GENOMICS 13, 741–760 (1992)

The β Globin Gene Cluster of the Prosimian Primate *Galago crassicaudatus:* Nucleotide Sequence Determination of the 41-kb Cluster and Comparative Sequence Analyses

DANILO A. TAGLE,*,[1] MICHAEL J. STANHOPE,† DAVID R. SIEMIENIAK,‡,[2] PHILIP BENSON,† MORRIS GOODMAN,† AND JERRY L. SLIGHTOM†,‡

*Departments of *Molecular Biology and Genetics and †Anatomy and Cell Biology, Wayne State University School of Medicine, Detroit, Michigan 48201; and ‡Molecular Biology Unit 7242, The Upjohn Company, Kalamazoo, Michigan 49007*

Received November 1, 1991; revised March 27, 1992

The nucleotide sequence of the β globin gene cluster of the prosimian *Galago crassicaudatus* has been determined. A total sequence spanning 41,101 bp contains and links together previously published sequences of the five galago β-like globin genes (5′–ε–γ-ψη–δ–β–3′). A computer-aided search for middle interspersed repetitive sequences identified 10 LINE (L1) elements, including a 5′ truncated repeat that is orthologous to the full-length L1 element found in the human ε–γ intergenic region. SINE elements that were identified included one *Alu* type I repeat, four *Alu* type II repeats, and two methionine tRNA-derived Monomer (type III) elements. *Alu* type II and Monomer sequences are unique to the galago genome. Structural analyses of the cluster sequence reveals that it is relatively A + T rich (about 62%) and regions with high G + C content are associated primarily with globin coding regions. Comparative analyses with the β globin cluster sequences of human, rabbit, and mouse reveal extensive sequence homologies in their genic regions, but only human, galago, and rabbit sequences share extensive intergenic sequence homologies. Divergence analyses of aligned intergenic and flanking sequences from orthologous human, galago, and rabbit sequences show a gradation in the rate of nucleotide sequence evolution along the cluster where sequences 5′ of the ε globin gene region show the least sequence divergence and sequences just 5′ of the β globin gene region show the greatest sequence divergence.

INTRODUCTION

The different developmentally expressed β-type globin genes of primates and other mammals can be traced back to a single progenitor gene, which tandemly duplicated some 150 to 200 millon years ago (MYA) in the early mammals. By about 110 to 135 MYA, the two gene lines from the tandem duplication had differentiated into an embryonically expressed locus (proto-ε) and an ontogenetically later expressed locus (proto-β) (Efstratiadis *et al.*, 1980; Czelusniak *et al.*, 1982; Koop and Goodman, 1988). Further tandem duplications of the embryonic 5′ locus and adult 3′ locus, in the early eutherian mammals (85–100 MYA), led to a genomic domain of five developmentally regulated loci (5′–ε–γ–η–δ–β–3′). In this β globin gene cluster, ε, γ and η originated from the 5′ proto-ε locus and were embryonically expressed, while δ and β originated from the 3′ proto-β locus and were expressed in the non-embryonic, later ontogenetic stages of life (Hardies *et al.*, 1984; Hardison, 1984; Hill *et al.*, 1984; Goodman *et al.*, 1984, 1987). This ancestral eutherian β globin gene cluster underwent varying degrees of change during the evolution of different eutherian orders. For example, lagomorph (rabbit) and rodent (mouse) β globin gene clusters lack the η globin locus whereas artiodactyl (goat, sheep and ox) clusters lack the γ globin locus. However, each primate β globin gene cluster so far examined contains sequences from all five loci (ε,γ, η, δ, β) in the ancestral 5′ to 3′ arrangement.

The β globin gene cluster in mammals ranges in length from approximately 20 kb in lemurs to about 90 kb in goats (Harris *et al.*, 1984, 1986; Townes *et al.*, 1984). The complete nucleotide sequence of the β globin gene cluster of human (Collins and Weissman, 1984; Li *et al.*, 1985), rabbit (Margot *et al.*, 1989), and mouse (Shehee *et al.*, 1989) have been determined, and these sequences reveal that only about 10% of the DNA in a mammalian β globin gene cluster encodes globin mRNAs.

