Alu Insertion Polymorphisms for the Study of Human Genomic Diversity

Astrid M. Roy-Engel,*[,1] Marion L. Carroll,[†,1] Erika Vogel,* Randall K. Garber,[†,‡] Son V. Nguyen,[†] Abdel-Halim Salem,[†,‡] Mark A. Batzer[†,‡,2] and Prescott L. Deininger*[,§,2]

Tulane Cancer Center, Department of Environmental Health Sciences, Tulane University Health Sciences Center, New Orleans, Louisiana 70112, [†]Departments of Pathology, Genetics, Biochemistry and Molecular Biology, Stanley S. Scott Cancer Center, Louisiana State University Health Sciences Center, New Orleans, Louisiana 70112, [‡]Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, Baton Rouge, Louisiana 70803 and [§]Laboratory of Molecular Genetics, Alton Ochsner Medical Foundation, New Orleans, Louisiana 70121

Manuscript received February 24, 2001
Accepted for publication June 8, 2001

ABSTRACT

Genomic database mining has been a very useful aid in the identification and retrieval of recently integrated Alu elements from the human genome. We analyzed Alu elements retrieved from the GenBank database and identified two new Alu subfamilies, Alu Yb9 and Alu Yc2, and further characterized Yc1 subfamily members. Some members of each of the three subfamilies have inserted in the human genome so recently that about a one-third of the analyzed elements are polymorphic for the presence/absence of the Alu repeat in diverse human populations. These newly identified Alu insertion polymorphisms will serve as identical-by-descent genetic markers for the study of human evolution and forensics. Three previously classified Alu Y elements linked with disease belong to the Yc1 subfamily, supporting the retroposition potential of this subfamily and demonstrating that the Alu Y subfamily currently has a very low amplification rate in the human genome.

ALU elements have been accumulating in the human genome throughout primate evolution, reaching a copy number of over a million per genome. However, most of these Alu copies are not identical and can be classified into several subfamilies (reviewed in Deininger and Batzer 1993). These different subfamilies of Alu elements were generated once mutations occurred within the "master" or "source" gene that actively retroposed at different rates and time periods of primate evolution (Deininger *et al.* 1992). Currently, the Alu retroposition rate is reduced by 100-fold from its peak early in primate evolution (Shen *et al.* 1991). The vast majority of the Alu elements present in the human genome inserted before the radiation of extant humans and are therefore observed in all individuals in the human population. However, almost all of the recently integrated Alu elements in the human genome are restricted to several closely related "young" subfamilies, with the majority being Ya5 and Yb8 subfamily members (Batzer *et al.* 1994, 1995). Several of these new subfamilies appear to originate from an Alu element that fortuitously inserted into a favorable region of the genome capable of supporting Alu retroposition. Subsequent or concurrent mutations in the new source element(s) result in groups of elements that are identifiable as new subfamilies.

Collectively, the Alu Y, Ya5, Ya5a2, Ya8, and Yb8 subfamilies comprise <10% of the Alu elements present within the human genome, with the Ya5/8 and Yb8 subfamilies together accounting for <0.5% of all Alu elements. Although the human genome contains >1,000,000 copies of Alu (~10% of the genome; Smit 1996), <0.5% are polymorphic. Due to their recent evolutionary introduction into the human genome, many of the young Alu elements are polymorphic between individuals and/or populations. There is an inverse correlation between the age of the Alu subfamily and the percentage of polymorphic elements it contains. Identification of evolutionarily recent Alu subfamilies and their polymorphic insertions is useful for human population studies, forensics, and DNA fingerprinting for two reasons: (i) There is no apparent specific mechanism to remove newly inserted Alu repeats, making inserts identical by descent; and (ii) the Alu insertions have a known ancestral state (Batzer and Deininger 1991; Batzer *et al.* 1994).

The availability of large quantities of human genomic DNA sequence provided by the Human Genome Project facilitates genomic database mining for recently integrated Alu elements. Through this approach we were able to identify the youngest Alu subfamily reported to date, termed (Ya5a2), and determined that the majority of its members are Alu insertion polymorphisms (Roy *et al.* 2000). We expanded our computational analyses to identify other Alu subfamilies derived from the Alu

Corresponding author: Prescott L. Deininger, Tulane Cancer Center, SL-66, Tulane University Medical Center, 1430 Tulane Ave., New Orleans, LA 70112. E-mail: pdeinin@tulane.edu

[1] These authors contributed equally to this work.

[2] These are equal senior authors.

Y and Yb8 subfamilies. Here, we present the analysis of three of the most recently formed Alu subfamilies and demonstrate their utility for the study of human genomic diversity.

MATERIALS AND METHODS

Computational analyses: Sequence alignments for the identification of Alu subfamilies were made using MegAlign software (DNAStar version 3.1.7 for Windows 3.2). Screening of the GenBank nonredundant (nr), the high throughput genome sequence (htgs), and the genomic survey sequence (gss) databases was performed using 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/). Database searches for Yb8 consensus Alus showed a common single-base variant termed Yb9. The databases were searched for matches to the 289 bases of the Yb9 consensus sequence (as inferred from the previous Yb8 analysis) or the 281 bases of the Alu Y consensus with the expected value (real) set at $-e\ 1.0e^{-150}$ and $-e\ 1.0e^{-140}$, respectively, in the advanced BLAST options. Only Alu Yb9 elements with all nine diagnostic mutations were selected. A similar type of search procedure was performed with the Yc1 and Yc2 consensus sequences or with an oligonucleotide query sequence complementary to the subfamily diagnostic base positions. Only Alu Yc1/Yc2 elements with 100% identity to the oligonucleotide query sequences or entire subfamily-specific consensus sequence were utilized for further analysis. To estimate the copy numbers of the Yb9 subfamily we searched the draft sequence of the human genome (Lander *et al.* 2001), using a subfamily-specific probe that contained the Yb9-specific mutation as well as the insertion in the Yb8 subfamily. A complete list of the Alu elements identified from the GenBank search is available from M. A. Batzer or P. L. Deininger.

DNA samples: Human DNA samples from the European, African-American, Alaskan Native, Egyptian, and Asian 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: Flanking unique DNA sequences adjacent to each Alu repeat were used to design primers for the Yb9, Yc1, and Yc2 Alu elements (Table 1). PCR primers and reactions were performed as previously described (Roy *et al.* 1999). The heterozygosity associated with each element was determined by the amplification of 20 individuals from each of four populations (African American, Alaskan Native or Asian, European, and Egyptian; 160 total chromosomes). The chromosomal location for elements identified from randomly sequenced anonymous large-insert clones was determined by PCR as previously described (Roy *et al.* 1999).

RESULTS

The Alu Yb9, Yc1, and Yc2 subfamilies: Analysis of a set of 243 Yb8 Alu elements retrieved from the GenBank database allowed us to identify a putative subfamily containing all the known Yb8 diagnostic mutations plus one new mutation, which is referred to as Yb9 in compliance with the standard Alu subfamily nomenclature (Batzer *et al.* 1996). The Yb9 consensus sequence is shown in Figure 1. Searches from the nr, the htgs, and gss retrieved a total of 56 Yb9 elements. Of these, 25 elements were retrieved from the nr database (30.4% of the human genome at the time), giving an estimated size of 82 members for the Yb9 subfamily. This estimate is also in good agreement with a search of the draft human genomic sequence (Lander *et al.* 2001) that identified 79 perfect matches with a Yb9 subfamily-specific query sequence.

