L1 Repeat Elements in the Human ϵ-$^{G}\gamma$-Globin Gene Intergenic Region: Sequence Analysis and Concerted Evolution within This Family[1]

Peter K. Rogan, Julian Pan,† and Sherman M. Weissman*,†*

*Departments of Molecular Biophysics and Biochemistry and †Human Genetics, Yale University School of Medicine

We have deduced the sequence of a composite long interspersed repeated DNA in primates and herein describe its relationship to a complex repeat element (L1Heg) located in the interval linking the human ϵ- and $^{G}\gamma$-globin genes. The main element of L1Heg is 3′ truncated and interrupted by the insertion of the 3′ end of a second L1 element. Transposition of L1Heg into this intergenic locus generated a 62-bp duplication of flanking sequences. In contrast, insertion of the second repeat may have been mediated by homology between donor and target sequences. The main repeat represents a novel class of abundant elements whose sequences have diverged from other rodent and primate LINES ~1.3 kb downstream from the 5′ terminus of L1Heg. Comparison of L1Heg with the sequences of two other related L1 members revealed a complex set of rearrangements confined within a region that resembles the long terminal repeats of other types of retroposons. The boundaries of conversion-like events were defined on the basis of the clustering of nucleotide sequence variants common to two or more nonallelic 3′ L1H elements. Several of these events are apparently initiated or resolved within a common 150-bp region that coincides with the 3′ terminus of a pan-mammalian open reading frame. This analysis showed that concerted genetic interactions and random drift both contribute appreciably to sequence variation within this set of L1H members.

Introduction

Short (~300 bp; Jelinek et al. 1980) and long (>2,000 bp; Singer 1982) middle repetitive DNA sequences are interspersed with single-copy DNA in the β-globin cluster. The full-length long interspersed sequences in primates (denoted *Kpn*I [Shafit-Zagardo et al. 1982*a*, 1982*b*] or L1H) and in rodents (denoted *Bam*HI [Meunier-Rotival and Bernardi 1984] or L1Md [Voliva et al. 1983]) can be >6 kb long, although 5′-truncated (Kole et al. 1983) or internally scrambled (Potter 1984) elements are common in mammalian genomes. The 3′ terminus of many repeats consists of a 20–30-bp deoxyadenosine-rich tract (Burton et al. 1986). Several overlapping, eponymic restriction-endonuclease fragments that were previously thought to constitute separate repetitive-sequence families are now known to form a contiguous pan-mammalian element—i.e., the 1.2-, 1.5-, and 1.8-kb primate *Kpn*I fragments; the 1.9- and 2.5-kb human *Hin*dIII fragments; the mouse 0.5- and 4.0-kb *Bam*HI, 1.5- and 1.7-kb *Bst*NI, 1.3-kb

1. Key words: long interspersed repetitive DNA, transposable elements, globin genes, gene conversion, phylogenetic partitioning.

Address for correspondence and reprints: Dr. Peter K. Rogan, NCI-Frederick Cancer Research Facility, P.O. Box B, Frederick, Maryland 21701-1240.

Mol. Biol. Evol. 4(4):327–342. 1987.

0737-4038/87/0404-0001$02.00

*Eco*RI, "MIF", and "R" repeat segments (Singer and Skowronski 1985; Burton et al. 1986; Mottez et al. 1986).

Intergenic sequences constitute 85% of the human β-globin gene cluster. The phenotypes of a variety of mutations are characterized by the removal of large segments of globin DNA (Collins and Weissman 1984). Phenotypically distinguishable, intergenic deletions with distinct termini define potential cis-acting regulatory sequences that may be necessary for normal hemoglobin switching. The segment between the fetal and embryonic genes has been removed as part of large deletions that silence the β gene, but no function has been associated with this sequence since no deletion endpoints map in this region (van der Ploeg et al. 1980; Kioussis et al. 1983; Vanin et al. 1983). As a basis for the study of functional aspects of the intergenic DNA between these β-like globin genes, we have determined the DNA sequence of the region between ε and $^{G}\gamma$.

More than half of this intergenic domain consists of a single, 7.6-kb long interspersed repeat element. A distinct 5′ terminus, poor overall sequence conservation, and internal rearrangement distinguish this full-length L1 element from other typical repeat sequences (Hattori et al. 1985; M. F. Singer, G. Grimaldi, R. E. Thayer, J. Skowronski, and S. Contente, personal communication). We have also identified a number of other sequences belonging to the β-globin cluster as being L1H related. Finally, we have examined the possible evolutionary dynamics of the repeat family by evaluating the relative contributions of random mutation and concerted interactions among a set of independently isolated, nonallelic 3′-terminal L1H sequences.

Material and Methods

Sequencing Strategy

Figure 1 presents a map of the β-globin cluster, showing the location of the sequenced region and the positions of the *Eco*RI restriction sites in the intergenic DNA sequence. Plasmid A36 (Forget et al. 1982) contains part of the $^{G}\gamma$ gene and 5.6 kb of upstream sequences. Phage Tεγ 51 (Kaufman et al. 1980) spans the ε gene and extends 7.6 kb downstream. Tεγ 51 and A36 are separated by a 263-bp *Eco*RI fragment that was independently cloned into M13mp8. The 4.1-kb (Tεγ A), 1.6-kb (Tεγ C), and 1.5-kb (Tεγ D) fragments released from an *Eco*RI digest of Tεγ 51 were subcloned into pBR322. Restriction fragments spanning these *Eco*RI sites were directly subcloned from Tεγ 51 into M13mp8. For sequencing, these plasmids and A36B (a 3.4-kb *Bam*HI-*Eco*RI fragment derived from the 5′ end of pA36) were sheared, repaired with S1 nuclease and Klenow DNA polymerase I, fractionated by polyethylene-glycol precipitation (0.7% PEG-8000 in 0.5 M NaCl), ligated to dephosphorylated, *Sma*I-digested M13mp8 (Deininger 1983), and transformed into the *Escherichia coli* host JM101. Phage hybridizing to purified intergenic DNA were selected for dideoxynucleotide chain termination sequencing (Sanger et al. 1977). The sequences of overlapping clones were then assembled into a consensus (Staden 1982) so that each base was independently determined a minimum of three times. More than 90% of the sequence was determined on both strands.

