Potential Gene Conversion and Source Genes for Recently Integrated Alu Elements

Astrid M. Roy,[1,6] Marion L. Carroll,[2,6] Son V. Nguyen,[2] Abdel-Halim Salem,[2] Michael Oldridge,[3] Andrew O. M. Wilkie,[3,4] Mark A. Batzer,[2,7] and Prescott L. Deininger[1,5,7,8]

[1]Tulane Cancer Center, Department of Environmental Health Sciences, Tulane University Medical Center, New Orleans, Louisiana 70112, USA; [2]Departments of Pathology, Biometry and Genetics, Biochemistry, and Molecular Biology, Stanley S. Scott Cancer Center, Neuroscience Center of Excellence, Louisiana State University Health Sciences Center, New Orleans, Louisiana 70112, USA; [3]Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX2 6HE, UK; [4]Oxford Craniofacial Unit, The Radcliffe Infirmary NHS Trust, Oxford OX2 6HE, UK; [5]Laboratory of Molecular Genetics, Alton Ochsner Medical Foundation, New Orleans, Louisiana 70121, USA

Alu elements comprise >10% of the human genome. We have used a computational biology approach to analyze the human genomic DNA sequence databases to determine the impact of gene conversion on the sequence diversity of recently integrated Alu elements and to identify Alu elements that were potentially retroposition competent. We analyzed 269 Alu Ya5 elements and identified 23 members of a new Alu subfamily termed Ya5a2 with an estimated copy number of 35 members, including the de novo Alu insertion in the *NF1* gene. Our analysis of Alu elements containing one to four (Ya1–Ya4) of the Ya5 subfamily-specific mutations suggests that gene conversion contributed as much as 10%–20% of the variation between recently integrated Alu elements. In addition, analysis of the middle A-rich region of the different Alu Ya5 members indicates a tendency toward expansion of this region and subsequent generation of simple sequence repeats. Mining the databases for putative retroposition-competent elements that share 100% nucleotide identity to the previously reported de novo Alu insertions linked to human diseases resulted in the retrieval of 13 exact matches to the *NF1* Alu repeat, three to the Alu element in *BRCA2*, and one to the Alu element in *FGFR2* (Apert syndrome). Transient transfections of the potential source gene for the Apert's Alu with its endogenous flanking genomic sequences demonstrated the transcriptional and presumptive transpositional competency of the element.

Alu elements belong to a class of retroposons termed SINEs. SINEs are Short INterspersed Elements usually ~100–300 bp in length commonly found in introns, 3′ untranslated regions of genes, and intergenic genomic regions (Deininger and Batzer 1993). Alu is the most abundant class of SINEs in primate genomes, reaching a copy number in excess of one million/haploid genome (Jelinek and Schmid 1982; Jurka et al. 1993, Smit 1999). Alu elements increase their genomic copy number by an amplification process termed retroposition (Rogers and Willison 1983; Weiner et al. 1986).

Alu elements appear to have arisen in the last 65 million years (Deininger and Daniels 1986). The human Alu family of repeats is composed of a small number of distinct subfamilies characterized by subfamily-specific diagnostic mutations (Slagel et al. 1987; Willard et al. 1987; Shen et al. 1991; Batzer et al. 1996b). The source Alu gene(s) for each of the subfamilies has been retropositionally active during different periods of primate evolution. The rate of Alu amplification (mostly Sx subfamily) appears to have reached its peak between 60 and 35 million years, and subsequently decreased several orders of magnitude to the present amplification rate (Shen et al. 1991). Only a limited number of SINEs, termed master or source genes, appear to be capable of retroposition (Deininger and Daniels 1986; Batzer et al. 1990; Deininger et al. 1992), although the critical factor(s) defining functional source genes are not understood. A variety of factors influence the retroposition process (Schmid and Maraia 1992). All of the recently integrated young Alu subfamilies appear to be retropositionally active. Almost all of the recently integrated Alu elements within the human genome belong to one of four closely related subfamilies (Y, Ya5, Ya8, and Yb8), with the majority being Ya5 and Yb8 subfamily members (Batzer et al. 1990, 1995; Deininger and Batzer, 1999).

Previously, analysis of individual Alu elements from the different subfamilies involved laborious procedures, such as cloning, library screening, and subsequent sequencing (Batzer et al. 1990, 1995; Arcot et al. 1995a). However, the availability of large-scale human

[6]These authors contributed equally to this work.
[7]These authors contributed equally to this work as senior authors.
[8]Corresponding author.
E-MAIL PDEININ@TCS.TULANE.EDU ; FAX (504) 588-5516.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.152300.

genomic DNA sequences as a result of the Human Genome Project facilitates genomic database mining for Alu elements (Roy et al. 1999). We have taken advantage of these databases and have analyzed a significant portion of the Alu Ya5 subfamily, as well as intermediates between the Ya5 subfamily and the ancestral Alu Y subfamily. In addition, we searched the databases for putative retroposition-competent source Alu genes that generated the de novo Alu inserts associated with a number of human diseases (Deininger and Batzer 1999).

RESULTS

Computational Analyses

To search for subfamilies unidentified previously within the Ya5 Alu subfamily, we selected all of the Alu family members that matched our Ya5 consensus query sequence from the human genome nonredundant (nr) database. Only Ya5 elements found randomly within other sequences were included in our analysis, thereby eliminating Alu elements that had been identified previously in directed Alu-specific projects. In addition, truncated Alu elements were eliminated from the analysis. Ya4 elements that did not contain the first Ya5-specific diagnostic mutation #11 (Fig. 1) (Shen et al. 1991), which is a CpG dinucleotide in the Ya5 subfamily, were considered as Ya5 Alu family members. We obtained a total of 269 matches to the Ya5 query sequence that met our criteria. Of these, 47 shared 100% nucleotide identity with the subfamily consensus sequence and 83 were near perfect matches (aside from a few CpG mutations).

Analysis of the 269 Ya5 Alu elements resulted in the initial identification of two subsets of potential subfamilies containing two diagnostic mutations each, one with six members and the other with four. These subfamiles will be referred to as Ya5a2 and Ya5b2, respectively, in compliance with the standard Alu subfamily nomenclature (Batzer et al. 1996a). Each consensus sequence with the two diagnostic mutations specific to each new Alu subfamily is shown in Figure 1. Interestingly, the de novo Alu Ya5 insert present within an intron of the *NF1* gene (Wallace et al. 1991) is an exact match to the Ya5a2 consensus. The nr database contained 16.0% of human DNA sequences for a total of 515,596,000 bases on the date of the search. The estimated size of the Ya5a2 subfamily is (3×10^9 bp/515,596,000 bp) $\times$ 6 unique Ya5a2 matches = 35 subfamily members. In comparison, the estimated size of the Ya5b2 subfamily is (3×10^9 bp/515,596,000 bp) $\times$ 4 unique Ya5b2 matches = 22 subfamily members. We utilized only the randomly found Ya5a2 elements for the calculations to avoid overestimating the size of the subfamilies. However, these numbers may be underestimations, because some specific polymorphic elements of these subfamilies may not be represented in the database.

