# Genome-wide analysis of the human *Alu* Yb-lineage

*Anthony B. Carter,[1†] Abdel-Halim Salem,[1,2†] Dale J. Hedges,[1] Catherine Nguyen Keegan,[3] Beth Kimball,[3] Jerilyn A. Walker,[1] W. Scott Watkins,[4] Lynn B. Jorde[4] and Mark A. Batzer[1,3]**

[1]Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA
[2]Department of Anatomy, Faculty of Medicine, Suez Canal University, Ismailia, Egypt
[3]Department of Pathology, Louisiana State University Health Sciences Center, 1901 Perdido Street, New Orleans, LA 70112, USA
[4]Department of Human Genetics, University of Utah Health Sciences Center, Salt Lake City, UT 84112, USA
**Correspondence to:* Tel: +1 225 578 7102; Fax: +1 225 578 7113; E-mail: mbatzer@lsu.edu
[†]These authors contributed equally to this research.

*Date received (in revised form): 10th December 2003*

## Abstract

The *Alu* Yb-lineage is a 'young' primarily human-specific group of short interspersed element (SINE) subfamilies that have integrated throughout the human genome. In this study, we have computationally screened the draft sequence of the human genome for *Alu* Yb-lineage subfamily members present on autosomal chromosomes. A total of 1,733 Yb *Alu* subfamily members have integrated into human autosomes. The average ages of Yb-lineage subfamilies, Yb7, Yb8 and Yb9, are estimated as 4.81, 2.39 and 2.32 million years, respectively. In order to determine the contribution of the *Alu* Yb-lineage to human genomic diversity, 1,202 loci were analysed using polymerase chain reaction (PCR)-based assays, which amplify the genomic regions containing individual Yb-lineage subfamily members. Approximately 20 per cent of the Yb-lineage *Alu* elements are polymorphic for insertion presence/absence in the human genome. Fewer than 0.5 per cent of the Yb loci also demonstrate insertions at orthologous positions in non-human primate genomes. Genomic sequencing of these unusual loci demonstrates that each of the orthologous loci from non-human primate genomes contains older Y, Sg and Sx *Alu* family members that have been altered, through various mechanisms, into Yb8 sequences. These data suggest that *Alu* Yb-lineage subfamily members are largely restricted to the human genome. The high copy number, level of insertion polymorphism and estimated age indicate that members of the *Alu* Yb elements will be useful in a wide range of genetic analyses.

***Keywords:*** *mobile elements, SINEs*

## Introduction

Short interspersed elements (SINEs) are a class of retroelements that are typically less than 500 nucleotides long and lack open reading frames (ORFs).[1–5] *Alu* elements are dimeric, primate-specific SINEs that have reached a copy number in excess of one million in the human genome.[6] *Alu* elements have reached this copy number via an RNA-mediated retroposition process that is dependent on the recruitment of an Line-1 (L1) protein possessing both reverse transcriptase and endonuclease activity.[7–9] Amplification of *Alu* subfamilies is thought to occur from a limited number of 'master' or 'source' genes that are retropositionally competent.[10,11] New *Alu* subfamilies are created when mutations occur in the source or master gene sequence and subsequently give rise to new lineages of elements that share the novel diagnostic mutation(s). Various *Alu* element subfamilies have dispersed throughout primate genomes at different time periods giving rise to a hierarchical series of elements that are of different genetic ages.[12] Since *Alu* elements first appeared in the primate genome, their amplification rate has fluctuated and is thought to be currently 100-fold slower than the peak rate 40 million years ago.[13,14]

The Yb-lineage, consisting of Yb7, Yb8 and Yb9, is the second largest 'young' group of evolutionarily-related *Alu* subfamilies in the human genome. In this study, the authors searched available genomic databases to annotate all *Alu* Yb elements present in the draft sequence of the human autosomal genome. In addition, a screen of diverse human genomes was conducted to identify those Yb elements that are polymorphic within the human population. In this paper, the authors determine the average age, genomic distribution and human diversity of 1,733 autosomal *Alu* Yb elements (see Figure 1).

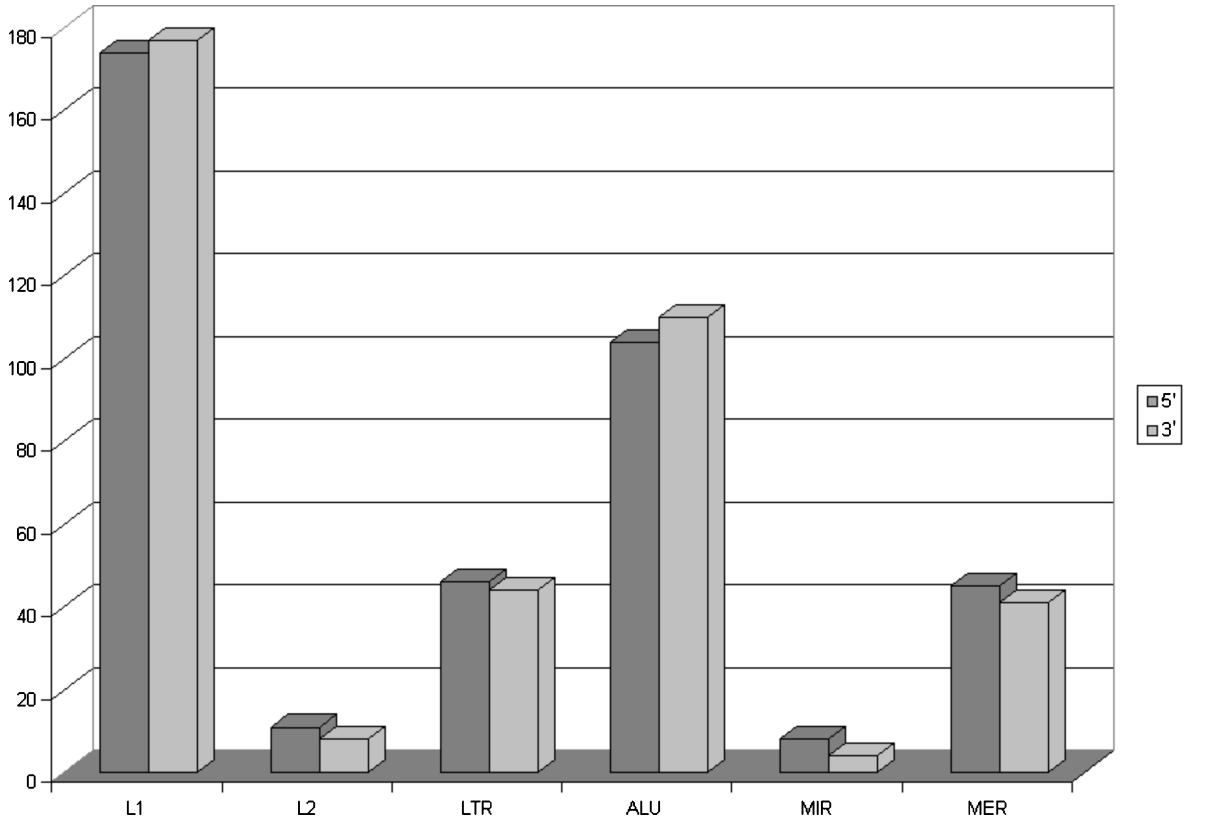


**Figure 1.** *Alu* Yb integrations within other human repeated sequences. *Alu* insertions within other known mobile elements were grouped according to the element in which they inserted. Mobile element categories included: LINE-1 (L1), LINE-2 (L2), long terminal repeats (LTR), *Alu* (*ALU*), mammalian-wide interspersed repeats (MIR) and medium reiteration frequency sequences (MERs).