Sequence Analysis

Internal repetitive units were identified by means of a dot matrix–representation system (White et al. 1984). Precise pairwise alignments of these repeats and other published sequences were carried out by means of an implementation of the Needle-

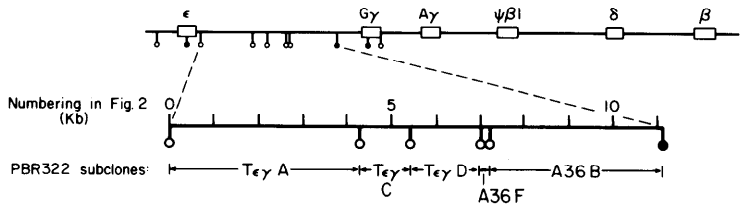

FIG. 1.—Restriction map and subcloning strategy for the intergenic ε-$^{G}\gamma$ globin region. The lower map is an expanded view of the sequenced region, showing the boundaries of the plasmid subclones derived from Tεγ 51 and pA36. ♁ = *Eco*RI; and ↓ = *Bam*HI.

man-Wunsch algorithm (Devereux et al. 1984). Multiple alignments were produced by iteratively rectifying the intergenic DNA sequence against as many as 13 related L1 sequences. This procedure was independent of which sequence was designated as the rectification template.

We distinguished concerted interactions from random mutation in multiple sequence alignments by means of the method of phylogenetic partitioning (Stephens 1985). An algorithm to identify clustered sequence variations common to two or more sequences has been programmed in PASCAL.

Southern Blotting and Copy-Number Determination

Human placental genomic DNA was prepared according to standard methods (Maniatis et al. 1982, pp. 280–282), digested with a variety of restriction endonucleases, electrophoresed in 0.85% agarose, transferred to nylon membrane (Biodyne), and hybridized to nick-translated plasmid or to primer-extended M13mp8 recombinants (Southern 1975). The abundance of repetitive DNA segments was qualitatively estimated by comparing autoradiographic signal intensities of (^{32}P)-labeled genomic DNA hybridized to Tεγ A, Tεγ C, Tεγ D, and subclones derived from these plasmids. The genomic copy number of M13mp8 subclones derived from A36B was determined by screening 7×10^4 phage from an amplified human genomic EMBL3 library (complexity, 8×10^7 kb; source, partial *Sau*3A digest of male leukocyte DNA [a gift from A. M. Frischauf]) and tabulating the frequency of positive plaques that hybridized to one or more probes (Benton and Davis 1977). Filters were washed three times in 0.3 M NaCl, 0.03 M sodium citrate (2 × SSC) at 65 C prior to autoradiographic exposure.

Isolation of Other L1H Elements

Homologues of [^{32}P]-labeled probe 1 were selected from human plasmid- and cosmid-DNA genomic libraries prepared from JY cells (Biro et al. 1981). Clone pA9 contained a 2.2-kb insert whose sequence was determined according to the method of Maxam and Gilbert (1980) by P. A. Biro. Clone cos5 (originally isolated by means of hybridization to a human major-histocompatibility-complex class I gene; R. Srivastava, personal communication) possesses an L1H-related 3.1-kb *Eco*RI fragment (cos5R). Self-ligated multimers of cos5R were sheared, repaired with DNA polymerase I Klenow enzyme, and ligated to *Sma*I-cleaved M13mp8. The cos5R sequence was determined according to the method of Sanger et al. (1977). Both cos5R and pA9 sequences are available from the authors on request.

Results

Figure 2 presents the DNA sequence beginning at the *Eco*RI site (position 1) 178 bp downstream from the polyadenylation site of the ε-globin gene and ending at the


TR
3′ terminus of the main element
Eco RI
3′ terminus of the internal element
Eco RI
Eco RI
5′ terminus of the internal element
5′ sequence homology breakpoint: with composite L1H sequence
with TφGI

FIG. 2—(*Continued*)

*Bam*HI site (position 11128) 2.2 kb upstream of the $^{G}\gamma$ gene. Sequences farther downstream were determined by Shen et al. (1981). The overall base composition was calculated to be 18% dC, 38% dT, 24% dA, and 20% dG.

Identification of an L1 Family Member

Cloned segments of the human *Kpn*I canonical repeat family hybridized to Tεγ C and Tεγ D (results not shown; Shafit-Zagardo et al. 1982*a*). Sequence comparisons showed that Tεγ A and A36 were also L1H related (Collins and Weissman 1984). We and others (Singer and Skowronski 1985) constructed a hypothetical map of a composite *Kpn*I repeat and deduced the organization of the intergenic element (denoted L1Heg; fig. 3).

L1Heg possesses duplicate copies of a 1.8-kb sequence related to the 3′-terminal regions of several other L1H repeat units (Collins and Weissman 1984). However, the duplicated region of the main L1Heg element (positions 1887–3711, fig. 2) lacks a 300-bp segment at the conventionally defined L1H 3′ end. In contrast, a recursively inserted, 5′-truncated L1H element (positions 4642–6751, fig. 2) has an intact 3′ terminus. The conventional 5′ → 3′ direction of both the main and recursively inserted L1H elements is in the opposite orientation of the ε- and $^{G}\gamma$-globin gene transcription units (fig. 3).

The duplicated sequences of the two L1H segments are weakly conserved (71% identity). The main L1Heg duplicated segment also exhibits low levels of sequence similarity with other human and African green monkey L1 members (results not shown). In contrast, the divergence of the internal L1H repeat from the known sequences of other L1H elements approximates the average dissimilarity in this set (15%

FIG. 2.—Nucleotide sequence beginning at the *Eco*RI site 178 bp downstream from the polyadenylation site of ε and ending at the *Bam*HI site 2.1 kb upstream of the $^{G}\gamma$ cap site. Positions of *Eco*RI sites are shown. A box surrounds the sequence of the recursively inserted L1H element. Imperfect direct repeats flanking the main L1Heg element are identified by TR and underlined with arrows. The sites at which L1Heg diverges from other elements are noted. Inverted termini and internal direct repeats within the ψ^{LTR} are overlined with broken and solid arrows, respectively.