Very few orthologous sequence data sets that compare noncoding DNAs from a series of species are presently available. Of those that do exist, the most extensive are from closely related simian primates; they involve intergenic sequences that flank the ψα globin genes (Sawada *et al.*, 1985), noncoding sequences that include and flank

Sequence data from this article has been deposited with the GenBank Data Library under Accession No. M73981.

[1] Present addresses: Department of Human Genetics, 4562 MSRB II, The University of Michigan Medical Center, Ann Arbor, MI 48109-0650.

[2] Howard Hughes Institute 4522 MSRB I, The University of Michigan Medical Center, Ann Arbor, MI 48109-0650.

0888-7543/92 $5.00

the $\gamma\eta$ globin gene (Bailey *et al.*, 1991; Fitch *et al.*, 1988; Koop *et al.*, 1986; Maeda *et al.*, 1988; Miyamoto *et al.*, 1987), and δ–β intergenic sequences (Savatier *et al.*, 1985, 1987). Consequently, much still remains to be learned about the sequence features, evolution, and functional constraints that act upon intergenic noncoding sequences. Comparative analyses of the β globin gene cluster sequences from distantly related species can be used to identify evolutionarily conserved DNA elements or phylogenetic footprints, which can provide insights into the evolution of gene clusters at the molecular level (Tagle *et al.*, 1988; Gumucio *et al.*, 1991).

In this context, we decided to determine and analyze the complete nucleotide sequence of the β globin gene cluster from a distantly related primate, the prosimian galago (*Galago crassicaudatus*). Fossil evidence indicates that the prosimian (galago)/simian (human) divergence time dates back to about 55 MYA (Fleagle, 1988). Here we report 41,101 bp of continuous sequence spanning the entire β globin cluster of the galago. This cluster contains five globin-like genes that are arranged in the order of their developmental expression: 5′–ϵ–γ–(embryonic)–$\psi\eta$–(nonexpressed)–δ–β–(fetal and postnatal)–3′ (Tagle, 1990; Tagle *et al.*, 1988, 1991). An analysis of the compositional and structural features of the complete cluster sequence of galago is presented. We have identified the DNA sequences that are orthologous (derived from the same DNA sequence in the last common ancestral species) among galago, human, rabbit, and mouse β globin gene clusters and the locations of short and long interspersed repetitive nuclear elements (SINEs and LINEs) within the galago cluster. Differences in the degree of sequence divergence among the orthologous noncoding regions of galago, human, and rabbit have also been determined.

MATERIALS AND METHODS

Cloning and nucleotide sequencing. Twelve overlapping recombinant Charon 35 phage clones that span the galago β globin gene cluster (Fig. 1) were previously isolated and described (Tagle, 1990; Tagle *et al.*, 1988, 1991). Phage clones λ Gcr 18.3, λ Gcr 11.9, λ Gcr 15.4, λ Gcr 16.1, λ Gcr 15.2A, and λ Gcr 15.2B were used to generate a complete and ordered set of *Eco*RI fragments (R1 to R22) that were subcloned into pUC-18 (Yanisch-Perron *et al.*, 1985). Plasmid subclones R5 to R22 were sequenced by the dideoxynucleotide chain termination method (Sanger *et al.*, 1977) and/or by the chemical cleavage method (Maxam and Gilbert, 1980). In the latter method, restriction fragments from R7, R9, R12, and parts of R8 and R10 were end-labeled and sequenced as described in Tagle *et al.* (1988) for the ϵ and γ globin genes and in Koop *et al.* (1989) for the $\psi\eta$ gene. The sequences of the galago δ and β globin genes have been presented by Tagle *et al.* (1991). These previously published galago β globin gene sequences represent only about 9 kb of the β globin gene cluster sequence presented here. The remaining intergenic sequences were obtained by the dideoxynucleotide chain termination method using initially universal forward and reverse vector primers and then by using synthetic oligomers that were designed from the previously determined sequence (referred to as oligomer walking). Nucleotide sequence readings ranged from 400 to 950 bp. Approximately 95% of the intergenic nucleotide sequences were determined on both DNA strands, and the remaining 5% were ary structures were resolved by sequencing at a higher temperature (55°C) using *Taq* DNA polymerase. Selected *Bam*HI- and *Hin*dIII-generated fragments (not shown) were also subcloned into pUC-18 and sequenced to verify nucleotide sequence contiguity across the cloning site junctions of certain *Eco*RI clones.

