Louisiana State University
LSU Scholarly Repository

Faculty Publications | Department of Biological Sciences

9-3-2004

Analysis of the human Alu Ya-lineage

Anthony C. Otieno
Ctr. Bio-Modular Microsystems

Anthony B. Carter
Ctr. Bio-Modular Microsystems

Dale J. Hedges
Ctr. Bio-Modular Microsystems

Jerilyn A. Walker
Ctr. Bio-Modular Microsystems

David A. Ray
Ctr. Bio-Modular Microsystems

See next page for additional authors

Follow this and additional works at: https://repository.lsu.edu/biosci_pubs

Recommended Citation

Otieno, A., Carter, A., Hedges, D., Walker, J., Ray, D., Garber, R., Anders, B., Stoilova, N., Laborde, M., Fowlkes, J., Huang, C., Perodeau, B., & Batzer, M. (2004). Analysis of the human Alu Ya-lineage. *Journal of Molecular Biology, 342* (1), 109-118. https://doi.org/10.1016/j.jmb.2004.07.016

This Article is brought to you for free and open access by the Department of Biological Sciences at LSU Scholarly Repository. It has been accepted for inclusion in Faculty Publications by an authorized administrator of LSU Scholarly Repository. For more information, please contact ir@lsu.edu.

Authors

Anthony C. Otieno, Anthony B. Carter, Dale J. Hedges, Jerilyn A. Walker, David A. Ray, Randall K. Garber, Bridget A. Anders, Nadica Stoilova, Meredith E. Laborde, Justin D. Fowlkes, Cheney H. Huang, Benjamin Perodeau, and Mark A. Batzer

This article is available at LSU Scholarly Repository: https://repository.lsu.edu/biosci_pubs/223

doi:10.1016/j.jmb.2004.07.016 J. Mol. Biol. (2004) 342, 109–118

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Analysis of the Human *Alu* Ya-lineage

Anthony C. Otieno†, Anthony B. Carter†, Dale J. Hedges Jerilyn A. Walker, David A. Ray, Randall K. Garber, Bridget A. Anders Nadica Stoilova, Meredith E. Laborde, Justin D. Fowlkes Cheney H. Huang, Benjamin Perodeau and Mark A. Batzer*

Department of Biological Sciences, Biological Computation and Visualization Center, Center for Bio-Modular Microsystems, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA

**Corresponding author*

The *Alu* Ya-lineage is a group of related, short interspersed elements (SINEs) found in primates. This lineage includes subfamilies Ya1–Ya5, Ya5a2 and others. Some of these subfamilies are still actively mobilizing in the human genome. We have analyzed 2482 elements that reside in the human genome draft sequence and focused our analyses on the 2318 human autosomal Ya *Alu* elements. A total of 1470 autosomal loci were subjected to polymerase chain reaction (PCR)-based assays that allow analysis of individual Ya-lineage *Alu* elements. About 22% (313/1452) of the Ya-lineage *Alu* elements were polymorphic for the insertion presence on human autosomes. Less than 0.01% (5/1452) of the Ya-lineage loci analyzed displayed insertions in orthologous loci in non-human primate genomes. DNA sequence analysis of the orthologous inserts showed that the orthologous loci contained older pre-existing Y, Sc or Sq *Alu* subfamily elements that were the result of parallel forward insertions or involved in gene conversion events in the human lineage. This study is the largest analysis of a group of "young", evolutionarily related human subfamilies. The size, evolutionary age and variable allele insertion frequencies of several of these subfamilies makes members of the Ya-lineage useful tools for human population studies and primate phylogenetics.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: short interspersed elements; retroposons

Introduction

SINEs (short interspersed elements) are retroposable genetic elements typically less than 500 nucleotides long that are interspersed ubiquitously throughout different genomes.[1,2] *Alu* elements are the most successful primate SINEs and have amplified to more than one million copies in primate genomes.[3] *Alu* elements mobilize by a mechanism termed target-primed reverse transcription, in which an *Alu* RNA transcript is reverse-transcribed into a DNA copy that subsequently re-integrates into a new genomic site.[4–9] As *Alu* elements do not produce any of the proteins needed to facilitate their movement or retroposition, they must instead capitalize on the ORF2 product of long interspersed elements (LINEs) that encodes the endonuclease and reverse transcriptase activities needed for mobilization.[6–8,10] This method of "borrowing" LINE mobilization factors has enabled *Alu* elements to reach high copy numbers in both human and non-human primate genomes.

Although there are many copies of *Alu* elements in the human genome, only a few are believed to be retropositionally competent.[11–14] The accumulation of new mutations within these active "master" or "source" genes, results in the creation, through evolutionary time, of new *Alu* subfamilies or lineages of elements with common diagnostic mutations.[11,15] Specific *Alu* subfamilies can be identified by their diagnostic mutations.[16–18] Some "young" *Alu* subfamilies have amplified so recently in humans that they are largely absent from the genomes of non-human primates, or are lineage-specific inserts within different primate taxa.[3,19]

† A.C.O. and A.B.C. contributed equally to this research.

Abbreviations used: SINEs, short interspersed elements; PCR, polymerase chain reaction; LINEs, long interspersed elements; MER, medium reiteration frequency sequences; LC, low-complexity sequences; SSRs, simple sequence repeats; HWE, Hardy–Weinberg equilibrium; my, million years.

E-mail address of the corresponding author: mbatzer@lsu.edu

0022-2836/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.

These evolutionarily recent *Alu* insertions are useful for the study of human population genetics and non-human primate phylogenetics.[20–27]

Young subfamilies typically have a large number of human-specific insertions, share a higher number of diagnostic point mutations, and contain some loci that are still polymorphic with respect to their presence or absence in diverse human populations. The *Alu* Ya-lineage is the largest young evolutionarily related group of *Alu* elements in the human genome. The Ya-lineage is comprised of Ya5 elements, that have all five diagnostic mutations as well as smaller subfamilies, which contain less than five of the diagnostic mutations. Subfamilies that contain other mutations in addition to having all five diagnostic mutations, such as Ya8 and Ya5a2, are also considered part of the Ya-lineage and have been characterized previously.[19,28] Here, we survey 2318 autosomal Ya-lineage *Alu* elements containing five or fewer diagnostic mutations from the human genome draft sequence and the human genomic diversity associated with these elements.