Using a different approach, we also retrieved one previously identified subfamily, Yc1 [formerly termed Sb0 (Jurka 1995)], and a new variant, Yc2. GenBank database searches for Alu Y elements that perfectly match the consensus sequence brought several Alu Y elements to our attention that share one or two specific mutations that differ from the Y consensus. Closer inspection facilitated the retrieval of the additional Alu subfamilies. BLAST searches using the consensus sequence for Alu Yc1 and Yc2 will also retrieve a large number of elements that are matches to the Alu Y subfamily as well, making the analysis of the elements identified in this manner impractical. Therefore, we selected only the elements of these subfamilies with 100% identity to the oligonucleotide query sequence that contained the subfamily-specific diagnostic bases. A total of 176 Yc1 (13 perfect matches to the entire subfamily consensus sequence) and 17 Yc2 (11 perfect matches to the entire subfamily consensus sequence) elements were retrieved. A count of all Yc1 elements retrieved by BLAST on a single initial search of the nr database yielded a total of 116 elements, giving an estimated copy number of 381 Yc1 elements in the human genome (the nr database contained 30.4% of the human genome sequence at the time of the search). Interestingly, three of the four elements previously classified as Alu Y elements linked to disease (Deininger and Batzer 1999) belong to the Alu Yc1 subfamily (Figure2): the *de novo* insertion in the C1 inhibitor gene (C1inh; Stoppa-Lyonnet *et al.* 1990), another *de novo* insertion in BRCA2 (BRCA2; Miki *et al.* 1996), and glycerol kinase deficiency (GK; Zhang *et al.* 2000).

About one-half of the 56 total Yb9 elements (29) shared 100% nucleotide identity with the subfamily consensus sequence. To get an approximation of the age of the Yb9 subfamily, we evaluated the number of non-CpG mutations present within the different Alu elements as previously described (Roy *et al.* 2000). A total of 19 CpG mutations, 25 non-CpG mutations, and two 5′ truncations occurred within the 56 Alu Yb9 subfamily members identified. Using a neutral rate of evolution for primate intervening DNA sequences of 0.15% per million years (Miyamoto *et al.* 1987) and the non-CpG mutation density of 0.1908% (25/13,104 bases using only non-CpG bases) within the 56 Yb9 Alu elements yield an estimated average age of 1.27 million years (myr). The age for the Yb9 subfamily members is predicted at a 95% confidence level in the range of 0.8–1.8 myr, given that the mutations were random and fit a binomial distribution. No analysis can be made for the

TABLE 1

PCR primers, chromosomal locations, and PCR product sizes

Name	Accession	Position	5′ primer sequence (5′–3′)	3′ primer sequence (5′–3′)	A.T[a]	Human diversity[b]	Chr.[c] loc.	Product size[d] Filled	Product size[d] Empty
				Alu Yb9					
Yb9NBC1	AC024091	26414–26105	AGTATCTTTAGATCCAGGGTGAAGC	TTCCAGTGGTAAGTCTATGGGAAT	60	FP	12	411	86
Yb9NBC2	AC024896	142649–142362	GCAGACAGTACCCACTTATTTTTGT	TGGTTCTATAAAGCATTTGTTTCTTC	55	FP	7	462	146
Yb9NBC3	AC005342	167963–167121	GAAGCTCACTCTGCCATGTG	CATGTTCGGTCCTGCTTACA	60	FP	12	527	200
Yb9NBC6	AC020900	61455–61742	GCAGACCGTATGTTCAATAAATGAC	CCACTTGGAAAAACACCCAAA	55	FP	5	493	153
Yb9NBC7	AC009062	351148–351457	CAGTAAATGATGGGAACAACCTTC	CTAAATGTCAAGCTATGCCACAGA	58	HF	16	403	83
Yb9NBC8	AC011967	156726–157013	TAACTTTAGTTTTCCATCGCACATT	ACACTAGTTTTACCCTTGTCAGCAC	60	LF	18	419	86
Yb9NBC9	AC022199	71329–71616	AGCTTCCCATTTCTGGTTTGTCTAT	GCCTTGTTAAACCCAACCTTTT	60	FP	17	453	120
Yb9NBC10	AC025961	22060–21773	GTTTTCCTGGTGTGCCCTAAATA	TTTACCTAACTCACAAGACCCAAAG	60	IF	4	524	197
Yb9NBC11	AC019189	172700–172987	AAAGACTTTTCAGGTTCTTGTAGCA	ATGCATGTCTATGCAAACTATCAAA	55	FP	1	392	74
Yb9NBC12	AC011170	158821–159108	AACACCTGAGAAAGGCTCATTG	GCTTTGGAAAATACCTAAGAAGCAC	55	LF	10	414	93
Yb9NBC13	AC003985	36492–36792	TCTAGTTTGGAGTGCCATGC	CTCCCAGTCATGCTTCCTGT	60	FP	7	510	167
Yb9NBC15	AC006036	24671–25059	CCAAGGTTCAGCTTTATGCTC	GCTCAAAACCGCTGAATTGT	60	FP	7	489	159
Yb9NBC16	AC024057	13965–13678	GAAGAAAGAATGGAAAGGGTAGTTG	ACCCCCATGACACTAATTTACCTAT	55	—	3	416	117
Yb9NBC17	AC012664	22598–22311	ACTTACCCAAAACGCATGATTC	AACGTAGATGCAGACAACACTCTTT	60	FP	2	709	391
Yb9NBC18	AC005751	35038–35343	CGTTTGAAAGCTACGTTACC	TCCCATGAGGTAGTGATGAT	60	FP	16	531	203
Yb9NBC21	AL136081	33392–33083	TTCATGTAGCCAAAACTACCTGTTC	TTAACAGCTTACAGTTTGGCAGAG	52	FP	6	425	107
Yb9NBC22	AL139193	160587–160874	TGCACAACTATACACCAGACACTG	TTGTCTCCCATCAGTAGAACCTAAG	55	LF	14	435	110
Yb9NBC23	AL356756	80321–80034	CAGGACTTTTATTGAATCCTCACCT	AAAGAGAGACATGGCCCAATTA	58	FP	14	412	83
Yb9NBC25	AC008558	90385–90099	GAGTTGTCAAATTTGGAATGGATAC	ACATCATTAAGCTCTTCCTGACATT	55	FP	5	496	159
Yb9NBC27	AC011966	13523–13810	CATGGATAACATCATAAGGCTTCAG	ACACTAGTTTTACCCTTGTCAGCAC	55	LF	15	482	149
Yb9NBC28	AC004808	25856–26143	AAAAAGGTGTGCCTGGATATTTA	CTGTGGCATAACTCAAACTGTAATG	55	FP	7	539	208
Yb9NBC29	AC005008	29772–30089	GTAATATGAGGTGATGGGGGTTACT	GGTGAAAGAAGAACCCCTAAGTTAT	60	LF	7	474	138
Yb9NBC30	AC003003	35922–36249	GAACCCCATCCATTCTCTTACA	GTGGCAAAATATTGGCGACT	60	IF	16	508	156
Yb9NBC31	AF107258	58225–58541	TTTCTCAGCACTATCCCTGT	GACAGTGAGTTGGCAGTACC	56	FP	21	457	130
Yb9NBC32	AL121582	154486–154199	CCTAACCCCTACATTTTACCATTTC	GTCATTTGCACTTGTCAAAGAGTGT	55	FP	20	469	141
Yb9NBC34	AL121841	28487–28800	CCAAGTTTCCTCTGCTGGAA	CACAAATACTCCCTGCCTCAG	55	FP	14	489	90
Yb9NBC35	AC040906	166712–167029	TTAACAGCTTACAGTTTGGCAGAG	TTCATGTAGCCAAAACTACCTGTTC	60	FP	6	427	109
Yb9NBC36	AF015725	15626–15909	AAGCAGTCATCATCCATTTT	ACCACAAAAATGCACTTACC	60	FP	21	521	201
Yb9NBC37	AB014460	3311–3621	CAAAATGGCCGTGTTCTTTT	GTGTCCACGGATCTTTGCAG	62	FP	16	458	142
Yb9NBC39	AC004542	12799–13104	AGAGGATCTTTGCAGGCACT	GTGCTGTGGGGTTAGGAAGA	55	—	22	509	176
Yb9NBC40	AP000237	60117–60404	AGTGAGTTGGCAGTACCCAAAT	CTCAGCACTATCCCTGTTCTTACAT	60	FP	21	450	124
Yb9NBC41	AC004140	4672–4851	TGTTTCCTCATCTGCCAACTT	AAAAGACTGTTGATGACCACTCAG	55	FP	7	761	389
Yb9NBC42	AC004945	152516–152833	GACATCTCCCTCCTTTCTCT	GAAAACCTGAACATGGGTAA	55	FP	7	521	177
Yb9NBC44	AC006561	13793–13506	GACTACAACATACCATCCTCAAAGG	GTATAGGAAACAGCGTGTTGTGAC	55	FP	12	426	106
Yb9NBC45	AL121978	17555–17268	GGAGAACCACTTGAACATGGAG	AGCCCTGCTATATCCAGCTCTT	60	—	6	486	167
Yb9NBC48	Z95114	30489–30202	GCTGCATACCAGACCTTTGTC	TTGTGCTGTAAGGCTGAGTAGG	60	FP	22	432	117
Yb9NBC49	AC005375	129358–129604	ATCCTTTAGATCAGAGGTCATCAAG	CAACAACTAATCTGCTTTCTGTCAC	58	FP	17	393	134
Yb9NBC50	AL109865	28485–28772	GTTCCACAAGTACAGGAGAAAATGT	GAAGCTCTTTAGGAAACCAAATCTC	55	IF	11	460	138