## Materials and methods

### Computational analyses

Screening of the National Center for Biotechnology Information's (NCBI's) GenBank non-redundant human genome genetic database and the University of California at Santa Cruz 2001 human genome draft sequence was performed using a local installation of BLAST (available at NCBI (http://www.ncbi.nlm.nih.gov/)) to locate all chromosomal locations of *Alu* Yb subfamily members.[16] As search criteria, a 21 base pair oligonucleotide (5′-ACTGCAGTCCGCAGTCCGGCC-3′) that is unique to all Yb *Alu* subfamily members was used to locate individual elements in the draft human genomic sequence. Only those elements that contained an exact sequence complement to the search oligonucleotide were retained for further analyses. Once a Yb *Alu* element was located, a 700–1,200 base pair fragment, which included the *Alu* and adjacent sequence, was placed into the University of Washington Genome Center's RepeatMasker Web server (http://repeatmasker.genome.washinton.edu/cgi-bin/RepeatMasker) for annotation of all identifiable repeat sequences. Precise locations of the Yb *Alu* insertions were determined during the annotation process. *Alu* Yb elements extracted from the draft human genomic sequence were aligned using MEGALIGN (DNASTAR V.5) to determine mutation density. Multiple sequence alignments of all the *Alu* Yb-lineage subfamily members can be found on the authors' website (http://batzerlab.lsu.edu).

### Cell lines and DNA samples

Cell lines used to isolate DNA samples were as follows: human (*Homo sapiens*, HeLa ATCCCCL2); common chimpanzee (*Pan troglodytes*, ATCCCRL1609); lowland gorilla, (*Gorilla gorilla*, AG05253B); orangutan (*Pongo pygmaeus*, ATCCCRL6301); green monkey (*Cercopithecus aethiops*, ATCCCCL70); owl monkey (*Aotus trivirgatus*, ATCCCRL1556); and pygmy chimpanzee (*Pan paniscus*, AG05253A). Human DNA from South American populations was purchased as part of the Human Variation Panel available from the Coriell Institute for Medical Research. Additional human DNA samples from the European, African-American and Asian population groups was isolated from peripheral blood lymphocytes available from previous studies.[15]

### Primer design and PCR amplification

Oligonucleotide primers for the PCR amplification of each *Alu* element were designed using the 700–1,200 base pair flanking unique sequence fragments and Primer3 software (Whitehead Institute for Biomedical Research, Cambridge, MA, USA; http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). The sequences of the oligonucleotide primers, annealing temperatures, polymerase chain reaction (PCR) product sizes and chromosomal locations for all Yb-lineage *Alu* elements can be found on the authors' website (address given above). The primers were subsequently screened against the GenBank non-redundant database to determine if they resided in a unique DNA sequence. PCR amplification was performed in 25 μl aliquots using 10–50 ng of target DNA, 200 nM of each oligonucleotide primer, 200 μM deoxynucleotide triphosphates (dNTPs) in 50 mM KCl, 1.5 mM $MgCl_2$, 10 mM tris-HCl buffer (pH 8.4) and 1 unit *Taq* DNA polymerase. Each sample was subjected to an initial denaturation step at 94°C for 150 seconds, followed by 32 PCR cycles of one minute of denaturation at 94°C, one minute at the annealing temperature and one minute of extension at 72°C, followed by a final extension step at 72°C for ten minutes.

### DNA sequencing

DNA sequencing was performed on gel-purified PCR products that had been cloned using the TOPO TA cloning vector (Invitrogen) using chain termination sequencing on an Applied Biosystems 3100 automated DNA sequencer.[17]

The sequence of the non-human primate Yb7AD231, Yb7AD84, Yb8AC1233, Yb8AC914a, Yb7AD234, Yb7AD52 and Yb8AC1737 orthologous loci have been assigned accession numbers (AY345942–AY345966). Sequence alignments for all of the Yb-lineage subfamily members were performed using MegAlign software (DNAStar version 3.1.7 for Windows 3.2).

### Statistical analyses

A comparison of Yb insertion distribution among human chromosomes was conducted using $\chi^2$ tests with one degree of freedom. The expected number of insertions for each chromosome was estimated based on the proportion of the total genomic sequence that the individual chromosome represented.[18] Hardy-Weinberg equilibrium tests were conducted using $\chi^2$ tests and a Markov chain method implemented using Arlequin software.[19,20]

## Results

### Yb element copy number and chromosomal distribution

A total of 1,733 Yb-lineage *Alu* elements were detected within autosomal chromosomes (Table 1). With the addition of 118 subfamily elements previously found on the sex chromosomes, 1,851 Yb elements have been recovered from the human draft sequence.[21] Twenty-four per cent (417) of the autosomal Yb elements were found to be integrated within other repeated

**Table 1.** *Alu* Yb element polymerase chain reaction (PCR) analysis summary

| | ***Alu* Yb elements** |
|---|---|
| **Loci analysed by PCR** | **1,202** |
| Fixed present | 962 |
| High frequency insertion polymorphisms | 18 |
| Intermediate frequency insertion polymorphisms | 181 |
| Low frequency insertion polymorphisms | 41 |
| **Total polymorphic** | 240 |
| Paralogues | 32 |
| **Loci not analysed by PCR** | **531** |
| Inserted in other repeats | 417 |
| No PCR results | 112 |
| End of contig | 2 |
| **Total autosomal elements analysed** | **1,733** |
| **Total sex chromosome elements analysed*** | **118** |

*From Callinan *et al.* (2003).[21]

sequences and were therefore not amenable to PCR. From a total of 1,314 *Alu* Yb elements, 112 produced inconclusive PCR results leaving 1,202 *Alu* Yb loci from which we were able to obtain PCR results.

Chi squared tests were performed on frequency data from all chromosomes to test the distribution of *Alu* Yb element insertions against a random insertion model in which the number of insertions on each chromosome was proportional to the size of the chromosome (Table 2). Individual chromosome distribution was assessed based on the size of the chromosome and a total number of 1,851 Yb *Alu* elements recovered from the draft sequence of the human genome.[15,22] Chromosomes 1, 2, 5, 6, 7 and Y were statistically different from the random insertion model at a 5 per cent significance level.