Results

Ya-lineage element copy number and chromosomal distribution

A total of 2482 Ya-lineage elements possessing one to five diagnostic mutations were recovered from the human genome draft sequence. The autosomes and sex chromosomes contain a total of 2318 and 164 elements, respectively.[29] A total of 615 autosomal elements integrated within other repeated sequences and were therefore not amenable to further polymerase chain reaction (PCR)-based analyses. Another 232 produced inconclusive PCR results and one element was located at the end of a sequencing contig (e.g. not enough flanking 3′ genomic sequence to develop an oligonucleotide primer). Of the 1470 autosomal elements that could be analyzed by PCR, 18 were inserted in paralogous sequences, 1139 were present on both chromosomes of all individuals tested (fixed present) and 313 were polymorphic for insertion presence/absence in diverse human populations (Table 1).

A χ^2 "goodness of fit" test was performed on the chromosomal distribution data to test a model of random insertion for the *Alu* Ya-lineage elements in which the number of expected insertions on each chromosome is proportional to the percentage of the genome that each chromosome represents (Table 2). The genome-wide, chromosomal distribution was assessed based on a total number of elements recovered from the human genome draft sequence. Human chromosomes 15, 18, 21 and 22 were statistically different from the random insertion model at the 5% significance level with expected numbers greater than the observed number of elements. Chromosomes 1, 6, 7, 12, 13, 14, and 19 were statistically different at the 5% significance level with observed numbers greater than what would be expected based upon a random insertion model.

Table 1. *Alu* Ya-lineage element PCR) analysis summary

	Alu Ya-lineage elements
Loci analyzed by PCR	1470
Fixed present	1139
High frequency insertion polymorphisms	15
Intermediate frequency insertion polymorphisms	264
Low frequency insertion polymorphisms	34
Total polymorphic	313
Paralog	18
Loci not analyzed by PCR	848
Inserted in other repeats	615
No PCR results	232
End of contig	1
Total autosomal elements analysed	2318
Total sex chromosome elements analyzed[a]	164

[a] From Callinan *et al.*[48]

Ya-lineage sequence attributes

The length of direct repeats flanking individual Ya-lineage *Alu* elements range from four to 23 base-pairs in length with an average length of 13 base-pairs. One hundred and twenty elements contain no detectable direct repeat sequences. *Alu* Ya-lineage element oligo(dA)-rich tails ranged from three to 115 base-pairs in length with an average of 27 base pairs. Approximately, 4.4% (102/2315) of the elements contain tails with simple sequence repeats ($\geq$4 consecutive units). Only three Ya elements did not have detectable oligo-(dA)-rich tails.

Recombination, incomplete reverse transcription or improper integration into the genome could cause sequence truncations in individual *Alu* elements according to the contemporary model of *Alu* genomic movement and integration.[7,9,30] There are 95% more truncations in the 5′ region of the Ya-lineage elements compared with the 3′ region, which is consistent with the current *Alu* retrotransposition model, since it posits that reverse transcription initiates at the 3′ end of the source or master *Alu* sequence.[31] A total of 280 autosomal Ya-lineage *Alu* elements were found to have collectively lost 9172 base-pairs of 5′ *Alu* sequence. Investigation of the 3′ ends showed that 583 base-pairs are missing from 15 autosomal elements.

Flanking genomic sequence content

Ya-lineage elements that have integrated directly adjacent to other human repeats and are not amenable to PCR, were analyzed for human repeat content. Of these elements, 47% (289/615) integrated within or next to LINE-1 elements, 22% (135/615) integrated next to evolutionarily older *Alu* elements and the remaining 31% (191/615) integrated next to LTR (long terminal repeats), MER (medium reiteration frequency sequences), LC (low-complexity sequences) and SSRs (simple

Table 2. Chromosomal distribution of autosomal Ya-lineage *Alu* elements

Chr.	Percentage of the human genome	Number of observed *Alu* elements	Number of expected *Alu* elements	S/NS[a]
1	8.01	213	184	S
2	7.93	206	182	NS
3	6.54	166	150	NS
4	6.28	151	144	NS
5	5.96	146	137	NS
6	5.59	152	129	S
7	5.16	154	119	S
8	4.80	99	110	NS
9	4.36	95	100	NS
10	4.41	91	101	NS
11	4.48	118	103	NS
12	4.37	125	101	S
13	3.65	114	84	S
14	3.32	104	76	S
15	3.17	48	73	S
16	2.99	62	69	NS
17	2.76	62	64	NS
18	2.56	43	59	S
19	1.95	60	45	S
20	2.06	52	47	NS
21	1.47	20	34	S
22	1.57	18	36	S
X	4.97	119	114	NS
Y	1.65	45	38	NS
Total no. elements[b]		2299		

[a] Statistically significant (S) or not statically significant (NS) at 5% level.
[b] Total no. of elements = fixed present + polymorphic + *Alu* elements with failed PCR results + *Alu* elements within other human repeats.

sequence repeats). A total of 28 autosomal Ya-lineage elements contain an independent, full-length *Alu* element either in the oligo(dA)-rich tail or immediately adjacent to it, such that both elements are contained within a single set of direct repeats; data are available on our webpage†. One thousand bases of 5′ and 3′ genomic sequence flanking each *Alu* element were analyzed for GC content. The mean GC content for the 5′ and 3′ flanking sequence was 39.1% and 39.2%, respectively. The total mean GC content (including *Alu* element) is 41.0% in these genomic regions.