(*continued*)

TABLE 1

(Continued)

Name	Accession	Position	5′ primer sequence (5′–3′)	3′ primer sequence (5′–3′)	A.T[a]	Human diversity[b]	Chr.[c] loc.	Product size[d]	
								Filled	Empty
Yb9NBC53	AQ382257	185–472	GGGACTGGGTATAAATGAGGTG	GGACCAATCCTACCTTGTATGG	55	HF	20	454	68
Yb9NBC54	AL050305	39695–40005	TAGGATGAGAATGAACTTTGAGATG	CCATTTATAACCAATGACGACAAAG	58	FP	X	492	172
Yb9NBC55	AQ076355	91–379	CTCAGATAAGGAAACTGAAACACAG	CCTATACCTTAAAACAAGCTTGGAC	60	FP	1	425	108
Yb9NBC58	AC022199	109996–109711	TTGACTGTAACGTCACTTTATTTGC	TGACTAGTGCTTTTGTGTATTGAGAA	60	FP	17	445	128
Yb9NBC59	AL121582	149776–150063	GTTTTCTCAGTCTCTTGCATTTTGG	GGGTGCAGAGACCAAAAACTT	55	FP	20	480	160
Yc1NBC1	AC011296	4067–3787	AGTACGTGAGGTTTCTATGCCTTG	GATTTGTCCATAATAGCCCCTAACT	60	IF	7	481	159
Yc1NBC2	AC006195	139237–139517	TCTCTCATGAACATAGATACAAA	CGTGCATTCTTGAGATAAAT	60	IF	7	443	102
Yc1NBC3	AC010072	48921–49201	GGATACCCCTTGGGAAAAGA	GAACACCATGTAACCCTCACC	63	FP	14	405	92
Yc1NBC6	AC004016	82266–81986	CAAACTCTGTGCACCTTGACA	CACTCGCATTATGGATTTTGG	65	FP	7	1009	677
Yc1NBC8	AC007298	28402–28682	CATCAAACCCCACACACTCA	TCCTTGGAGCCACATGTTTT	63	FP	12	463	115
Yc1NBC9	AL121603	31558–31838	GCCCAGCTGGAAATAGCTT	AGAAATTCTGCATGTGTCTCAG	63	IF	14	490	159
Yc1NBC11	AF123462	93456–93176	GGGAATGTTCATAGGACATGG	TGGAACATGCCAGAAAGAGA	63	FP	14	778	437
Yc1NBC13	AL122006	69774–70054	TCCCAAGTCCCATCCCTTAGAA	GCCATTCCTCACCAGCCATT	60	FP	1	504	165
Yc1NBC14	AL031734	146718–146998	TGAGGTCCTGTGACTTGGTG	TCGCAAAGCATTTCTCAAAG	60	FP	1	464	149
Yc1NBC15	AL031650	85392–85112	GGAATGGCATAGGAAGTGGA	ACCAAATGAAAGGGGAGACA	63	FP	20	418	112
Yc1NBC20	AP001696	246018–246298	GCAAGTAATGAAAGGATTCTAGGG	AGAGCTGCCCTATTTCTT	60	FP	21	486	163
Yc1NBC23	AC004626	28992–29271	TGCTCAGGTTAGGGATGTTAATGC	TTCTGAGCTGCTGGGGGACT	60	LF	16	445	120
Yc1NBC24	AL137013	69320–69041	TCAAAGGGGAATACTGGGGAAA	GGGGAAATGACAATCAAGTGGAA	60	FP	X	408	88
Yc1NBC25	AC018637	72620–72340	GGGCAGGGGATGTAGGGACT	TGCCCCTGTTTCATCTGTGC	60	FP	7	432	108
Yc1NBC26	AC027279	127822–128102	TCACTTCAGAAGGGGAAAAA	TGTGGTCGCCTGGACTTCAA	60	FP	16	472	165
Yc1NBC28	AC017019	30139–29859	TGGTGAGTTCCTGGTCTTGCTG	TGCTCACTCTTTGGGTCCACAC	60	—	Y	414	99
Yc1NBC30	AL157756	37868–38148	GCCGCTAGCCTTTGTGTAAA	CAAAGTCATCTCTGACCCCCAGA	60	FP	14	497	177
Yc1NBC31	AC008062	103843–103563	TTCTGTAAAAGCCTGTTAGGTCCA	CAGCATTTCACTGTCAGCATTGG	60	LF	7	443	110
Yc1NBC32	AC005866	37960–37680	GCGAGGGCAAGCAGACAATAA	GTGGAGCTCACCCCTTCAGA	55	FP	12	425	114
Yc1NBC33	AL132994	40508–40788	CTTTATGGGTCTTACAGTAGAA	TGCATATGTAGCCTCTGATTC	60	FP	14	500	186
Yc1NBC34	AL136382	87933–87653	CCCACAACCCTTTCCCAGAG	CAGCAACCTGGATGGAGTGG	60	—	1	477	165
Yc1NBC35	AC004638	32778–33058	CCCATTCTCCATGCCGTGAT	TGCAAGGCATTGGGGATACA	60	IF	16	481	162
Yc1NBC36	AL121903	24409–24129	CACAGGAACTATTCCCCACAA	GCGAAATTCTTGAAGGAAAAACTGG	60	FP	20	437	101
Yc1NBC37	AL049562	25982–25702	CGTGCATTCCTTCTCATCACA	GGCACTTTTACCTAAAGAGCTTACA	60	FP	X	406	88
Yc1NBC38	AC000118	10509–10789	TCCAAACCTCTCTTGTTTGGAT	TGAAAGGATTTATGCCAGGTG	60	FP	7	435	113
Yc1NBC39	AP001695	126848–127128	TCCAGGAAGCAAACAGAATTCAGAG	TCAACCCCACTCTGATGCTCAA	60	FP	21	623	291
Yc1NBC45	AF218891	1964–1684	TGGCCACATTGGAATTCAAAACTAT	TTCCTGCTGCTAAGTGACACATGA	60	FP	20	401	94
Yc1NBC46	Z86061	56824–57103	CTTTGAAGCATGCAAGGAAAGG	CAGTTTCCAATCTTAGGGACTTGA	60	FP	X	489	172
Yc1NBC48	AC007094	66892–66612	TTCACCACAATTAATAGGAAGGTTT	CAAACGCAGCACAGGATCTGA	60	FP	7	700	392
Yc1NBC49	AC011493	52071–51791	TGTGCTGTTACTATGGAGCCCTAC	CTGGGGAGACATCCCCTTCC	60	—	19	413	94
Yc1NBC50	AC010382	50258–49978	GGTATGGGGCCAAATTTAATCCA	TCCAAGAGAAGCCAAACCTACAGA	60	IF	5	406	101
Yc1NBC51	AC009415	123638–123918	TCATACAAAAGACAGGCTTTGC	CAAGGGAACAGATTCAGAAGAAACA	55	LF	7	521	208
Yc1NBC52	AC002429	141029–140749	GCTTTTGACACATCCCCAGGT	CACAAGATTTGGGGCCAAGAG	62	FP	7	429	111
Yc1NBC53	AC004848	43020–42740	AAAGCTATCAACCATGCCAACA	GAAAATGCTATTTTGGGGAATG	62	IF	7	505	186