## Yb sequence attributes

The vast majority of Yb *Alu* loci that were annotated in this study contained direct repeats ranging from 4–22 base pairs in

**Table 2.** Distribution of all autosomal *Alu* Yb family members

| Chromosome | Percentage of the human genome | Number of observed *Alu* elements | Number of expected *Alu* elements | S/NS* |
|---|---|---|---|---|
| 1 | 8.01 | 123 | 148 | S |
| 2 | 7.93 | 121 | 147 | S |
| 3 | 6.54 | 119 | 121 | NS |
| 4 | 6.28 | 123 | 116 | NS |
| 5 | 5.96 | 141 | 110 | S |
| 6 | 5.59 | 125 | 104 | S |
| 7 | 5.16 | 128 | 96 | S |
| 8 | 4.80 | 93 | 89 | NS |
| 9 | 4.36 | 75 | 81 | NS |
| 10 | 4.41 | 66 | 82 | NS |
| 11 | 4.48 | 82 | 83 | NS |
| 12 | 4.37 | 78 | 81 | NS |
| 13 | 3.65 | 76 | 68 | NS |
| 14 | 3.32 | 61 | 61 | NS |
| 15 | 3.17 | 51 | 59 | NS |
| 16 | 2.99 | 44 | 55 | NS |
| 17 | 2.76 | 51 | 51 | NS |
| 18 | 2.56 | 47 | 47 | NS |
| 19 | 1.95 | 27 | 36 | NS |
| 20 | 2.06 | 41 | 38 | NS |
| 21 | 1.47 | 35 | 27 | NS |
| 22 | 1.57 | 26 | 29 | NS |
| X | 4.97 | 98 | 92 | NS |
| Y | 1.65 | 20 | 31 | S |
| **Total number of elements** | | **1,851** | | |

*Statistically significant (S) or not statistically significant (NS).

length, which are generated during the integration process. A total of 55 elements, however, had no discernable direct repeats. These may have either been the result of unorthodox integrations or, alternatively, mutations and/or rearrangements subsequent to integration that resulted in the loss of repeat sequences.

The oligo-(dA)-rich tails and internal A-rich regions of *Alu* elements have previously been shown to serve as seeds for the genesis of simple sequence repeats.[23–28] The oligo-(dA)-rich tails of each *Alu* element ranged from 3–144 base pairs in length. Approximately 5 per cent (91/1,733) of the Yb *Alu* family members had simple sequence repeats in their tails. Only one element, Yb8AC1733, did not possess an oligo-(dA) tail.

Incomplete reverse transcription or improper integration into the genome occasionally truncates individual *Alu* subfamily members. The authors found that the 5′ regions of the *Alu* elements were more susceptible to truncations than the 3′ regions, which is consistent with the current *Alu* retrotransposition model, as it proposes that reverse transcription initiates at the 3′ end of the *Alu* sequence.[29] A total of 244 Yb *Alu* elements were found to have collectively lost 17,033 base pairs of 5′ *Alu* sequence. Analysis of the 3′ ends of Yb elements showed no appreciable sequence truncations.

Examination of the GC content of 1 kb of 5′ and 3′ flanking genomic sequence from each of the 1,733 autosomal Yb *Alu* elements indicates that their integration is specific to regions where GC content approximates 39 per cent. The flanking regions of Yb *Alu* elements that have integrated into other known human repeats were also analysed for repeat content (Figure 1). Approximately 40 per cent of Yb *Alu* repeats integrated within L1 elements and 26 per cent integrated within other, older pre-existing *Alu* elements. Twelve of the Yb elements contained an independent, full-length *Alu* element either in the oligo-(dA)-rich tail or immediately adjacent to the Yb *Alu* family member, such that both elements shared a single set of direct repeats. Data are available on the authors' website.

## Human diversity

The human genetic diversity associated with each *Alu* Yb locus was estimated using individuals from four diverse populations (African-American, Asian, European and Egyptian). Amplification of human autosomal loci revealed that 20 per cent of Yb-lineage elements were polymorphic for insertion presence/absence. The heterozygosity of each polymorphic *Alu* element was also calculated. Allele frequencies for each locus, as well as the associated heterozygosity calculations, are available in tabular form on the authors' website.

Individual chromosomal insertion polymorphism rates were found to be as low as 13 per cent (chromosome 4) and as high as 32 per cent (chromosome 20) (Table 3). The average polymorphism rate for all 22 autosomes was 20 per cent. All of the insertion polymorphisms were subsequently categorised as high (HF), intermediate (IF) or low frequency (LF) as previously described in Carroll *et al.*[15] Allele frequency was classified as:

fixed present (FP)
low frequency (LF)
intermediate (IF)
high frequency (HF) insertion polymorphism.

Fixed present: every individual tested had the *Alu* element in both chromosomes. Low frequency insertion polymorphism: the element is absent in all individuals tested, except for one or two homozygous or heterozygous individuals. Intermediate frequency insertion polymorphism: the *Alu* element is variable as to its presence or absence in at least one population. High frequency insertion polymorphism: the element is present in all individuals in the populations

**Table 3.** Distribution of autosomal *Alu* Yb insertion polymorphisms

| Chr | Number polymorphic | Total on chr | % Polymorphism |
|---|---|---|---|
| 1 | 15 | 88 | 17 |
| 2 | 9 | 80 | 11 |
| 3 | 12 | 77 | 16 |
| 4 | 11 | 86 | 13 |
| 5 | 21 | 105 | 20 |
| 6 | 19 | 94 | 20 |
| 7 | 19 | 84 | 23 |
| 8 | 17 | 58 | 29 |
| 9 | 11 | 49 | 22 |
| 10 | 8 | 47 | 17 |
| 11 | 7 | 44 | 16 |
| 12 | 10 | 53 | 19 |
| 13 | 11 | 58 | 19 |
| 14 | 10 | 48 | 21 |
| 15 | 9 | 36 | 25 |
| 16 | 9 | 30 | 30 |
| 17 | 7 | 36 | 19 |
| 18 | 7 | 29 | 24 |
| 19 | 4 | 13 | 31 |
| 20 | 10 | 31 | 32 |
| 21 | 8 | 27 | 30 |
| 22 | 5 | 25 | 25 |

chr, chromosome.

tested, except for one or two heterozygous or absent individuals. These categories comprise 8 per cent (18), 75 per cent (181) and 17 per cent (41) of the polymorphisms, respectively. The genome-wide distribution of various frequency classes of the polymorphic Yb *Alu* elements is shown in Figure 2. The physical positions of the Yb *Alu* insertion polymorphisms on both autosomal and sex chromosomes were determined using the BLAST-like Alignment Tool (BLAT) (Table 3).

A total of 888 $\chi^2$ analyses were performed on polymorphic *Alu* Yb elements to determine if they were in Hardy-Weinberg equilibrium. Fifty-seven deviations from Hardy–Weinberg were recorded ($p < 0.05$). Thirty-five of these deviations were the result of low expected values, however. A total of 44 deviations from Hardy–Weinberg would be expected by chance alone at the $p = 0.05$ significance level. A Markov chain model was then applied to the data using the Arlequin population data analysis software.[19,20] Out of 888 comparisons, the test yielded only eight that were significant ($p < 0.01$), which is the number that would have been expected by chance alone. The results of both tests suggest that Yb *Alu* insertion polymorphisms as a whole do not significantly depart from Hardy–Weinberg equilibrium.