Paralogous insertions

Paralogous insertions are *Alu* elements that have inserted into duplicate genomic loci and, consequently, contain identical or nearly identical flanking genomic sequence. Computational searches for paralogous elements were performed using direct repeats and flanking oligonucleotide primer sequences as search criteria. However, this approach did not yield all existing paralogous *Alu* elements, with some additional elements recovered during the PCR stage. Our analysis yielded a total of 18 autosomal paralogs (Table 1). Analysis of a monochromosomal hybrid cell line DNA panel was used to determine the chromosomal location of the duplicated *Alu* elements (see Materials and Methods). Eleven *Alu* elements were found in regions that duplicated on the same chromosome. Six are in regions that duplicated onto two different chromosomes and one element is in a genomic region that duplicated onto three different chromosomes.

† http://batzerlab.lsu.edu

Insertion polymorphisms

Four major continental populations (African American, Asian, European and South American) were analyzed to determine the *Alu* Ya-lineage-associated human genomic diversity. A total of 313 polymorphic *Alu* elements were identified on the human autosomes and nine have been previously reported on the X chromosome (Table 1).[29] It is very likely that more than 313 polymorphic Ya-lineage elements exist in the human genome, since the draft sequence of the human genome is a composite derived from only a few individuals. PCR amplification of human autosomal Ya loci revealed an overall polymorphism rate of 22% (313/1452). A total of 78% (244 elements) of the polymorphic Ya-lineage loci had all five of the subfamily-specific diagnostic base mutations. Average heterozygosity and allele frequency data for the autosomal polymorphic loci were calculated and are available on our website. Individual autosomal chromosome insertion polymorphism rates ranged from 11% (chromosome 9) to 40% (chromosome 21). A table of genome-wide human insertion polymorphisms is available on our website†.

Alu elements were categorized as either polymorphic or fixed present (FP). Fixed present is defined as when every individual tested has the *Alu*

element on both chromosomes. Polymorphic elements were further classified as high (HF), intermediate (IF) or low frequency (LF). The following frequency classifications have been previously established.[32] Low frequency insertion polymorphisms are those exhibiting insertion frequencies of less than 30%. Intermediate frequency insertion polymorphisms are those loci where the *Alu* element is present at frequencies ranging from 30% to 70%. High frequency insertion polymorphisms are characterized by greater than 70% insertion frequency. High, intermediate and low frequency categories comprise 4.70% (15), 84.3% (264) and 11.0% (34) of the total autosomal *Alu* polymorphisms, respectively. Autosomal and X-chromosome positions of all Ya-lineage polymorphic elements were determined using BLAT screening (the BLAST-like alignment tool)† and Ensembl Human Genome Server‡ (Figure 1A and B).[33,34]

A total of 901 χ^2 goodness of fit tests were performed and yielded a total of 42 deviations from Hardy–Weinberg Equilibrium (HWE) ($p < 0.05$). Approximately, 45 deviations would be expected by chance alone at the 5% significance level. A total of 16 of the 42 deviations were the result of poor-quality PCR amplification. Due to the fact that a large number of statistical tests were performed and none of the significant departures cluster by locus or population, we believe that the remaining 26 deviations represent normal statistical fluctuation and conclude that the Ya-lineage *Alu* insertion polymorphisms do not deviate from HWE. In addition, a Markov–Chain based analysis was applied to the data using the population data analysis software Arlequin.[35,36] Out of 901 comparisons, the test suggested that only 17 were significant, which is lower than what would be expected by chance alone at the 5% significance level. Thus, the results of both tests suggest that Ya-lineage *Alu* insertion polymorphisms as a whole do not significantly depart from HWE.

Analysis of polymorphic *Alu* insertion loci

Overall heterozygosity values for polymorphic *Alu* insertion loci in African American, Asian, European and South American populations were calculated as 0.35, 0.32, 0.34 and 0.33, respectively. Pairwise χ^2 tests of independence were performed between population-specific genotype distributions for each of the loci. The percentage of pairwise tests that showed a significant difference (at the 5% significance level) were as follows: African American *versus* Asian (45%); African American *versus* European (39%); African American *versus* South American (32%); Asian *versus* European (31%); Asian *versus* South American (23%) and European *versus* South American (15%).

We examined the distributions of all the polymorphic loci further to identify loci that showed variability in only one population. Sixteen loci were polymorphic only in the African American population and monomorphic (present) in all others. Eleven loci were polymorphic only in the African American population while being monomorphic absent for all others. The European population contained five such loci, two of which were monomorphic present for the other populations and three of which were monomorphic absent from other populations. The South American population contained two *Alu* loci that were polymorphic in that population with one being monomorphic fixed for the additional populations and one being monomorphic absent for the other populations. The Asian population contained no examples of population-specific allele variability. Lastly, we calculated the total polymorphism rate for each major population. The level of polymorphism in African Americans was 19%, Asians were 13%, Europeans were 15% and South Americans were 13%.

Evolutionary age estimates

The *Alu* Ya-lineage is made up of several groups or subfamilies with varying number of diagnostic mutations.[32] Here, we refer to these constituent groups as *subfamily-Ya1*, *Ya2*, *Ya3*, *Ya4*, *Ya5*, *Ya5a2* and *Ya8* in which each is named according to the number of diagnostic mutations contained within the consensus sequence.[37] Evolutionary age estimates of *subfamlilies-Ya5a2* and *Ya8* have been reported previously.[19,28] The *Ya5* subfamily is the largest group comprising greater than 75% of the entire Ya-lineage. CpG dinucleotide and non-CpG nucleotide mutation densities and neutral mutation rates of 0.90%/million years for CpG bases and 0.15%/million years for non-CpG bases were used to calculate the average evolutionary age of the Ya5 subfamily as reported.[12,32,38,39] The autosomal Ya5 elements were used in both CpG and non-CpG mutation calculations to determine age estimates. The CpG-based age estimates yielded an average age of 2.27 my (million years). Non-CpG age estimates yielded an average age estimate of 2.56 my for the Ya5 *Alu* elements.

