J. Mol. Biol. (1995) **247**, 418–427

JMB

Dispersion and Insertion Polymorphism in Two Small Subfamilies of Recently Amplified Human *Alu* Repeats

Mark A. Batzer[1]*, Carol M. Rubin[2], Utha Hellmann-Blumberg[3] Michelle Alegria-Hartman[1], Esther P. Leeflang[3], Joshua D. Stern[2] Hernan A. Bazan[4], Tamim H. Shaikh[4], Prescott L. Deininger[4,5] and Carl W. Schmid[2,3]

[1]*Human Genome Center L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory P.O. Box 808, Livermore CA 94551, U.S.A.*

[2]*Section of Molecular and Cell Biology University of California Davis, CA 95616, U.S.A.*

[3]*Department of Chemistry, University of California Davis, CA 95616, U.S.A.*

[4]*Department of Biochemistry and Molecular Biology Louisiana State University Medical Center, 1901 Perdido Street, New Orleans LA 70112, U.S.A.*

[5]*Laboratory of Molecular Genetics, Alton Ochsner Medical Foundation, New Orleans, LA 70121, U.S.A*

**Corresponding author*

Newly isolated members of two recently propagated (young) *Alu* subfamilies were examined for sequence diversity and insertion polymorphism in primate genomes. The smaller subfamily (termed HS-2) is comprised of approximately 5 to 25 members, while the larger (termed Sb2) includes approximately 125 to 600 members. Individual members of these *Alu* subfamilies share distinguishing sets of diagnostic mutations, are well-conserved relative to each other, and have expanded in the human lineage. At least one member from each subfamily is known to be polymorphic in humans. Three newly characterized HS-2 *Alu* family members as well as three Sb2 *Alu* repeats are monomorphic (fixed) in humans. The existence of a number of *Alu* subfamilies that have amplified in parallel within the human genome provides compelling evidence for the simultaneous activity of multiple dispersed *Alu* source genes.

Keywords: insertion polymorphism; recent *Alu* subfamilies

Introduction

The *Alu* family of short interspersed elements (SINEs) is one of the most abundant repeats found in mammalian genomes (for recent reviews, see Schmid & Maraia, 1992; Deininger & Batzer, 1993). *Alu* elements, approximately 300 base-pairs (bp) long and present at a copy number in excess of 500,000, are thought to be derived from the 7SL RNA gene and have amplified within primate genomes over the last 65 million years. Each *Alu* element is dimeric in structure, composed of two tandemly arranged halves. *Alu* repeats are typically flanked by short intact direct repeats, and contain a 3' oligo(dA)-rich tail. The left monomer contains an internal RNA polymerase III promoter which is thought to be important in the mobilization of *Alu* repeats. *Alu* elements are thought to have spread throughout the genome by an RNA-mediated transposition process termed retroposition.

M.A.B. and C.M.R. contributed equally to the work described in this paper.

Present address: E. P. Leeflang, Division of Biological Sciences, University of Southern California, Los Angeles, CA 90089-1340, U.S.A.

Abbreviations used: SINE, short interspersed element; PCR, polymerase chain reaction.

0022–2836/95/130418–10 $08.00/0

Alu repeats may be divided into subfamilies or clades of related sequences based upon mutations from the *Alu* consensus sequence that are held in common among members (reviewed by Schmid & Maraia, 1992; Deininger & Batzer, 1993). Nucleotide sequence divergence among *Alu* repeats increases with the age of different subfamilies; therefore, *Alu* repeats seem to have appeared in primate genomes at different times during evolution. The subfamily structure of *Alu* repeats has led to the hypothesis that a small set of *Alu* sequences retain the ability to produce new copies (the ''master'' gene hypothesis; Deininger & Slagel, 1988; Batzer *et al*., 1990; Batzer & Deininger, 1991; Shen *et al*., 1991; Deininger *et al*., 1992). As an alternative, the simultaneous expansion of multiple *Alu* source genes (Matera *et al*., 1990*b*; Leeflang *et al*., 1992; Jurka 1993; Hutchinson *et al*., 1993), or relay of active genes (Britten *et al*., 1988) have also been proposed. Here, we describe the analysis of two young subfamilies of *Alu* repeats which appear to have amplified very recently in human evolutionary history. We also discuss the evidence for simultaneously active *Alu* source genes and the implications for evaluation of alternative *Alu* evolution models.

Since *Alu* subfamily nomenclature is unresolved we will use the following designations in this paper. CS designates the subfamily called ''CS'' by Shen *et al*. (1991) and ''Precise'' by Britten *et al*. (1988) and Matera *et al*. (1990*b*). HS/PV designates the subfamily called ''HS'' by Batzer *et al*. (1990), Batzer & Deininger (1991) and Shen *et al*., 1991 as well as ''PV'' by Matera *et al*. (1990*a*). HS-2 designates the subfamily descended from HS/PV, which incorporates diagnostic mutations at 123, 134, and 166 (Batzer *et al*., 1990; Matera *et al*., 1990*b*; Batzer & Deininger, 1991; Shen *et al*., 1991). Sb2 designates the subfamily descended from the CS *Alu* subfamily which incorporates an eight nucleotide insertion at position 252, as well as seven additional single-nucleotide mutations (Jurka, 1993; Hutchinson *et al*., 1993). A comparison of all of the subfamilies described here except for Sb2 can be found in Deininger & Batzer (1993).

Results

DNA sequence analysis

The oldest *Alu* repeats display greater than 10% divergence from the *Alu* consensus sequence and are estimated to have appeared about 65 million years ago (Shen *et al*., 1991). In contrast, HS-2 sequences show little variation from their consensus and terminate in pure oligo(dA)-rich tails (Figure 1), both of which are indicative of their relative youth (Batzer *et al*., 1990; Matera *et al*., 1990*b*). With the exception of HS C37, which sustained a 14-bp deletion, sequence identity with the HS-2 subfamily consensus sequence is greater than 99% (Figure 1). If CpG positions, which mutate at approximately nine times the rate of non-CpG positions (Bird, 1980), and the deletion are eliminated, the average divergence from the HS-2 consensus is $2/1686 \times 100\% = 0.1\%$. Although the number of changes is too low to be statistically accurate, a neutral rate of evolution of 0.15%/million years (Miyamoto *et al*., 1987) suggests an average age of 660,000 years for HS-2 subfamily members. The low level of nucleotide substitutions is an indication of the recent expansion of these elements within the human genome.

Sb2 *Alu* repeats also quite closely match the subfamily consensus sequence (Figure 1), sharing from 97.9 to 100% nucleotide identity. The Sb2 *Alu* family members listed in Figure 1 show a non-CpG divergence of 0.4%, or an average age of 2.7 million years. Hutchinson *et al*. (1993) previously reported an average age of 4.1 million years for another group of Sb2 *Alu* repeats. Sb2 *Alu* family member D1 is the first known example of an *Alu* repeat which exactly matches its subfamily consensus sequence. As expected from this high degree of nucleotide identity, D1 appears to have arisen very recently in primate evolution, showing a high degree of insertion polymorphism among human populations (Figure 1; see below). However, the presence of the D1 *Alu* repeat in all three diverse population groups surveyed, and absence from the genomes of non-human primates, suggests that it may have arisen just prior to the radiation of modern humans.