## Evolutionary age estimates

Average age estimates of the Yb7, Yb8 and Yb9 subfamilies were calculated using the CpG dinucleotide and non–CpG nucleotide mutation densities as described in Carroll *et al.*[15] Consensus sequences for each subfamily are shown in Figure 3. A total of 157 elements (42,233 nucleotide bases) from the annotated autosomal Yb7 elements were used in both CpG and non-CpG mutation calculations in determining Yb7 age estimates. There were 169 CpG mutations (out of 7,222 nucleotide bases analysed) and 253 non-CpG mutations (out of 35,011 non-CpG bases analysed), yielding a mutation density of 0.0230 and 0.0070, respectively. A total of 994 Yb8 autosomal elements (267,557 nucleotide bases) were used in CpG and non-CpG age estimates. There were 1,050 CpG mutations (out of 45,724 nucleotide bases analysed) and 798 non-CpG mutations (out of 221,833 non-CpG bases analysed), yielding a mutation density of 0.0229 and 0.00359, respectively. A total of 63 elements (16,947 nucleotide bases) from the autosomal Yb9 elements were used in both CpG and non-CpG mutation calculations in determining Yb9 age estimates. A total of 47 CpG mutations (out of 2,898 nucleotide bases analysed) and 49 non-CpG mutations (out of 14,049 non-CpG bases analysed) yielded a mutation density of 0.0160 and 0.0030, respectively. Using neutral mutation rates of 1.46 per cent/million years for CpG bases and 0.15 per cent/million years for non-CpG bases, the average ages of the Yb7, Yb8 and Yb9 subfamilies were calculated.[15,30–32] CpG and non-CpG age estimations for Yb7 subfamily were 1.6 million years and 4.81 million years, respectively. For the Yb8 subfamily, CpG- and non-CpG-based age estimates were 1.57 million years and 2.39 million years, respectively. For the Yb9 subfamily, CpG and non-CpG age estimates were 1.11 million years and 2.32 million years, respectively. These age estimations are in good agreement with previous estimates for the *Alu* Yb subfamily.[15]

## *Alu* Yb-lineage origin and orthologous insertions

To determine the approximate time of insertion of each Yb locus during primate evolution, orthologous loci from several non-human primates were PCR amplified to detect the presence or absence of individual *Alu* inserts. Non-human primate PCR amplification resulted in the recovery of seven loci that appeared to contain a Yb *Alu* element (Table 4). Analysis of the DNA sequences from the non-human primate PCR products, however, showed that the orthologous loci contained *Alu* elements from older Y, Sx, Sg or Sc subfamilies (see below).

Nine Yb-lineage *Alu* elements yielded PCR results indicating the presence of an *Alu* filled site at orthologous positions in non-human primate genomes. Two of these *Alu* elements (Yb8NBC185 and Yb8NBC253) have previously been sequenced and analysed. These appear to be the result of gene conversion and parallel insertion.[33] The remaining seven (Yb8AC1233, Yb7AD234, Yb7AD52, Yb8AC1737, Yb8AC914a, Yb7AD231 and Yb7AD84) newly reported Yb-lineage elements also produced a filled site amplicon in at least one non-human primate genome (Table 4 and Figure 4). DNA sequences obtained from these loci demonstrated that they do not contain authentic Yb-lineage *Alu* elements. Further examination of both human and non-human primate sequences was undertaken to aid in reconstructing the history of these loci. Individual alignments corresponding to six of the seven anomalous loci with their non-human primate orthologous sequences are available on the authors' website. The seventh locus had no associated insertion at the orthologous locations, as outlined below.

For Yb7AD234, the PCR results suggested that filled alleles were present in both the human and the owl monkey, but not in other non-human primates. Examination of the orthologous sequences showed that a parallel independent *Alu* insertion event had most likely occurred, in which a human *Alu* Yb element and an owl monkey *Alu* Sc element independently inserted immediately adjacent to a *Alu* Sg element common to both genomes. *Alu* Sg elements are approximately 35 million years old and are found throughout most of the primate order.[22] This is consistent with the *Alu* Sg element representing the ancestral state. In humans, a full-length Yb element integrated adjacent to the existing *Alu* Sg. A subsequent non-homologous recombination event probably resulted in a chimeric *Alu* Yb7 element and one remaining monomer from the ancestral *Alu* Sg repeat. Presumably, the other recombinant allele was lost from

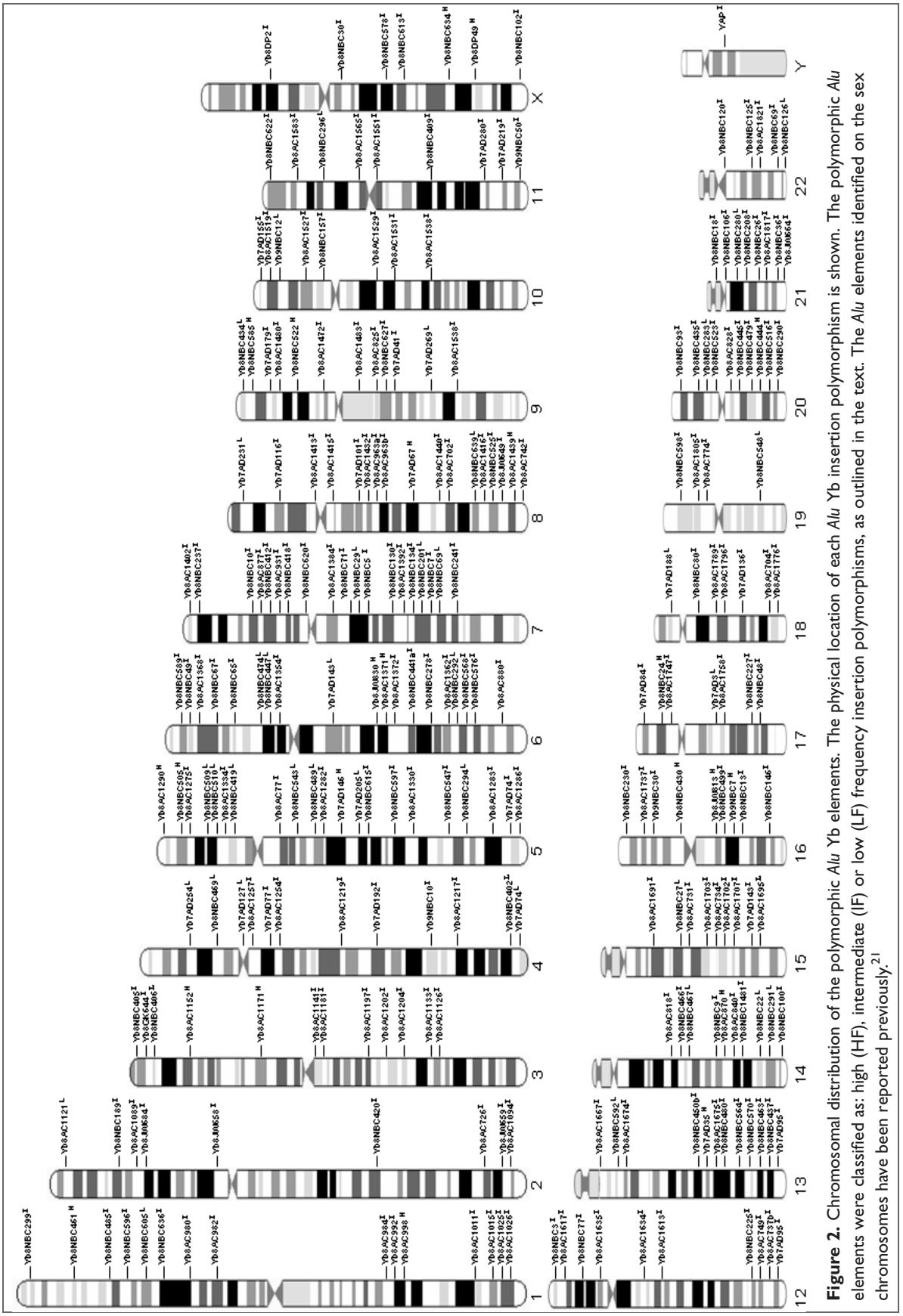

**Figure 2.** Chromosomal distribution of the polymorphic *Alu* Yb elements. The physical location of each *Alu* Yb insertion polymorphism is shown. The polymorphic *Alu* elements were classified as: high (HF), intermediate (IF) or low (LF) frequency insertion polymorphisms, as outlined in the text. The *Alu* elements identified on the sex chromosomes have been reported previously.[21]