To derive a second estimate of the copy numbers of the Ya5a2 and Ya5b2 Alu subfamilies, we used their consensus sequences as queries for the high throughput genome sequence (htgs) and genomic survey sequence (gss) databases. Seventeen additional Alu Ya5a2 elements were found in these searches. Of the 23 total Ya5a2 elements, 13 shared 100% nucleotide identity with the subfamily consensus sequence. No additional Ya5b2 elements were found in the other databases, therefore the Ya5b2 subfamily was not subjected to further analysis. Three additional potential subfamilies, Ya5a1 (five members), Ya5b1 (four members), and Ya5c1 (four members) with only one specific diagnostic mutation were identified (Fig. 1). Because of the small copy number, and the possibility that some

```
                .         .         .         .         .         .
Ya5   GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGA 60
Ya5a2 ............................................................ 60
Ya5b2 ............................................................ 60
Ya5a1 ............................................................ 60
Ya5b1 ............................................................ 60
Ya5c1 ............................................................ 60

                .         .       11.    12   .         .         .
Ya5   TCACGAGGTCAGGAGATCGAGACCATCCCGGCTAAAACGGTGAAACCCCGTCTCTACTAA 120
Ya5a2 ...........A................................................ 120
Ya5b2 .........................A...C.............................. 120
Ya5a1 ....................G....................................... 120
Ya5b1 ............................................................ 120
Ya5c1 ............................................................ 120

                .         .    13   .         .         .   14    .
Ya5   AAATACAAAAAA-TTAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGGGAG 179
Ya5a2 ...........A................................................ 180
Ya5b2 ...........-................................................ 179
Ya5a1 ...........-................................................ 179
Ya5b1 ...........-.........................G...................... 179
Ya5c1 ...........-................................................ 179

                .         .         .         .         .      15 .
Ya5   GCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCCCG 239
Ya5a2 ............................................................ 240
Ya5b2 ............................................................ 239
Ya5a1 ............................................................ 239
Ya5b1 ............................................................ 239
Ya5c1 ...........................................................G..... 239

                .         .         .         .
Ya5   CCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTC 281
Ya5a2 ........................................... 282
Ya5b2 ........................................... 281
Ya5a1 ........................................... 281
Ya5b1 ........................................... 281
Ya5c1 ........................................... 281
```

Figure 1 Consensus sequence alignment of Ya5, and the potential new subfamily members identified. Nucleotide substitutions at each position are indicated with the appropriate nucleotide. Deletions are marked by dashes (-). The Ya5 diagnostic nucleotides are indicated in bold with the corresponding diagnostic number above as defined by Shen et al. (1991).

of those represent parallel mutations rather than new subfamilies, no further analyses were performed.

To determine the age of the Ya5a2 subfamily, we divided the nucleotide substitutions within the elements into those that have occurred in CpG dinucleotides and those that have occurred in non-CpG positions. The distinction between types of mutations is made because the CpG dinucleotides mutate at a rate that is ~10 times faster than non-CpG (Labuda and Striker 1989; Batzer et al. 1990), as a result of the deamination of 5-methylcytosine (Bird 1980). A total of five non-CpG mutations and seven CpG mutations occurred within the 23 Alu Ya5a2 subfamily members identified. By use of a neutral rate of evolution for primate-intervening DNA sequences of 0.15%/one-million years (Miyamoto et al. 1987) and the non-CpG mutation rate of 0.092% (5/5382 bases using only non-CpG bases) within the 23 Ya5a2 Alu elements, yields an estimated average age of 0.62 million years for the Ya5a2 subfamily members with a predicted 95% confidence level in the range of 0.28–1.08 million years, given that the mutations were random and fit a binomial distribution. The Ya5a2 subfamily appears to be much younger than Ya5, Ya8, or Yb8 Alu subfamilies with estimated ages of 2.8 million years (Batzer et al. 1990), 2.75 million years (Roy et al. 1999), and 2.7 million years (Batzer et al. 1995), respectively (Fig. 2).

Determination of the number of elements that perfectly match the subfamily consensus sequence can also give an indirect estimate of Alu subfamily age and recent rate of mobilization. Recently transposed Alu elements share higher levels of nucleotide identity with their source copies because they have not resided in the genome long enough to accumulate random mutations. In contrast, older Alu elements that have resided in the genome for longer periods of time tend to have less nucleotide identity with their source genes as a result of the accumulation of random mutations subsequent to integration into the genome. We compared our search results for the Ya5a2 subfamily with parallel searches from the Ya8 and Ya5 Alu subfamilies. Our BLAST searches from the nr database yielded one perfect match of 12 elements for Ya8, 47 of 269 for Ya5, and 3 of 6 for Ya5a2 (Fig. 2). Searching all three databases (nr, gss, and htgs) yielded 5 perfect matches of 27 for Ya8 and 13 of 23 for Ya5a2. These results are in good agreement with the previous estimates, indicating that Ya5a2 is the youngest Alu subfamily reported to date, as it also has the highest proportion of elements that share 100% nucleotide identity with the consensus sequence.

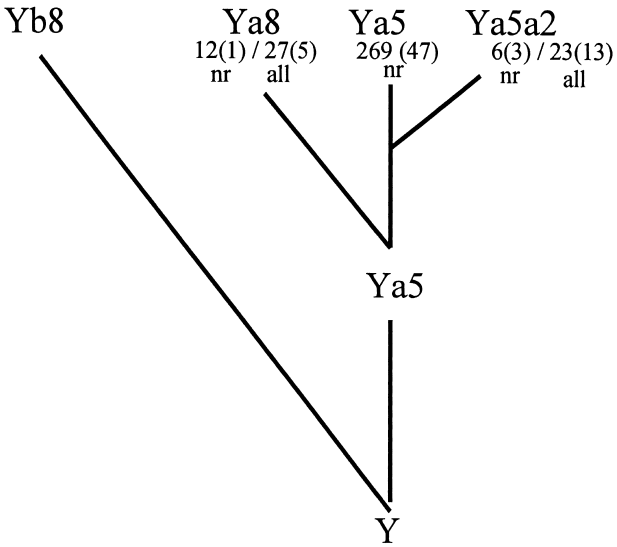

Figure 2 Schematic for the evolution of recently integrated Alu subfamilies. The origin of the Ya5a2 Alu subfamily is shown after the divergence of Ya5 and Yb8 elements. The total number of elements found in the nr-database (perfect matches in parenthesis) are shown first separated by a slash from the total number of elements found in all three databases (nr, gss, htgs). For the Ya5 elements only the nr-database results are shown.

Table 1. Alu Middle A-Rich Region

Ya5-middle A rich region	A_n 4	5	6	7	8	9	10	11
T($\mathbf{A}_n$)TACA$_6$TT[a]	0	269[c]	9	1	0	1	—	—
TA$_5$TAC($\mathbf{A}_n$)TT[b]	0	2	269[c]	37[d]	11	7	3	0

[a] n = 5 in Ya5 consensus.
[b] n = 6 in Ya5 consensus.
[c] Data from the non-redundant database only.
[d] All 23 Ya5a2 members are included.