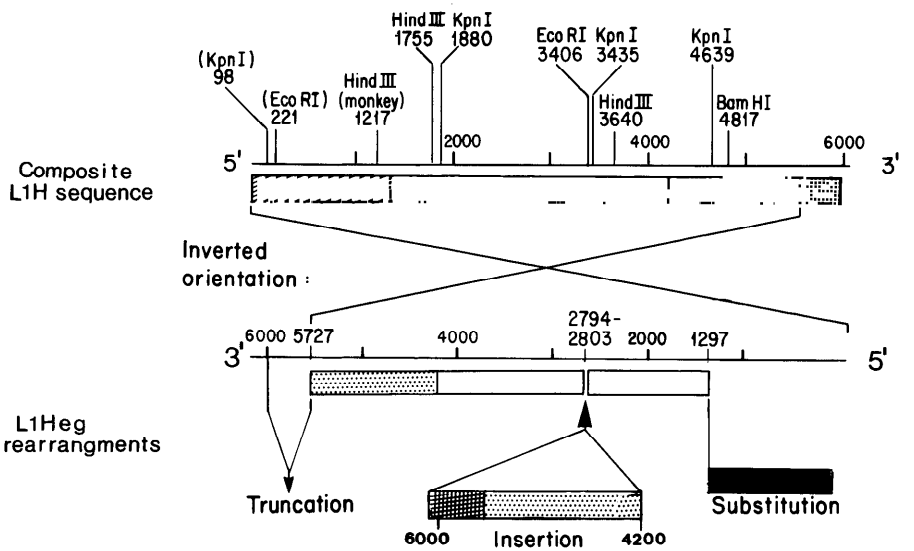

FIG. 3.—Comparison of L1Heg and composite L1H element organization. The composite L1H sequence, constructed by aligning overlapping L1H sequences (Collins and Weissman 1984), shows the positions of canonical restriction sites. Blocks of L1 sequences are schematically coded as shaded regions. Rearrangements noted in L1Heg are mapped onto coordinates of the composite sequence. Juxtapositions of sequences that are homologous to segments of the L1 composite are keyed with identical shading patterns.

mismatch). Thus, the main repeat may represent either a highly mutated L1 variant or a representative of an ancient lineage.

L1Heg Possesses a Novel 5′ Terminus

The 5′ region of L1Heg contains a 486-bp sequence (denoted ψ^{LTR}; fig. 2; see Jagadeeswaran et al. 1982) that exhibits a number of general features typical of the structure of a long terminal repeat sequence of transposable DNA elements (Finnegan et al. 1978; Calos and Miller 1980; Farabaugh and Fink 1980; Swanstrom et al. 1981). A 17-bp, markedly GC-rich sequence, starting with the 5′ dinucleotide TG and terminating with the 3′ dinucleotide CA, is present as a near-perfect inverted repeat at both ends of the ψ^{LTR}. Two similar sequences beginning at positions 9101 and 9154 might be remnants of internal direct repeats that have since diverged. The ψ^{LTR} also exhibits a local bias in dG (28%) and dC (28%) that is not characteristic of adjacent AT-rich L1H sequences. In spite of these structural similarities, the ψ^{LTR} has no known relationship to any eukaryotic sequence, lacks consensus promoter or polyadenylation sequences, and is presumably nonfunctional.

The sequence of the 5′ terminus of L1Heg diverges from another major class of primate repetitive elements beyond position 8279 (Grimaldi et al. 1984; Potter 1984; Hattori et al. 1985) and from rodent L1 families beyond position 8212 (Soares et al. 1985; Loeb et al. 1986). We wished to define which sequence features were conserved among L1H members carrying the L1Heg 5′ motif.

The sequences of two human genomic clones, cos5R and pA9, selected with the ψ^{LTR} probe are >87% identical with L1Heg and span the breakpoint where the rodent and other primate L1 members diverge (fig. 4B). Except for a 50-bp insertion in pA9 at position 9000 of L1Heg, pA9 and L1Heg are in register until position 9450. Cos5R

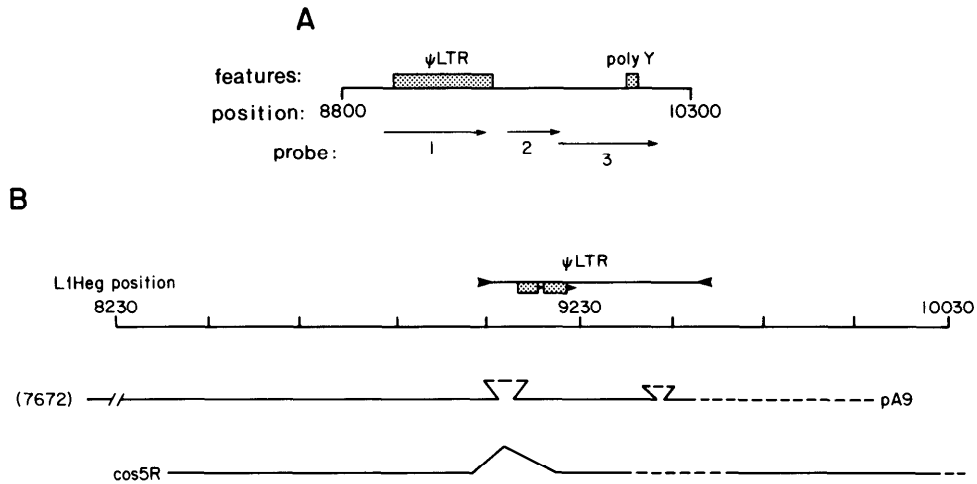

FIG. 4.—Characterization of the L1Heg 5′ terminus. A, M13mp8 recombinants 1–3 were used to define the 5′ end of L1Heg (table 1). Numbering corresponds to the sequence given in fig. 2. ψ^{LTR} (positions 9027–9501) and poly Y (positions 10043–10076) refer, respectively, to a long terminal repeat–like sequence and an oligopyrimidine stretch in this region. B, Schematic sequence comparison of L1Heg with independently isolated L1H elements (pA9 and cos5R). Dotted lines represent regions of unknown origin/relationship, and solid lines are regions with >85% identity. ≶ = Insertion; and _/_ = deletion.

and L1Heg maintain registry through position 10020, although the ψ^{LTR} is extensively rearranged, e.g., a 150-bp substitution of apparently unrelated DNA is adjacent to a 200-bp deletion.

The 5′ end of L1Heg was located between positions 9500 and 10300 by comparing the genomic representation of adjacent, nonoverlapping L1Heg subclones. Probes 1–3 are ordered in a 3′ → 5′ orientation with respect to L1Heg (fig. 4A). Hybridization of restriction-endonuclease digests of human DNA with probes 1 and 2 resulted in diffuse patterns similar to those obtained using the highly reiterated Tεγ C– and Tεγ D–labeled recombinants (results not shown). Clone 3, which detected ~50 discrete bands, contains a pyrimidine tract (poly Y; position 10043) and a poly(CA) repeat (position 10225; see Sun et al. 1984). These contrasting blotting patterns implied either that the majority of L1Heg relatives are 5′-truncated derivatives of a small population of longer elements or that L1Heg is linked to a member of a distinct repetitive-sequence family.