Each of the HS-2 and Sb2 *Alu* repeats is flanked by short perfect direct repeats that range in size from 8 to 16 bp. The perfect direct repeats are considered to be an indication of a recent *Alu* retroposition/insertion event. Individual Sb2 and HS-2 subfamily members also contain oligo(dA)-rich tails 11 to 29 bp in length. The oligo(dA)-rich tails arise from the source or ''master'' gene during self-priming or through an as yet undefined post-transcriptional polyadenylation mechanism, since *Alu* sequences do not contain any known signals for post-transcriptional polyadenylation. It is also interesting to note that the 3' end of the D1 Sb2 repeat contains the beginning of a microsatellite with the nucleotide sequence $(A_7T)_3$. Since this *Alu* repeat is an exact match to the subfamily consensus sequence and is flanked by perfect direct repeats the microsatellite is presumably the result of a single mutation in the oligo(dA)-rich region followed by intra-allelic recombination, one of the predominant modes of microsatellite evolution (Levinson & Gutman, 1987).

Characterization of the hybrid sequence 5F4

Alu family member 5F4 is an unusual sequence which contains diagnostic features of two *Alu* subfamilies, Sb2 and HS/PV (Figure 1). Positions 57, 64, 144, 211, 236, 248 and the insertion at 252 display the Sb2 diagnostic mutations, while positions 89, 96, and 98 display HS/PV mutations. The hybrid nature of this *Alu* repeat is probably not an artifactual recombination event occurring during library construction because the direct repeats are maintained at both ends of the *Alu* sequence. The nucleotide sequence of 5F4 also does not appear to be the result of the accumulation of random mutations seen in older *Alu* repeats, since there is only one non-diagnostic mutation from the Sb2 consensus sequence. In addition, polymerase chain reaction (PCR) analysis of the 5F4 locus in humans revealed a PCR product of the expected fragment length. Some of the potential mechanisms for creation of this hybrid are outlined in the Discussion.

Copy number determination

Subfamily copy numbers have been determined by several methods. To determine the copy number of HS / PV and HS-2 *Alu* repeats, a randomly sheared total human genomic library was probed with oligonucleotide HS / PV1 (Materials and Methods) and members of the HS / PV subfamily were isolated and subjected to DNA sequence analysis. Only two of these 150 clones were HS-2 subfamily members, as indicated by the presence of three additional diagnostic mutations. Therefore, approximately 1% of the HS / PV *Alu* sequences are also HS-2 subfamily members. The total number of HS / PV sequences has been estimated at 500 (Batzer *et al*., 1990) to 2500 (Matera *et al*., 1990*a*), so the total number of HS-2 subfamily members estimated by this method is between 5 and 25.

HS/PV and HS-2 copy numbers were also determined independently by diagnostic restriction digestion using the restriction sites outlined in Figure 1. A 197 bp restriction fragment (referred to hereinafter as TT) was released by double digestion of total human DNA with *Taq*I and *Tth*111I. This fragment includes virtually all of the well-conserved HS/PV and HS-2 subfamilies, as well as other sequences (Hellmann-Blumberg *et al*., 1993). To determine the fraction of this fragment contributed by *Alu* sequences, it was end-labeled and digested with additional diagnostic enzymes (Figure 2). Digestion with *Alu*I produced two prominent

```
                                                                                          BspH I        Taq I
                                        10         20         30         40         50         60         70         80
HS/PV Consensus                 GGCCGGGCGC GGTGGCTCAC GCCTGTAATC CCAGCACTTT GGGAGGCCGA GGCGGGCGGA TCACGAGGTC AGGAGATCGA
HSC4N5       AAGGAATCTCA        ...T...... .......... .......... .......... .......... .......... .......... ..........
30B          AAAAGAAAACTAAAC    .......... .......... .......... .......... .......... .......... .......... ..........
1C           GGAATGAGACTG       .......... .......... .......... .......... .......... .......... .......... ..........
HSC37        AAAAACTCAGTGA      .......... .......... .......... .......... .......... .......... .......... ..........

CS Consensus                    .......... .......... .......... .......... .......... .......... .......... ..........
A3:          GAGTGAGACAG        .......... .......... .......... .......... .......... ......T... ...T...... ..........
A6:          CCTGGCCTCCC        -........T .......... .......... .......... .......... ...A..T... ...T...... ..........
B2:          AGAAAGAAGGTCAGAA   .......... .......... .......... ......T... ........A. ......T... ...T...... ..........
D1:          AAAAATGG           .......... .......... .......... .......... .......... ......T... ...T...... ..........
5F4:         AGAAAATACTCTTTTTTA.......... .......... .......... .......... ........A. ......T... ...T...... ..........

                    (Bfa I)               (Bfa I,Bfa I,Bfa I)                   (Bfa I)           (Bfa I)   Alu I
                                                                                                            Bfa I
                                                                                      oligo 1047  A TCAGGATCGA
                        90        100        110        120        130        140        150        160        170
HS/PV Consensus   GACCATCCCG GCTAAAACGG TGAAACCCCG TCTCTACTAA AAATACAAAA AATTAGCCGG GCGTAGTGGC GGGCGCCTGT AGTCCCAGCT
HSC4N5            .......... .......... .........A .......... ..C....... ...-...... .......... .......... .....T....
30B               .......... .......... .......... .......... ..C....... ...-...... .......... .......... .....T....
1C                .......... .......... .......... .......... ..C....... ...-...... .......... .......... .....T....
HSC37             .......... .......... .......... .......... ..C....... ...-...... .......... .......... .....T....

CS Consensus      ........T. .....C.... .......... .......... .......... .......... ....G..... .......... ..........
A3:               .......... .......A.. .......... .......... .......... .......... ...C...... .......... ..........
A6:               A......... .......A.. ........T. .......... .......... .......... ...C...... .......... ..........
B2:               .......... .......A.. .......... .......... .......... .......... ...C...... .......... ..........
D1:               .......... .......A.. .......... .......... .......... .......... ...C...... .......... ..........
5F4:              ........C. .....A.C.. .......... .......... .......... .......... ...C...... .......... ..........

                                                                  Alu I     (Alu I)(Alu I)
                  TGAACCCTC 3' end oligo 1047       Psp AI                         BstY I
                                                                                oligo 655   ACTGCAGTC CGCAGTCCGG
                       180        190        200        210        220        230        240        250
HS/PV Consensus   ACTTGGGAGG CTGAGGCAGG AGAATGGCGT GAACCCGGGA GGCGGAGCTT GCAGTGAGCC GAGATCCCGC CACTGCACTC C-------AG
HSC4N5            .......... .......... .......... .......... .......... .........- .......... .......... .-------..
30B               .......... .......... .......... .......... .......... .......... ........A. .......... .-------..
1C                .......... .......... .......... .......... .......... .......... ......G... .......... .-------..
HSC37             .......... .......... .......... .......--- ---------- -......... .......... .......... .-------..

CS Consensus      ...C...... .......... .......... .......... G......... .......... .....CG... .......... .-------..
A3:               .......... .......... .......... .......... A......... .......... .....T.... .......G.. .GCAGTCCG.
A6:               ....A..... .......... .......... .......... A......... .......... .....T.... .......G.. .ACAGTCCG.
B2:               .......... .......... .......... .......... A......T.. .......... .....T.... ..A....G.. .GCAGTCCG.
D1:               .......... .......... .......... .......... A......... .......... .....T.... .......G.. .GCAGTCCG.
5F4:              .......... .......... .......... .......... A......... .......... .....T.... .......G.. .GCAGTCCG.

                                Tth 111 I/Asp I
                      260        270        280
HS/PV Consensus   CCTGGGCGAC AGAGCGAGAC TCCGTCTC
HSC4N5            .......... .......... ...A....     A25 AAGGAATCTCA
30B               .......... .......... ........     A23 AAAAGAAAACTAAAC
1C                .......... .......... ........     A29 GGAATGAGACTG
HSC37             .......... .......... ........     A11 AAAAACTCAGTGA

                  CCT  end oligo 655
CS Consensus      .......... .......... ........
A3:               .......... .......... ........     A23 GAGTGAGACAG
A6:               .......... .....A.... ........     A16 CCTGGCCTCCC
B2:               .......... .......... ........     A22 GAAAGAAAGAAGGTCAGAA
D1:               .......... .......... ........     (A7T)3A17 AAAAATGG
5F4:              .......... .......... ........     A26 AGAAAATACTCTTTTTT
```