Stability of the Middle A-Rich Region in Alu Ya5 Members

The oligo-dA-rich tails and middle A-rich regions of Alu elements have been shown previously to serve as nuclei for the genesis of simple sequence repeats (Arcot et al. 1995b). In the autosomal recessive neurodegenerative disease, Friedreich ataxia, the most common mutation, is the hyperexpansion of a GAA within the middle A-rich region of an Sx Alu element (Montermini et al. 1997). Because these regions appear unstable, we analyzed the middle A-rich region of Alu elements retrieved from the databases to detect expansions/contractions of this sequence.

To evaluate potential expansions/contractions, we performed a BLAST query of three databases (nr, htgs, and gss) using the Alu Ya5 consensus sequence with varying numbers of A nucleotides within the middle A-rich region (TA$_n$TACA$_n$TT). Our results demonstrate that the majority of the elements identified matched the consensus sequence. However, there is a trend for an A expansion at both positions (Table 1). In contrast,

very few sequence contractions were detected for any of the positions.

Human Genomic Variation

To determine the human genomic variation associated with the Ya5a2 Alu subfamily members, we selected the 13 Ya5a2 elements identical to the subfamily consensus sequence as well as 2 others and determined the degree of fixation associated with the elements using PCR-based assays of a panel of diverse human DNA samples with the primers shown in Table 2. The panel is composed of 20 individuals of European origin, African-Americans, Greenland natives, and Egyptians for a total of 80 individuals (160 chromosomes). The Alu elements were classified as fixed absent, fixed present, and high, intermediate, or low frequency insertion polymorphisms (see Table 3 for definitions). By use of this approach, 3 of the 14 elements tested (Ya5NBC206, Ya5NBC207, and Ya5NBC235) were always present in the human genomes that were surveyed, suggesting that these elements became fixed in the genome prior to the radiation of modern humans from Africa. Five of the elements (Ya5NBC208, Ya5NBC240, Ya5NBC241, Ya5NBC242, and Ya5NBC220) are intermediate frequency Alu insertion polymorphisms. The remaining six elements are low-frequency Alu insertion polymorphisms (Table 3). The population-specific genotypes and levels of heterozygosity for each element are shown in Table 4. The high proportion of polymorphic elements is in good agreement with our other observations, indicating that the Ya5a2 subfamily is younger than any of the other Alu subfamilies identified previously in the human genome.

Gene Conversion and Alu Sequence Diversity

In our query of the human genome (nr) database, 91 of the Alu elements identified contain one to four of the five Ya5 diagnostic nucleotides (Fig. 1). Of these 91 intermediate elements, 4 are Ya1, 1 Ya2, 7 Ya3, and 79 Ya4 Alu elements (Fig. 3). Surprisingly, not all of the Alu elements with different numbers of subfamily mutations had the same combination of mutations. To facilitate identification of the individual elements with different diagnostic mutation combinations, the diagnostic nucleotides were numbered consecutively in order of abundance (Ya3.1, Ya3.2, etc., see Fig. 3). Seventeen Alu elements (Ya4.4) did not contain the first diagnostic mutation (#11), but were still classified as Ya5 for the analyses outlined above.

Previous evolutionary analyses of the Ya5 founder element with different primate DNA samples demonstrated the sequential accumulation of the Ya5 diagnostic mutations with diagnostic positions #13/#14 first, followed by #12/#16, and finally position #11 (Shaikh and Deininger 1996). Our data are not consistent with a sequential order in the accumulation of the diagnostic mutations. The elements classified as Ya1, Ya2, Ya3.4, Ya3.5, and Ya4.4 (26 total) fit the proposed order (Fig. 3). However, the remaining 65 elements represent almost every other permutated order. Several mechanisms could explain the occurrence for mosaic

Table 2. Alu Ya5a2 PCR Primers, Chromosomal Locations, and PCR Product Sizes

Name	5′ Primer sequence (5′–3′)	3′ Primer sequence (5′A–3′)	A.T.[a]	Chromosome[b]	Product size[c]	
					filled	empty
Ya5NBC206	TCCTTAGCTATCTCACAAGCTACAT	ACACATTTCCTTCAAGAGGTCAAAG	60°C	4	734	424
Ya5NBC207	CAGTTTTATACACTGGCCTGTTTTC	TTGTAGGAGAAAGAGGGGAAATACT	50°C	6	443	122
Ya5NBC208	AATACCTTGTACATCTTCACCCCTA	TCTCTCTGCTGCACAGTTTGTT	50°C	14	441	115
Ya5NBC240	CAGGAGATAAATATGTTCGGAGAGT	TAACTGGGACAGTGAGTTTTACCTG	55°C	9	505	202
Ya5NBC241	GGTTCCAATAGAGAGCAACAGAA	ACCTTAAGCTTTCCCCCAGA	55°C	15	392	66
Ya5NBC242	AACAAAATTCCCTTTCCTCCA	GGCAATCTGACCTTGGGTAA	55°C	7	503	192
Ya5NBC7	TGATGGATATTTGGGTTGGTTC	GGACTGTAAACTAGTTCAACCATTGTG	60°C	7	522	216
Ya5NBC205	ACATGAAGGGCCGACTGTAT	TGCTGCTGCATTATCAACTG	50°C	21	435	81
Ya5NBC209	GTCTATGGGAAGATGAAGAATAGGA	GATGGAGTCACTCATGTGAAAAGTA	55°C	14	447	116
Ya5NBC239	CAGCTGAGAACTGTCACAAATAGAA	ATCAATGACTGACTTGTGCTGAGT	55°C	9	531	198
Ya5NBC243	CCATGATTCGTCATTCACCA	AGGAGACCTGCCAATGAATG	60°C	21	406	86
Ya5NBC220	AAATCAAGCTGCCATACCTCA	GAAACCATCCTTCACAGTGG	60°C	1	463	141
Ya5NBC235	CCCAAGGCACTTGCTGTTA	CCCTTCGAGAAAGAGGAAGG	50°C	2	391	76
Ya5NBC244	CCTATGGCTGAAACTTCTGAAACT	ATATCTTGGTCCACTAGACAAGCAC	60°C	18	453	130
Ya5NBC237[d]	CCCATGGAGGGTCTTTCCTA	CTGGAAACCATCCTTCACAGT	60°C	1	410	88

[a]Amplification of each locus required 2.5 min at 94°C initial denaturing, and 32 cycles for 1 min 94°C, 1-min annealing temperature (A.T.) and 1-min elongation at 72°C. A final extension time of 10 min at 72°C was also used.
[b]Chromosomal location determined from accession information or by PCR analysis of NIGMS monochromosomal hybrid cell line DNA samples.
[c]Empty product sizes calculated by removing the Alu element and one direct repeat from the filled sites that were identified.
[d]Alu Ya5a2 element of the *FGFR2* gene.