We screened a human genomic λ library and scored phage that hybridized with at least two probes (Heller et al. 1984). Table 1 indicates the abundance of clones 1 and 2 (2.5×10^3 copies) and 3 (1.7×10^2 copies) in the human genome. Clones 1 and 2 cohybridize in a nonrandom manner to the same recombinants and are, therefore, linked at most genomic loci. As a control, we have also demonstrated that clones 1 and 2 are frequently associated (though are less abundant) with L1H sequences derived from the common 3′ terminus. Clone 3 is not usually associated with either clone 1 or clone 2 and thus represents a distinct repetitive family. This places the 5′ end of the 7.6-kb L1H repeat ~3.3 kb upstream of the $^{G}\gamma$ gene.

Duplication of Flanking Target Sequence

A sequence unit containing the entire L1Heg element and an independent, low-abundance repeat (clone 3, fig. 4A) is bounded on both sides by a pair of imperfect

Table 1
Distribution of L1Heg Repetitive DNA in Randomly Selected Genomic Clones

A. Genomic Frequency of Intergenic ε-^{G}γ Recombinant Probes[a]

Probe	Size (kb)	Total Hybridized	Genome	
			Copy Equivalents	Fraction
1	0.42	1,273 (of 71,150)	2,683	4.5×10^{-4}
2	0.22	1,101 (of 71,150)	2,320	1.9×10^{-4}
3	0.43	81 (of 71,150)	170	1.4×10^{-5}
1 + 2	0.64	174 (of 8,593)	2,843	6.0×10^{-4}
TεγD	1.51	286 (of 8,593)	4,988	2.5×10^{-3}

B. Number of Clones Hybridizing with Two Repetitive DNA Fragments

Probes	Observed Total	Expected Total[b]	χ^2	*P*-Value
1 and 2	1,019	63	14,585	<0.001
2 and 3	10	2	32	>0.99
1 and 3	12	5	12	>0.99
[1 + 2] and TεγD ...	163	6	3,869	<0.001

[a] Probes are derived from the putative 5′ terminus of L1Heg. Their positional relationship to sequence features in L1Heg are shown in fig. 4A.

[b] The expected frequency is the number of λ phage that would hybridize to both probes if their targets were independently distributed (in a Poisson distribution) in the human genome.

direct-repeat sequences (84% identity; positions 1807–1869 and 10528–10588, fig. 5B). A synthetic oligonucleotide containing the 3′ direct repeat hybridized to two bands in a genomic Southern blot of *Eco*RI-digested DNA (results not shown). These

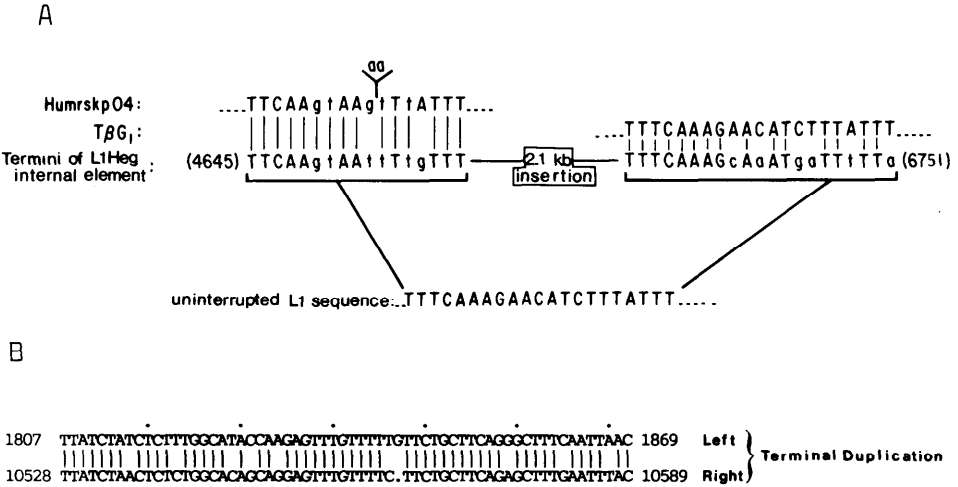

FIG. 5.—A, Imperfect direct repeat sequences flanking the inserted L1H element. Nucleotides homologous to an uninterrupted target sequence (Manuelidis 1982) in L1Heg or at paralogous positions in other nonallelic L1H elements (Humrskp04, Digiovanni et al. 1983; TβG1, Hattori et al. 1985) are shown in uppercase characters. Nonmatching nucleotides are in lowercase characters. Identical sites conserved between L1Heg and paralogous sequences are connected by vertical match lines. B, Imperfect direct repeat sequences flanking the entire L1Heg element. Sequences are aligned to optimize identity and are numbered using fig. 2 coordinates.

flanking sequences may be remnants of an unusually large target-site duplication that accompanied the original L1Heg transposition event. Thus, members of the L1Heg and clone-3 repetitive-sequence families may have been linked prior to genomic insertion.

Genomic Fixation of L1H Subfamilies

Relationships between nonallelic L1 repeat elements can be elucidated by examining the distribution of nucleotide sequence differences among family members. Regions of sequence that have participated in concerted events will share a cluster of sequence variations determined by a particular conversion-like event (CLE) (Stephens 1985). In a large family of dispersed repeats, a statistically clustered block of identical sequence variants (defined as a set of congruent sites) common to a subset of elements is defined as a phylogenetic partition. A partition can represent (1) a conserved domain common to a set of repeats that have been amplified from a common founder sequence and/or (2) a region homogenized by a series of concerted genetic interactions (such events are designated as CLE).

Each "unique" phylogenetic partition, indicated by the open boxes in figure 6, represents a defined sequence region that distinguishes the designated L1 element from the nine other sequences in the sample. Five of the nine unique partitions (in sequences C–E, G, and H) have a common 5′ end between positions 160 and 280. Four such partitions (in sequences A, B, D, and E) have an additional 3′ boundary between positions 350 and 390. The overall percentage of sequence identity between these two boundaries is approximately equal to the flanking partitionable regions, so that sequences within this interstitial segment are not preferentially conserved. Rather, there may be a tendency to initiate or resolve concerted genetic events within this region (see Discussion). However, since at least one of nine unique partitions (in sequence F) traverses this region, this is not an obligate boundary for concerted action.

In the absence of selective amplification of a few founder elements, the likelihood of observing both donor and recipient in a CLE is vanishingly small, since the data set represents <0.1% of all genomic repeats. The detection of several such (multiple) partition classes (▥, ▧, ▨, and ▩ / ▬, fig. 6) is consistent with the sequential amplification of different subsets. The endpoints of these partition classes are evenly distributed throughout the L1H 3′ end, and, in three (▧, ▨, and ▩ / ▬) of the four cases, overlap unique partitions. None of these partitions contains more than three congruent sites or spans more than a 50-bp interval; this could be a consequence of secondary CLE (Abastado et al. 1984), parallel or backward mutation that results in an apparent contraction of the region boundaries defined by the original amplification or conversion event(s).