Computer-aided analyses of nucleotide sequences. Base composition and other sequence features [such as open reading frames (ORFs), strand asymmetry, subsequence breakdown, and repetitive elements] of the galago β globin gene cluster were identified and analyzed using the DNA analysis computer programs available from The Genetics Computer Group package (GCG; Madison, WI; Devereux *et al.*, 1984) and the MacVector Package Version 6.6 from IBI Technologies (New Haven, CN). In addition to the ORF search, the galago cluster sequence was submitted to GRAIL (Gene Recognition and Analysis Internet Link; Oak Ridge, TN; Uberbacher and Mural, 1991) for additional searches of potential protein coding regions. The location of shared nucleotide sequence identities between the galago β globin gene cluster sequences and human (Collins and Weissman, 1984; Li *et al.*, 1985), rabbit (Margot *et al.*, 1989), or mouse (Shehee *et al.*, 1989) β globin gene cluster sequences were first identified by dotplot comparisons using the GCG computer program COMPARE. Homology plots of the galago globin gene cluster sequence with itself were used to locate repeated regions. In addition, identity searches using SINE (Daniels and Deininger, 1983, 1985, 1991; Daniels *et al.*, 1983) and L1 (Hattori *et al.*, 1986; Scott *et al.*, 1987) consensus sequences were used to further identify and define the boundaries of these repetitive elements.

Pairwise alignment among human, galago, and rabbit β globin gene cluster sequences in their matching regions were obtained using the alignment programs of Smith and Waterman (1981), Wilbur and Lipman (1983), Lipman and Pearson (1985), and Needleman and Wunsch (1970). These aligned sequences are available in diskettes from the authors upon request. In all cases, gaps were inserted into the alignments to increase sequence identities. Due to the conserved nature of the orthologous gene loci, alignments in these regions were used as anchor points to align the more diverged intergenic regions. The Local Alignment Diagrammer program (LAD; Schwartz *et al.*, 1991) was used to display the aligned sequences as plots. Each plot depicts pairwise interspecies alignments computed by the SIM program (Huang *et al.*, 1990) with a score of 1 for matches, −1.5 for mismatches, −6 for opening a gap, and −0.2 for each symbol in the gap. An alignment is displayed only if its score exceeds a threshold (τ), chosen so that the probability is 0.05 that random sequences matching the given sequences in length and nucleotide composition have a gap-free alignment scoring of at least τ.

Pairwise divergence values for the aligned sequences were calculated following the method of Nei and Gojobori (1986). Nucleotide substitutions (both transitions and transversions) and gaps (regardless of length) were counted as single events. Divergence values were corrected for hidden, superimposed substitutions by the method of Jukes and Cantor (1969).

RESULTS AND DISCUSSION

Organization and Nucleotide Sequence of the Galago β Globin Gene Cluster

A schematic diagram showing the organization of the galago β globin gene cluster is shown in Fig. 1A. The 12 overlapping recombinant phage clones used to reconstruct the structure of the cluster are also shown. Restriction enzyme site maps of these 12 clones revealed the extent of overlaps for each clone. Southern blot analysis of *Eco*RI restriction digests of the phage clones localized the five linked β-like globin genes and was confirmed by