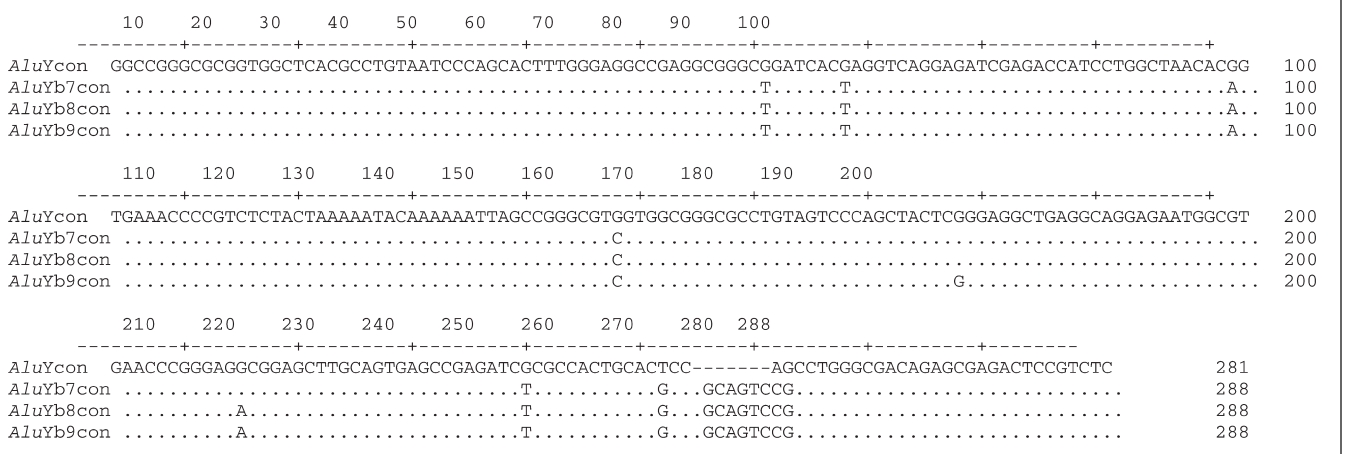


**Figure 3.** Sequence alignment of the *Alu* Yb7, Yb8 and Yb9 subfamilies. The consensus sequence for the *Alu* Y subfamily is shown at the top. The sequences of the *Alu* Yb7, Yb8 and Yb9 subfamilies are shown below. The dots below represent the same nucleotides as the consensus sequence. Mutations are shown as the correct base for each of the subfamilies.

the population, or this was an intrachromosomal recombination event. Within the owl monkey lineage, a full-length *Alu* Sc element integrated into the flanking 3′ region of the *Alu* Sg element, leaving the adjacent *Alu* element completely intact. In this case, an independent insertion of the *Alu* Yb and *Alu* Sc elements in the human and owl monkey lineages provides the most parsimonious explanation for the extant sequences.

The human genomic locus containing *Alu* Yb8AC1233 has a complete *Alu* Yb element, while the orthologous loci in pygmy chimpanzee, common chimpanzee, gorilla and orangutan contain an older, full-length *Alu* Y element. Upon closer examination of the human *Alu* Yb8 element sequence, it was apparent that the second monomer of the element shared several mutations with the non-human primate *Alu* Y element. This hybrid human element probably resulted from the insertion of the Yb8 element within the ancestral *Alu* Y locus, followed by a non-homologous recombination event that created a chimeric sequence.

The *Alu* Yb locus Yb8AC1737 yielded non-human primate PCR results, indicating that this element resided in the genomes of humans, pygmy chimpanzees, common chimpanzees, gorillas and orangutans (Table 4). *Alu* Yb8A1737 integrated into a chromosomal region that had previously undergone a segmental duplication event. Amplification of all human templates generates filled and empty site products for the Yb8AC1737 element. Humans have a duplicated region which has an *Alu* Sx element (top band in Figure 4) and a second allele (bottom band in Figure 4) which contain a Yb element that is associated with an adjacent 473 base pair deletion 3′ of the element. As in the case of Yb8AC1233, examination of the human sequence demonstrated similarities with the older *Alu* element present in the ancestral state. In this case, multiple shared mutations are located the 5′ end of the element. It is likely that in the chromosomal segment that represents the smaller allele, following the insertion of an *Alu* Yb8 element downstream of the ancestral *Alu* Sx, a non-homologous recombination event occurred which deleted the intervening sequence and created the chimeric element.

PCR results from Yb7AD231 initially indicated that the *Alu* element was present in non-human primate loci and absent in human loci. Examination of the orthologous sequences, however, showed that an *Alu* Sg element in non-human primates was replaced by an *Alu* Yb in humans. In conjunction with this replacement event, the human locus is missing 375 base pairs 3′ of the Yb7AD231 element that are found in non-human primates. As in the above cases, comparison of human and non-human primate sequences revealed that insertion of an *Alu* Yb8 near to the ancestral *Alu* Sg in the human lineage resulted in a non-homologous recombination event which removed the intervening sequence and created a chimeric element.

Investigation of orthologous insertions at the Yb8AC914a locus revealed a more complex history. While an *Alu* Yb element resides at the human locus, an *Alu* Sx element is present at orthologous pygmy chimpanzee, common chimpanzee and gorilla loci. It appears that this human genomic region underwent a segmental duplication event some time before the separation of humans and great apes, as evidenced by the appearance of two alleles of different sizes in all extant humans examined and in multiple non-human primate species (Figure 4). The sequences of these products show that the upper band of both humans and great apes is made up of two adjacent *Alu* Sx elements separated by a 327 base pair stretch of sequence. The lower band in humans and great apes is a single *Alu* Yb and a single *Alu* Sx element, respectively, with the intervening 327 base pair sequence absent. While it is difficult

**Table 4.** Presence or absence of *Alu* Yb inserts in non-human primate orthologous loci

| *Alu* element | Human | Common chimpanzee | Pygmy chimpanzee | Gorilla | Orangutan | Green monkey | Owl monkey | Types |
|---|---|---|---|---|---|---|---|---|
| Yb7AD231 | +(Yb) | +(Sg) | +(Sg) | +(Sg) | +(Sg) | 0 | 0 | GC & Del |
| Yb7AD84 | +(Yb) | +(Y) | +(Y) | +(Y) | +(Y) | +(Y) | 0 | GC |
| Yb8AC1233 | +(Yb) | +(Y) | +(Y) | +(Y) | +(Y) | +(Y) | 0 | GC |
| Yb8AC914a | +(Yb) | 0 | +(Sx) | +(Sx) | 0 | 0 | 0 | GC |
| Yb7AD234 | +(Yb) | – | – | – | – | 0 | +(Sc) | Ind |
| Yb7AD52 | +(Yb) | Non-repetitive sequence | Non-repetitive sequence | Non-repetitive sequence | Non-repetitive sequence | 0 | 0 | Del |
| Yb8AC1737 | +(Yb) | +(Sx) | +(Sx) | +(Sx) | +(Sx) | 0 | 0 | GC & Del |