The random mutation frequency in this set of repeats can be estimated by counting sequence patterns that cannot be partitioned and occur only once in the alignment. These sites (14% of all nucleotide positions) are distributed evenly throughout the sample. Each pattern often contains three or more different nucleotides, supporting the possibility that these sites have not been homogenized by CLE. Since the abundance of these sites (100) is comparable with the number of partitionable loci, we suggest that random mutation and CLE have both contributed appreciably to L1H sequence diversity.

Discussion

The DNA sequence between the ε and $^{G}\gamma$ genes of the human β-globin complex contains a full-length, rearranged long interspersed repeat element flanked by 62-bp

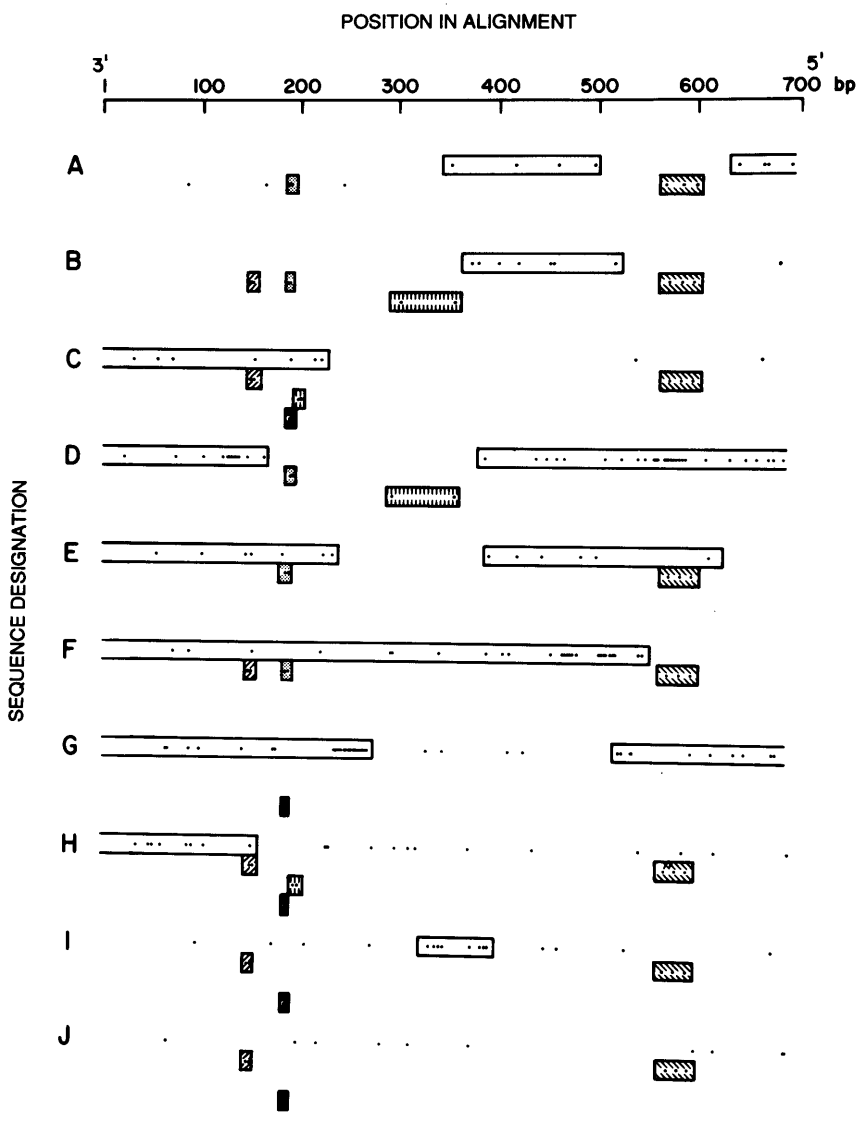

FIG. 6.—Phylogenetic partitioning of L1 repeat sequences. A set of nucleotide loci constituting a common partition pattern are considered clustered when $P(d \leqslant d_o) <= .025$ and $P(g \geqslant g_o) <= .05$. Ten aligned independent L1 sequences representing 700 bp derived from the 3′ terminus (sequences A–J) have been sorted into unique (□) or multiple (▨, ▨, ▥, and ▬ / ▩) partition classes designated by a box around each clustered interval. Multiple partitions (▬ and ▩) occur at the same sites. Sequences are derived from the following sources: A, L1Heg internal repeat; B, Sun et al. 1984 (kpnt); C, Potter 1984; D, Sun et al. 1984 (kpne); E, Digiovanni et al. 1983 (lg03); F, Digiovanni et al. 1983 (lg08); G, Digiovanni et al. 1983 (lg04); H, Lerman et al. 1983; I, Larhammar et al. 1985; and J, Hattori et al. 1985. The first test assesses the likelihood that partitioned sites are significantly clustered (d_o = the observed distance between any two copartitioned nucleotide sites), assuming that they could be located anywhere in the sequence with equal probability (d = the interval length separating two randomly distributed sites in the test sequence). The second test determines whether a clustered partition is bounded by a nonpartitionable region by evaluating the probability that a random-length interval lacking partitionable sites (g) is as long as or longer than the longest observed interval (g_o).

direct repeats. The most striking features of this family member are the insertion of a second L1 element into the 1.9-kb *Hin*dIII region of the main repeat and the substitution of a novel, highly reiterated sequence at the consensus 5′ end.

Sequence duplication flanking the human L1Heg retroposon, the reduced spacing between the ε and γ genes, and the presence of a single copy of the target site in the prosimian intergenic region (~6 kb; see Barrie et al. 1981; Rogan 1987; D. Tagle, personal communication) suggests that initial transposition of the main element into this site may have occurred in anthropoids after the divergence of the prosimian and simian lineages. The accumulation of mismatched nucleotides (16%) in the flanking 62-bp duplications presumably reflects the ancient age of the initial insertion event (fig. 5B). The poor sequence conservation of the main repeat relative to the internal duplication and other L1H family members implies that the L1H internal insertion was an independent, more recent event.