Table 3. Alu Ya5a2 (*NF1*)-Associated Human Genomic Diversity

Ya5a2 elements	Accession no. (duplicates)	Position	Allele frequency[a]
Ya5NBC206	AC004057	76767–77048	fixed present
Ya5NBC207	AL118555 (AL132992)	9981–9700 (40728–41009)	fixed present
Ya5NBC208	AL109919	70170–69889	intermediate
Ya5NBC220	AC007611	136715–136434	intermediate
Ya5NBC240	AC133410 (AL135841)	34800–35081 (49829–49548)	intermediate
Ya5NBC241	AC018924	144017–144298	intermediate
Ya5NBC242	AC009517	161301–161582	intermediate
Ya5NBC7	AC004848	24522–24241	low
Ya5NBC205	AL011328	204488–204207	low
Ya5NBC209	AC00808	147056–146775	low
Ya5NBC239	AL133284	115867–115586	low
Ya5NBC244	AC026839	64885–64604	low
Ya5NBC243	AJ011929	151192–151473	low
Ya5NBC235[b]	AQ748733	458–739	fixed present
Ya5NBC237[c]	AL031274	33175–33501	intermediate

[a]Allele frequency was classified as fixed present, fixed absent, low, intermediate, or high frequency insertion polymorphism. (Fixed present) every individual tested had the Alu element in both chromosomes; (low frequency insertion polymorphism) the absence of the element from 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 tested, except for one or two heterozygous or absent individuals.
[b]Several Ns.
[c]Ya5NBC237 is the exact match to the *FGFR2* Alu insertion.

Alu elements, which are addressed in the discussion section. However, we believe the most likely explanation for the existence of these mosaic elements is through gene conversion events. A limited amount of gene conversion between Yb8 Alu elements has been reported previously (Batzer et al. 1995; Kass et al. 1995). In theory, gene conversion may change the sequence of all or part of any Alu element in either an evolutionarily forward (Ya5 subfamily in this case) or backward (Y subfamily) direction by changing the di-

Table 4. Alu Ya5a2-Associated Human Genomic Diversity

	African American				Greenland natives				European				Egyptian				
Elements	genotype[a]			*f*Alu[b]	genotypes			*f*Alu	genotypes			*f*Alu	genotypes			*f*Alu	het.[c]
Ya5NBC206	20	0	0	1.000	20	0	0	1.000	20	0	0	1.000	20	0	0	1.000	0.000
Ya5NBC207	20	0	0	1.000	20	0	0	1.000	20	0	0	1.000	20	0	0	1.000	0.000
Ya5NBC208	4	1	7	0.375	3	0	4	0.429	13	0	6	0.684	7	0	5	0.583	0.482
Ya5NBC236	5	6	2	0.615	5	8	6	0.474	15	5	0	0.875	6	8	1	0.667	0.422
Ya5NBC240	5	1	9	0.367	11	0	4	0.733	5	1	10	0.344	5	3	3	0.591	0.464
Ya5NBC241	3	9	5	0.441	6	11	2	0.605	0	7	11	0.194	3	8	4	0.467	0.459
Ya5NBC242	2	13	1	0.531	7	4	3	0.643	3	4	11	0.278	3	3	1	0.643	0.474
Ya5NBC7	0	0	19	0.000	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0.000
Ya5NBC205	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0.000
Ya5NBC209	0	1	17	0.028	0	0	17	0.000	0	0	19	0.000	0	0	19	0.000	0.000
Ya5NBC239	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0.000
Ya5NBC243	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0.000
Ya5NBC220	0	14	5	0.368	1	15	2	0.472	0	18	1	0.474	0	9	2	0.409	0.502
Ya5NBC244	0	0	12	1.000	—	—	—	—	0	0	10	0.000	0	0	8	0.000	0.000
Ya5NBC235	20	0	0	1.000	20	0	0	1.000	20	0	0	1.000	20	0	0	1.000	0.000
Ya5NBC237[d]	18	1	0	0.974	15	4	0	0.895	20	0	0	1.000	18	1	0	0.974	0.075

[a]Genotypes: +/+ Alu, +/− Alu, −/− Alu.
[b]Frequency of the presence of the Alu.
[c]Average heterozygosity.
[d]Ya5NBC237 is the exact match to the *FGFR2* Alu insertion.
— not determined.

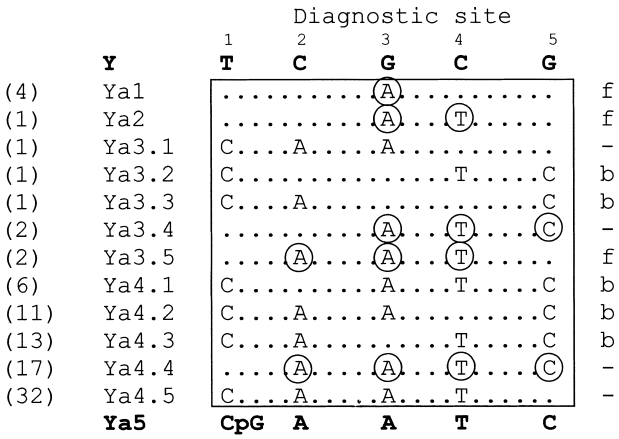

Figure 3 Evolution of the diagnostic nucleotide positions from Y to Ya5 Alu elements. Alignment of the five Alu Ya5 diagnostic nucleotides as defined by Shen et al. (1991) and the different Ya1, Ya2, Ya3, and Ya4 elements found in the nr database. For easy reference, individual elements containing different combinations of the diagnostic mutations were numbered consecutively in order of abundance (Ya3.1, Ya3.2, etc.). Ya4.4 elements were considered as Ya5 elements in the first Ya5 subfamily analysis in this paper. The total number of elements found for each subgroup is indicated at *left* in parenthesis. Potential forward (f) or backward (b) gene conversions are indicated at *right*. The previously reported order of appearance of Ya5 diagnostic mutations (Shaikh and Deininger 1996) is indicated below. Elements with diagnostic mutations that follow the stepwise hierarchical accumulation are circled.

agnostic mutations. In addition, double gene conversions would be extremely rare, making the direction of the gene conversion clear in some elements. We classified the 91 mosaic Alu element sequences as gene converted forward (f), backward (b), or could not be determined (-), (see Fig. 3) If the Alu elements that fit the proposed sequential evolution are ignored in the analysis, all of the other elements may be classified as backward gene conversion (32 total) or could not be determined (33 total), and none were clearly gene-converted forward. Therefore, backward gene conversion may have contributed to between 10% and 20% (32 to 65/269 Ya5 + [91–17] Ya1–Ya4) of the Alu Ya5 sequence diversity. Interestingly, evaluation of the five random Ya5a2 non-CpG mutations shows that one mutation in position #13 is a backward mutation to the Y subfamily, another putative example of a reverse gene conversion.

In Search of Retroposition-Competent Alu Repeats

Sixteen different Alu insertions have been linked to human diseases (Deininger and Batzer 1999). Four belong to the Alu Y subfamily, one to the Ya4 subfamily, eight to the Ya5 subfamily, and three to the Yb8 subfamily. Closer inspection of the nucleotide sequences of these Alu elements show that they have some mutations that are different from their respective subfamily consensus sequences. Because these Alu insertions are very recent in origin, they are likely to be identical to their source genes aside from rare mutations introduced during reverse transcription of the Alu element. Therefore, sequence database queries utilizing each Alu element along with its individual mutations (away from the subfamily consensus sequence) may facilitate the identification of the source Alu element that generated the copy. This strategy is similar to that used previously in the identification of active LINE elements from the human genome (Dombroski et al. 1993).