Figure 1. HS-2 and Sb2 *Alu* repeat sequences and diagnostic oligonucleotide probes. HS-2 sequences are compared with the HS / PV subfamily consensus sequence, and Sb2 sequences are compared with the CS subfamily consensus sequence. Direct repeats flanking each *Alu* element are underlined. Deletions are denoted with a (-) relative to the other sequences. Mutations and insertions are denoted with the appropriate base, while (·) represent the same nucleotide that is present in the consensus sequence. Diagnostic oligonucleotide primers are shown above the consensus sequences. Recognition sites for restriction enzymes are also shown, with cryptic restriction sites in parentheses.

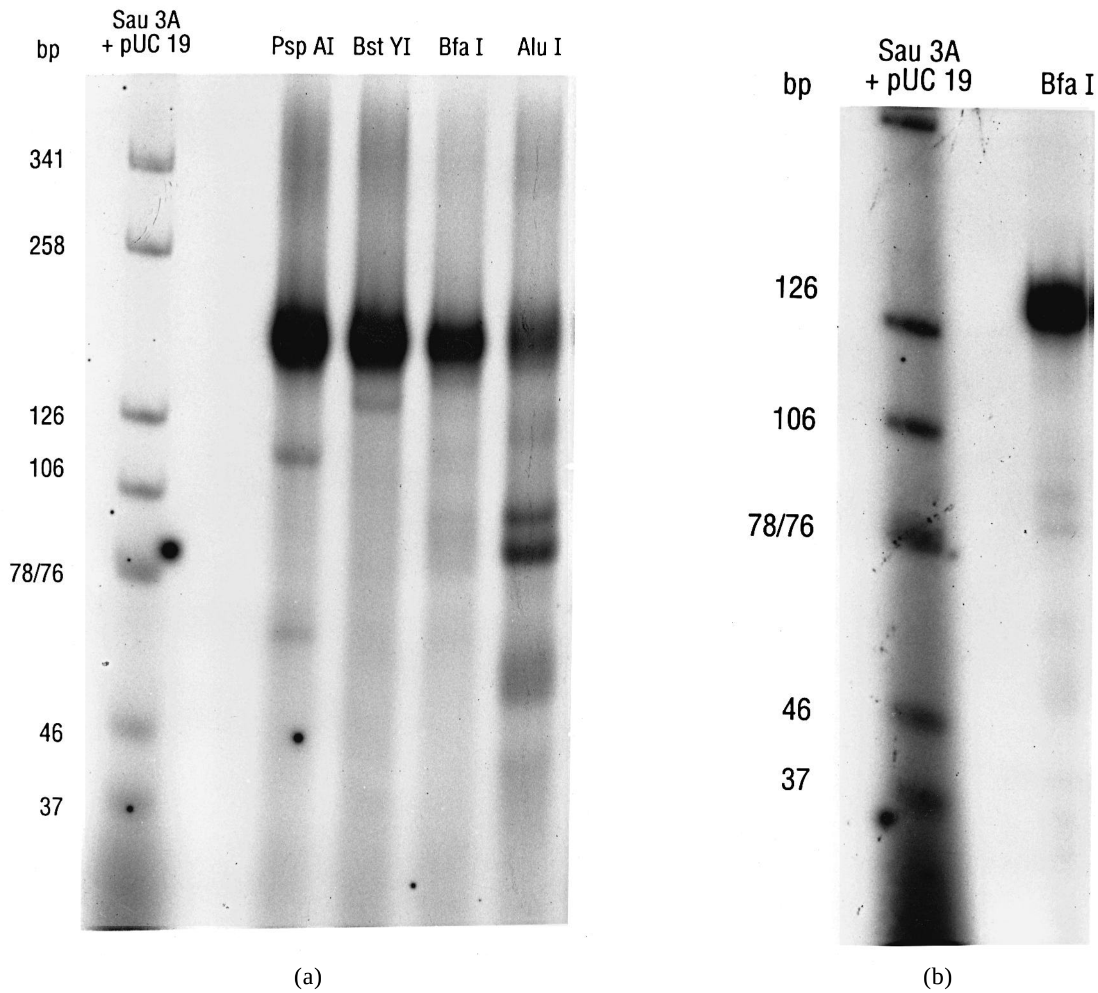


Figure 2. HS-2 *Alu* subfamily quantification. (a) *Psp*AI and *Alu*I digestions of total genomic DNA yield estimations of total *Alu* family content in the *Taq-Tth* restriction fragment (TT). *Bst*YI digestion is diagnostic for the HS/PV subfamily, while *Bfa*I digestion is diagnostic for the HS-2 subfamily. (b) *Bfa*I digestion of the 155 bp fragment released by restricting the TT band with *Bst*YI. The amount of 89 bp restriction fragment released is an estimation of the copy number of HS-2 subfamily members, since it contains both HS/PV and HS-2 diagnostic restriction sites. The fragment length marker was *Sau*3A digested pUC19 with fragment sizes listed in base-pairs.

restriction fragments which resulted from the consensus *Alu*I sites at positions 168 and 217, and fainter fragments from the cryptic sites at positions 228 and 233 (Figure 2(a)). Phosphorimager quantitation revealed that the TT restriction fragment was comprised of 50% *Alu* sequence-related material. When this restriction fragment was digested with *Bst*YI (a diagnostic enzyme for the parent HS/PV subfamily), prominent bands of 155 and 42 bp were released (Figure 2(a)). The 155 bp fragment, enriched for HS/PV sequences, was isolated from the gel and restricted with *Bfa*I to determine what proportion of the *Bst*YI fragment also possessed a *Bfa*I site at position 145, which is diagnostic for HS-2 *Alu* family members (Figure 2(b)). The amount of 89 bp fragment released indicates the number of HS-2 sequences relative to their parent HS/PV sequences (Figure 1). Three trials yielded an average of 1% of the TT restriction fragment that also contained the diagnostic *Bst*YI and *Bfa*I sites. Assuming the true copy number for HS/PV *Alu* repeats is 500 to 2500, analysis by restriction digestion confirms the hybridization and sequencing estimate of 5 to 25 HS-2 *Alu* family members within the human genome (Table 1).