We can estimate the age of the main L1Heg element on the basis of the following mechanistic assumptions: (1) At the time of insertion, the "full-length" main element was an intact derivative of a founder retroposon. (2) The coding potential of the main repeat was inactivated no later than the date of the recursive-insertion event. (3) The truncated, internal L1H repeat was not selected for a functional gene product. (4) The unselected primate mutation rate is 1.3×10^{-9}/site/year (Britten 1986). (5) The synonymous and replacement rates for L1 long open reading frames are 1.3×10^{-9} and 1.0×10^{-9}/site/year, respectively (calculated from a set of aligned L1H sequences according to the method of Kimura 1981). The solutions to rate equations developed from an evolutionary model analogous to that of Miyata and Yasunaga (1981) but based on the above assumptions suggest that the main L1H element was introduced at this locus 71–81 Myr ago. If it is assumed (on the basis of the overall divergence from the other sequences in fig. 6) that the recursively inserted L1H member is 32 Myr old, the main L1 element is ≥35 Myr older than the inserted sequence. Thus, the initial L1 insertion at this locus prior to the catharrine/platyrrhine divergence preceded the second L1 insertion, which may have occurred in the catharrine lineage (Ciochon and Fleagle 1985).

The duplicated structure of L1Heg arose from the insertion of a 5′-truncated 1.8-kb L1H segment into the main element. Both the 5′ and 3′ ends of the inserted repeat have some similarity to a 15-bp target sequence in an intact L1 element (Manuelidis 1982) and to paralogous sequences of other L1H members (fig. 5A; see Hattori et al. 1985; Digiovanni et al. 1983). This is not characteristic of a classical transposition event (Shapiro 1979; Harshey and Bukhari 1981) and implies that the insertion may have been mediated by homology-dependent interactions (Voliva et al. 1984; Jones and Potter 1985; Schindler and Rush 1985).

The 300-bp deletion at the 3′ end of the main L1Heg segment is unusual because most family members do not exhibit 3′ truncation. Either a homology-directed excision of this sequence subsequent to L1Heg insertion or internally primed reverse transcription (Van Arsdell et al. 1981) is consistent with the loss of these sequences.

The presence of a novel, abundant 5′-terminus L1Heg suggests (1) that the abundance of full-length long interspersed repetitive elements may have been underestimated in primates and (2) that the propagation of these repeats does not require a particular 5′-terminal sequence. The saltatory amplification of L1 subsets apparently correlates with the recruitment of different 5′-terminal sequences. Both mouse (Loeb et al. 1986; Mottez et al. 1986) and rat (Soares et al. 1985) 5′-terminal structures resemble tandemly arranged RNA polymerase II promoters that could maintain the transcriptional viability

of L1 elements after retroposition. In contrast, rearrangements at the 5′ termini of L1Heg, pA9, and cos5R are consistent with the possibility that this class of elements either (1) may not carry internal transcriptional regulatory sequences or (2), like the *Drosophila* retroposon 17.6 (Inouye et al. 1986), has accumulated defects that impair mobility.

Hardies et al. (1986) suggest that the heterogeneity of target-site duplications flanking rodent L1 members argues in favor of concerted evolution in the L1 family via amplification of a limited number of templates rather than via recombination between existing repeats. However, classes of rodent L1 repeats can be distinguished from one another on the basis of multiple short conversion tracts (Jubier-Maurin et al. 1985). We delimit the boundaries of CLE within L1H sequences by directly searching for the expected products of CLE.

CLEs have occurred often and/or numerous founder elements have been amplified during the evolution of this sequence family. There are, for example, nine "unique" phylogenetic partitions with an average CLE length of 240 bp in the 10 L1H specimens. The five "multiple" partitions are much shorter, perhaps because subsequent mutational events and CLEs have obscured the original common sequence motif. Frequent DNA-mediated conversion (Scherer and Davis 1980; Ernst et al. 1982; Jinks-Robertson and Petes 1985) or RNA template switching via reverse transcriptase (Gilboa et al. 1981; Weiner et al. 1986) at nonallelic loci could account for this concerted behavior.

CLEs might preferentially initiate or be resolved within a 150-bp region defined by boundaries common to several independent unique partitions. Soares et al. (1985) suggest that the 3′ terminus of a pan-mammalian unidentified open reading frame lies within the nonpartitionable region. This suggestion raises the intriguing possibility that potential selection for a functional gene product in a limited number of repeats is accompanied by preference for genetic-exchange events initiating and terminating outside the protein-coding domain.

It is difficult to imagine how the *majority* of L1H founder elements could selectively maintain the coding potential of a single (set of) gene product(s) in the face of high overall rates of mutation and back-conversion by daughter copies. Instead, different classes of L1 convertants may have been selected for retention of multiple coding functions, although few, if any, elements have maintained this capability. For example, an L1-encoded gene product could have catalyzed the retrotransposition of these repeated sequences (Hattori et al. 1986). Thus, multiple L1 repeat categories may have evolved (1) elements capable of mobilizing all repeats that retain sequences necessary for transposition, (2) elements that preferentially act on members of the same class of convertants, (3) members that may continue to be transposed, though they may encode a gene product possessing a different function or a "pseudogene," and (4) degenerate repeats, unable to transpose or express any gene product.

Degenerate elements might continue to disperse, however, through intrachromosomal illegitimate recombination. The human β-globin gene cluster contains seven distinct elements in a 67-kb interval. Other clustered multigene families—such as the human α-interferon, U1, U2, and class I major histocompatibility loci (Rogan 1987)—also possess a high concentration of L1 DNA. The γ-globin and α-interferon duplication units are each bounded by homologous L1H segments. This genomic organization could have been generated, for example, via unequal intrachromosomal recombination across a pair of misaligned L1H segments situated both upstream and downstream of a single γ gene. In fact, the single fetal globin gene of *Ateles geoffroyi,* a cattharine

primate, is flanked on both sides by sequences derived from the 3′ terminus of L1H (Rogan 1987).

In summary, we have described a long interspersed repeat element, situated between the human ε- and $^{G}\gamma$-globin genes, that represents a novel, abundant class of L1 units. The main repeat is interrupted by a distinct 5′-truncated member that may have inserted via a homology-mediated mechanism. In addition, we have examined the concerted interactions at the 3′ terminus of the family by surveying potential CLEs within a set of nonallelic sequences. A majority of these events are initiated or resolved within a common 150-bp interval that coincides with the 3′ terminus of an evolutionarily conserved open reading frame.

Acknowledgments

We are grateful to Michael Liskay, Alan Weiner, and Elisabetta Ullu for critically reading this manuscript. Discussions with Jeff Strathern and Steve Hughes were particularly helpful to P.K.R. in appreciating the potential role of RNA intermediates in CLEs. This work was supported by National Institutes of Health grant AM28376-05 to S.M.W.

LITERATURE CITED

ABASTADO, J. P., B. CAMI, T. DINH, J. IGOLEN, and P. KOURILSKY. 1984. Processing of complex heteroduplexes in *Escherichia coli* and COS-1 monkey cells. Proc. Natl. Acad. Sci. USA **81**: 5792–5796.