A database query using the sequence of the individual Alu elements responsible for each disease to mine three databases (nr, htgs, and gss) identified exact complements to four of the disease-associated Alu repeats. Thirteen of the identified elements were exact matches to the *NF1* Alu insertion (Ya5a2 subfamily, Table 3; Wallace et al. 1991); three were exact matches to the *BRCA2* Alu element (Miki et al. 1996) (accession nos. AL121964, AL136319, and AL135778); one matched the *FGFR2* Alu repeat (Oldridge et al. 1999) (accession no. AL031274); and one matched the Alu repeat in the *IL2RG* gene (Lester et al. 1997) (accession no. AC010888).

Potential Source Gene for the Ya5 Insert in *FGFR2*

As mentioned above, our BLAST query only detected one exact match (accession no. AL031274 or Ya5NBC237) to the Ya5 Alu found in the *FGFR2* gene that caused Apert syndrome. We estimated the level of human genomic variation associated with Ya5NBC237 using the same human DNA panel and determined that it was an intermediate frequency Alu insertion polymorphism (Table 4).

Mobilization-competent Alu elements must be capable of transcription, the first step in the retroposition process. To evaluate Alu Ya5NBC237 as a potential source gene for the de novo insert in the patient with Apert syndrome, we determined its transcription capability. Constructs with the genetic loci containing the Ya5NBC237 Alu and the de novo Apert syndrome Alu element were made. Transcription levels from the two constructs were evaluated by Northern blot analysis relative to a control plasmid in which the Alu element is flanked immediately upstream by vector sequence.

Transient transfections (Fig. 4) of the constructs into rodent cell line C6 (rat glial tumor) were performed. Although the Alu element in the control plasmid has an intact internal Pol III promoter, Alu transcripts are barely detectable from the control plasmid. In contrast, the transcription from the Apert's Alu element and its potential source gene were elevated three- to fourfold, as expected for putative mobilization-competent Alu repeats. This result suggests that the genomic flanking sequence of Ya5NBC237 probably makes the Alu transcription competent, one of the several requirements of a source gene. The same results

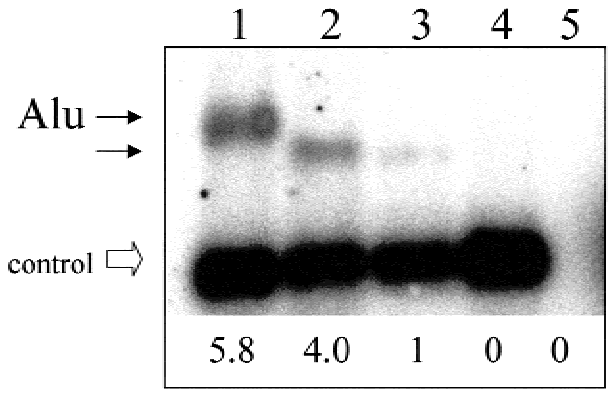

Figure 4 Evaluation of transcriptional capability of the potential *FGFR2* source Ya5 Alu element. The transcriptional efficiency of the de novo *FGFR2* Alu repeat and its putative source gene were evaluated by Northern blot analysis from transient transfection studies. The following constructs were evaluated: (lane *1*) p^{-290}Ap, (lane *2*) $p\text{-}^{416}$Ya5NBC237, and (lane *3*) p^{NP}Ya5NBC237. Lanes *4* and *5* are internal control only, and no DNA controls, respectively. Small arrows indicate the Alu transcripts and the open arrow indicates the internal control transcript. The ratio of the Alu transcript/control transcript (numbers below) was normalized to the p^{NP}Ya5NBC237 transcription ratio, which was assigned the arbitrary value of 1.

were obtained from transfections in the human embryonic kidney cell line 293 (data not shown).

DISCUSSION

Our computational and experimental analyses of the Ya5 subfamily of Alu repeats provides an overall picture of the most active of the recently integrated young Alu subfamilies from the human genome. The analysis of Alu Ya5 repeats allowed us to address a number of questions about the biology of these elements, such as the potential impact of gene conversion events, and the identification of Alu family members from the human genome that may be capable of retroposition.

Alu elements spread throughout the genome by retroposition in the last 65 million years. The master/source gene model (Batzer et al. 1990; Shen et al. 1991; Deininger et al. 1992) posits that a very small subset of the >1,000,000 Alu elements within the human genome are capable of high levels of retroposition; although a much larger number may make a few copies. The formation of Alu subfamilies may be explained by the sequential accumulation of mutations within the active source gene(s) followed by proliferation of the mutated source elements. A number of studies indicate that relatively few source Alu genes have played a dominant role in the amplification and evolution of Alu elements (Shen et al. 1991; Deininger et al. 1992; Deininger and Batzer 1993; Kapitonov and Jurka 1996). Although retroposition is the primary mode of SINE mobilization and sequence evolution through mutations in the source gene(s), our analysis suggests that gene conversion and genetic instability of Alu-based simple sequence repeats have also had a significant impact on the sequence architecture of this major family of human genomic sequences.

There are several alternatives that could explain the occurrence of mosaic Alu elements. First, some of the mosaic Alu elements with a single mutation could be explained by the occurrence of parallel mutations. However, this seems unlikely unless there were selection for these specific mutations, possibly through a post-transcriptional selection process (Sinnett et al. 1992). It is also difficult to envision a selection process that would only select for mutations at adjacent diagnostic positions, such as we see here. Also, recombination between different Alu elements could have generated some of these intermediate Alu elements that contain a mosaic of diagnostic mutations. However, in many cases, multiple recombination events would be required to obtain this outcome, making it highly unlikely. Although there are alternative mechanisms, we believe gene conversion is the most likely explanation for the occurrence of mosaic Alu elements.

The mechanisms of genome-wide gene conversion between mobile elements are not well understood in humans (see Kass et al. 1995, and references therein). Our data show that even the very short, dispersed Alu elements appear to be capable of high levels of gene conversion, which usually involve only short sequence stretches. In addition, our data show that reverse or backward gene conversions may be more favored. It seems likely that higher levels of the Y element copy number (Shen et al. 1991) or transcription (Shaikh et al. 1997) may play a role in determining the directionality of the gene conversion events. Although older Alu subfamilies, such as J and Sx are present in higher copy numbers in the genome, they diverged greatly from their consensus sequences due to mutations that have accumulated throughout evolution. Gene conversion would not be favored between such divergent sequences. However, Alu Y elements tend to be more conserved (better matches to Ya5) and with high copy number (Batzer et al. 1995). Therefore, both abundance (genomic copy number and/or transcript levels) and sequence identity appear to be influential in the Alu gene conversion events observed.