(*continued*)

TABLE 1

(Continued)

Name	Accession	Position	5′ primer sequence (5′–3′)	3′ primer sequence (5′–3′)	A.T[a]	Human diversity[b]	Chr.[c] loc.	Product size[d]	
								Filled	Empty
Yc1NBC56	AC006017	155231–154951	TCTGTAAAAGTGCTTCACAT	GGGGGTGTGATATTCGTGCTG	55	—	7	593	287
Yc1NBC58	AL133367	83515–83795	TGCTGCATCAATCAGCCAGA	TCCCAGTCCTTGGCAACCAT	65	FP	14	427	118
Yc1NBC59	AC006213	58483–58763	ACCCTCCCCTCCTTCTGTGG	CCCTGCAGAACGCTGGAAAA	60	FP	19	428	93
Yc1NBC60	AL136319	30378–30658	GAAACCGCCAAGATTCTCACC	TCTCCATCATGATTCCCAACTGA	60	IF	10	522	205
Yc1NBC63	AL121964	57663–57943	GGTACTCAGTAACACATCAAGA	AAGCTGGGTGGTGGGTTCAC	60	IF	6	502	181
Yc1NBC64	AL121904	25022–25262	CAGATCCTGGTTCGTAGGAGGTG	CAAGCTGTGTATTCTTGATACTGC	60	IF	20	600	292
Yc1NBC65	AL049643	46216–45936	TTGGCTGAGGATATCAGATGTGT	TCCAGTGCTTAAGAGTAAAGCAAGC	60	FP	X	456	152
Yc1NBC66	AJ006998	11416–11136	GGCCTAGCAAGGTCTTTTGC	TGATGAGTGTACAAGCCACACTT	60	FP	21	422	110
Yc1NBC69	AB020859	19030–19310	CCCACATTTATCAGTACCTACA	CCTTGCAGAATAGCAATGAT	55	IF	8	524	210
Yc1NBC70	AL133238	24939–24661	AGCAATTGTGGAGCCAGGAA	GAGGGTGCTTAGTGGGAGCAAA	60	FP	14	452	137
Yc1RG60	AC019215	161766–162046	TCCCACATTTCAGTGTGAATTT	GGCATTTGGGATAGTTCCTG	60	HF	8	474	159
Yc1RG62	AC007428	139021–139301	GCTCAACATGCATAACCTTGAAC	ATTTCCAAAGAAACCCCTGACT	60	FP	?	522	216
Yc1RG83	AC009004	751–1030	CTGGCTGGGAGATTTTGTTAAA	GTTGGAAACAGTGTATTGCCTGTA	60	FP	19	724	397
Yc1RG64	AC009289	65992–66272	TCCAGTCATCTTAATGTGCCTTAG	GGATAGACCTTGCCTTTCTGATT	60	FP	14	380	67
Yc1RG65	AC019181	63269–63549	GCAGCCTGGATATCAATTAAGG	ATGGGTTAAAACTCCTAGCACTG	60	FP	2	735	413
Yc1RG66	AC009506	7323–7615	CTTTTTCTCAGACTGTGCTTGC	GCAACAAGACAAAACAGCAACTG	60	FP	1	419	109
Yc1RG67	AC008039	178981–179192	AAACTACCTTCCCCAGACTCC	CCCTAAGGACTTTATAATGGGACT	60	FP	7	382	125
Yc1RG68	AC008039	164672–164954	ATGGTGTCCACAAGAAACTGAG	GGAAGGCTCCATTATAGGTCTTG	60	IF	7	480	166
Yc1RG70	AC006323	3461–3741	CTCTGCAGCATGACAAATCAAT	CAGCATCTAAAGCACTCACTTCA	60	FP	17	504	178
Yc1RG71	AC011450	98261–98574	TACTGAAGACCAGTGGGCACAA	TTCCACTCACCTTACCCAGATTA	60	FP	19	435	73
Yc1RG73	AC007739	154145–154426	ATTGCCAAGAACCTTGTGTTTC	GGGGTTGAGAAAAAGTTCAGTG	60	FP	2	463	143
Yc1RG74	AC006038	73850–74014	AACTACCGTAGAATGGGCAAAAT	GAATGGATGGAAAACCAACATAA	60	FP	2	415	226
Yc1RG77	AC005783	19041–19327	GAGAAAGAGCTGCAAGGATGTC	CAAGTAAGGCCCAAATGAGGT	60	FP	19	401	84
Yc1RG78	AC002044	13430–13712	CTCCAGGATCTGCTTTCATCTTA	TCATGGTAACTAGCACAAGATCC	60	FP	16	431	119
Yc1RG79	AC004690	35856–36140	GGGTCTATCATCACCTTAATTTTGA	TGGTTTTTAGATGCCAAACACTAT	60	FP	7	497	158
Yc1RG80	AC004485	74445–74724	CAGAAATTGGTCCTTACAGTTTCC	AGAGGTGAACAGTTATTGCCTGA	60	FP	7	482	134
Yc1RG81	AF088219	1767000–176982	CACACAGCAGCAGTTACAAAAAC	CTTCTAGGCTTTAGTTGGGGAAG	60	FP	17	535	354
Yc1RG82	AF088219	99726–100005	CCTGGACCTTTAGGCATTTTT	CAGTCATCTCATCTTCACAGCAC	60	FP	17	388	91
Yc1RG83	AC005026	82038–82232	GCAGTAATGGTGGCCTGTTATAG	GGAAACTGTTAATGCTTCCCTCT	60	FP	7	389	153
Yc1RG84	AF131217	50031–50317	CCACTTGCCACTCCTATGGTAT	AAAATGCACAGGAATAGCGTTC	60	FP	21	387	60
Yc1RG86	AC005412	78372–78652	ATTGGGTGACCAGTTGTATTGAC	CTTCTGGAGGGGAACTGTTTTTA	60	FP	17	499	188
Yc1RG87	AC008071	84205–84487	GAACATGTGAACACATTGCTAGG	AATGTACCTTCAAAGTCACACAGC	60	FP	7	427	92
Yc1RG88	AC006305	13802–14086	GGTGACTCTCAACCCTAACTTCA	GTGGATTCCCGACAGAAGTATT	60	FP	18	395	74
Yc1RG90	AC004671	68017–68298	CCTTAATAATTTCCCCCGGATT	GCTGTAGGGGCTAAAATACCAAC	60	FP	12	398	100
Yc1RG91	AC005288	37818–38107	AATGGGTGAAAAGAGGTAGAAGG	TGTGTCCTTAACAAAGAGGATGG	60	FP	17	700	391
Yc1RG92	AC004675	78485–78767	ACACTCTATGCAGGCAGTCATCT	CCTGGACCTTTAGGCATTTTT	60	FP	17	402	85
Yc1RG93	AC005324	137294–137574	GGGATTCAAGTGTGGGTAGAAT	AAGGAAGGCAATATGATGTGG	60	LF	17	377	63
Yc1RG95	AL049537	38717–38997	ACCTAACAGATGACCTGCTGAAA	GAGGTAGAGAAAGGCAAGCATTC	60	IF	20	701	390
Yc1RG96	AF042091	61095–61379	ACACACAACCTGAAAACTCAACC	CCACACCAGCATGTTATTTGAT	60	FP	21	457	128