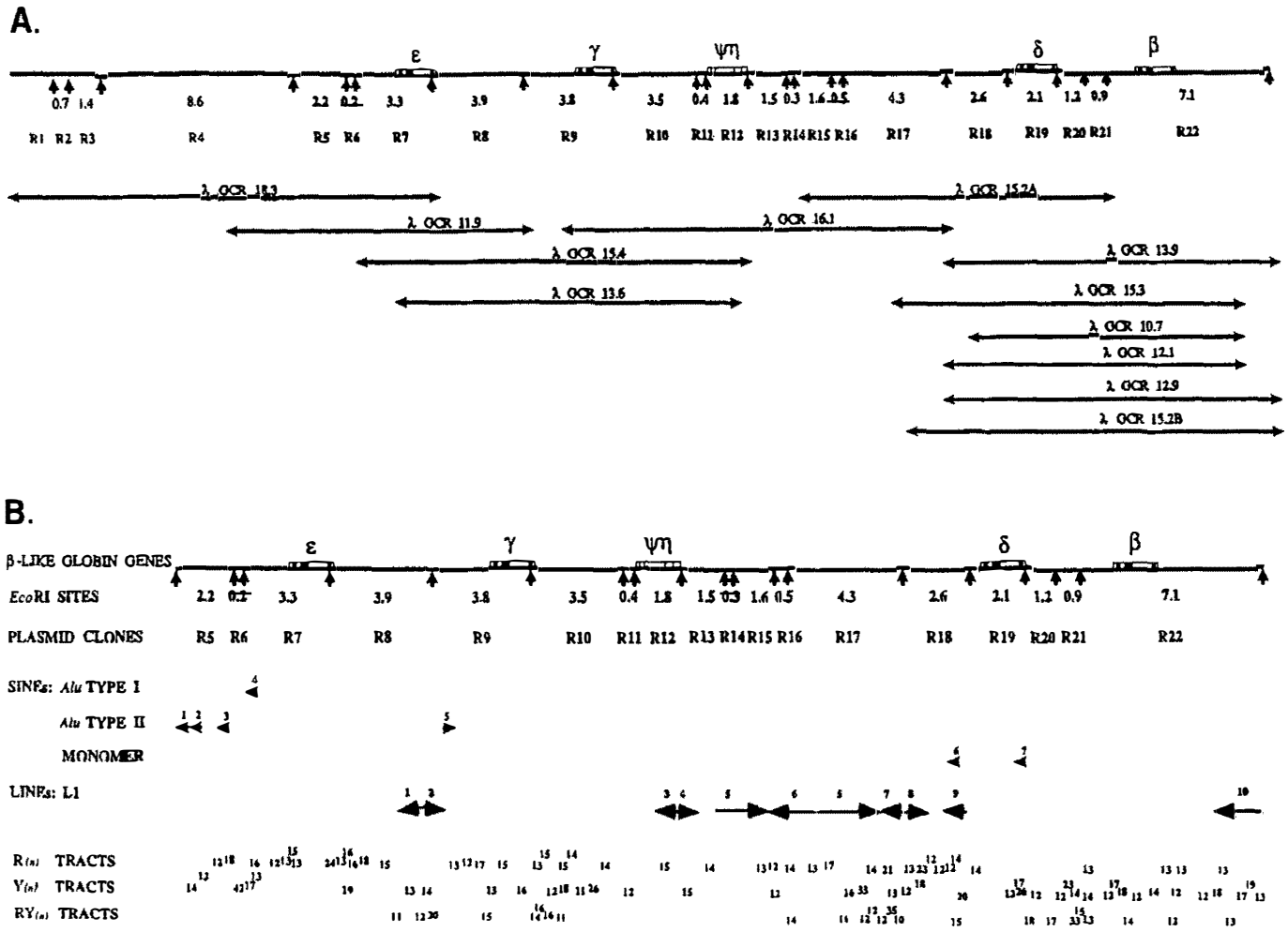


FIG. 1. (A) Overview of the galago β globin gene cluster. The top most line shows the organization of the galago β globin gene cluster. The β-like globin genes are denoted by the rectangles and are labeled on top according to their orthologous relationship with known mammalian globin genes. The approximate position of the three exons separated by two introns and their 5′ and 3′ untranslated region are indicated inside the rectangles where the filled areas denote the coding regions. *Eco*RI restrictions sites along the cluster are indicated by arrows below the line and the size of each *Eco*RI fragment is listed (kilobases) between the arrows. The ordered set of *Eco*RI fragments that were subcloned into pUC-18 and numbered 1 through 22 are shown below the cluster map. The 12 overlapping recombinant λ clones used to reconstruct the cluster organization are also shown below the cluster map. The length and position of the λ phage clones correspond directly to the galago linkage map. (B) Synopsis of sequence features of the galago β globin gene cluster. The extent of the galago β globin gene cluster region that was sequenced (*Eco*RI fragments 5 to 22) is indicated in the top line. The β-like globin genes are represented by the rectangles on the cluster map. The position, direction, and length of the interspersed repeat elements (SINEs and LINEs) are shown below the cluster map. Left-pointing arrowheads indicate that the repeat is oriented in the 3′ to 5′ direction relative to the direction of globin transcription. Right-pointing arrowheads for the repeats indicate a 5′ to 3′ direction. The locations of purine $(R)_n$ and pyrimidine $(Y)_n$ tracts (where $n \geq 10$) and of alternating purine/pyrimidine $(RY)_n$ (where $n \geq 5$) tracts are indicated by numbers corresponding to their lengths.

firmed by sequence orthology with other known primate and mammalian globin genes previously presented and characterized in Tagle *et al.* (1988) for the ε and γ globin genes, Tagle *et al.* (1991) for the δ and β globin genes, and Koop *et al.* (1989) and Bailey *et al.* (1991) for the ψη globin gene locus.