Alu Ya-lineage origin and orthologous insertions

Non-human primate DNA was subjected to PCR analysis with the same primers designed to detect individual human *Alu* insertion loci. This resulted in the recovery of five non-human primate loci that appeared to contain Ya-lineage *Alu* elements. DNA sequence analysis of these loci however, showed that these orthologous loci contain older pre-existing *Alu* elements from other subfamilies or other non-repetitive genomic sequences (Table 3). Only one genuine Ya5 is known to exist in an

† http://genome.ucsc.edu/cgi-bin/hgBlat?hgsid=5329687

‡ http://www.ensembl.org/

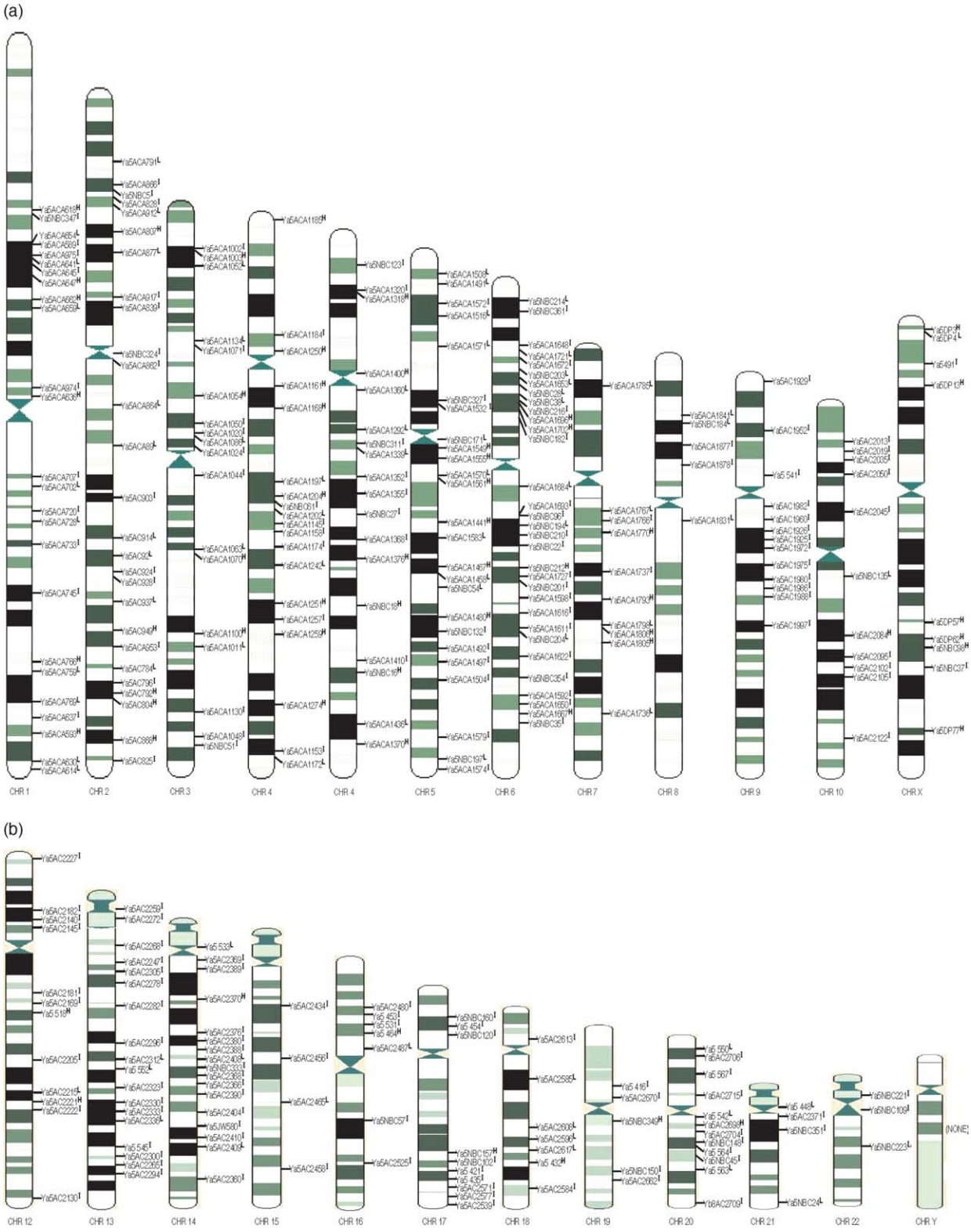

Figure 1. Chromosomal distribution of the polymorphic *Alu* Ya-lineage elements. The physical location of each *Alu* Ya-lineage insertion polymorphism is shown. The polymorphic *Alu* elements were classified as: high (H), intermediate (I) or low (L) frequency insertion polymorphisms, as outlined in the text. The Ya-lineage *Alu* elements located on the sex chromosomes have been reported previously.[48]

orthologous locus which resides in a number of primate lineages.[13,14]

To date, eight Ya-lineage loci have yielded PCR results indicative of the presence of an *Alu*-filled site at orthologous positions in non-human primate genomes. All eight human loci contain Ya5 subfamily members. Three of these elements (Ya5NBC42, Ya5NBC91 and Ya5NBC188) have

Table 3. Presence or absence of *Alu* Ya insertions in non-human primate orthologous loci

ALU ELEMENT	HUMAN	PYGMY CHIMP	COMMON CHIMP	GORILLA	ORANGUTAN	GREEN MONKEY	SPIDER MONKEY	TYPE
Ya5ACA1808	+(Ya5)	+(*Alu* Y) Plus 318 base pairs of genomic sequence	+(*Alu* Y) Plus 318 base pairs of genomic sequence	+(*Alu* Y) Plus 318 base pairs of genomic sequence	0	0	0	GC & Del
Ya5ACA1267	+(Ya5)	—	—	—	—	0	Insertion of ERV of 629 base pairs	Ind
Ya5ACA1786	+(Ya5)	+(*Alu* Sq)	+(*Alu* Sq)	0	0	0	0	GC
Ya5ACA1792	+(Ya5)	—	—	—	—	—	+(*Alu* Sc)	Ind
Ya5ACA2578	+(Ya5)	Non-repetitive sequence	Non-repetitive sequence	Non-repetitive sequence	Non-repetitive sequence	0	0	Del

