" using an algorithm for computing optimal paths in a directed acyclic graph (Corman *et al.*, 1990). In the meta-alignments, 1 SIM alignment follows another only if its starting positions in the 2 sequences follow the ending positions of the other alignment. The chaining was done so as to maximize the number of matches in the meta-alignment. For the rabbit and human α-globin genes, the divergence was determined from the aligned sequences and corrected for multiple substitutions at a single site (Jukes & Cantor, 1969) to obtain the number of substitutions per site. The time of divergence between rabbit and human was taken as the time of the mammalian radiation, about 80 million years ago (Romero-Herrera *et al.*, 1973).

3. Results

(a) *Nucleotide sequence of the rabbit α-like globin gene cluster*

A diagram of the portion of the rabbit α-like globin gene cluster isolated in 38 kb of cloned DNA (Cheng *et al.*, 1986, 1987, 1988) is shown in Figure 2. Analysis of a population of laboratory rabbits by genomic blot-hybridization shows that the gene

result of additional duplications of the ζ-ζ-θ gene set (Cheng & Hardison, 1988). The homology blocks containing ζ and θ genes (Z blocks and T blocks, respectively) are bounded by a characteristic junction sequence called a J sequence (Fig. 2). In this paper they will be referred to by the name of the gene that they follow, e.g. Jζ1 is 3′ to gene ζ1. As will be explained below, the J sequences extend from the C repeat at the 5′ end through a sequence homologous to the 3′ portions of the α-globin gene (Cheng *et al.*, 1987).

New sequence data (Fig. 1) were combined with previously published sequences to make a contiguous sequence of 10,621 bp, beginning 2273 bp 5′ to ζ1, extending through the α-globin gene and ending 204 bp 3′ to the polyadenylation signal of the θ1 pseudogene. It is available from the GenBank database under accession number M35026. This three-gene set contains homologs to each of the three α-like globin genes found in mammalian species, and it contains most of the DNA in the basic set of genes that has duplicated to evolve this gene cluster.

(b) *Short, interspersed C repeats*

The rabbit α-like globin gene cluster contains at

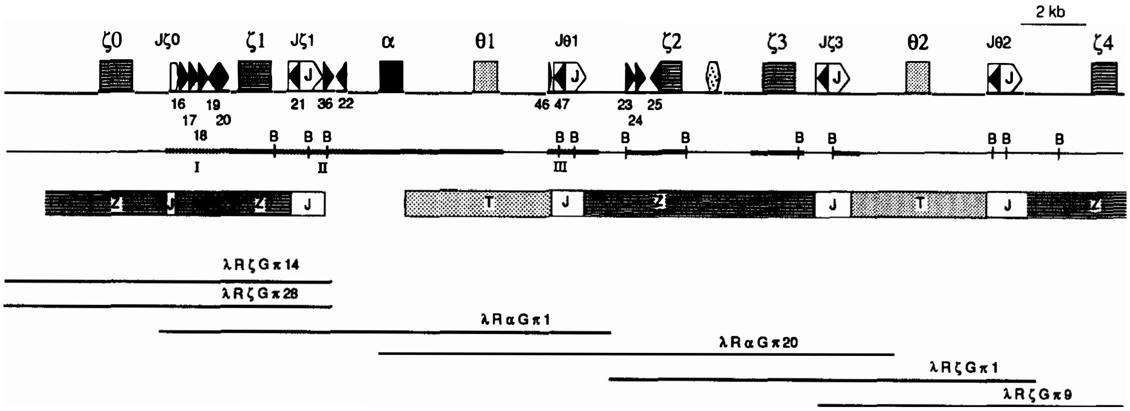

Figure 2. Organization of the rabbit α-like globin gene cluster. Related genes are represented by boxes with the same shading, C repeats are shown as filled triangles and J sequences are shown as open, pointed boxes. A repetitive sequence found by hybridization experiments (Cheng *et al.*, 1987) is shown as a stippled polygon between genes ζ2 and ζ3; current data do not exclude the possibility that it is one or several divergent C repeats. *Bam*HI (B) sites within the gene cluster are indicated on the 2nd line, and sequenced segments are shown as thick regions on this line. The new sequences reported in this paper are shaded on the 2nd line. Boxes below the gene map show the T homology blocks containing θ genes and Z homology blocks containing ζ genes separated by junction, or J, sequences. Horizontal lines below the homology blocks indicate the positions of λ clones isolated from this region of the rabbit genome by Cheng *et al.* (1987, 1988).

spersed repeat in the rabbit genome (Cheng *et al.*, 1984; Hardison & Printz, 1985). These repeats tend to insert into or nearby one another in groups (Krane *et al.*, 1991); this is readily apparent in the 5′ flanks of genes ζ1 and ζ2 (Fig. 2), although the segment from α through θ1 has remained free of C repeats. A total of eight C repeats are in the ζ1-α-θ1 sequence, comprising 2·7 of the 10·6 kb, or 25% of the contiguous sequence. At least seven additional C repeats have also been detected by sequence analysis and hybridization studies in the remainder of the gene cluster (Fig. 2), and it is likely that the unidentified repeats between ζ2 and ζ3 include more C repeats and possibly a J sequence. In contrast, no members of the predominant long interspersed family of repeated DNA, L1Oc (Demers *et al.*, 1989), have been found in the α-like globin gene cluster either by sequence determination or by hybridization studies (Cheng *et al.*, 1987).

(c) *Base composition*

The base composition of the ζ1-α-θ1 DNA sequence is 62% G+C and 38% A+T; this is essentially the reverse of the values reported for the whole rabbit genome, 44% G+C and 56% A+T (Sober, 1968), or for the rabbit β-like globin gene cluster, 39% G+C and 61% A+T (Margot *et al.*, 1989). The high G+C content of the α-like globin gene cluster is uniform over most of the gene cluster (Fig. 3(a)), and regions of greater than 65% G+C content can be found throughout the sequence, not just in close association with functional genes. This is in striking contrast to the β-like globin gene cluster, which has a very low G+C content throughout, exemplified by the δ-β region shown in Figure G+C-rich C repeat, C14, from a recently transposing subfamily (Krane *et al.*, 1991). Seven segments of the rabbit α-like globin gene cluster are noticeably high in A+T content (>60%) relative to the remainder of the cluster (Fig. 3(a)), but in five of these cases the A+T richness is derived from the poly(A) and $(CT)_n$ tracts found at the 3′ end of C repeat sequences (Krane *et al.*, 1991) that have transposed into this region of the rabbit genome. As previously noted (Cheng *et al.*, 1986), the longest A+T-rich stretch (between α and θ1) is flanked by 10 bp-long inverted repeats, suggesting that it too may have entered this gene cluster by a transposition event.

As expected for a sequence with a low A+T content, many open reading frames are observed on both strands (Fig. 4). Some of the open reading frames correspond to the exons of the α-globin gene, the only functional gene in this region, but most do not correspond to regions that encode known polypeptides. This illustrates the difficulty in identifying potential genes by mapping open reading frames in G+C-rich sequences.

(d) *CpG islands*

Similar to the situation in human (Bird *et al.*, 1987; Fischel-Ghodsian *et al.*, 1987), the rabbit α-like globin gene cluster contains CpG islands, whereas the β-like globin gene cluster does not. The rabbit α-globin gene has many CpG dinucleotides in its 5′ flank and internally (Fig. 3(a)). This abundance of CpGs is much greater than is seen for gene ζ1, which is equally rich in G+C content; thus, the cluster of CpG dinucleotides in the α-globin gene does not simply result from a high G+C content.

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | sales@docketalarm.com | 1-866-77-FASTCASE