The nucleotide sequences of the remaining intergenic regions of the galago β globin gene cluster were determined from the ordered set of *Eco*RI plasmid subclones R5 through R22 that span the galago cluster (Fig. 1B). The complete sequence of the galago β globin gene cluster is presented in Fig. 2. In general, the organization and position of the β-type genes in the galago globin gene (Efstratiadis *et al.*, 1980; Hardies *et al.*, 1984; Hardison, 1984; Harris *et al.*, 1984; Hill *et al.*, 1984; Goodman *et al.*, 1984, 1987). A general scheme that depicts some of the major elements and molecular events that have occurred in the evolution of the mammalian β globin gene cluster is presented in Fig. 3. The entire sequenced region of the galago β globin gene cluster spans 41,101 bp and includes sequences starting 4.3 kb 5′ of the ε gene, intergenic DNAs spanning 6.1, 3.8, 11.1, and 3.1 kb between the ε and γ, γ and ψη, ψη and δ, and δ and β globin genes, respectively, and extending 4.7 kb 3′ of the β globin gene.

Galago Globin Genes

four expressed genes exhibit the basic exon–intron structure of globin genes consisting of three exons separated by two intervening sequences (reviewed in Collins and Weissman, 1984), and each is structurally able to encode the 146-residue globin polypeptide (Tagle *et al.*, 1988, 1991). Codon usage for these expressed galago β globin genes was analyzed (Tagle, 1990), and there appears to be the same codon usage bias for the amino acids, Leu (the codon CTG is preferred 51/67), Val (the codon GTG is preferred 39/63), and Gly (the codon GGC is preferred 22/47), as that found for human globin genes. This codon bias has also been observed for the genes of other mammalian and avian species that are available in GenBank (Wada *et al.*, 1991). The preference for C and G in the third codon position is consistent with the prevalence of most mammalian and avian genes in GC-rich genomic regions (Wada *et al.*, 1991).