(*continued*)

TABLE 1

(Continued)

Name	Accession	Position	5′ primer sequence (5′–3′)	3′ primer sequence (5′–3′)	A.T[a]	Human diversity[b]	Chr.[c] loc.	Product size[d]	
								Filled	Empty
Yc1RG97	AF042090	42069–42352	AAGTGCACACTTTGACGTTCAC	CCTTGATTGGCATTCAGGTTTA	60	HF	21	441	88
Yc1RG98	U92032	3903–4188	TCTTATCTCTGACACCTGACACG	AAAGAACCCAGAGCTATGACAGA	60	FP	6	442	113
Yc1RG99	AL022163	85835–86116	AAAGCACTTGGTACAGAATCAGC	CCATGGCGAAGTTAATGAGAAGT	60	IF	X	390	64
Yc1RG100	AL354872	86112–86401	ACTTCCATGAGCTAGTGGCTGTA	GATCTCTAAACGATAAAGGCTCAC	60	LF	1	474	143
Yc1RG101	AL031662	26328–26613	CCAGCCAAAGAGATTACCAAAA	GTCCAGTCCATTTCTCAAAGAAG	60	HF	20	541	235
Yc1RG102	AL158040	201136–201376	CTGCCTTTGTAGTAAATGTCAAGG	GTAGACATTCGCTCCACCTTTAT	60	FP	10	414	110
Yc1RG103	AL158157	101226–101505	GGCATTTGCATTCTGATGCTTA	GACATGTTAGAGAAAAGGTGACATC	60	HF	9	383	79
Yc1RG104	AL157384	87495–87786	CTGGAAGGGATCTTTTCTTATGG	CCCTTTTCTGATCCTATTCTCCA	60	FP	9	438	130
Yc1RG107	AL358293	139195–139492	GTTTGATCAGCTGTCCTCAGACT	TGAATGAATTTTTGAGTTGGTGA	60	FP	14	399	76
Yc1RG108	AL035458	29348–29629	GTTATATGAACAAAGCCCGGTA	GACCAAAAGAACCGAGAAGAAAC	60	FP	20	381	71
Yc1RG109	AL137794	36815–37094	GCTAGAATTCATAATGGAACCATCC	TCCAGTTGAGTTGGAGTGATTT	60	FP	1	502	188
Yc1RG110	AL109824	732–1012	CTAGGGTTAAGGAGTCCCTTGG	GTGACCTAGGCCAGAGGTTAATG	60	FP	20	395	85
Yc1RG113	AL163278	90774–91055	CTGTACCGCTAAGAGCTTCTGTG	GATATCTCAGCAGAATGGCAGAC	60	FP	21	376	76
Yc1RG114	Z98051	36444–36724	ATCAGGCATACAGTCTGAAAAGC	AATCTTGGTTAGTGTGAGTCAACC	60	FP	X	426	110
Yc1RG115	Z98046	60991–61271	GTTCTGCTGTTTGGATCTGGAAT	GTGGTGAAGGTACAGACTCATCC	60	FP	X	392	72
Yc1RG116	AL078621	142330–142621	GGTTAAAAGAACACATGGGATGG	GAAAGGTGGGTGTCTAAATGCTA	60	FP	22	419	99
Yc1RG117	AL096861	42260–42540	GAATAACCCAAACTTGGTAGGTG	TGCAATAAAGAGTGTTCCTCTCC	60	FP	X	490	166
Yc1RG118	Z71183	21436–21716	TACACAGACCAATGGGGAAAGTA	TCCAGATCCATGACATAACACT	60	FP	22	389	89
Yc1RG120	AL023283	61027–61306	TCTGCTCTTGCTATACACTGCTG	GAAGGCAGTGAATGAGACACTCT	60	FP	6	499	194
Yc1RG121	AL109760	24171–24451	CATGGACATTTGGGAGAATGTA	CGCCCCTATAATTACTCAGCAG	60	FP	4	398	92
Yc1RG123	AL023882	16690–16970	CACACACACACACAAAATTAGCC	GTGAGTCTTGAAACGGCTTTTAC	60	LF	16	563	234
Yc1RG124	AL022397	18401–18681	AAATCACTGTACCAACCCTGTCA	GCAAACACCACTGAAGCATAAA	60	FP	1	397	79
Yc1RG125	X76070	298–578	TGTTCTCTTCCTGTCTCCATTTC	CTGTTTCTATGATCTTGAAGGATGG	60	IF	2	415	97
Yc1RG126	AP001752	250076–250356	CCCTGTAGTAATGGCTCAGTGAA	GGCGATTTAGGCATAGACATAGA	60	FP	21	415	91
Yc2NBC1	AC002430	108794–109074	ACATAGTGGGCATTCAAGAG	CTTAATGTTTCCATTTCTCCA	55	IF	7	467	131
Yc2NBC5	AC007384	128277–128557	GAAGGAATACAGGGAGGAAT	CTCCCAAACAACTTAAAACC	55	IF	7	461	125
Yc2NBC9	Z98051	36444–36724	GAAAAGCCTGATACTTTTGG	CTTGGTTAGTGTGAGTCAACC	55	FP	X	407	91
Yc2NBC11	Z69666	9696–9416	CGACAAGTGACTAACCTTACG	CTCCTCCAATGATCTATGTGT	55	FP	16	409	82
Yc2NBC13	AC007882	150095–149815	TGGGATAATGATTTGTCTCC	AACATGTGGCAGATGATGA	60	FP	16	407	89
Yc2NBC15	AC007541	129217–129497	GGTAAGGCAAAACCAAGTAA	GTTTTGAGGAAGCTGATGAC	55	FP	12	410	92
Yc2NBC17	AC005541	74313–74593	ATCAAATGGCAGCCTTACT	GGTTTTCCATTCCTGAGTTA	60	FP	7	401	82
Yc2NBC19	AL022163	81833–82113	GCTTAAAGCACTTGGTACAGA	TGGCGAAGTTAATGAGAAGT	55	HF	X	393	67

Perfect matches to the consensus are in italics.

[a] Amplification of each locus required 2 hr 30 min at 94° initial denaturing and 32 cycles for 1 min 94°, 1 annealing temperature (A.T.), and 1 min elongation at 72°. A final extension time of 10 min at 72° was also used.

[b] Allele frequency was classified as fixed present (FP), low (LF), intermediate (IF), or high frequency (HF) 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. —, indeterminable.

[c] Chromosomal location determined from accession information or by PCR analysis of NIGMS monochromosomal hybrid cell line DNA samples.

[d] Empty product sizes calculated by removing the Alu element and one direct repeat from the filled sites that were identified.