Copy numbers of CS, HS/PV, and Sb2 subfamilies were also determined by Southern hybridization with highly specific probes. Human DNA digested with *Hae*III and *Hin*fI releases a 225 bp *Alu* consensus fragment. The intensity of this band when probed with subfamily-specific oligonucleotides was compared with a dilution series of *Alu* subfamily standards (data not shown). For Sb2, a value of 0.005% of total human DNA, or approximately 500 copies, was obtained. The HS/PV probe showed 0.02% of total DNA, or 2000 copies. The CS probe hybridized to several bands because of the relative divergence of members of this older subfamily. Quantification of the two main *Alu* consensus bands at 225 and 270 bp gave 0.37% of total DNA, or 37,000 copies. These values agree well with previous estimates (Matera *et al*., 1990*a*; Willard *et al*., 1987). A summary of the copy numbers and method of determination for each subfamily of *Alu* repeats reported here is shown in Table 1.

Phylogenetic distribution of *Alu* repeats

The total number of Sb2 *Alu* family members located within the genomes of chimpanzees and humans was compared using the following procedure. *Alu* repeat containing restriction fragments released by double digesting chimpanzee and human DNA samples with either *Bsp*HI / *Asp*I or *Taq*I / *Asp*I are shown in Figure 3. The Southern blot probed with the Sb2-specific oligonucleotide shows the expected 210 and 200 bp fragments only in the human DNA, but not in the chimpanzee DNA, indicating that this subfamily has not greatly expanded in chimpanzees (Figure 3(a)). In contrast, when the blot is hybridized with a non-specific *Alu* probe, the intensity of the chimpanzee lanes is approximately equal to the human lanes (Figure 3(b)).

In order to ascertain the distribution of individual *Alu* repeats in human and non-human primates we interrogated individual *Alu* insertion loci for the presence or absence of each *Alu* sequence as described (Batzer & Deininger, 1991; Batzer *et al*., 1991). Using this analysis, we have previously determined that the TPA 25 HS-2 family member is highly polymorphic in the human population (Batzer & Deininger, 1991; Batzer *et al*., 1991) and that the distribution of this element varies with the geographic origin of the population subgroup being analyzed (Perna *et al*., 1992; Batzer *et al*., 1994). The HS C4N5 HS-2 family member was previously classified as monomorphic (fixed for the presence of the *Alu* repeat) in 79 unrelated individuals (Batzer *et al*., 1991). PV 71, a previously reported HS-2 subfamily member (Matera *et al*., 1990*b*), was also present throughout the human population as assayed by Southern blot hybridization (Hellmann-Blumberg, unpublished results). All of the other HS-2 *Alu* repeats shown in Figure 1 were present (monomorphic for the presence of each *Alu* repeat) in every human tested, and absent from the orthologous positions in the genomes of non-human primates (data not shown).

The Sb2 *Alu* family members shown in Figure 1 were also surveyed using the same PCR-based method. Sb2 repeats A3, A6, and B2 were found to be present (monomorphic for the presence of the *Alu* sequence) in all human populations tested, and absent from orthologous positions in non-human primate genomes (data not shown). *Alu* family member 5F4 is not present in non-human primates, but its status in humans remains uncertain since the pre-integration site for this element appears to be an unidentified repetitive element. The fifth Sb2 repeat, D1, was absent from non-human primate genomes and found at a frequency of 40% or higher in U.S. Caucasians and African-Americans and only at a frequency of 3% in Asian individuals (Table 2). Within each population group, the D1 *Alu* repeat was in Hardy–Weinberg equilibrium as judged by a chi-square test for goodness of fit.

Chromosomal assignment of *Alu* repeats

To determine the location of each *Alu* family member we analyzed human / rodent hybrid cell line DNA panels by PCR as described (Batzer & Deininger, 1991; Batzer *et al*., 1991). Human cells used to construct the hybrids may have one of three genotypes: homozygous for the presence of the *Alu* repeat (+ +), homozygous for the absence of the *Alu* sequence (– –), or heterozygous (+ –). Amplification of hybrid cell line DNA samples will produce a large fragment (400 to 700 bp) if the *Alu* repeat is present and / or a smaller fragment (100 to 350 bp) if it is absent. Amplification of the hybrid cell line panel DNA samples produced DNA fragments 400 to 700 bp in length for all of the loci, which indicate the presence of each *Alu* repeat. Two of the HS-2 *Alu* family members, 30B and HS C4N5, are located on

Table 1

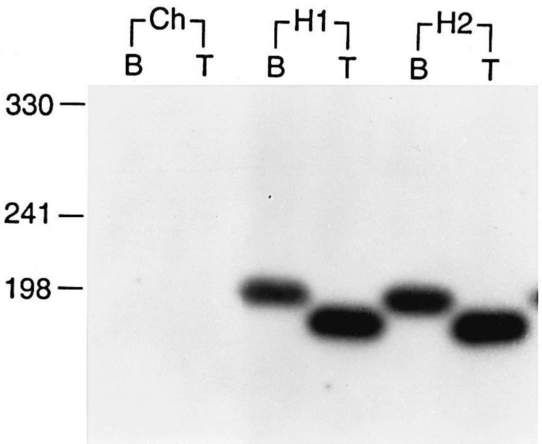
Alu subfamily copy number estimates and methods of detection

Subfamily	Library screening	Restriction digestion	Southern blotting	Database statistics
CS	nd	11,000–22,000[a]	37,000	17,000–34,000
Sb2	nd	nd	500	nd
HS / PV	500–2000	13,000–26,000[b]	2000	420–850
HS-2	5–25	5–25	nd	20–40

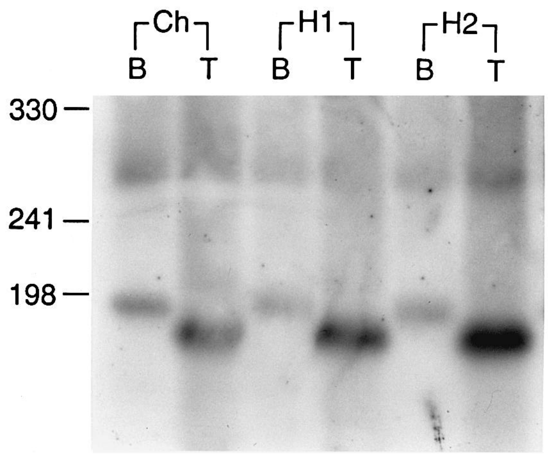
nd, not determined.
Database statistics were taken from Jurka & Milosavljevic (1991).
The estimates refer to the total number of *Alu* repeats within each subfamily in the human genome.

[a] Estimate is based on *Psp*AI digestion. This site includes a frequently mutated CpG and the estimate is undoubtedly low.

[b] Estimate is based on *Bst*YI restriction of the TT band to release a 155 bp fragment (see Materials and Methods). This fragment was difficult to resolve from the much more prominent parental restriction fragment for phosphorimager analysis, resulting in an overestimation of its intensity. *Bst*YI recognizes PuGATCPy, restricting some *Alu* repeats which are not *bona fide* HS / PV subfamily members and also resulting in an overestimate of copy numbers.