BARRIE, P. A., A. J. JEFFREYS, and A. F. SCOTT. 1981. Evolution of the β globin gene cluster in man and the primates. J. Mol. Biol. **149**:319–336.

BENTON, W. D., and R. W. DAVIS. 1977. Screening of λgt recombinant clones by hybridization to single plaques in situ. Science **196**:180–182.

BIRO, P. A., D. PEREIRA, A. K. SOOD, B. DE MARTINVILLE, U. FRANKE, and S. M. WEISSMAN. 1981. The structure of the human major histocompatibility locus. Pp. 315–332 *in* C. JANEWAY, E. SERCARTZ, and H. WIGZELL, eds. Immunoglobulin idiotypes: ICN-UCLA Symposium on Molecular and Cell Biology. Vol. **20.** Academic Press, New York.

BRITTEN, R. J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. Science. **231**:1393–1398.

BURTON, F. H., D. D. LOEB, C. F. VOLIVA, S. L. MARTIN, M. H. EDGELL, and C. A. HUTCHISON III. 1986. Conservation throughout mammalian radiation and extensive protein-encoding capacity of the highly repeated DNA long interspersed sequence one. J. Mol. Biol. **187**:291–304.

CALOS, M. P., and J. H. MILLER. 1980. Transposable elements. Cell **20**:579–595.

CIOCHON, R. L., and R. G. FLEAGLE. 1985. Primate evolution and human origins. Benjamin/Cummings, New York.

COLLINS, F., and S. M. WEISSMAN. 1984. The molecular genetics of human hemoglobin. Prog. Nucleic Acids Res. Mol. Biol. **31**:315–462.

DEININGER, P. 1983. Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. Anal. Biochem. **129**:216–223.

DEVEREUX, J., P. HAEBERLI, and O. SMITHIES. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:387–395.

DIGIOVANNI, L., S. HAYNES, R. MISRA, and W. R. JELINEK. 1983. *Kpn* I family of long-dispersed repeated DNA sequences in man: evidence for entry into genomic DNA or DNA copies of poly(A) terminated *Kpn* I RNAs. Proc. Natl. Acad. Sci. USA **80**:6533–6537.

ERNST, J. F., J. W. STEWART, and F. SHERMAN. 1982. Formation of composite iso-cytochromes C by recombination between non-allelic genes in yeast. J. Mol. Biol. **161**:373–394.

FARABAUGH, P. J., and G. R. FINK. 1980. Insertion of the eukaryotic transposable element Tyl creates a 5-base pair duplication. Nature. **286**:352–356.

FINNEGAN, D., G. RUBIN, M. YOUNG, and D. HOGNESS. 1978. Repeated gene families in *Drosophila melanogaster.* Cold Spring Harbor Symp. Quant. Biol. **42**:1053–1063.

FORGET, B. G., D. TUAN, P. A. BIRO, P. JAGADEESWARAN, and S. M. WEISSMAN. 1982. Structural features of the DNA flanking the human non-α globin genes: implications in the control of fetal hemoglobin switching. Trans. Assoc. Am. Phys. **94**:204–210.

GILBOA, E., S. MITRA, S. GOFF, and D. BALTIMORE. 1981. A detailed model of reverse transcription and tests of crucial aspects. Cell **18**:93–105.

GRIMALDI, G., J. SKOWRONSKI, and M. F. SINGER. 1984. Defining the beginning and end of *Kpn* I family segments. EMBO J. **3**:1953–1959.

HARDIES, S. C., S. L. MARTIN, C. F. VOLIVA, C. A. HUTCHISON III, and M. H. EDGELL. 1986. An analysis of replacement and synonymous changes in the rodent L1 repeat family. Mol. Biol. Evol. **3**:109–125.

HARSHEY, R. M., and A. BUKHARI. 1981. A mechanism of DNA transposition. Proc. Natl. Acad. Sci. USA **78**:1090–1094.

HATTORI, M., S. HIDAKA, and Y. SAKAKI. 1985. Sequence analysis of a KpnI family member near the 3′ end of human β-globin gene. Nucleic Acids Res. **13**:7813–7827.

HATTORI, M., S. KUHARA, O. TAKENAKA, and Y. SAKAKI. 1986. The L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. Nature **321**:625–628.

HELLER, D., M. JACKSON, and L. LEINWAND. 1984. Organization and expression of non-Alu family interspersed repetitive DNA sequences in the mouse genome. J. Mol. Biol. **173**:419–436.

INOUYE, S., K. HATTORI, S. YUKI, and K. SAIGO. 1986. Structural variations in the *Drosophila* retrotransposon, 17.6. Nucleic Acids Res. **14**:4765–4778.

JAGADEESWARAN, P., P. BIRO, D. TUAN, J. PAN, B. FORGET, and S. M. WEISSMAN. 1982. Interspersed repetitive sequences of the human genome: are they transposons? Pp. 29–35 *in* B. BONNE-TAMIR, T. COHEN, and R. GOODMAN, eds. Human genetics. Part A: The unfolding genome. Alan R. Liss, New York.

JELINEK, W., T. P. TOOMEY, L. LEINWAND, C. M. DUNCAN, P. BIRO, P. CHOUDHARY, S. M. WEISSMAN, C. RUBIN, C. HOUCK, P. L. DEININGER, and C. W. SCHMID. 1980. Ubiquitous, interspersed repeated sequences in mammalian genomes. Proc. Natl. Acad. Sci. USA **77**: 1398–1401.

JINKS-ROBERTSON, S., and T. D. PETES. 1985. High-frequency meiotic gene conversion between repeated genes on non-homologous chromosomes in yeast. Proc. Natl. Acad. Sci. USA **82**: 3350–3354.

JONES, R. S., and S. S. POTTER. 1985. L1 sequences in HeLa extrachromosomal circular DNA: evidence for circularization by homologous recombination. Proc. Natl. Acad. Sci. USA **82**: 1989–1993.

JUBIER-MAURIN, V., B. DOD, M. BELLIS, M. PIECHACZYK, and G. ROIZES. 1985. Comparative study of the L1 family in the genus *Mus:* possible role of retroposition and conversion events in its concerted evolution. J. Mol. Biol. **184**:547–564.

KAUFMAN, R. E., P. J. KRETSCHMET, J. W. ADAMS, H. C. COON, W. F. ANDERSON, and A. W. NIENHUIS. 1980. Cloning and characterization of DNA sequences surrounding the human γ, δ and β globin genes. Proc. Natl. Acad. Sci. USA **77**:4229–4233.

KIMURA, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78**:454–458.

KIOUSSIS, D., E. VANIN, T. DELANGE, R. A. FLAVELL, and F. G. GROSVELD. 1983. β-Globin gene inactivation by DNA translocation in β-thalassaemia. Nature **306**:662–667.