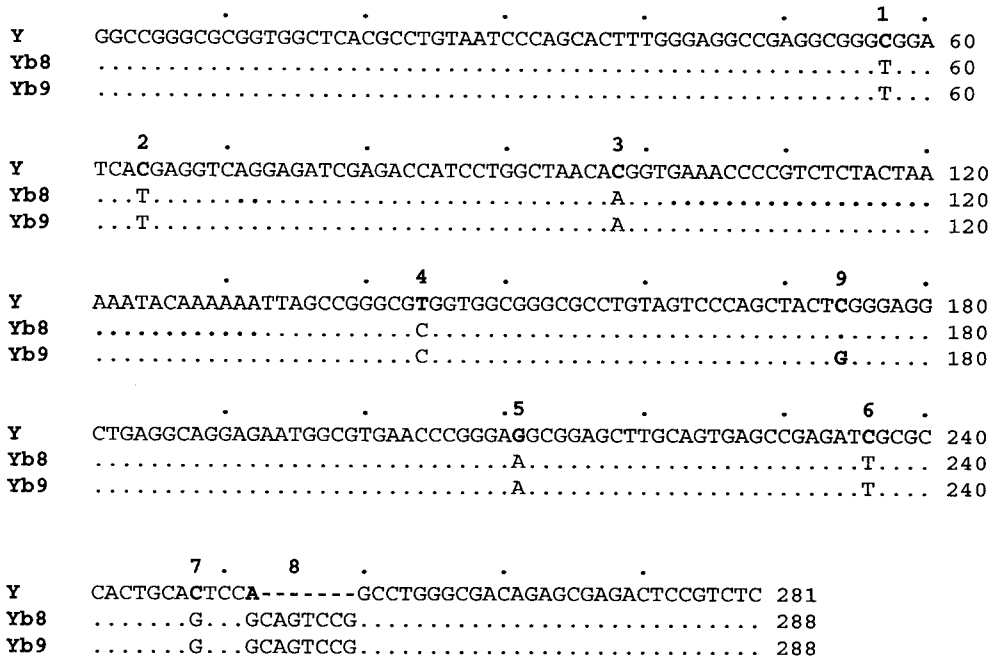

FIGURE 1.—Consensus sequence alignment of Y, Yb8, and the potential new subfamily Yb9 identified. Nucleotide substitutions at each position are indicated with the appropriate nucleotide. Deletions are marked by dashes (-). The Yb8 and Yb9 diagnostic nucleotides are indicated in boldface type with the corresponding diagnostic numbers above.

Yc1 and Yc2 Alu elements, because only subfamily members with perfect identity to the subfamily consensus sequence or one mismatch were isolated from the database using one of the database screening procedures.

Phylogenetic distribution and human genomic diversity of the new subfamilies: Amplification of the Yb9, Yc1, and Yc2 elements from nonhuman primate genomes facilitated the analysis of the phylogenetic distribution of these elements, using PCR and the oligonucleotide primers in Table 1. Almost all of the elements evaluated were absent from the genomes of the nonhuman primates, suggesting that these elements dispersed and were fixed in the human genome after the human and African ape divergence.

We performed a PCR analysis on a panel of human DNA samples to determine the levels of human diversity associated with the Alu elements from these new subfamilies, using the oligonucleotide primers shown in Table 1. The panel consists of 20 individuals of European origin, African-Americans, Asians, and Egyptians for a total of 80 individuals (160 chromosomes). We were able to analyze 28 out of the 56 Yb9 elements, 97 out of 176 Yc1 elements, and 8 out of 17 Yc2 Alu elements, using this approach. Several factors did not allow for analysis of all the elements. Mainly, we were unable to design appropriate primers due to insufficient flanking unique DNA sequences or because the element analyzed resided within another type of repeat as described previously (BATZER *et al.* 1991). The Alu elements were classified as fixed present and high, intermediate, or low frequency insertion polymorphisms (see Table 1 for definitions). In general, we observed that approximately one-fourth to one-third of the elements analyzed had some degree of insertion polymorphism (Yb9 with 10/28, Yc1 with 24/97, and Yc2 with 3/8). The population-specific genotypes and levels of heterozygosity for each element are shown in Table 2. The high proportion of polymorphic elements in these Alu subfamilies is in good agreement with our previous observations, indicating that these subfamilies are very recent in origin and still actively retroposing within the human genome.

DISCUSSION

From our subset of AluYb8 and Y elements, we were able to retrieve three Alu subfamilies termed Yb9, Yc1, and Yc2. A schematic of the evolutionary relationship of these subfamilies with the previously defined Alu subfamilies is shown in Figure 3. Alu subfamilies arise as a result of mutations occurring in an existing master element or new source elements capable of significant amplification. In this case, the new subfamilies are presumably examples of Alu subfamilies that may have originated from the rare instances when an Alu element fortuitously becomes both transcriptionally and retropositionally active, therefore allowing it to be another Alu source gene.

The young Alu subfamilies are currently active with respect to retroposition, whereas the older Alu subfamilies typically are not. The old Alu subfamilies (Sx, J, and Sg1), which comprise the vast majority (>1,000,000 copies) of the Alu elements present in the human genome, appear completely inactive as none of their members have been associated with *de novo* Alu inserts that result in human diseases (Table 3). When noting the ratio of reported Alu insertions associated with diseases and the estimated size of the Alu subfamily, the younger

```
                .         .         .         .         .         .
Y       GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGA 60
Yc1     ............................................................ 60
Yc2     ............................................................ 60
C1inh   ............................................................ 60
BRCA2   ............................................................ 60
GK      ----------------------------------------.............T... 60

                .         .         .      2  .         .         .
Y       TCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCTACTAA 120
Yc1     ..........................................................  120
Yc2     ..........................................A.................  120
C1inh   ..........................................T.................  120
BRCA2   ........................T.................................  120
GK      ..........................................................  120

                .         .         .  1      .         .         .
Y       AAA-----TACAAAAAATTAGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCG 175
Yc1     ...-----.........................A.........................  175
Yc2     ...-----.........................A.........................  175
C1inh   ...AAAAA.........................A.........................  180
BRCA2   ...-----.........................A.........................  175
GK      ...-----.........................A.........................  175

                .         .         .         .         .         .
Y       GGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGAT 235
Yc1     ..........................................................  235
Yc2     ..........................................................  235
C1inh   ..............G...........................CG..............  240
BRCA2   ..........................................................  235
GK      ..............G...........................................  235

                .         .         .         .
Y       CGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTC 281
Yc1     ..............................................  281
Yc2     ..............................................  281
C1inh   ..............................................  286
BRCA2   ..............................................  281
GK      ..............................................  281
```

FIGURE 2.—Consensus sequence alignment of Y, Yc1, Yc2, and three Alu Yc1 elements associated with disease. The diseases linked withYc1 Alu elements are the angioedema caused by a *de novo* insertion in the C1 inhibitor gene (C1inh; STOPPA-LYONNET *et al.* 1990), breast cancer with another *de novo* insertion in BRCA2 (BRCA2; MIKI *et al.* 1996), and glycerol kinase deficiency (GK; ZHANG *et al.* 2000). Nucleotide substitutions at each position are indicated with the appropriate nucleotide. Deletions are marked by dashes (-). The diagnostic nucleotides are indicated in boldface type with the corresponding diagnostic numbers above.

subfamilies Ya5, Yb8, and Yc1 currently appear to be ~1000 times more active than the Alu Y subfamily with 7/2640, 3/1852, and 3/400 compared to 1/200,000 (Table 3). The Alu Ya5a2 subfamily appears to have even a higher current retroposition rate (1/40), but the very young age and small size of the subfamily may be an influencing factor. In general, two independent observations support the current mobility of these young Alu subfamilies within the human genome. First, there are examples of Alu inserts that have caused disease that belong to these young subfamilies. Second, the subfamilies have a high proportion of Alu insertion polymorphisms between individuals/populations (Table 3), indicating the recent proliferative/amplification activity of these Alu elements in the human genome.

Alu elements that are polymorphic for insertion presence/absence have previously proven useful for the study of human population genetics and forensics (BATZER *et al.* 1991, 1994; PERNA *et al.* 1992; NOVICK *et al.* 1993; HAMMER 1994; TISHKOFF *et al.* 1996; STONEKING *et al.* 1997; MAJUMDER *et al.* 1999; COMAS *et al.* 2000; JORDE *et al.* 2000; WATKINS *et al.* 2001). The identification of very young Alu subfamilies with a high proportion of polymorphic members provides new sources of Alu insertion polymorphisms for the study of human population genetics. However, it is important to note that an exhaustive analysis of these small subfamilies will only generate a relatively small number of new Alu insertion polymorphisms.