(a)



(b)

Figure 3. Comparative distribution of Sb2 *Alu* subfamily members. Total chimpanzee (Ch) or human (H1 or H2) DNA was digested with *Bsp*HI/*Asp*I (B) or *Taq*I/*Asp*I (T). *Bsp*HI is a diagnostic site for Sb2 *Alu* repeats, while *Taq*I restricts all subfamilies of *Alu* repeats. (a) The membrane was probed with Sb2-specific oligonucleotide 655 to detect Sb2 *Alu* family members. (b) The filter was hybridized with a non-specific *Alu* probe to detect all subfamilies of *Alu* repeats. Fragment lengths are shown in base-pairs.

chromosome 19, which is especially rich in *Alu* repeats (Korenberg & Rykowski, 1988). The remaining HS-2 and Sb2 *Alu* repeats are dispersed over a variety of human chromosomes (Table 3).

Discussion

Four different *Alu* subfamilies (CS, HS/PV, HS-2 and Sb2) have expanded simultaneously in humans following their divergence from non-human primates, as demonstrated by insertion presence/absence polymorphism within human populations and/or their absence from orthologous loci in non-human primates. The expansion and inter-relationships of these four subfamilies of *Alu* repeats is outlined in Figure 4. The HS/PV subfamily has expanded independently in humans and chimpanzees (Leeflang *et al*., 1993*a*,*b*), and five HS/PV sequences are known to be polymorphic in humans (Matera *et al*., 1990*b*; Batzer & Deininger, 1991; Batzer *et al*., 1994; Kass *et al*., 1994). Polymorphisms involving CS and HS-2 *Alu* elements have also been reported (reviewed in Schmid & Maraia, 1992 and Deininger & Batzer, 1993; Blonden *et al*., 1994). One particular Sb2 *Alu* repeat segregates with the Huntington disease gene in two families (Hutchinson *et al*., 1993), while a second is located on the Y chromosome of some humans (Hammer, 1994). The continued retroposition activity of the HS/PV and Sb2 subfamilies is illustrated by the recent sporadic (*de novo*) appearance of HS/PV (Wallace *et al*., 1991; Vidaud *et al*., 1993), or Sb2 (Muratani *et al*., 1991) elements in individuals.

The simultaneous expansion of different, yet ancestrally related, *Alu* repeats in the human genome is relevant to possible models of *Alu* family reproduction (see Introduction). Clearly more than one *Alu* sequence is currently capable of retroposition in humans. The formation of approximately 2000 new *Alu* sequences within the human genome over the last five million years (400/million years) is consistent with an amplification of about one *Alu* family member per 100 human births (Deininger & Batzer, 1993). This rate of amplification is currently quite slow in comparison to the amplification rate of *Alu* sequences earlier in primate evolution of 20,000/million years (Shen *et al*., 1991). Thus, the current amplifications of several distinct subfamilies support a ‘‘multiple dispersed source’’ model, in which some retroposed *Alu* repeats (and/or the original ancestral master gene), by fortuitous combination of internal nucleotide sequence and transcriptionally favored location within the genome, retain the ability to propagate (Schmid & Maraia, 1992; Deininger & Batzer, 1993; Jurka, 1993; Hutchinson *et al*., 1993; Brookfield, 1993). However, these recent amplifications represent only a fraction of a percent of the total *Alu* sequences within the human genome which were made over a relatively modest time period (approximately five million years) and themselves are derived from a single, very specific subset of older *Alu* repeats. Thus, these data are still consistent with the possibility that there was one, or a very limited number of master genes, earlier in primate evolution that were either much more powerful (proficient at amplification) or long-lived than the currently active source genes which have produced only 0.4% of all the *Alu* repeats located within the human genome. Even at that point there were probably a number of these relatively weak dispersed source genes, but their amplification rate and longevity were too short to make a long-term impact on *Alu* amplification (Deininger & Batzer, 1993). Alternatively, if there were multiple powerful elements capable of amplification, all but one must have been inactivated about 35 million years ago, allowing fixation down to the CS subfamily alone since all of the more recent *Alu* subfamilies appear to be derived from the CS subfamily lineage as outlined in Figure 4.

Table 2

Distribution of the D1 Sb2 *Alu* family member in humans

Population	Genotype		Observed individuals	Expected individuals	Allele frequencies	
U.S. Caucasians	+	+	7	4.72	+	0.41
	+	–	9	13.56		
	–	–	12	9.72	–	0.59
African-Americans	+	+	8	5.83	+	0.45
	+	–	10	14.34		
	–	–	11	8.83	–	0.55
Asians	+	+	0	0.02	+	0.03
	+	–	1	0.97		
	–	–	15	15.01	–	0.97

The presence and absence of the *Alu* repeat are denoted + and –, respectively. Expected values are based upon Hardy–Weinberg equilibrium.

Alu family member D1, identified in this paper, is present in the genomes of approximately 40% of Caucasians and African-Americans, but is largely absent from Asian individuals. The variable distribution of the D1 *Alu* repeat suggests that it will make a useful marker for the study of human population genetics. The allele frequencies of the D1 insertion suggest that it arose early in human evolution, but was not fixed in the genome at the time modern humans began to radiate. The distribution of the D1 repeat in Asians is probably the result of the low frequency of the insertion in the individuals that gave rise to modern Asians, gene flow from outside Asia back into that geographic region, or genetic drift. Further studies on additional African, Pacific and New World (Amerindian) populations will be required to trace the exact evolutionary history of this *Alu* repeat.

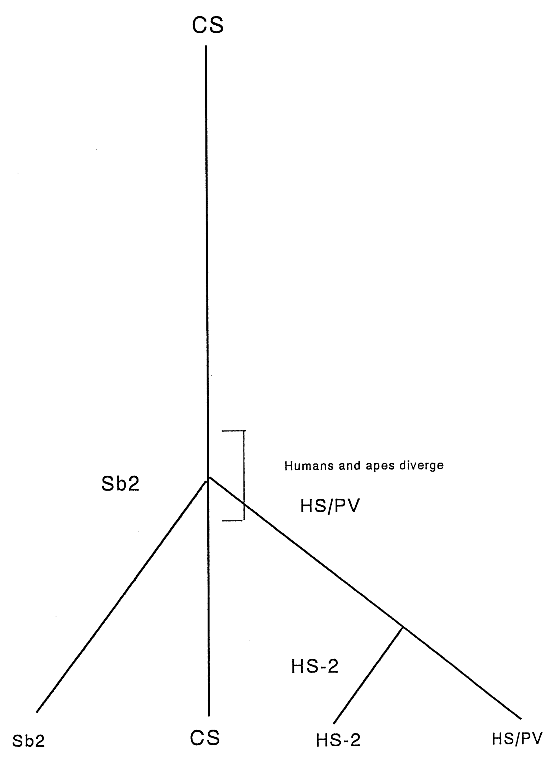


Figure 4. Evolutionary tree of *Alu* subfamily relationships. Primate evolutionary history is shown along the vertical axis with older *Alu* repeats located at the top of the diagram. The *Alu* repeat subfamilies currently propagating in the human genome are shown at the bottom. HS/PV repeats have been found in humans as well as a small number in chimpanzee and gorilla genomes. The branch points for the Sb2 and HS/PV subfamilies correspond approximately to when these subfamilies began to amplify. The time frame of the human African-ape divergence is also shown.