There are multiple examples of gene conversion events in literature. Genetic exchange between exogenous and different endogenous mouse L1 elements has been demonstrated previously to readily occur (Belmaaza et al. 1990). Kass et al. (1995) reported previously a gene conversion event in which one of the oldest Alu family members was converted to one of the youngest Alu subfamilies, Yb8. In addition, a partially converted Yb8 Alu element was also reported previously by Batzer et al. (1995). In yeast, some types of

mobile elements spread through the genome by gene converting pre-existing elements (Hoff et al. 1998). When we combine this type of mobilization in the yeast genome with the Alu gene conversions reported previously, as well as those in this paper, one could argue that gene conversion may represent a second type of amplification mechanism for short interspersed elements in the human genome. These observations suggest that evolutionary studies of all types of interspersed elements that ignore gene conversion events may lead to biased conclusions.

Variations in the length of the middle A-rich region and oligo-dA-rich tails of Alu elements are not uncommon (Economou et al. 1990; Arcot et al. 1995b; Jurka and Pethiyagoda 1995). Microsatellite repeats have been found to be associated with the 3′ oligo (dA) tails and the middle A-rich region of Alu elements. In the case of Friedreich ataxia, the most common mutation is the hyperexpansion of a GAA trinucleotide repeat within the middle A-rich region of an Sx Alu (Montermini et al. 1997). However, microsatellites in the middle of Alu elements are not as common because of the much shorter initial length of the middle A-rich region. Arcot et al. (1995b) reported previously that only about one-fourth of the Alu elements containing $(AC)_n$ repeats had them as a part of their middle A-rich region. The one specific example they studied in detail had an evolutionary expansion of the A-rich region (orangutan and gibbon) before the genesis of the AC repeat; suggesting the requirement for an initial expansion. Interestingly, our large-scale analysis of the middle A-rich regions of Ya5 elements demonstrates a trend toward expansion of the A region, providing additional support for this region of the Alu elements to act as a potential nucleus for the genesis of simple sequence repeats.

From our subset of 269 AluYa5 elements, we were able to identify a new Alu subfamily termed Ya5a2. The estimated average age of 0.62 million years (0.28–1.08 million years with 95% confidence) makes Ya5a2 the youngest subfamily of Alu repeats identified in the human genome to date. It is as abundant as the Ya8 subfamily (Roy et al. 1999) and its higher level of insertion polymorphism suggests a higher level of current retroposition. The Ya5a2 subfamily may have originated from a Ya5 Alu element that inserted in a genomic region that favored transcription and corresponding retroposition activity of the element, thereby generating a source gene. The subsequent accumulation of the two specific mutations facilitated the differentiation of the copies made by the Ya5a2 source gene from the larger background of several hundred genomic Ya5 Alu family members. As new Alu elements integrate into the genome in favorable genomic locations, they can occasionally remain retropositionally competent and generate copies of themselves. However, the frequency of fortuitous insertions of new Alu elements into favorable genomic locations for subsequent mobilization is still a rare event because the continuity of the hierarchical subfamily sequence structure of the Alu elements is largely conserved throughout primate evolution.

Alu elements that are polymorphic for insertion presence/absence have been proven previously to be useful for the study of human population genetics and forensics (Batzer et al. 1991; Jorde et al. 2000; Perna et al. 1992; Batzer et al. 1994; Tishkoff et al. 1996; Stoneking et al. 1997). The identification of a very young Alu subfamily with a high proportion of polymorphic members provides a new source of Alu insertion polymorphisms for the study of human population genetics. However, it is important to note that theYa5a2 subfamily is extremely small (~35 copies total in a background of >1,000,000) comparable with Ya8, so that an exhaustive analysis of a single human genome would only generate ~20 polymorphic Ya5a2 elements.

Because our analysis of Alu elements related to the Apert's insertion only included ~40% of the human genome (both finished and draft sequence included), there are possibly one or two other perfect complements in the human genome that have not yet been sequenced and may be the actual source gene for these elements. The transcriptional potential of this element would be consistent with its role as the potential source Alu gene. This confirms the existence of minor active source genes that differ from the source gene that generated almost all of the Alu elements present in the human genome today. In addition, the de novo Apert's Alu element was also transcriptionally active. There are two possible explanations for this result. First, the transcriptional capacity of the elements was evaluated by transient transfections in tissue culture. This system does not reflect the influence of chromatin structure and methylation patterns (position effects) on the transcription and presumably retroposition potential of the two Alu repeats. Alternatively, the de novo Apert's Alu element may have inserted in a region of the *FGFR2* gene that fortuitously enhances its own transcription capability. Although further studies will be required to make more definitive statements in this regard, the transcriptional capability of Ya5NBC237 is consistent with one of the many requirements a source gene possesses, making it a plausible candidate source gene for the de novo Apert's insertion.

In summary, the computational analyses of a subset of recently integrated Alu elements demonstrate that Alu sequence evolution is affected by a number of dynamic events. New retroposition-competent Alu source genes, gene conversion, and genetic instability each play an important role in Alu sequence evolution and proliferation within the human genome.

METHODS

Computational Analyses

Screening of the GenBank nr, the htgs, and the gss databases were performed by use of the Advanced Basic Local Alignment Search Tool 2.0 (BLAST) (Altschul et al. 1990) available from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). For the Ya5 subfamily analysis, the database was searched for matches to the 281 bases of the Ya5 consensus sequence with the following advanced options: -e 1.0 e-120, -b 1000, and -v 1000. A region composed of 500 bases of flanking DNA sequence directly adjacent to the sequences identified from the databases that matched the initial GenBank BLAST query were subjected to annotation by use of either RepeatMasker2 from the University of Washington Genome Center server (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) or Censor from the Genetic Information Research Institute (http://www.girinst.org/Censor_Server-Data_Entry_Forms.html) (Jurka et al. 1996). These programs annotate the repeat sequence content of DNA sequences from humans and rodents. The sequences were then subjected to more detailed analysis by use of MegAlign (DNAStar version 3.1.7 for Windows 3.2). The following parameters were used to select the Ya5 elements to be analyzed: (1) Ya5 had to have all five diagnostic nucleotides (except for the first position, as it is a highly mutable CpG). (2) No truncated Alu elements were included in the analysis. (3) No Alu elements identified as a result of directed cloning strategies designed to identify Alu repeats were included (only those randomly found within larger data sequence). (4) Duplicate Alu elements were eliminated on the basis of flanking sequences. The consensus sequences of the Yb8 and Ya8 subfamilies were used for parallel searches of the three GenBank databases mentioned above. A complete list of the Alu elements identified from the GenBank search is available from M.A.B. or P.L.D. and at http://www.genome.org/cgi/doi/10.1101/gr152300.

To search for putative source genes of the Alu elements that have been associated previously with different diseases, the three GenBank databases were searched by use of the sequence of each individual repeat to identify exact complements (Deininger and Batzer 1999).

DNA Samples

Human DNA samples from the European, African-American, Egyptian, and Greenland native population groups were isolated from peripheral blood lymphocytes (Ausubel et al. 1996) that were available from previous studies (Roy et al. 1999).