+, Polymerase chain reaction (PCR) product indicates presence of *Alu* insert; –, small PCR product indicates absence of an *Alu* insert; 0, no PCR product of the locus was observed; GC, gene conversion; Ind, independent insertion; Del, *Alu*-mediated deletion.

to establish the exact history of the events leading to the extant alleles, it appears most likely that, within one of the duplicated regions, a non-homologous recombination between the two *Alu* Sx elements occurred, resulting in the removal of the intervening 327 base pair sequence. This event appears to have taken place some time before the separation of the lineages leading to humans and other great apes, as the two allele sizes are present in humans, chimps, gorillas and organgutans (Figure 4). The other, larger, allele did not undergo this recombination. In the human lineage, an additional event occurred; this resulted in the replacement, in the smaller allele, of the remaining *Alu* Sx element with an *Alu* Yb8 element. This may have been the result of an *Alu* Yb8 insertion in the *Alu* Sx tail followed by a non-homologous recombination event that removed the *Alu* Sx. Unlike other analogous replacement events (see above), however, no clear signature of the older *Alu* Sx remains in the human *Alu* Yb sequence, and the *Alu* Yb8 element is not chimeric. As a consequence, it may represent a more 'pure' gene conversion event, where the donor sequence has removed all evidence of the target sequence from the genome.

The *Alu* Yb7AD84 locus contained an *Alu* Yb element adjacent to the second monomer of an older *Alu* Y element. Orthologous loci in non-human primates contained a similar arrangement of *Alu* elements, but with an *Alu* Y in the place of the Yb8. Inspection of alignments between the *Alu* Yb8 element and its non-human orthologues revealed that the first monomer of the *Alu* Yb8 shared several mutations with the ancestral Y element, indicating that recombination had occurred, creating a chimeric element and replacing the ancestral Y monomer with a Yb8 element in the human lineage.

Finally, *Alu* Yb7AD52 contained a complete Yb7 element within humans and a larger than predicted product size (based on human sequence) in non-human primates. The non-human primate sequences showed that no insertion was present at orthologous loci, but rather 320 base pairs of sequence adjacent to the *Alu* Yb7 had been deleted within humans. While it is unclear if this deletion occurred simultaneously with the *Alu* insertion, evidence of L1 and *Alu* insertion associated genomic deletions raises the possibility that the *Alu* Yb7AD52 deletion resulted during the integration process.[34]

### Paralogous insertions

Computational searches for paralogous Yb *Alu* elements were performed using direct repeats and PCR primer sequences as search criteria, because *Alu* elements of the same evolutionary age have conserved core sequences but unique 5′ and 3′ flanking sequences. Typically, direct repeats and oligonucleotide primers both reside in adjacent unique flanking regions and are, therefore, unique to individual *Alu* insertions. The authors' analysis showed that there were 32 autosomal paralogous Yb elements (Table 1).

## Discussion

In this study, the authors expand on previously published data to include all human genomic loci containing Yb-lineage *Alu* elements. Two previously published datasets included annotations of 118 Yb elements on the sex chromosomes and 244 Yb elements on the autosomes.[15,21] The analysis has recovered 1,489 unique *Alu* Yb loci, making a total of 1,851 Yb-lineage elements that have now been annotated in the human genome draft sequence (Table 1). The number of Yb-lineage *Alu* elements recovered from the draft sequence is in good agreement with previously published estimates of subfamily

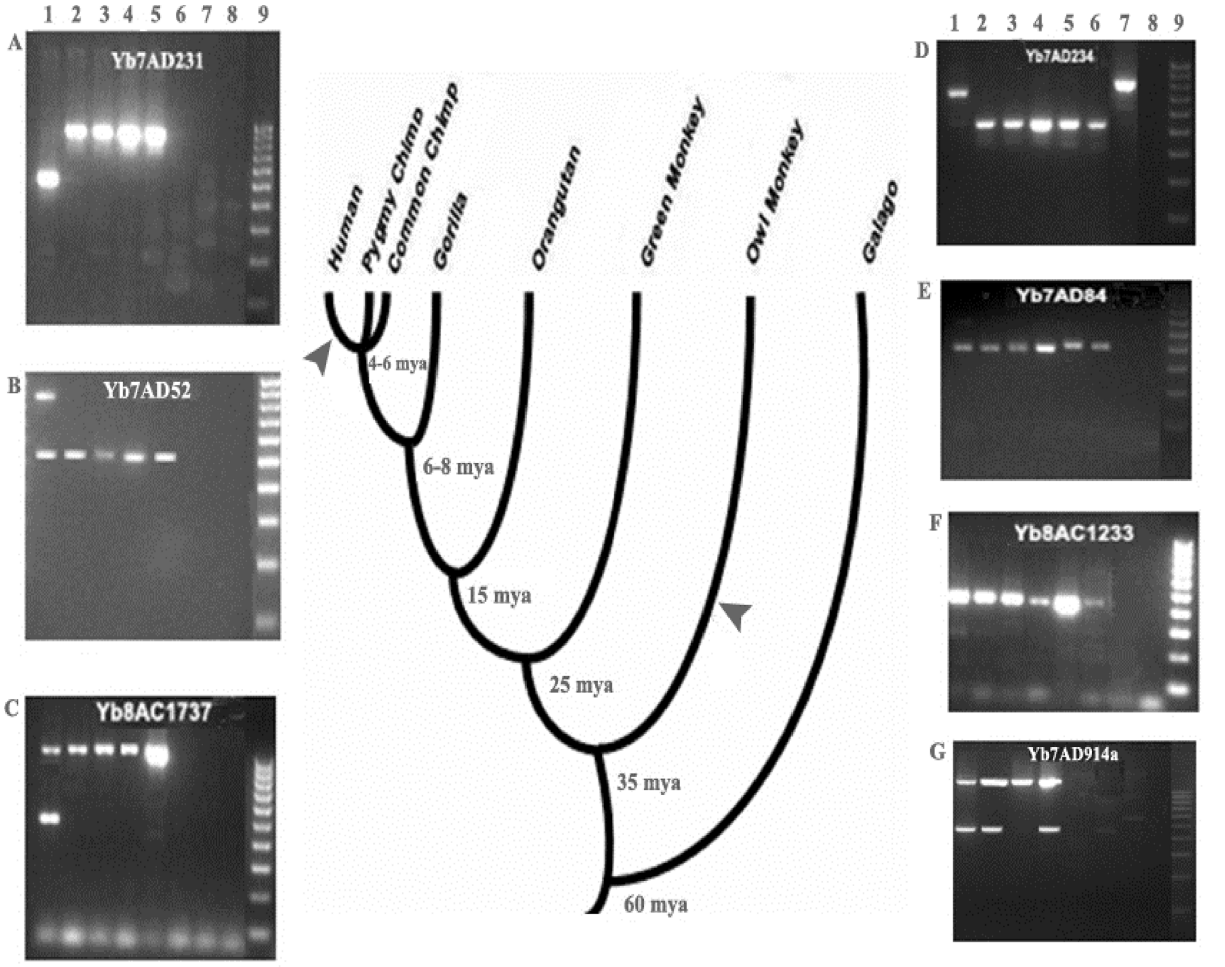


**Figure 4.** *Alu*-mediated non-homologous recombinations, parallel independent insertions and gene conversion. Estimated evolutionary time periods between the different primate speciation events are indicated in millions of years (mya).[44] Arrows on the phylogenetic tree denote the potential time period for the parallel independent insertion of *Alu* Yb7AD234 in the human lineage and *Alu* Sc in the owl monkey lineage. The lanes are: 1, human; 2, pygmy chimpanzee; 3, common chimpanzee; 4, gorilla; 5, orangutan; 6, green monkey; 7, owl monkey; 8, double-distilled water (control); 9, 10 base pair molecular marker.