Master element *vs.* source gene: Alu elements have been proposed to fit an evolutionary model where the copies arose from "master" genes (DEININGER and SLAGEL 1988; LABUDA and STRIKER 1989; SHEN *et al.* 1991; DEININGER *et al.* 1992). A master gene can be defined as an element that is highly active during a long period, therefore generating a lot of copies of itself. However, we demonstrated that recently inserted Alu elements (*de novo*) belong to a variety of Alu subfamilies, indicating the simultaneous presence of multiple active elements in the human genome. These active elements that have a low rate of amplification and are only active for a very short period of time should not be classified as master genes. To distinguish between them, we suggest the use of the nomenclature of "master gene" when

TABLE 2

Alu Yb9, Yc1, and Yc2 associated human genomic diversity

	African American					Asian/Alaska native					European					Egyptian					
	Genotypes					Genotypes					Genotypes					Genotypes					
Elements	+/+	+/−	−/−	*f*Alu	Het[a]	+/+	+/−	−/−	*f*Alu	Het[a]	+/+	+/−	−/−	*f*Alu	Het[a]	+/+	+/−	−/−	*f*Alu	Het[a]	Avg het[b]
Yb9NBC8	0	0	20	0.000	0.000	0	0	19	0.000	0.000	0	0	17	0.000	0.000	0	0	12	0.000	0.000	0.000
Yb9NBC7	19	0	0	1.000	0.000	19	0	0	1.000	0.000	17	3	0	0.925	0.142	16	0	0	1.000	0.000	—
Yb9NBC10	3	1	4	0.438	0.525	2	0	14	0.125	0.226	3	0	14	0.176	0.299	6	0	9	0.400	0.497	0.369
Yb9NBC12	1	6	12	0.211	0.341	0	14	5	0.368	0.478	0	9	8	0.265	0.401	0	9	5	0.321	0.452	0.418
Yb9NBC22	0	0	14	0.000	0.000	0	0	15	0.000	0.000	0	0	15	0.000	0.000	0	0	13	0.000	0.000	0.000
Yb9NBC27	0	0	15	0.000	0.000	0	0	12	0.000	0.000	0	0	18	0.000	0.000	0	0	11	0.000	0.000	0.000
Yb9NBC29	0	1	9	0.050	0.100	0	7	12	0.184	0.309	0	2	8	0.100	0.189	0	0	3	0.000	0.000	0.199
Yb9NBC30	2	1	11	0.179	0.304	0	3	16	0.079	0.149	0	6	11	0.176	0.299	1	3	14	0.019	0.246	0.250
Yb9NBC50	0	0	15	0.000	0.000	0	6	7	0.231	0.369	1	0	15	0.063	0.121	1	3	14	0.139	0.246	0.184
Yb9NBC53	13	0	2	0.867	0.239	20	0	0	1.000	0.000	15	0	1	0.938	0.121	15	0	2	0.882	0.214	0.144
Yc1NBC1	1	7	12	0.225	0.073	0	2	18	0.050	0.062	0	10	10	0.250	0.068	0	7	13	0.175	0.078	0.070
Yc1NBC2	1	13	6	0.375	0.038	0	15	5	0.375	0.038	1	15	4	0.425	0.023	0	10	10	0.250	0.068	0.042
Yc1NBC9	4	13	3	0.525	0.008	3	13	4	0.475	0.008	3	8	9	0.350	0.045	0	0	14	0.000	0.000	0.015
Yc1NBC23	0	0	18	0.000	0.000	0	0	19	0.000	0.000	0	0	18	0.000	0.000	0	0	19	0.000	0.000	0.000
Yc1NBC31	0	0	18	0.000	0.000	0	0	19	0.000	0.000	0	0	19	0.000	0.000	0	0	19	0.000	0.000	0.000
Yc1NBC35	1	6	7	0.286	0.073	2	10	8	0.350	0.045	2	13	2	0.500	0.000	1	12	2	0.467	0.012	0.032
Yc1NBC50	0	2	18	0.050	0.062	14	4	0	0.889	0.081	4	9	5	0.472	0.009	5	2	10	0.353	0.048	0.050
Yc1NBC51	0	4	18	0.091	0.169	0	0	18	0.000	0.000	0	0	20	0.000	0.000	0	0	9	0.000	0.000	—
Yc1NBC53	8	7	1	0.719	0.070	3	12	1	0.563	0.022	1	13	2	0.469	0.011	4	11	2	0.559	0.020	0.031
Yc1NBC60	6	9	3	0.583	0.027	6	9	5	0.525	0.008	5	11	4	0.252	0.008	2	7	10	0.289	0.062	0.026
Yc1NBC63	0	0	0	—	—	1	5	8	0.250	0.082	3	6	10	0.316	0.056	0	3	10	0.115	0.096	0.078
Yc1NBC64	0	0	5	0.000	0.000	0	5	8	0.192	0.323	0	5	12	0.147	0.258	0	6	12	0.167	0.286	0.216
Yc1NBC69	0	0	13	0.000	0.000	8	4	5	0.588	0.030	2	4	7	0.308	0.070	3	7	5	0.433	0.024	0.031
Yc1RG60	16	0	4	0.800	0.328	19	0	0	1.000	0.000	14	5	1	0.825	0.296	18	0	0	1.000	0.000	0.156
Yc1RG68	1	4	14	0.158	0.273	6	6	8	0.450	0.508	3	7	10	0.325	0.450	3	3	14	0.225	0.358	0.397
Yc1RG93	0	0	20	0.000	0.000	0	0	20	0.000	0.000	0	0	19	0.000	0.000	0	0	14	0.000	0.000	0.000
Yc1RG95	2	17	1	0.525	0.512	4	15	0	0.605	0.491	0	20	0	0.500	0.513	6	12	0	0.67	0.457	0.493
Yc1RG97	19	1	0	0.975	0.050	19	0	0	1.000	0.000	20	0	0	1.000	0.000	18	0	0	1.000	0.000	0.013
Yc1RG99	19	1	0	0.975	0.050	6	14	0	0.650	0.467	8	11	1	0.675	0.450	14	4	1	0.842	0.273	0.310
Yc1RG100	0	0	18	0.000	0.000	0	0	19	0.000	0.000	0	0	18	0.000	0.000	0	0	16	0.000	0.000	0.000

(*continued*)

TABLE 2

(Continued)

Elements	African American					Asian/Alaska native					European					Egyptian					Avg het[b]
	Genotypes			*f*Alu	Het[a]	Genotypes			*f*Alu	Het[a]	Genotypes			*f*Alu	Het[a]	Genotypes			*f*Alu	Het[a]	
	+/+	+/−	−/−			+/+	+/−	−/−			+/+	+/−	−/−			+/+	+/−	−/−			
Yc1RG101	20	0	0	1.000	0.000	17	0	0	1.000	0.000	17	2	0	0.947	0.102	16	0	0	1.000	0.000	0.026
Yc1RG103	16	2	2	0.850	0.262	19	0	0	1.000	0.000	18	0	0	1.000	0.000	15	0	0	1.000	0.000	0.065
Yc1RG123	0	0	20	0.000	0.000	0	0	20	0.000	0.000	0	0	20	0.000	0.000	0	0	20	0.000	0.000	0.000
Yc1RG125	0	16	4	0.400	0.492	0	9	11	0.225	0.358	0	17	3	0.425	0.501	0	19	0	0.500	0.514	0.466
Yc2NBC1	1	4	3	0.375	0.061	3	6	5	0.429	0.027	13	1	0	0.964	0.065	10	3	1	0.821	0.093	0.061
Yc2NBC5	3	10	4	0.471	0.010	3	10	1	0.400	0.031	17	3	0	0.925	0.071	13	4	0	0.882	0.085	0.049
Yc2NBC19	15	3	0	0.917	0.077	18	0	0	1.000	0.000	7	10	3	0.600	0.03	14	4	1	0.842	0.081	0.047

[a] This is the unbiased heterozygosity.
[b] Average heterozygosity is the average of the population heterozygosity.
Elements in italics were screened using DNA collected from Alaska natives rather than from the Asian population.