Oligonucleotide Primer Design and PCR Amplification

A region composed of ~500 bases of flanking unique DNA sequences adjacent to each Alu repeat were used to design primers for 14 Ya5a2 Alu elements (13 exact matches to consensus, Table 2). PCR primers were designed with the Primer3 software (Whitehead Institute for Biomedical Research) (http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). The resultant PCR primers were screened against the GenBank nr database for the presence of repetitive elements by use of the BLAST program, and primers that resided within known repetitive elements were discarded and new primers were designed. PCR amplification was carried out in 25-μL reactions with 50–100 ng of target DNA, 40 pM of each oligonucleotide primer, 200 μM dNTPs in 50 mM KCl, 1.5 mM $MgCl_2$, 10 mM Tris-HCl (pH 8.4), and Taq DNA polymerase (1.25 units) as recommended by the supplier (Life Technologies). Each sample was subjected to the following amplification cycle: an initial denaturation of 2:30 min at 94°C, 1 min of denaturation at 94°C, 1 min at the annealing temperature, 1 min of extension at 72°C, repeated for 32 cycles, followed by a final extension at 72°C for 10 min. Twenty microliters of each sample was fractionated on a 2% agarose gel with 0.25 μg/ml ethidium bromide. PCR products were directly visualized by UV fluorescence. The human genomic diversity associated with each element was determined by the amplification of 20 individuals from each of 4 populations (African American, Greenland native, European, and Egyptian; 160 total chromosomes). The chromosomal location for elements identified from randomly sequenced large-insert clones was determined by PCR analysis of National Institute of General Medical Sciences (NIGMS) human/rodent somatic cell hybrid mapping panels 1 and 2 (Coriell Institute for Medical Research, Camden, NJ).

Construction of Plasmids

The following constructs were made: p^{-416}Ya5NBC237 (416 bp upstream genomic – Alu – 223 bases downstream); p^{-290}Ya5Ap (290 bp upstream genomic – Alu – 293 bases); and p^{NP}Ya5NBC237 (no upstream vector flank–Alu – 223 bases). Unless otherwise noted, PCR was performed in 20-μL reactions by use of an MJ Research PTC 200 thermal cycler with the following conditions: 1X Promega buffer, 1.5 mM $MgCl_2$, 200 μM dNTPs, 0.25 μM primers, 1.5 units of *Taq* polymerase (Promega) at 94°C for 2 min; 94°C for 20 sec, 55°C (annealing temperature) for 20 sec, 72°C for 1 min, for 30 cycles; 72° C for 3 min. To PCR amplify and clone the 864-bp fragment containing the de novo Alu Ya5 from Apert syndrome patient 1 (accession no. AF097344), the following primers were used: forward, 5′-GGTGTGGCCAAAGTGGAGGATGTGTAC-3′ and reverse, 5′-TTATTCAAGGATAAAAGGGGCCATTTC-3′ with an annealing temperature of 50°C; and for the 920-bp fragment containing AluYa5NBC237 (accession no. AL031274) the primers used were: forward, 5′-TTATTCCATTG GTCCTTTCCACCAG-3′ and reverse, 5′-CAGGCAGGGAGG TACTTGTCTCTTG-3′ with an annealing temperature of 55°C.

For the p^{NP}Ya5NBC237, PCR amplification from the clone was done with the same reverse primer and the FAlu5 primer 5′-GGCCGGGCGCGGTGGCTCA-3′.

The final PCR product of the complete construct was cloned into pGEMTeasy Vector System I (Promega). Constructs were subjected to DNA sequence analysis to verify their sequence context. Purified plasmids from the constructs were prepared by alkaline lysis of bacterial cells followed by banding in a CsCl gradient twice. DNA concentrations were determined spectrophotometrically by use of A_{260} and verified by visual examination of ethidium bromide-stained agarose gels.

Alu Transcription in Cell Lines and RNA Analysis

Transient transfections were carried out in the rodent cell line C6 glioma (ATCC CCL107). Monolayers were grown to 50%–70% confluency and transfected with 3 μg of the construct-containing plasmid and 1 μg of control plasmid (p^{7SL}BC1) by use of LipofectAmine Plus (GIBCO Life Sciences) following the manufacturer's recommended protocol. Total RNA was isolated 16–20 h post-transfection.

RNA was extracted from cell lines utilizing the Trizol Reagent (Life Technologies, Inc.) according to the manufactur-

er's protocol. Equal amounts of RNA were fractionated on a 2% agarose–formaldehyde gel and then transferred to a nylon membrane, Hybond-N (Amersham). Northern blots were hybridized utilizing the following end-labeled oligonucleotide probes: unique-1 5′-TGTGTGTGCCAGTTACCTTG-3′ (complementary to the 3′ end of the control plasmid) and AluYA5-1 5′-ACCGTTTTAGCCGGGAATGGTC-3′ (complementary to Ya5 Alu RNA, but not to 7SL) in 5× SSC, 5× Denhardt's, 1% SDS, and 100 µg/mL herring sperm DNA. Oligonucleotides were end labeled by incorporating [γ-^{32}P]ATP (Amersham) with T4 polynucleotide kinase (New England BioLabs), and subsequently separated from free label by filtration through a Sephadex G-50 column. Blots were washed three times at 45°C with a low stringency buffer (2× SSC and 1% SDS) and subjected to autoradiography or quantified with a FujiFilm FLA-2000 fluorescent image analyzer (Fuji Photo Film Co. LTD). Statistical analysis was performed with the Jandel SigmaStat Statistical Software Version 2, (Jandel Corporation).

ACKNOWLEDGMENTS

A.M.R. was supported by a Brown Foundation fellowship from the Tulane Cancer Center. This research was supported by National Institutes of Health RO1 GM45668 to P.L.D., Department of the Army DAMD17-98-1-8119 to P.L.D. and M.A.B., and award no. 1999-IJ-CX-K009 from the Office of Justice Programs, National Institute of Justice, Department of Justice to M.A.B. Opinions expressed in this document are those of the authors and do not necessarily represent the official position of the U.S. Department of Justice.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Arcot, S.S., Shaikh, T.H., Kim, J., Bennett, L., Alegria-Hartman, M., Nelson, D.O., Deininger, P.L., and Batzer. M.A. 1995a. Sequence diversity and chromosomal distribution of "young" Alu repeats. *Gene* **163:** 273–278.

Arcot, S.S., Wang, Z., Weber, J., Deininger, P.L., and Batzer, M.A. 1995b. *Alu* repeats: A source for the genesis of primate microsatellites. *Genomics* **29:** 136–144.

Ausubel, F.M., Brent, R. Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K. 1996. *Current protocols in molecular biology*, John Wiley & Sons, Inc. Canada.

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

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

Batzer, M.A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D.H., Shaikh, T.H., Novick, G.E., Ioannou, P.A., Scheer, W.D., Herrera, R.J.,and Deininger, P.L. 1994. African origin of human-specific polymorphic Alu insertions. *Proc. Natl. Acad. Sci.* **91:** 12288–12292.