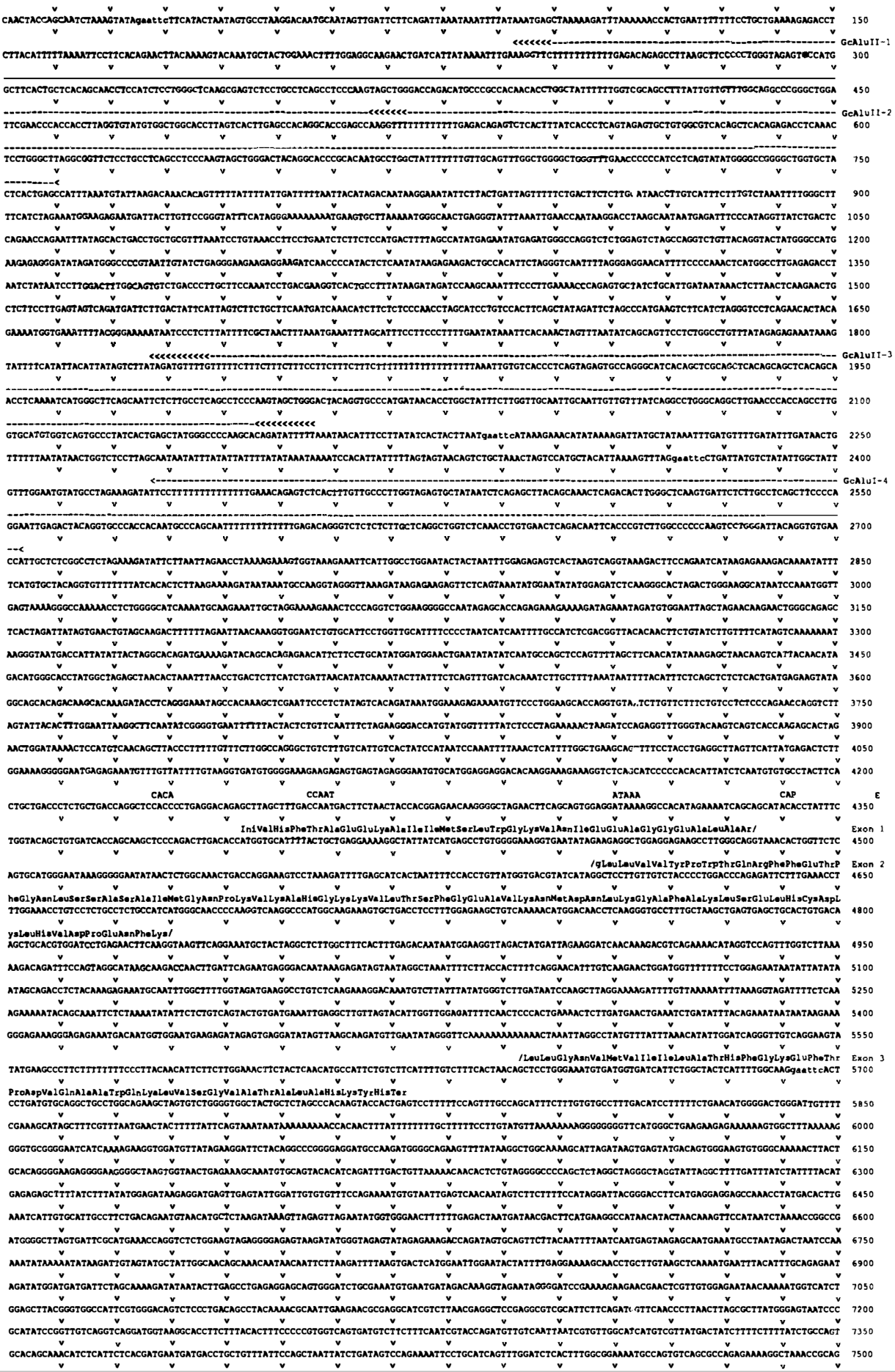
Like all other primate $\psi\eta$ globin genes studied thus far, structural anomalies have maintained the galago $\psi\eta$ locus as an nonfunctional gene or pseudogene. These anomalies include two deletions (that resulted in a frame shift and six in-frame termination codons) and the loss of consensus intron splice junction sequences for introns 1 and 2 (Tagle, 1990; Bailey *et al.*, 1991). In addition to having deleterious mutations in its coding and 5′ regulatory regions (Tagle, 1990; Bailey *et al.*, 1991), the galago $\psi\eta$ gene is truncated by the insertion of L1 elements in intron 2 (discussed below).