Given that there are approximately one million *Alu* repeats in the human genome, identifiable recombination events that have led to altered subfamily characteristics between *Alu* family members are remarkably rare. Merrit *et al*. (1990) found evidence of *Alu* sequence conversion during a gene duplication event in the α_1-acid glycoprotein genes. Martignetti & Brosius (1993) described a similar event as the probable origin of the BC200β pseudogene. However, the α and β-globin gene clusters in humans and great-apes, despite gene conversion and duplication in the coding regions, include 14 *Alu* repeats that do not display any evidence of gene conversion (Sawada & Schmid, 1986; Koop *et al*., 1986). With the background of these findings, the discovery of a second Sb2 *Alu* repeat which may have undergone gene conversion event is intriguing. Kass *et al*. (1995) have recently shown that the Sb2 *Alu* repeat in the low density lipoprotein receptor (LDLR) gene (Yamamoto *et al*., 1984) has converted from a very old *Alu* repeat in non-human primates to an Sb2 (young) *Alu* repeat in humans. Here, we have reported a new hybrid *Alu* repeat (5F4) which contains mutations characteristic of two young *Alu* repeat subfamilies (HS-2 and Sb2). The majority of previously reported *Alu* repeats is readily identifiable as belonging to a single specific subfamily. The 5F4 *Alu* repeat could have originated as either an Sb2 or HS/PV subfamily member that subsequently acquired the mutations characteristic of the second subfamily by chance. The hybrid *Alu* repeat (5F4) may also be the amplification product of a short-lived master gene that is intermediate in sequence structure, or the result of an HS/PV and Sb2 recombination/gene conversion event.

The first mechanism outlined above requires the concerted acquisition of three or more point mutations by chance. Given that two of the mutations which create the hybrid are relatively rare A⟨-⟩C transversions, and that there are 21 highly mutable CpG dinucleotides in the 5F4 sequence, only one of which has converted (to CpA), the hybrid does not appear to result from exceptional mutability. In the second case, the 5F4 *Alu* repeat may in fact be the product from the amplification of a source gene which is a hybrid in nucleotide sequence structure between the HS-2 and Sb2 subfamilies of *Alu* sequences. If this hybrid subfamily had amplified to any appreciable extent then several additional members of this new hybrid subfamily would have been identified through library screening and DNA sequence analysis of the Sb2 subfamily members. This does not appear to be the case based upon DNA sequence analysis of the remaining Sb2 *Alu* family members reported here as well as those described previously (Jurka, 1993; Hutchinson *et al*., 1993). The third alternative explains the anomalous nucleotide sequence of 5F4 by heteroduplex formation among, or integration and loss involving, two or more *Alu* repeats (Kass *et al*., 1995). In this model, the sequence was corrected to the Sb2 sequence at positions 57 and 64 and to the HS / PV sequence at positions 89, 96 and 98. Similar models for gene conversion have been invoked to explain mutation patterns in yeast Ty elements (Roeder & Fink, 1983) and for the low density lipoprotein receptor Sb2 *Alu* repeat (Kass *et al*., 1995). *Alu*-mediated recombination / gene conversion events may be more common than previously supposed. In fact, the paucity of previously reported gene conversion events involving *Alu* repeats may simply reflect the difficulty involved in the analysis of older *Alu* repeats in general, as a result of the increased number of random mutations in these elements as compared to younger subfamilies of *Alu* sequences. Alternatively, Sb2 *Alu* repeats may possess some property which enables them to gene convert more efficiently than members of other subfamilies.

The identification of *Alu* subfamilies permits critical evaluation of competing models for *Alu* SINE propagation. The features that new *Alu* repeats share (e.g. CpG richness, long oligo(dA)-rich tails) and the favorable environment of the HS / PV founder (Leeflang *et al*., 1993*b*) provide clues to the requirements for *Alu* reproduction. Polymorphisms, useful in determining primate phylogeny, are found only in recently expanded subfamilies (e.g. Sb2 clone D1), and *Alu*-mediated recombination / gene conversion events may be more common for recently inserted *Alu* repeats, or may be more readily recognized.

Materials and Methods

Cell lines and DNA samples

Cell lines used in this study were: *Homo sapiens*, HeLa (ATCC CCL2); *Pan troglodytes*, Wes (ATCC CRL1609); *Gorilla gorilla*, Ggo-1 (primary gorilla fibroblasts) provided by Dr Stephen J. O'Brien; *Cercopithecus aethiops*, CV1 (ATCC CCL70), and *Aotus trivirgatus*, OMK (637-69 ATCC CRL1556). DNA samples from five individual chimpanzees (*Pan troglodytes*), one gorilla (*Gorilla gorilla*), three orangutans (*Pongo pygmaeus*), one macaque (*Macaca fascicularis*), and one marmoset (*Leontopithecus saguinus*) were obtained from BIOS Laboratories. Chromosomal locations were determined by polymerase chain reaction (PCR) amplification of NIGMS human / rodent somatic cell hybrid mapping panels 1 and 2 (Coriell Institute for Medical Research). *Pongo pygmaeus* DNA was also provided by Drs Morris Goodman and Jerry Slightom. Cell lines were maintained as directed by the source and DNA isolations were performed as described (Ausabel *et al*., 1987). Additional human DNA samples were isolated from peripheral lymphocytes (Ausabel *et al*., 1987) available from previous studies. The U.S. Caucasians were from northern European ancestry. The African-American group was collected in New Orleans, Louisiana. The Asian group was comprised of Chinese and Vietnamese individuals.

Library preparation and screening

Two HS-2 *Alu* repeats were isolated from a randomly sheared total human genomic library constructed in bacteriophage λ ZAP II (Batzer *et al*., 1990; Batzer & Deininger, 1991). The library was plated and screened using Magnagraph nylon membranes (Micron Separations Inc.) and (γ-^{32}P) end-labeled (6×10^8 cts / min per μg) oligonucleotide HS / PV1 5'-CACCGTTTTAGC-CGGGATGG-3' as a probe (Ausabel *et al*., 1987), with stringent washes at 65°C in 6 × SSC / 0.05% sodium pyrophosphate (Batzer *et al*., 1990). Excision subcloning and other procedures were as reported (Batzer *et al*., 1992). HS-2 clones 1C and 30B as well as all of the Sb2 *Alu* family members reported here were selected from a human library (Stratagene) constructed by partially digesting DNA with *Sau*3AI and ligating fragments into bacteriophage λ DASH. Approximately six human genome equivalents were plated and the plaques were transferred to nitrocellulose membranes (Schleicher & Schuell) by standard methods (Sambrook *et al*., 1989). For the identification of Sb2 *Alu* family members, filters were hybridized with [γ-^{32}P]ATP end-labeled oligonucleotide 655 (Figure 1) and washed with 5 × SSPE (Sambrook *et al*., 1989) twice at room temperature and once at 60°C. Positive plaques were mapped and fragments containing sequences which hybridized to oligo 655 were subcloned into pUC19 (Sambrook *et al*., 1989). HS-2 *Alu* repeats from the λ DASH library were similarly identified by hybridization with oligonucleotide 1047 (Figure 1).

Southern hybridization

Twelve μg of DNA from several humans and a chimpanzee were digested with enzymes diagnostic for the Sb2 consensus sequence (Figure 1), *Bsp*HI / *Asp*I, or *Taq*I / *Asp*I, fractionated by agarose gel electrophoresis, and immobilized on nitrocellulose (Schleicher & Schuell) by standard methods (Sambrook *et al*., 1989). The membrane was hybridized at 42°C with [γ-^{32}P]ATP end-labeled oligonucleotide 655 (Figure 1) and washed at 60°C. Subsequently, the membrane was stripped and hybridized with a non-specific, full-length *Alu* probe labeled by [α-^{32}P]dATP incorporation during polymerase chain reaction (PCR) amplification.