Batzer, M.A., Rubin, C.M., Hellmann-Blumberg, U., Alegria-Hartman, M., Leeflang, E.P., Stern, J.D., Bazan, H.A., Shaikh, T.H., Deininger, P.L., and Schmid, C.W. 1995. Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. *J. Mol. Biol.* **247:** 418–427.

Batzer, M.A., Deininger, P.L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C.M., Schmid, C.W., Zietkiewicz, E., and Zuckerkandl, E. 1996a. Standardized nomenclature for *Alu* repeats. *J. Mol. Evol.* **42:** 3–6.

Batzer, M.A., Arcot, S.S., Phinney, J.W., Alegria-Hartman, M., Kass, D.H., Milligan, S.M., Kimpton, C., Gill, P., Hochmeister, M., Ioannou, P.A., et al. 1996b. Genetic variation of recent Alu insertion in human populations. *J. Mol. Evol.* **42:** 22–29.

Belmaaza, A., Wallenburg, J.C., Brouillette, S., Gusew, N., and Chartrand, P. 1990. Genetic exchange between endogenous and exogenous LINE-1 repetitive elements in mouse cells. *Nucleic Acids Res.* **18:** 6385–6391.

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

Deininger, P.L. and Daniels, G. 1986. The recent evolution of mammalian repetitive DNA elements. *Trends Genet.* **2:** 76–80.

Deininger, P.L. and Batzer, M.A. 1993. Evolution of retroposons. In *Evolutionary biology* (ed. M.K. Heckht), pp. 157–196. Plenum Publishing, New York, NY.

———. 1999. Alu repeats and human disease. *Mol. Genet. Metab.* **67:** 183–193.

Deininger, P.L., Batzer, M.A., Hutchison, I.C., and Edgell, M. 1992. Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8:** 307–312.

Dombroski, B.A., Scott, A.F., and Kazazian, Jr., H.H. 1993. Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc. Natl. Acad. Sci.* **90:** 6513–6517.

Economou, E.P., Bergen, A.W., Warren, A.C., and Antonarakis, S.E. 1990. The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. *Proc. Natl. Acad. Sci.* **87:** 2951–2954.

Hoff, E.F., Levin, H.L., and Boeke, J.D. 1998. Schizosaccharomyces pombe retrotransposon Tf2 mobilizes primarily through homologous cDNA recombination. *Mol. Cell. Biol.* **18:** 6839–6852.

Jelinek, W.R. and Schmid, C.W. 1982. Repetitive sequences in eukaryotic DNA and their expression. *Annu. Rev. Biochem.* **51:** 813–844.

Jurka, J. and Smith, T. 1988. A fundamental division in the Alu family of repeated sequences. *Proc. Natl. Acad. Sci.* **85:** 4775–4778.

Jurka, J. and Pethiyagoda, C. 1995. Simple repetitive DNA sequences from primates: Compilation and analysis. *J. Mol. Evol.* **40:** 120–126.

Jurka, J., Kaplan, D.J., Duncan, C.H., Walichiewicz, J., Milosavljevic, A., Murali, G., and Solus, J.F. 1993. Identification and characterization of new human medium reiteration frequency repeats. *Nucleic Acids Res.* **21:** 1273–1279.

Jurka, J., Klonowski, P., Dagman, V., and Pelton, P. 1996. CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20:** 119–121.

Jorde, L.B., Watkins, W.S., Bamshad, M.J., Dixon, M.E., Ricker, C.E., Seielstad, M.T., and Batzer, M.A. 2000. The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosome data. *Am. J. Hum. Genet.* **66:** 979–988.

Kapitonov, V. and Jurka, J. 1996. The age of Alu subfamilies. *J. Mol. Evol.* **42:** 59–65.

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

Labuda, D. and Striker, G. 1989. Sequence conservation in Alu evolution. *Nucleic Acids Res.* **17:** 2477–2491.

Lester, T., McMahon, C., VanRegemorter, N., Jones, A., and Genet, S. 1997. X-linked immunodeficiency caused by insertion of Alu repeat sequences. *J. Med. Gen. Suppl.* **34:** S81.

Miki, Y., Katagiri, T., Kasumi, F., Yoshimoto, T., and Nakamura, Y. 1996. Mutation analysis in the BRCA2 gene in primary breast cancers. *Nat. Genet.* **13:** 245–247.

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* **238:** 369–373.

Montermini, L., Andermann, E., Labuda, M., Richter, A., Pandolfo, M., Cavalcanti, F., Pianese, L., Iodice, L., Farina, G., Monticelli, A., et al. 1997. The Friedreich ataxia GAA triplet repeat: Premutation and normal alleles. *Hum. Mol. Genet.* **6:** 1261–1266.

Oldridge, M., Zackai, E.H., McDonald-McGinn, D.M., Iseki, S., Morriss-Kay, G.M., Twigg, S.R., Johnson, D., Wall, S.A., Jiang, W., et al. 1999. De novo Alu-element insertions in FGFR2 identify a distinct pathological basis for Apert syndrome. *Am. J. Hum. Genet.* **64:** 446–461.

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

Rogers, J.R. and Willison, K.R. 1983. A major rearrangement in the H-2 complex of mouse t haplotypes. *Nature* **304:** 549–552.

Roy, A.M., Carroll, M.L., Kass, D.H., Nguyen, S.V., Salem, A-H., Batzer, M.A., and Deininger, P.L. 1999. Recently integrated human Alu repeats: Finding needles in the haystack. *Genetica* **107:** 149–161.

Schmid, C.W. and Maraia, R. 1992. Transcriptional regulation and transpositional selection of active SINE sequences. *Curr. Opin. Genet. Dev.* **2:** 874–882.

Shaikh, T.H. and Deininger. P.L. 1996. The role and amplification of the HS Alu subfamily founder gene. *J. Mol. Evol.* **42:** 15–21.

Shaikh, T.H., Roy, A.M., Kim, J., Batzer, M.A., and Deininger, P.L. 1997. cDNAs derived from primary and small cytoplasmic Alu (scAlu) transcripts. *J. Mol. Biol.* **271:** 222–234.

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

Sinnett, D., Richer, C., Deragon, J.M., and 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.

Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H., and Deininger, P.L. 1987. Clustering and sub-family relationships of the Alu family in the human genome. *Mol. Biol. Evol.* **4:** 19–29.

Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9:** 657–663.

Stoneking, M., Fontius, J.J., Clifford, S.L., Soodyall, H., Arcot, S.S, Saha, N., Jenkins, T., Tahir, M.A., Deininger, P.L., and Batzer. M.A. 1997. Alu insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. *Genome Res.* **7:** 1061–1071.

Tishkoff, S.A., Ruano, G., Kidd, J.R., and Kidd, K.K. 1996. Distribution and frequency of a polymorphic Alu insertion at the plasminogen activator locus in humans. *Hum. Genet.* **97:** 759–764.

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

Weiner, A., Deininger, P.L., and Efstradiatis, A. 1986. The Reverse flow of genetic information: Pseudogenes and transposable elements derived from nonviral cellular RNA. *Annu. Rev. Biochem.* **55:** 631–661.

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

Received June 14, 2000; accepted in revised form August 9, 2000.