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

previously been sequenced in non-human primates and have been shown to contain evolutionarily older *Alu* elements. The authors report that the results seen can be explained by the two evolutionary mechanisms, gene conversion or parallel independent insertions.[21] The remaining five (Ya5ACA1808, Ya5ACA1267, Ya5ACA1786, Ya5ACA1792 and Ya5AC2578) are newly reported Ya-lineage elements that yield a filled site amplicon in at least one non-human primate genome. DNA sequence analysis of these non-human primate loci determined that these *Alu* insertions were not authentic Ya-lineage insertions but rather *Alu* Y, Sc or Sq elements. Comparisons between human and non-human primate DNA sequences showed that three evolutionary mechanisms generated these results: gene conversion, *Alu*-mediated deletion or independent parallel insertion (Table 3). These forms of non-traditional *Alu* sequence evolution have been reported previously.[28,40–46] Elements Ya5ACA1267 and Ya5ACA1792 demonstrate insertions at human and spider monkey loci, but no insertion at additional non-human primate loci. The most parsimonious explanation in these cases would be two independent insertion events, as competing explanations would require either multiple deletion events among several primate lineages, or both deletion and insertion events occurring on the lineage leading to humans. For locus Ya5ACA1267 this is further evidenced by the fact that the two insertions occurred at slightly different locations. The case is less clear for Ya5ACA1792, as subsequent sequence rearrangements/deletions in the spider monkey lineage have obscured the insertion point. Locus Ya5AC1808 appears to have resulted when a young Ya5 element inserted near a pre-existing Y element. A non-homologous recombination between the two elements then occurred, which resulted in the partial conversion of the AluY element into an AluYa5 and deletion of approximately 300 bp intervening sequence. Alignments of human and non-human primate orthologous sequences of the five newly reported unusual loci can be found on our website. Locus Ya5ACA1786 appears to be an authentic gene conversion, with an older Sq element sequence being converted to an AluYa5 sequence.

Discussion

Here we report 2318 unique autosomal *Alu* Ya-lineage loci resulting in a total of 2482 Ya-lineage members possessing five or fewer diagnostic mutations that have been recovered from the draft sequence of the human genome. The number of these Ya-lineage *Alu* elements recovered from the draft sequence compares favorably to previously published estimates of the size of this *Alu* subfamily.[28,32,45,47] A total of 1625 elements have been analyzed *via* PCR-based assays on the autosomes and sex chromosomes.[32,48] With a polymorphism rate of 22% for the *Alu* Ya-lineage, 510 polymorphic *Alu* repeats would be expected from the 2318 autosomal elements analyzed. We would expect 323 elements from the loci analyzed by PCR (1470 elements) to be polymorphic. A total of 313 Ya-lineage autosomal *Alu* insertion polymorphisms have been recovered in this study (Table 1). The present study only recovered those polymorphic elements that have inserted alleles present in the genomes of the few individuals whose DNA constitutes the human genome draft sequence. As a consequence, approximately 50% of the actual loci that exist in human populations will be missed.[49] In addition, a number of polymorphisms may have been missed as a result of our inability to examine them using PCR assays because they either inserted in paralogous loci, inserted next to or within other human repetitive elements or simply landed in a genomic region that was not amplifiable by PCR.

Separate Ya-lineage subfamilies emerge as a result of an accumulation of diagnostic mutations occurring within source or master *Alu* genes over the course of primate evolution. The result is a series of evolutionary subfamilies that make up the entire Ya-lineage. The number of subfamily members differs between the different *Alu* subfamilies. The Ya5 subfamily is comprised of those

elements that have five diagnostic mutations. The Ya5 average age calculated using non-CpG and CpG mutations was estimated to be 2.56 my and 2.27 my, respectively. The Ya5 *Alu* subfamily constitutes approximately 75% (1857 elements) of the entire Ya-lineage. The second largest young *Alu* lineage, the Yb, is similar in subfamily structure to the Ya-lineage (Figure 2A and B).[15] The Yb8 subfamily has an evolutionary age of approximately 2.39 my and makes up 57% (1055/1851) of the entire Yb-lineage. Assuming the Ya5 subfamily had a linear rate of amplification, the age of the oldest individual member can be calculated as twice the average calculated evolutionary age. Both age estimates show that the oldest Ya5 elements integrated into the primate lineage approximately 5.12 (2.56×2 for non-CpG) and 4.54 (2.27×2 for CpG) million years ago assuming a linear amplification rate. This time corresponds to the time of the human and African ape divergence of four to six million years ago.[19,25,50] Thus, we expect to potentially see some Ya5 *Alu* elements in non-human primate genomes.

Previous studies have shown that Ya-lineage elements exists in gorilla, chimpanzee and orangutan genomes.[13,14,16,51,52] At least two have been found in gorilla in which one is unique to the gorilla genome and the other is believed to be a progenitor or "founder" locus (previously identified as EPL locus) of Ya5 *Alu* elements.[14,52] This element contains all five diagnostic mutations and resides in chimpanzee and human orthologous positions, making it the only known Ya-lineage locus shared among the three species.[53] The founder (EPL) locus has also been traced to the orangutan genome, but this element contained a number of additional mutations within that lineage.[14,52] The founder locus is designated as Ya5ACA1363 in this study and is located on chromosome 5. In addition, there are a number of Ya5 elements that are unique to the chimpanzee genome.[13,51,54,55] These data, taken with the evolutionary age estimates reported here suggest that the first Ya-lineage elements integrated into primate genomes before the divergence of humans and other hominids and that the subfamily remained active in the chimpanzee and gorilla lineages after the speciation for some unknown period of time. However, human–chimpanzee comparisons demonstrate that, at least in the chimpanzee genome, the activity of the Ya5 family has been considerably lower than that of humans.[49]

Population analysis of polymorphic loci

Polymorphic loci were analyzed to detect differences in average heterozygosity, genotype distribution and allele frequency between four major human populations (African American, Asian, European and South American). If African populations contain more diversity relative to the other three populations, we would expect the African American population to have an overall average heterozygosity value closer to the theoretical maximum of 0.5. The overall average heterozygosity value for African American, Asian, European and South American populations was calculated as 0.35, 0.32, 0.34 and 0.33, respectively.