Table 3

Oligonucleotides for PCR amplification, annealing temperatures and chromosomal locations

Alu repeat	5' Flanking primer	3' Flanking primer	Annealing temperature (°C)	Chromosomal location
A. *HS-2* Alu *repeats*				
HS C37	CTACATGATGTGGGGTGGGCCTGCT	CTTTGGGAGTCCAGCCCACTGTGAA	56	6
PV 30B	GGAAAAGAGTATGGCTGTCT	AACCCAGAAGTGGAATTACA	60	19
PV 1C	TAAGCCCCATAAGGAATGAGACTG	TGTTAGGTACTTTGCTTGGTGCTG	60	12
TPA 25	GTAAGAGTTCCGTAACAGGACAGCT	CCCCACCCTAGGAGAACTTCTCTTT	58	8
HS C4N5	CATCCTTGGCAACTAGTTCCACTCI	ATCATAGACACGGTGTCCTGATCAI	50	19
B. *Sb2* Alu *repeats*				
A3	CCCCAAAGATAGTCAGGTTCTAA	CCTCCTCCATTCTCACTCAATC	60	14
A6	ACTACTCACCAGCAAAACACCTG	GCAGCTATAGCCTTATGAAAACA	60	5
B2	TGGGAAGAGGGTTCTAGTTT	ACTGAAGGACATTAGAGGAC	59	6
D1	TGCTGATGCCCAGGGTTAGTAAA	TTTCTGCTATGCTCTTCCCTCTC	67	3
5F4	CAAATTTTCCTTCAAGAAAATAAAAC	ACATGTATGGTATGTAAAAATTTAG	55	11

All of the oligonucleotides are listed in the 5' > 3' orientation.
Oligonucleotide primers for TPA 25 and HS C4N5 were reported (Batzer & Deininger, 1991).
The chromosomal locations were determined by PCR amplification of hybrid cell line DNA panels as described in Materials and Methods.

DNA sequence analysis

DNA was sequenced by standard dideoxy procedures using Sequenase (U.S. Biochemicals) and [α-^{35}S]dATP with plasmid templates and internal HS *Alu*-specific primers (Batzer *et al*., 1990) according to the manufacturer's (U.S. Biochemical) protocol, or with a GIBCO BRL dsDNA Cycle Sequencing kit and universal primers. Sequences were aligned and analyzed using GeneWorks (Intelligenetics) and GCG (Devereux *et al*., 1984). The DNA sequences isolated in this study have been assigned GenBank accession numbers U02507 (HS C37), U02531 (1C), U02532 (30B), U12580 (A3), U12581 (A6), U12582 (B2), U12583 (D1), and U12584 (5F4).

Copy number quantification by restriction digestion

In order to determine the HS/PV copy number a series of restriction digests were performed and quantified. Genomic DNA (600 μg) was digested with *Taq*I and *Tth*111I. This digestion produced a 197 bp restriction fragment (referred to hereafter as TT), which was isolated by agarose gel electrophoresis, dephosphorylated with calf intestinal phosphatase, end-labeled with [γ-^{32}P]ATP, and precipitated to remove unincorporated label (Sambrook *et al*., 1989). The TT restriction fragment was digested with *Alu*I and chromatographed on a non-denaturing polyacrylamide gel to ascertain what proportion of the fragment resulted from *Alu* repeats. The TT restriction fragment was also digested with *Psp*AI, *Bst*YI, or *Bfa*I, which represent diagnostic restriction sites for CS, HS/PV and HS-2 *Alu* subfamilies respectively, or a combination of *Bst*YI and *Bfa*I to determine the copy number of the HS-2 subfamily. The digested DNA was fractionated on a 12% (w/v) non-denaturing polyacrylamide gel and exposed to Kodak XAR-5 film at –70°C. Gels were then dried and analyzed with a Fujix BAS 1000 phosphorimager for quantification.

PCR amplification

PCR amplification was carried out in 100 μl reactions using 100 ng of target DNA, 750 ng of each oligonucleotide, 200 μM dNTPs in 50 mM KCl, 1.5 mM $MgCl_2$, 10 mM Tris-HCl (pH 8.4) and AmpliTaq DNA polymerase (3.0 units) according to the supplier's (Roche Molecular Diagnostics) instructions. Each sample was subjected to the following amplification cycle; one minute at 94°C (denature), two minutes at the annealing temperature, and two minutes at 72°C (extension), for 30 cycles using the oligonucleotide primers and annealing temperatures listed in Table 3. Twenty μl of each sample was then fractionated on a 2% (w/v) agarose gel with 0.5 μg/ml ethidium bromide. The PCR products were directly visualized using UV fluorescence.

Acknowledgements

We thank Dr Jerzy Jurka for a preprint of his paper describing the Sb2 *Alu* family and Ms Catherine Winter for technical assistance. This work was supported by NIH grants HG 00340 and HG 00770 (P.L.D.), GM 21346 (C.W.S.), the University of California Agricultural Experiment Station (C.W.S.), U.S. Department of Energy LDRD 94-LW-103 (M.A.B.), and CMC 07 from the University of California Center for Molecular Cytometry (M.A.B. and C.W.S.). Work by M.A.B. was conducted under the auspices of the U.S. Department of Energy at Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

References

Ausabel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith J. A. & Struhl, K. (1987). *Current Protocols in Molecular Biology*, John Wiley & Sons, New York.

Batzer, M. A. & Deininger, P. L. (1991). A human-specific subfamily of *Alu* sequences. *Genomics*, **9**, 481–487.

Batzer, M. A., Kilroy, G. E., Richard, P. E., Shaikh, T. H., Desselle, T. D., Hoppens, C. L. & Deininger, P. L. (1990). Structure and variability of recently inserted *Alu* family members. *Nucl. Acids Res.* **18**, 6793–6798.

Batzer, M. A., Gudi, V. A., Mena, J. C., Foltz, D. W., Herrera, R. J. & Deininger, P. L. (1991). Amplification dynamics of human-specific *Alu* family members. *Nucl. Acids Res.* **19**, 3619–3623.

Batzer, M. A., Bazan, H. A., Kim, J., Morrow, S. L., Shaikh, T. H., Arcot, S. S. & Deininger, P. L. (1992). Large-scale subcloning of bacteriophage λ ZAP clones. *BioTechniques*, **12**, 370–371.

Batzer, M. A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D. H., Shaikh, T. H., Novick, G. E., Ioannou, P. A., Scheer, W. D., Herrera, R. J. & Deininger, P. L.

(1994). African origin of human-specific polymorphic *Alu* insertions. *Proc. Nat. Acad. Sci., U.S.A.*, **91**, 12,288–12,292.

Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucl. Acids Res.* **8**, 1499–1504.

Blonden, L. A. J., Terwindt, G. M., Den Dunnen, J. T. & Van Ommen, G.-J. B. (1994). A polymorphic STS in intron 44 of the dystrophin gene. *Human Genet.* **93**, 479–480.