size.[15,35–37] Also 1,307 Yb-lineage loci have been analysed via PCR-based assays on the autosomes and sex chromosomes.[15,21]

The overall proportion of *Alu* insertion polymorphisms for Yb-lineage subfamilies in human populations was 20 per cent. Based on all 1,851 annotated Yb elements, if one assumes that the insertion polymorphism rate for the entire Yb-lineage is 20 per cent, one would expect to see approximately 370 Yb polymorphic *Alu* repeats. To date, 247 Yb *Alu* insertion polymorphisms (239 autosomal and eight sex chromosome) have been recovered. There are numerous reasons for the difference between the observed and expected numbers of polymorphic Yb insertions that have been recovered.
The present study only recovered those polymorphic *Alu* Yb-lineage elements that have inserted alleles present in the genomes of the few individuals whose DNA constitutes the human genome draft sequence. As a consequence, a fraction of the polymorphic *Alu* insertion loci, typically those of low insertion frequencies, will not be identified through computational screening of the draft human genomic sequence. In addition, a number of polymorphisms may have been missed as a result of researchers' inability to examine them using PCR assays because they inserted in paralogous loci. Some of the Yb elements were not amenable to PCR because they had inserted into other, pre-existing repetitive elements in the genome, or simply did not amplify in the PCR analysis.

The emergence of separate Yb-lineage subfamilies is the result of an accumulation of diagnostic mutations occurring within source genes over the course of primate evolution. The total number of *Alu* subfamily members differs greatly between the Yb7, Yb8 and Yb9 subfamilies. Subfamilies Yb7 and Yb9 contain 158 and 63 subfamily members, respectively. The Yb8 subfamily comprises 994 elements, which is

approximately 56 per cent of the entire Yb-lineage. There are multiple scenarios that could account for the observed unequal copy numbers. There could have been a higher rate of amplification for the Yb8 subfamily with respect to the *Alu* Yb7 and *Alu* Yb9 families. Alternatively, the Yb8 source gene may have simply been active for a much longer period of time than the Yb7 and Yb9 sources. As the Yb7 subfamily is demonstrably older, however, it is more likely that it has been less transpositionally active or that it mutated early on to become the Yb8 source gene. The relatively young age estimate for the *Alu* Yb9 subfamily suggests it is more recent in origin, a fact that could also account for its lower copy number.

The Yb8 subfamily makes up over 50 per cent of the Yb-lineage *Alu* elements. Because the Yb8 subfamily provides the largest dataset, the authors used non-CpG-based and CpG-based average age estimates to calculate the number of *Alu* repeats that should be present at orthologous loci in non-human primates. Assuming that the Yb *Alu* elements had a linear rate of amplification, the age of the oldest individual Yb *Alu* repeats can be calculated as twice the average subfamily age. The average age calculated using non-CpG and CpG mutations was calculated to be 2.39 and 1.57 million years, respectively. In this study, the non-CpG-based age estimate indicates that the oldest *Alu* Yb8 subfamily members integrated into the primate lineage approximately 4.78 (2.39 × 2) million years ago. This is near the time of the human and African ape divergence, which is thought to have occurred 4–6 million years ago. Assuming that humans and chimpanzees diverged 4 million years ago, our non-CpG age estimate of 4.78 million years for the Yb subfamily would lead us to expect that roughly 16 per cent of *Alu* Yb8 insertions would be present at non-human primate loci; however, no authentic orthologous Yb insertions have been recovered. This suggests that either the 4 million year date of the human-chimpanzee divergence is too recent, or that the authors' age estimate is too old. The CpG-based age estimates, however, place the oldest Yb *Alu* elements at 3.14 (1.57 × 2) million years old. This is subsequent to the generally accepted time range of human–African ape divergence, so one should expect to see no Yb8 elements in non-human primate genomes. This result is in good agreement with current data.[18,36,38–43] The disparities between the CpG- and non-CpG-based subfamily age estimates are appreciable and systematic. They may be attributable to a number of factors. The well-established distribution of *Alu* elements within genic regions may affect their susceptibility to CpG-based methylation, resulting in an altered mutation density. Alternatively, ongoing sequence exchanges between non-homologous *Alu* elements may also contribute to deviations from published values. Further examination of the CpG methylation rates in retroposons, which take into account genomic location, rates of gene conversion and additional factors, will be necessary in order better to address the observed differences in mutation densities.

## Acknowledgments

This research was supported by the Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000-05)-05, (2000-05)-01 and (2001-06)-02 (MAB); National Science Foundation grants BCS-0218338 (MAB) and BCS-0218370 (LBJ) and National Institutes of Health RO1 GM59290 (M.A.B. and L.B.J.).