For each polymorphic locus, we performed

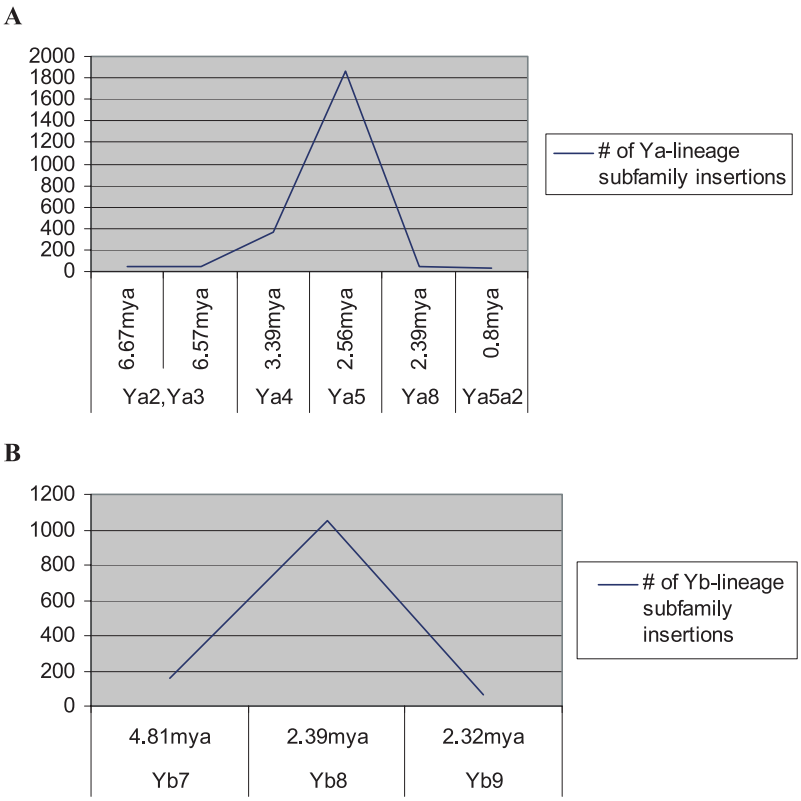

Figure 2. *Alu* subfamily copy numbers and average ages. A, *Alu* Ya-lineage expansion throughout evolutionary time. Subfamily-Ya5 constitutes approximately 75% of the entire lineage and has an average evolutionary age of 2.56 my. Additional subfamilies with greater than five diagnostic mutations are part of the Ya-lineage but were not examined here.[19,28] my, million years. B, Yb-lineage subfamily amplification throughout evolutionary time. Yb8 subfamily constitutes approximately 57% of the entire Yb-lineage and has an average evolutionary age 2.39 my.[15]

pairwise χ^2 "test of independence" tests between applicable, population-specific genotype distributions. The percentage of pairwise tests that showed a significant difference (at the 5% significance level) between genotype distributions were calculated: African American *versus* Asian (45%); African American *versus* European (39%); African American *versus* South American (32%); Asian *versus* European (31%); Asian *versus* South Americans (23%) and European *versus* South American (15%). In order to assess why there were a higher percentage of statistically different African American pairwise comparisons, we extracted the average heterozygosity values from the polymorphic loci that showed statistical differences between the African American population and one of the other three populations then recalculated the overall average heterozygosity value for each population: African American (0.41), Asian (0.24), European (0.27) and South American (0.26). These values indicate that the African American genotype distributions that were significantly different from other populations were so because they contained a higher degree of heterozygosity.

Materials and Methods

Computational analyses

Screening of the National Center for Biotechnology Information's (NCBI) Genbank non-redundant human genome database and the University of California Santa Cruz August, 2001 human genome draft sequence was performed using a local installation of BLAST (basic local alignment search tool), available at NCBI† to identify all *Alu* Ya-lineage elements in the human genome.[56] A 16 base-pair oligonucleotide (5′CCATCCCGGCTAAAAC3′) that is an exact complement to all *Alu* Ya-lineage elements was used to query the human genome draft sequence. A 700–1200 base-pair fragment that included the *Alu* element and adjacent genomic DNA sequences were extracted for individual insertion sites and placed into the University of Washington Genome Center's RepeatMasker Web server‡ to annotate repeat sequence content as described.[15] Subsequently, the *Alu* Ya-lineage sequences were aligned using MEGALIGN (DNASTAR V.5) to determine mutation density and element authenticity.

Cell lines and DNA samples

Cell lines used to isolate DNA samples were as follows: human (*Homo sapiens*, HeLa ATCC-CCL-2); common chimpanzee (*Pan troglodytes*, CCR-AG06939); pygmy chimpanzee (*Pan paniscus*, CCR-AG05253); lowland gorilla (*Gorilla gorilla*, CCR-AG05251); orangutan (*Pongo pygmaeus*, CCR-AG12256 OR CCR-GM06213); owl monkey (*Aotus trivirgatus*, ATCC-CRL-1556); spider monkey (*Ateles geoffroyi*, NG053052); green monkey (*Cercopithecus aethiops*, ATCC-CCL-70). Human DNA from South American populations (HD17 and HD18) 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 were isolated from peripheral blood lymphocytes available from previous studies.[32] DNA from the human-rodent somatic cell hybrid panel, used for paralog analysis, was obtained from the NIGMS Human Genetic Mutant Cell Repository at Coriell Institute, Camden, NJ (panel 2).