Britten, R. J., Baron, W. F., Stout, D. B. & Davidson, E. H. (1988). Sources and evolution of human *Alu* repeated sequences. *Proc. Nat. Acad. Sci., U.S.A.* **85**, 4770–4774.

Brookfield, J. F. Y. (1993). The generation of sequence similarity in SINEs and LINEs (letter). *Trends Genet.* **9**, 38.

Deininger, P. L. & Batzer, M. A. (1993). Evolution of retroposons. In *Evolutionary Biology* (Hect, M. K., MacIntyre, R. J. & Clegg, M. T., eds), vol. 27, pp. 157–196, Plenum Press, New York.

Deininger, P. L. & Slagel, V. K. (1988). Recently amplified *Alu* family members share a common parental *Alu* sequence. *Mol. Cell. Biol.* **8**, 4566–4569.

Deininger, P. L., Batzer, M. A., Hutchison, C. A., III & Edgell, M. H. (1992). Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8**, 307–311.

Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res.* **12**, 387–395.

Hammer, M. F. (1994). A recent insertion of an *Alu* element on the Y chromosome is a useful marker for human population studies. *Mol. Biol. Evol.* **11**, 749–761.

Hellmann-Blumberg, U., Hintz, M. F., Gatewood, J. M. & Schmid, C. W. (1993). Developmental differences in methylation of human *Alu* repeats. *Mol. Cell. Biol.* **13**, 4523–4530.

Hutchinson, G. B., Andrew, S. E., McDonald, H., Goldberg, Y. P., Grahm, R., Rommens, J. M. & Hayden, M. R. (1993). An *Alu* element retroposition in two families with Huntington disease defines a new active *Alu* subfamily. *Nucl. Acids Res.* **21**, 3379–3383.

Jurka, J. (1993). A new subfamily of recently retroposed *Alu* repeats. *Nucl. Acids Res.* **21**, 2252.

Jurka, J. & Milosavljevic, A. (1991). Reconstruction and analysis of human *Alu* genes. *J. Mol. Evol.* **32**, 105–21.

Kass, D. H., Aleman, C., Batzer, M. A. & Deininger, P. L. (1994). Identification of a human specific *Alu* insertion in the factor XIIIB gene, *Genetica*, **94**, 1–8.

Kass, D. H., Batzer, M. A. & Deininger, P. L. (1995). Gene conversion as a secondary mechanism of SINE evolution, *Mol. Cell. Biol*. **15**, 19–25.

Koop, B. F., Miyamoto, M. M., Embury, J. E., Goodman, M., Czelusniak, J. & Slightom, J. L. (1986). Nucleotide sequence and evolution of the orangutan ϵ globin gene region and surrounding *Alu* repeats. *J. Mol. Evol.* **24**, 94–102.

Korenberg, J. R. & Rykowski, M. C. (1988). Human genome organization: Alu, Lines, and the molecular structure of metaphase chromosome bands. *Cell*, **53**, 391–400.

Leeflang, E. P., Liu, W.-M., Hashimoto, C., Choudary, P. V. & Schmid, C. W. (1992). Phylogenetic evidence for multiple *Alu* source genes. *J. Mol. Evol.* **35**, 7–16.

Leeflang, E. P., Chesnokov, I. N. & Schmid, C. W. (1993*a*). Mobility of short interspersed repeats within the chimpanzee lineage. *J. Mol. Evol.* **37**, 566–572.

Leeflang, E. P., Liu, W.-M., Chesnokov, I. N. & Schmid, C. W. (1993*b*). Phylogenetic isolation of a human *Alu* founder gene: drift to new subfamily identity. *J. Mol. Evol.* **37**, 559–565.

Levinson, G. & Gutman, G. A. (1987). Slipped-strand mispairing: major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**, 203–221.

Martignetti, J. A. & Brosius, J. (1993). BC200 RNA: a neural RNA polymerase III product encoded by a monomeric *Alu* element. *Proc. Nat. Acad. Sci., U.S.A.* **90**, 11563–11567.

Matera, A. G., Hellmann, U. & Schmid, C. W. (1990*a*). A transpositionally and transcriptionally competent *Alu* subfamily. *Mol. Cell. Biol.* **10**, 5424–5432.

Matera, A. G., Hellmann, U., Hintz, M. F. & Schmid, C. W. (1990*b*). Recently transposed *Alu* repeats result from multiple source genes. *Nucl. Acids Res.* **18**, 6019–6023.

Merrit, C. M., Easteal, S. & Board, P. G. (1990). Evolution of human α_1-acid glycoprotein genes and surrounding *Alu* repeats. *Genomics*, **6**, 659–665.

Miyamoto, M. M., Slightom, J. L. & Goodman, M. (1987). Pylogenetic relations of humans and African apes from DNA sequences in the pseudo-η-globin region. *Science*, **238**, 369–373.

Muratani, K., Hada, T., Yamamoto, Y., Kaneko, T., Shigeto, Y., Ohue, T., Furuyama, J. & Higashino, K. (1991). Inactivation of the cholinesterase gene by *Alu* insertion: plausible mechanism for human gene transposition. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 11315–11319.

Perna, N. T., Batzer, M. A., Deininger, P. L. & Stoneking, M. (1992). *Alu* insertion polymorphism: a new type of marker for human population studies. *Human Biol.* **64**, 641–648.

Roeder, G. S. & Fink, G. R. (1983). Transposable elements in yeast. In *Mobile Genetic Elements* (Shapiro, J. A., ed.), pp. 299–328, Academic Press, New York.

Sambrook, J., Fritsch, E. F. & T. Maniatis. (1989). *Molecular Cloning: A Laboratory Manual*. 2nd edit., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sawada, I. & Schmid, C. W. (1986). Primate evolution of the α-globin gene cluster and its *Alu*-like repeats. *J. Mol. Biol.* **192**, 693–709.

Schmid, C. W. & Maraia, R. (1992). Transcriptional and transpositional selection of active SINE sequences. *Curr. Opin. Genet. Develop.* **2**, 874–882.

Shen, M. R., Batzer, M. A. & Deininger, P. L. (1991). Evolution of the ''master'' *Alu* gene(s). *J. Mol. Evol.* **33**, 311–320.

Vidaud, D., Vidaud, M., Bahnak, B. R., Siguret, V., Sanchez, S. G., Laurin, Y., Meyer, D., Goossens, M. & Lavergne, J. M. (1993). Hemophilia B due to *a de novo* insertion of a human-specific *Alu* subfamily member within the coding region of the factor IX gene. *Eur. J. Human Genet.* **1**, 30–36.

Wallace, M. R., Andersen, L. B., Saulino, A. M., Gregory P. E., Glover, T. W. & Collins, F. S. (1991). A *de novo Alu* insertion results in neurofibromatosis type 1. *Nature (London)*, **353**, 864–866.

Willard, C., Nguyen, H. T. & Schmid, C. W. (1987). Existence of at least three distinct *Alu* subfamilies. *J. Mol. Evol.* **26**, 180–186.

Yamamoto, T., Davis, C. G., Brown, M. S., Schneider, W. J., Casey, M. L., Goldstein, J. L. & Russell, D. W. (1984). The human LDL receptor: a cysteine-rich protein with multiple *Alu* sequences in its mRNA. *Cell*, **39**, 27–38.

Edited by J. Karn

(Received 14 September 1994; accepted 13 December 1994)