## References

1. Deininger, P.L. and Batzer, M.A. (1993), 'Evolution of retroposons', *Evolutionary Biology* Vol. 157, pp. 196.
2. Okada, N. (1991), 'SINEs', *Curr. Opin. Genet. Dev.* Vol. 1, pp. 498–504.
3. Schmid, C.W. (1996), '*Alu*: structure, origin, evolution, significance and function of one-tenth of human DNA', *Prog. Nucleic Acid Res. Mol. Biol.* Vol. 53, pp. 283–319.
4. Smit, A.F. (1999), 'Interspersed repeats and other mementos of transposable elements in mammalian genomes', *Curr. Opin. Genet. Dev.* Vol. 9, pp. 657–663.
5. Houck, C.M., Rinehart, F.P. and Schmid, C.W. (1979), 'A ubiquitous family of repeated DNA sequences in the human genome', *J. Mol. Biol.* Vol. 132, pp. 289–306.
6. Ullu, E. and Tschudi, C. (1984), '*Alu* sequences are processed 7SL RNA genes', *Nature* Vol. 312, pp. 171–172.
7. Mathias, S.L., Scott, A.F., Kazazian, H.H. Jr. *et al.* (1991), 'Reverse transcriptase encoded by a human transposable element', *Science* Vol. 254, pp. 1808–1810.
8. Deragon, J.M., Sinnett, D. and Labuda, D. (1990), 'Reverse transcriptase activity from human embryonal carcinoma cells NTera2D1', *Embo J.* Vol. 9, pp. 3363–3368.
9. Kajikawa, M. and Okada, N. (2002), 'LINEs mobilize SINEs in the eel through a shared 3′ sequence', *Cell* Vol. 111, pp. 433–444.
10. Jurka, J. (1997), 'Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons', *Proc. Natl. Acad. Sci. USA* Vol. 94, pp. 1872–1877.
11. Deininger, P.L., Batzer, M.A., Hutchison, C.A., 3rd *et al.* (1992), 'Master genes in mammalian repetitive DNA amplification', *Trends Genet.* Vol. 8, pp. 307–311.
12. Deininger, P.L. and Batzer, M.A. (2002), 'Mammalian retroelements', *Genome Res.* Vol. 12, pp. 1455–1465.
13. Paulson, K.E. and Schmid, C.W. (1986), 'Transcriptional inactivity of *Alu* repeats in HeLa cells', *Nucleic Acids Res.* Vol. 14, pp. 6145–6158.
14. Shen, M.R., Batzer, M.A. and Deininger, P.L. (1991), 'Evolution of the master *Alu* gene(s)', *J. Mol. Evol.* Vol. 33, pp. 311–320.
15. Carroll, M.L., Roy-Engel, A.M., Nguyen, S.V., *et al.* (2001), 'Large-scale analysis of the *Alu* Yb5 and Yb8 subfamilies and their contribution to human genomic diversity', *J. Mol. Biol.* Vol. 311, pp. 17–40.
16. Altschul, S.F., Gish, W., Milter, W. *et al.* (1990), 'Basic local alignment search tool', *J. Mol. Biol.* Vol. 215, pp. 403–410.
17. Sanger, F., Nicklen, S. and Coulson, A.R. (1977), 'DNA sequencing with chain-terminating inhibitors', *Proc. Natl. Acad. Sci. USA* Vol. 74, pp. 5463–5467.
18. Arcot, S.S., Adamson, A.W., Risch, G.W. *et al.* (1998), 'High-resolution cartography of recently integrated human chromosome 19-specific *Alu* fossils', *J. Mol. Biol.* Vol. 281, pp. 843–856.
19. Schneider, S., Roessli, D. and Excoffier, L. (2000), 'Arlequin: A software for population genetics data analysis', Ver, 2.000.
20. Guo, S.W. and Thomson, E.A. (1992), 'A Monte Carlo method for combined segregation and linkage analysis', *Am. J. Hum. Genet.* Vol. 51, pp. 1111–1126.
21. Callinan, P.A., Hedges, D.J., Salem, A.-H. *et al.* (2003), 'Comprehensive analysis of *Alu* associated diversity on the human sex chromosomes', *Gene* Vol. 317, pp. 103–110.
22. Batzer, M.A. and Deininger, P.L. (2002), '*Alu* repeats and human genomic diversity', *Nat. Rev. Genet.* Vol. 3, pp. 370–379.

23. Arcot, S.S., Wang, Z., Weber, J.L. *et al.* (1995), '*Alu* repeats: A source for the genesis of primate microsatellites', *Genomics* Vol. 29, pp. 136–144.
24. Economou, E.P., Bergen, A.W., Warren, A.C. *et al.* (1990), 'The polydeoxyadenylate tract of *Alu* repetitive elements is polymorphic in the human genome', *Proc. Natl. Acad. Sci., USA* Vol. 87, pp. 2951–2954.
25. Smit, A.F., Toon, G., Riggs, A. *et al.* (1995), 'Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences', *J. Mol. Biol.* Vol. 246, pp. 401–417.
26. Zuliani, G. and Hobbs, H.H. (1990), 'A high frequency of length polymorphisms in repeated sequences adjacent to *Alu* sequences', *Am. J. Hum. Genet.* Vol. 46, pp. 963–969.
27. Toth, G., Gaspari, Z. and Jurka, J. (2000), 'Microsatellites in different eukaryotic genomes: Survey and analysis', *Genome Res.* Vol. 10, pp. 967–981.
28. Beckman, J.S. and Weber, J.L. (1992), 'Survey of human and rat microsatellites', *Genomics* Vol. 12, pp. 627–631.
29. Jurka, J. and Klonowski, P. (1996), 'Integration of retroposable elements in mammals: Selection of target sites', *J. Mol. Evol.* Vol. 43, pp. 685–689.
30. Batzer, M.A., Kilroy, G.E., Richard, P.E. *et al.* (1990), 'Structure and variability of recently inserted *Alu* family members', *Nucleic Acids Res.* Vol. 18, pp. 6793–6798.
31. Labuda, D. and Striker, G. (1989), 'Sequence conservation in *Alu* evolution', *Nucleic Acids Res.* Vol. 17, pp. 2477–2491.
32. Miyamoto, M.M., Slightom, J.L. and Goodman, M. (1987), 'Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region', *Science* Vol. 238, pp. 369–373.
33. Roy-Engel, A.M., Carroll, M.L., El-Sawy, M. *et al.* (2002), 'Non-traditional *Alu* evolution and primate genomic diversity', *J. Mol. Biol.* Vol. 316, pp. 1033–1040.
34. Gilbert, N., Lutz-Prigge, S. and Moran, J.V. (2002), 'Genomic deletions created upon LINE-1 retrotransposition', *Cell* Vol. 110, pp. 315–325.
35. Batzer, M.A., Stoneking, M., Alegria-Hartman, M. *et al.* (1994), 'African origin of human-specific polymorphic *Alu* insertions', *Proc. Natl. Acad. Sci. USA* Vol. 91, pp. 12288–12292.
36. Batzer, M.A., Rubin, C.M., Hellmann-Blumberg, U. *et al.* (1995), 'Dispersion and insertion polymorphism in two small subfamilies of recently amplified human *Alu* repeats', *J. Mol. Biol.* Vol. 247, pp. 418–427.
37. Roy, A.M., Carroll, M.L., Nguyen, S.V. *et al.* (2000), 'Potential gene conversion and source genes for recently integrated *Alu* elements', *Genome Res.* Vol. 10, pp. 1485–1495.
38. Kass, D.H., Raynor, M.E. and Williams, T.M. (2000), 'Evolutionary history of B1 retroposons in the genus Mus', *J. Mol. Evol.* Vol. 51, pp. 256–264.
39. Cantrell, M.A., Filanoski, B.J., Ingermann, A.R. *et al.* (2001), 'An ancient retrovirus-like element contains hot spots for SINE insertion', *Genetics* Vol. 158, pp. 769–777.
40. Kass, D.H., Batzer, M.A. and Deininger, P.L. (1995), 'Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution', *Mol. Cell Biol.* Vol. 15, pp. 19–25.
41. Maeda, N., Wu, C.I., Bliska, J. *et al.* (1988), 'Molecular evolution of intergenic DNA in higher primates: Pattern of DNA changes, molecular clock, and evolution of repetitive sequences', *Mol. Biol. Evol.* Vol. 5, pp. 1–20.
42. Shen, M.R., Brosius, J. and Deininger, P.L. (1997), 'BC1 RNA, the transcript from a master gene for ID element amplification, is able to prime its own reverse transcription', *Nucleic Acids Res.* Vol. 25, pp. 1641–1648.
43. Salem, A.-H., Kilroy, G.E., Watkins, W.S., *et al.* (2003), 'Recently integrated *Alu* elements and human genomic diversity', *Mol. Biol. Evol.* Vol. 20, pp. 1349–1361.
44. Goodman, M., Porter, C.A., Czelusniak, J. *et al.* (1998), 'Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence', *Mol. Phylogenet. Evol.* Vol. 9, pp. 585–598.