Primer design and PCR amplification

Oligonucleotide primers for the PCR amplification of each *Alu* element were designed using the 700–1200 base-pair flanking unique sequence fragments and Primer3 software (Whitehead Institute of Biomedical Research, Cambridge, MA, USA)§. The sequences of the oligonucleotide primers, annealing temperatures, PCR product sizes and chromosomal locations for all autosomal Ya-lineage elements can be found on our website‖. (The primers were subsequently screened against the GenBank non-redundant database to verify that they were unique DNA sequence. PCR amplification was performed in 25 μl reactions using 10–50 ng of target DNA, 200 nM of each oligonucleotide primer, 200 μM dNTPs in 1×PCR Buffer II (AppliedBiosystems, Inc.), 1.5 mM $MgC1_2$ and 1 unit Taq DNA polymerase. Each sample was subjected to an initial denaturation step of 94 °C for 150 seconds, followed by 32 cycles of PCR at one minute of denaturation at 94 °C, one minute at the annealing temperature, one minute of extension at 72 °C, followed by a final extension step at 72 °C for ten minutes.

DNA sequence analysis

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.[57]

Statistical analyses

Hardy–Weinberg equilibrium tests using χ^2 goodness of fit analysis (using one degree of freedom) and a Markov–Chain method (implemented in Arlequin) were performed on polymorphic Ya-lineage elements.[35,36] A comparison of Ya-lineage insertion distribution among all human chromosomes was conducted using χ^2 goodness of fit tests (using one degree of freedom). The expected number of insertions for each chromosome was estimated based on the total genomic sequence that the individual chromosome represented.[42] Pairwise χ^2 test of independence tests were performed between the genotype distributions of polymorphic elements from four major populations.

Genbank accession numbers

The sequences of the orthologous non-human primate *Alu* insertion loci (bonobo, common chimpanzee, gorilla, orangutan, green monkey and spider monkey) have been assigned Genbank accession numbers (AY604157–AY604167).

† http://www.ncbi.nlm.nih.gov/
‡ http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker
§ http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
‖ http://batzerlab.lsu.edu

Acknowledgements

This research was supported by an award from the Technical Support Working Group (M.A.B.), National Science Foundation BCS-0218338 (M.A.B.), National Science Foundation EPS-0346411 (M.A.B.), and the State of Louisiana Board of Regents Support Fund (M.A.B.). Bridget Anders was supported by a National Science Foundation REU supplement to award BCS-0218338 (M.A.B.). B. P. was supported by National Institutes of Health P20 RR16456 from the BRIN program of the National Center for Research Resources. N. S., M. L. and C. H. were supported by a Howard Hughes Medical Institute grant through the Undergraduate Biological Sciences Education program to Louisiana State University.

References

1. Deininger, P. L. & Batzer, M. A. (1993). Evolution of retroposons. *Evol. Biol.* **27**, 157–196.
2. Shedlock, A. M. & Okada, N. (2000). SINE insertions: powerful tools for molecular systematics. *Bioessays*, **22**, 148–160.
3. Batzer, M. A. & Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Rev. Genet.* **3**, 370–379.
4. Sinnett, D., Richer, C., Deragon, J. M. & Labuda, D. (1992). Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences. *J. Mol. Biol.* **226**, 689–706.
5. Boeke, J. D. (1997). LINEs and Alus–the polyA connection. *Nature Genetics*, **16**, 6–7.
6. Kajikawa, M. & Okada, N. (2002). LINEs mobilize SINEs in the eel through a shared 3′ sequence. *Cell*, **111**, 433–444.
7. Feng, Q., Moran, J. V., Kazazian, H. H., Jr. & Boeke, J. D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, **87**, 905–916.
8. Mathias, S. L., Scott, A. F., Kazazian, H. H., Jr., Boeke, J. D. & Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. *Science*, **254**, 1808–1810.
9. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.
10. Dewannieux, M., Esnault, C. & Heidmann, T. (2003). INE-mediated retrotransposition of marked Alu sequences. *Nature Genet.* **35**, 41–48.
11. Deininger, P. L., Batzer, M. A., Hutchison, C. A., 3rd. & Edgell, M. H. (1992). Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8**, 307–311.
12. 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.
13. 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.
14. Leeflang, E. P., Liu, W. M., Chesnokov, I. N. & Schmid, C. W. (1993). Phylogenetic isolation of a human Alu founder gene: drift to new subfamily identity [corrected]. *J. Mol. Evol.* **37**, 559–565.
15. Carter, *et al*. (2004). Genome-wide analysis of the human Alu Yb-lineage. *Hum. Genome*, **1**, 167–178.
16. Shen, M. R., Batzer, M. A. & Deininger, P. L. (1991). Evolution of the master Alu gene(s). *J. Mol. Evol.* **33**, 311–320.
17. Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H. & Deininger, P. (1987). Clustering and subfamily relationships of the Alu family in the human genome. *Mol. Biol. Evol.* **4**, 19–29.
18. Willard, C., Nguyen, H. T. & Schmid, C. W. (1987). Existence of at least three distinct Alu subfamilies. *J. Mol. Evol.* **26**, 180–186.
19. Roy, A. M., Carroll, M. L., Kass, D. H., Nguyen, S. V., Salem, A. H., Batzer, M. A. & Deininger, P. L. (1999). Recently integrated human Alu repeats: finding needles in the haystack. *Genetica*, **107**, 149–161.
20. Salem, A. H., Ray, D. A., Xing, J., Callinan, P. A., Myers, J. S., Hedges, D. J. *et al*. (2003). Alu elements and hominid phylogenetics. *Proc. Natl Acad. Sci. USA*, **100**, 12787–12791.
21. Roy-Engel, A. M., Carroll, M. L., El-Sawy, M., Salem, A. H., Garber, R. K., Nguyen, S. V. *et al*. (2002). Non-traditional Alu evolution and primate genomic diversity. *J. Mol. Biol.* **316**, 1033–1040.
22. Batzer, M. A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D. H., Shaikh, T. H. *et al*. (1994). African origin of human-specific polymorphic Alu insertions. *Proc. Natl Acad. Sci. USA*, **91**, 12288–12292.
23. Stoneking, M., Fontius, J. J., Clifford, S. L., Soodyall, H., Arcot, S. S., Saha, N. *et al*. (1997). Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.* **7**, 1061–1071.
24. Sherry, S. T., Harpending, H. C., Batzer, M. A. & Stoneking, M. (1997). Alu evolution in human populations: using the coalescent to estimate effective population size. *Genetics*, **147**, 1977–1982.
25. Roy-Engel, A. M., Carroll, M. L., Vogel, E., Garber, R. K., Nguyen, S. V., Salem, A. H. *et al*. (2001). Alu insertion polymorphisms for the study of human genomic diversity. *Genetics*, **159**, 279–290.
26. Batzer, *et al*. (1996). Genetic variation of recent Alu insertions in human populations. *J. Mol. Evol.* **42**, 22–29.
27. Jurka, J. & Pethiyagoda, C. (1995). Simple repetitive DNA sequences from primates: compilation and analysis. *J. Mol. Evol.* **40**, 120–126.
28. Roy, A. M., Carroll, M. L., Nguyen, S. V., Salem, A. H., Oldridge, M., Wilkie, A. O. *et al*. (2000). Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res.* **10**, 1485–1495.
29. Callinan, P. A., Hedges, D. J., Salem, A.-H., Xing, J., Walker, J. A., Garber, R. K. *et al*. (2003). Comprehensive analysis of Alu associated diversity on the human sex chromosomes. *Gene*, **317**, 103–110.
30. Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D. & Kazazian, H. H., Jr. (1996). High frequency retrotransposition in cultured mammalian cells. *Cell*, **87**, 917–927.
31. Jurka, J. & Klonowski, P. (1996). Integration of retroposable elements in mammals: selection of target sites. *J. Mol. Evol.* **43**, 685–689.
32. Carroll, M. L. *et al*. (2001). Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J. Mol. Biol.* **311**, 17–40.