KOLE, L. B., S. R. HAYNES, and W. R. JELINEK. 1983. Discrete and heterogeneous high molecular weight RNAs complementary to a long dispersed repeat family (a possible transposon) of human DNA. J. Mol. Biol. **165**:257–286.

LARHAMMAR, D., B. SERVENIUS, L. RASK, and P. A. PETERSON. 1985. Characterization of an HLA dr-beta pseudogene. Proc. Natl. Acad. Sci. USA **82**:1475–1479.

LERMAN, M., R. THAYER, and M. F. SINGER. 1983. *Kpn*I family of long interspersed repeated DNA sequences in primates: polymorphism of family members and evidence for transcription. Proc. Natl. Acad. Sci. USA **80**:3966–3970.

LOEB, D. D., R. W. PADGETT, S. C. HARDIES, W. R. SHEHEE, M. B. COMER, M. H. EDGELL, and C. A. HUTCHISON III. 1986. The sequence of a large L1Md element reveals a tandemly repeated 5′ end and several features found in retroposons. Mol. Cell. Biol. **6**:168–182.

MANIATIS, T., E. FRITSCH, and J. SAMBROOK. 1982. Molecular cloning. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

MANUELIDIS, L. 1982. Nucleotide sequence definition of a major human repeated DNA, the Hind III 1.9 kb family. Nucleic Acids Res. **10**:3211–3219.

MAXAM, A. M., and W. GILBERT. 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65**:499–560.

MEUNIER-ROTIVAL, M., and G. BERNARDI. 1984. The Bam repeats of the mouse genome belong in several superfamilies the longest of which is over 9 kb in size. Nucleic Acids Res. **12**: 1593–1608.

MIYATA, T., and T. YASUNAGA. 1981. Rapidly evolving mouse α-globin-related pseudogene and its evolutionary history. Proc. Natl. Acad. Sci. USA **78**:450–453.

MOTTEZ, E., P. ROGAN, and L. MANUELIDIS. 1986. Conservation in the 5′ region of the long interspersed mouse L1 repeat: implications of comparative sequence analysis. Nucleic Acids Res. **14**:3119–3135.

POTTER, S. 1984. Rearranged sequences of a human *Kpn*I element. Proc. Natl. Acad. Sci. USA **81**:1012–1016.

ROGAN, P. K. 1987. A study of a long interspersed DNA repeat family common to rodents and primates. Ph.D. diss., Yale University.

SANGER, F., S. NICKLEN, and A. R. COULSON. 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:6463–6467.

SCHERER, S., and R. W. DAVIS. 1980. Recombination of dispersed repeated DNA sequences in yeast. Science **209**:1380–1384.

SCHINDLER, C. W., and M. G. RUSH. 1985. The *Kpn*I family of long interspersed nucleotide sequences is present on discrete sizes of circular DNA in monkey BSC-1 cells. J. Mol. Biol. **181**:161–173.

SHAFIT-ZAGARDO, B., F. L. BROWN, J. J. MAIO, and J. W. ADAMS. 1982*a*. *Kpn*I families of long, interspersed repetitive DNAs associated with the human β-globin gene cluster. Gene **20**:397–407.

SHAFIT-ZAGARDO, B., J. J. MAIO, and F. L. BROWN. 1982*b*. KpnI families of long, interspersed repetitive DNAs in human and other primate genomes. Nucleic Acids Res. **10**:3175–3193.

SHAPIRO, J. A. 1979. Molecular model for the transposition and replication of bacteriophage Mu and other transposable elements. Proc. Natl. Acad. Sci. USA **76**:1933–1937.

SHEN, S., J. L. SLIGHTOM, and O. SMITHIES. 1981. A history of the human fetal globin gene duplication. Cell **26**:191–205.

SINGER, M. F. 1982. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell **28**:433–434.

SINGER, M. F., and J. SKOWRONSKI. 1985. Making sense out of LINEs: long interspersed sequences in mammalian genomes. Trends Biochem. Sci. **10**:119–122.

SOARES, M. B., E. SCHON, and A. EFSTRADIATIS. 1985. Rat LINE1: the origin and evolution of long interspersed middle repetitive DNA elements. J. Mol. Evol. **22**:117–133.

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

STADEN, R. 1982. Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Res. **10**:4731–4751.

STEPHENS, J. C. 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2**:539–556.

SUN, L., K. E. PAULSON, C. W. SCHMID, L. KADYK, and L. LEINWAND. 1984. Non-Alu interspersed repeats in human DNA and their transcriptional activity. Nucleic Acids Res. **12**: 2669–2690.

SWANSTROM, R., W. DELORBE, J. M. BISHOP, and H. VARMUS. 1981. Nucleotide sequence of cloned unintegrated avian sarcoma virus DNA: viral DNA contains direct and inverted repeats similar to those in transposable elements. Proc. Natl. Acad. Sci. USA **78**:124–128.

VAN ARSDELL, S. W., R. A. DENISON, L. B. BERNSTEIN, A. M. WEINER, T. MANSER, and R. F. GESTELAND. 1981. Direct repeats flank three small nuclear RNA pseudogenes in the human genome. Cell **26**:11–22.

VAN DER PLOEG, L., A. KONINGS, M. OORT, D. ROOS, L. BERNINI, and R. A. FLAVELL. 1980. γ-β Thalassaemia studies showing that deletion of the γ- and δ-genes influences β-globin gene expression in man. Nature **283**:637–642.

VANIN, E., P. HENTHORN, D. KIOUSSIS, F. GROSVELD, and O. SMITHIES. 1983. Unexpected relationships between four large deletions in the human β-globin gene cluster. Cell **35**:701–709.

VOLIVA, C. F., C. L. JAHN, M. B. COMER, C. A. HUTCHISON III, and M. H. EDGELL. 1983. The L1Md long interspersed repeat family in the mouse: almost all examples are truncated at one end. Nucleic Acids Res. **11**:8847–8859.

VOLIVA, C. F., S. L. MARTIN, C. A. HUTCHISON III, and M. H. EDGELL. 1984. Dispersal process associated with the L1 family of interspersed repetitive DNA sequences. J. Mol. Biol. **178**: 795–813.

WEINER, A. M., P. L. DEININGER, and EFSTRADIATIS. 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem. **55**:631–661.

WHITE, C. T., S. C. HARDIES, C. A. HUTCHISON III, and M. H. EDGELL. 1984. The diagonal-traverse homology search algorithm for locating similarities between two sequences. Nucleic Acids Res. **12**:751–766.

ROY J. BRITTEN, reviewing editor

Received August 4, 1986; revision received January 5, 1987.