FIGURE 3.—Schematic diagram of the evolution of recently integrated Alu subfamilies. All the origins of the young Alu subfamilies are shown. The origins of the Yb9, Yc1, and Yc2 Alu subfamilies are shown after the divergence of the Yb8 and the Y subfamily, respectively. The size of the font is relative to the number of elements within each subfamily, the largest representing 100,000–200,000 copies; medium, 1000–2000 copies; and the smallest, 50–500 copies. The total number of elements from each subfamily linked to disease is indicated to the right. The proportion of polymorphic elements within each family is represented by the following: ±, rarely polymorphic elements are found; +, low percentage of polymorphic elements; ++, ~50% the elements are polymorphic; and +++, most of the elements are polymorphic.

referring to the highly active genes for long evolutionary periods of time, like the Alu element that generated the majority (>90%) of the Alu elements currently present in the genome today. For those copies, or daughters, that acquired the ability to retropose we propose the use of the term "source genes." However, some of the elements classified as source genes may be potential master genes, and only the progression of time will allow the appropriate distinction to be made.

Evolutionary reduction in the Alu retroposition rate: Our data indicate the existence of several currently active Alu elements that belong to different subfamilies within the human genome. However, the present amplification rate of Alu elements has drastically decreased from when it reached its peak between 35 and 60 million years ago (mostly Sx subfamily). The majority of the Alu elements present in the genome of extant humans inserted during this peak amplification period. There are multiple reasons that could explain the reduction in the amplification rate of Alu elements. First, mutations within or near the master Alu element could reduce its retroposition activity or even totally abolish it by a variety of mechanisms (DEININGER and BATZER 1993; SCHMID 1996). Alternatively, mutations within the master gene or in the LINE elements that affect the ability to "parasitize" LINE element-encoded enzymes necessary for retroposition could also reduce the Alu amplification rate. Furthermore, the host may have also evolved cellular

TABLE 3

Young Alu subfamilies copy number, inserts linked to disease, and polymorphism

Alu subfamily	Estimated copy number	Inserted linked with disease[a]	General subfamily polymorphism[b] (%)	
J, Sx, Sg1	>1,000,000	0	−	
Y	>200,000	1	±	
Ya5	2640	7	+	26
Ya5a2	40	1	+++	80[c]
Ya8	70	0	++	50
Yb8	1852	3	+	20
Yb9	80	0	+	36
Yc1	400	3	+	25[c]
Yc2	ND	0	+	37.5[c]

ND, not determined.

[a] Previously published Alu elements linked with disease (Deininger and Batzer 1999).

[b] The proportion of polymorphic elements within each family is represented by the following: ±, rarely polymorphic elements are found; +, low percentage of polymorphic elements; ++, ~50% the elements are polymorphic; and +++, most of the elements are polymorphic.

[c] Percentage polymorphism was determined using a selected subgroup introducing a bias.

mechanisms to reduce Alu proliferation. Finally, the availability of suitable genomic "insertion sites" may be reduced, since most evolutionarily neutral or positive sites are presumably already "filled" with different types of preexisting repeats. Alternatively, new Alu insertions may result in unacceptable local levels of unequal homologous recombination (Deininger and Batzer 1999).

AMR was supported by a Brown Foundation fellowship from the Tulane Cancer Center. This research was supported by National Institutes of Health RO1 GM45668 (P.L.D.); Department of the Army DAMD17-98-1-8119 to (P.L.D. and M.A.B.); Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000-05)-05 and HEF (2000-05)-01 (M.A.B. and P.L.D.); and award 1999-IJ-CX-K009 from the Office of Justice Programs, National Institute of Justice, Department of Justice (M.A.B.). Points of view in this document are those of the authors and do not necessarily represent the official position of the U.S. Department of Justice.

LITERATURE CITED

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

Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G., Seidman *et al.*, 1996 *Current Protocols In Molecular Biology*. John Wiley & Sons, New York.

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

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

Batzer, M. A., M. Stoneking, M. Alegria-Hartman, H. Bazan, D. H. Kass *et al.*, 1994 African origin of human-specific polymorphic Alu insertions. Proc. Natl. Acad. Sci. USA **91:** 12288–12292.

Batzer, M. A., C. M. Rubin, U. Hellmann-Blumberg, M. Alegria-Hartman, E. P. Leeflang *et al.*, 1995 Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. J. Mol. Biol. **247:** 418–427.

Batzer, M. A., P. L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda *et al.*, 1996 Standardized nomenclature for *Alu* repeats. J. Mol. Evol. **42:** 3–6.

Comas, D., F. Calafell, N. Benchemsi, A. Helal, G. Lefranc *et al.*, 2000 Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. Hum. Genet. **107:** 312–319.

Deininger, P. L., and M. A. Batzer, 1993 Evolution of retroposons, pp. 157–196 in *Evolutionary Biology*, edited by M. K. Heckht, R. J. MacIntyre and M. T. Clegg. Plenum Publishing, New York.

Deininger, P. L., and M. A. Batzer, 1999 Alu repeats and human disease. Mol. Genet. Metab. **67:** 183–193.

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

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

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

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

Jurka, J., 1995 Origin and evolution of Alu repetitive elements, pp. 25–42 in *The Impact of Short Interspersed Elements (SINEs) on the Host Genome*, edited by R. J. Maraia. R. G. Landes Company, Austin, Texas.

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

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860–921.

Majumder, P. P., B. Roy, S. Banerjee, M. Chakraborty, B. Dey *et al.*, 1999 Human-specific insertion/deletion polymorphisms in Indian populations and their possible evolutionary implications. Eur. J. Hum. Genet. **7:** 435–446.

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

Miyamoto, M. M., J. L. Slightom and M. Goodman, 1987 Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. Science **238:** 369–373.

Novick, G. E., T. Gonzalez, J. Garrison, C. C. Novick, M. A. Batzer *et al.*, 1993 The use of polymorphic Alu insertions in human DNA fingerprinting. Exper. Suppl. **67:** 283–291.

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

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

Roy, A. M., M. L. Carroll, S. V. Nguyen, A.-H. Salem, M. Oldridge *et al.*, 2000 Potential gene conversion and source gene(s) for recently integrated Alu elements. Genome Res. **10:** 1485–1495.

Schmid, C. W., 1996 Alu: structure, origin, evolution, significance and function of one-tenth of human DNA. Prog. Nucleic Acid Res. Mol. Biol. **53:** 283–319.

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

Smit, A. F., 1996 The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. **6:** 743–748.

Stoneking, M., J. J. Fontius, S. L. Clifford, H. Soodyall, S. S. Arcot *et al.*, 1997 Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. Genome Res. **7:** 1061–1071.

Stoppa-Lyonnet, D., P. E. Carter, T. Meo and M. Tosi, 1990 Clusters of intragenic Alu repeats predispose the human C1 inhibitor locus to deleterious rearrangements. Proc. Natl. Acad. Sci. USA **87:** 1551–1555.

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

Watkins, W. S., C. E. Ricker, M. J. Bamshad, M. L. Carroll, S. V. Nguyen *et al.*, 2001 Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. Am. J. Hum. Genet. **68:** 738–752.

Zhang, Y., K. M. Dipple, E. Vilain, B. L. Huang, G. Finlayson *et al.*, 2000 AluY insertion (IVS4–52ins316alu) in the glycerol kinase gene from an individual with benign glycerol kinase deficiency. Hum. Mutat. **15:** 316–323.

Communicating editor: Y.-X. Fu