Search for Other Protein Coding Regions

The galago β globin gene cluster sequence was searched by computer for open reading frames that are at least 30 amino acids in length [enough to detect the smallest globin exon (exon 1) of 91 bp] and that follow Fickett's criteria (Fickett, 1982) of G + C base composition (window of 200 with a probability of 0.92 or greater). This conventional search routine identified a total of 354 ORFs (Fig. 4). Of these, 11 of the 16 ORFs that are each greater than 300 bp in length are associated with a functional globin gene (ϵ, γ, δ, or β). The longest ORF in the cluster is found in the β globin gene, consists of 576 bp that begin 34 bp into intron 1, and extends through 225 bp into intron 2. The second most extensive ORF (489 bp) is in the δ globin gene. It also starts 34 bp into intron 1 and extends through 141 bp into intron 2, and the similarity of this ORF with that of the β globin genes is due to concerted evolution between the two loci (Tagle *et al.*, 1991). Of the remaining five ORFs greater than 300 bp, three are associated with L1 elements (positions 20,796 to 21,128; 20,897 to 21,196 and 27,828 to 27,493 of Fig. 2). Another ORF is associated with the *Alu* element (GcAluII-3), which shows an ORF (positions 1847 to 2168 of Fig. 2) throughout its entirety and may represent a relatively recent insertion event (discussed below). The remaining ORF at positions 32,071 to 32,391 (on the complementary strand, strand not presented in Fig. 2) appears not to be associated with any known structural feature of the β globin gene cluster sequence (i.e., with globin genes, repetitive elements, or simple repeat sequences). A search of the translated sequences of this ORF against GenBank did not reveal homology with any known gene.

While the above search routine of examining the positional and compositional bias of the sequence for ORFs correctly identified all the expressed galago β-like globin genes, the background noise was too high. The galago sequence was analyzed by GRAIL for protein coding regions. This search routine combines a set of seven sensor algorithms that measure important attributes of coding DNA versus noncoding DNA (such as statistical frame bias, Fickett's base composition, dinucleotide frequencies, and coding six-tuple word preferences) on a sliding window of 100 bases (Uberbacher and Mural, 1991). These sensors were applied to the known coding and noncoding human DNA sequences that are in GenBank, and the output was used to train a neural network to dissect potential protein coding regions from a set of unknown sequence, such as the galago globin cluster sequence, based on these "learned" attributes. The results of this analysis on the galago cluster sequence is shown in Fig. 4. Both sense and antisense strands were searched for potential exons, but only the sense strand showed significant peaks. Peaks with scores of greater than 0.8 were ranked as excellent candidates for coding sequences. The results of this analysis correctly predicted 75% or at least one of the exons of each transcribed galago globin gene (Fig. 4). The results of this analysis indicate that GRAIL, trained on human gene sequences, can also be used to identify protein coding regions of other mammalian species. Of the transcribed exons greater than 100 bp (exons 2 and 3), 88% were identified. Only exon 3 of the ϵ globin gene was missed. However, only about 50% of the exons less than 100 bp were found. As expected, the search did not identify the η globin pseudogene as one of the protein coding regions but did identify segments of two L1 elements (L1Gc-5 and L1Gc-6; see Figs. 1B and 4) as potential exons. Two peaks (positions 7011 to 7091 and 33,804 to 33,851) were also identified as potential coding sequences where no

FIG. 2. The complete nucleotide sequence of the galago β globin gene cluster. The total sequence of 41,101 bp was obtained from 13 *Eco*RI subcloned fragments. The *Eco*RI recognition sites are indicated in lowercase letters. The coding regions of the expressed β-like globin genes are indicated on top of the nucleotide sequence by their encoded amino acid sequences. Exons 1 and 2 of the $\psi\eta$ gene are indicated by asterisks above the nucleotide sequence. The promoter elements CACA, CCAAT, and TATA boxes as well as the putative CAP sites are labeled as such above their nucleotide sequence. SINE elements are indicated by an overline above their nucleotide sequence. LINE elements are indicated by a double overline above their corresponding nucleotide sequence. For both families of interspersed repeats, left and right arrows indicate the 3′ to 5′ or 5′ to 3′ directionality of the repeats, respectively. A series of arrowheads are used to indicate where short direct repeats flank insertion



DOCKET ALARM

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.