33. Clamp, M. *et al*. (2003). Ensembl 2002: accommodating comparative genomics. *Nucl. Acids Res.* **31**, 38–42.
34. Kent, W. J. (2002). LAT: the blast-like alignment tool. *Genome Res.* **12**, 656–664.
35. Schneider, S., Roessli, D., and Excoffier, L. (2000). Arlequin: a software for population genetics data analysis. Ver 2.000.
36. Guo, S. *et al*. (1992). A Monte Carlo method for combined segregation and linkage analysis. *Am. J. Hum. Genet.* **51**, 1111–1126.
37. Batzer, M. A., Deininger, P. L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C. M. *et al*. (1996). Standardized nomenclature for Alu repeats. *J. Mol. Evol.* **42**, 3–6.
38. Labuda, D. & Striker, G. (1989). Sequence conservation in Alu evolution. *Nucl. Acids Res.* **17**, 2477–2491.
39. Miyamoto, M. M., Slightom, J. L. & Goodman, M. (1987). Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. *Science*, **238**, 369–373.
40. Kass, D. H., Raynor, M. E. & Williams, T. M. (2000). Evolutionary history of B1 retroposons in the genus Mus. *J. Mol. Evol.* **51**, 256–264.
41. Nikaido, M., Rooney, A. P. & Okada, N. (1999). Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: hippopotamuses are the closest extant relatives of whales. *Proc. Natl Acad. Sci. USA*, **96**, 10261–10266.
42. Arcot, S. S., Adamson, A. W., Risch, G. W., LaFleur, J., Robichaux, M. B., Lamerdin, J. E. *et al*. (1998). High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. *J. Mol. Biol.* **281**, 843–856.
43. Cantrell, M. A., Filanoski, B. J., Ingermann, A. R., Olsson, K., DiLuglio, N., Lister, Z. & Wichman, H. A. (2001). An ancient retrovirus-like element contains hot spots for SINE insertion. *Genetics*, **158**, 769–777.
44. Kass, D. H., Batzer, M. A. & Deininger, P. L. (1995). Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. *Mol. Cell. Biol.* **15**, 19–25.
45. Batzer, M. A., Rubin, C. M., Hellmann-Blumberg, U., Alegria-Hartman, M., Leeflang, E. P., Stern, J. D. *et al*. (1995). Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. *J. Mol. Biol.* **247**, 418–427.
46. Maeda, N., Wu, C. I., Bliska, J. & Reneke, J. (1988). Molecular evolution of intergenic DNA in higher primates: pattern of DNA changes, molecular clock, and evolution of repetitive sequences. *Mol. Biol. Evol.* **5**, 1–20.
47. Batzer, M. A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D. H., Shaikh, T. H. *et al*. (1994). African origin of human-specific polymorphic Alu insertions. *Proc. Natl Acad. Sci. USA*, **91**, 12288–12292.
48. Callinan, P. A., Hedges, D. J., Salem, A.-H., Xing, J., Walker, J. A., Garber, R. K. *et al*. (2003). Comprehensive analysis of Alu associated diversity on the human sex chromosomes. *Gene*, **317**, 103–110.
49. Hedges, D. J., Callinan, P. A., Cordaux, R., Xing, J., Barnes, E. & Batzer, M. A. (2004). Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res.* **14**, 1068–1075.
50. Batzer, M. A. & Deininger, P. L. (1991). A human-specific subfamily of Alu sequences. *Genomics*, **9**, 481–487.
51. Leeflang, E. P., Chesnokov, I. N. & Schmid, C. W. (1993). Mobility of short interspersed repeats within the chimpanzee lineage. *J. Mol. Evol.* **37**, 566–572.
52. Shaikh, T. H. & Deininger, P. L. (1996). The role and amplification of the HS Alu subfamily founder gene. *J. Mol. Evol.* **42**, 15–21.
53. Matera, A. G., Hellmann, U. & Schmid, C. W. (1990). A transpositionally and transcriptionally competent Alu subfamily. *Mol. Cell Biol.* **10**, 5424–5432.
54. Schmid, C. W. (1998). Does SINE evolution preclude Alu function?. *Nucl. Acids Res.* **26**, 4541–4550.
55. Schmid, C. W. (1996). Alu: structure, origin, evolution, significance and function of one-tenth of human DNA. *Prog. Nucl. Acid Res. Mol. Biol.* **53**, 283–319.
56. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
57. Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, **74**, 5463–5467.

Edited by J. Karn

(Received 10 June 2004; received in revised form 8 July 2004; accepted 12 July 2004)