doi:10.1016/S0022-2836(03)00032-9 *J. Mol. Biol.* (2003) **326**, 1127–1146

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

LINE-1 preTa Elements in the Human Genome

Abdel-Halim Salem[1,2]†, Jeremy S. Myers[1]†, Anthony C. Otieno[1]† W. Scott Watkins[3], Lynn B. Jorde[3] and Mark A. Batzer[1]*

[1]*Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA*

[2]*Department of Anatomy Faculty of Medicine, Suez Canal University, Ismailia Egypt*

[3]*Department of Human Genetics, University of Utah Health Sciences Center, Salt Lake City, UT 84112, USA*

**Corresponding author*

The preTa subfamily of long interspersed elements (LINEs) is characterized by a three base-pair "ACG" sequence in the 3′ untranslated region, contains approximately 400 members in the human genome, and has low level of nucleotide divergence with an estimated average age of 2.34 million years old suggesting that expansion of the L1 preTa subfamily occurred just after the divergence of humans and African apes. We have identified 362 preTa L1 elements from the draft human genomic sequence, investigated the genomic characteristics of preTa L1 insertions, and screened individual elements across diverse human populations and various non-human primate species using polymerase chain reaction (PCR) assays to determine the phylogenetic origin and levels of human genomic diversity associated with the L1 elements. All of the preTa L1 elements analyzed by PCR were absent from the orthologous positions in non-human primate genomes with 33 (14%) of the L1 elements being polymorphic with respect to insertion presence or absence in the human genome. The newly identified L1 insertion polymorphisms will prove useful as identical by descent genetic markers for the study of human population genetics. We provide evidence that preTa L1 elements show an integration site preference for genomic regions with low GC content. Computational analysis of the preTa L1 elements revealed that 29% of the elements amenable to complete sequence analysis have apparently escaped 5′ truncation and are essentially full-length (approximately 6 kb). In all, 29 have two intact open reading frames and may be capable of retrotransposition.

Keywords: retrotransposon; insertion polymorphism; mobile elements

Introduction

Computational analysis of the draft sequence of the human genome indicates that repetitive sequences comprise 45–50% of the human genome mass, 17% of which consists of long interspersed elements (LINE-1s or L1s).[1–3] L1 elements are restricted to mammals, having expanded as a repeated DNA sequence family over the last 150 million years.[4] Full-length L1 elements are approximately 6 kb long and propagate *via* an RNA intermediate in a process known as retrotransposition. L1 retrotransposition likely occurs by a mechanism termed target primed reverse transcription (TPRT).[5] This mechanism of mobilization provides two useful landmarks for the identification of young L1 inserts: an endonuclease related cleavage site[6–8] and direct repeats or target site duplications flanking newly integrated elements.[9]

L1 retrotransposons have had a significant impact on the human genome through a variety of different mechanisms. *De novo* insertions disrupting open reading frames and splice sites have resulted in a number of human diseases,[10–12] new L1 integrations have been shown to have the potential to alter gene expression,[13,14] and once in the genome L1 elements provide regions of sequence identity blanketing the genome, that can be exploited during recombination.[15] L1 elements also generate sequence duplications by transducing adjacent genomic sequences at their 3′ end, thereby "shuffling" genomic sequence.[16–18] More recently, it has been suggested that L1 elements have paradoxical roles in genomic stability by serving both as molecular band aids, repairing double-stranded breaks in mammalian cells and

† These authors contributed equally to this work.

Abbreviations used: LINE, long interspersed elements; L1, LINE-1; Ta, transcribed, subset a.

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

as suspects for the generation of genomic deletions.[19–21] Thus, L1 elements exert a significant influence on the architecture of the human genome and provide dynamic units capable of ongoing change.

As a result of the limited amplification potential of the diverse L1 gene family, a series of discrete L1 subfamilies exist within the human genome.[4,22] L1 elements have expanded at different times during mammalian evolution, producing subfamilies of various ages.[4,22] Depending on the amplification period of the L1 subfamily, some L1 elements may be unique to a single lineage, species, or even a single population. Such is the case with the L1Hs (human specific) Ta (transcribed, subset a)[23] subfamily, which has been shown to be present only in the human lineage.[24]

Even though there are approximately 500,000 L1 elements in the human genome only a limited subset of 30–60 L1 elements appear to be capable of retrotransposition.[25,26] *De novo* L1 insertions resulting in human disease are largely a product of L1Hs Ta integrations, which have been shown to be the youngest most active L1 subfamily found in the human genome.[24,27,28] However, at least one L1 insert (JH-28) in exon 14 of the factor VIII gene resulting in hemophilia A, was the result of a preTa insertion, providing the first proof that preTa L1 elements are also currently capable of retrotransposition.[12] Previous studies have shown that some members of the preTa L1 subfamily have inserted so recently in the human genome that they are polymorphic with respect to insertion presence/absence,[27,29] all of which makes preTa L1 elements a likely source of identical-by-descent mobile element-based variation for the study of human population genetics.

Members of the L1 preTa subfamily share a common three base-pair diagnostic sequence within the 3′ untranslated region (UTR), which separates them from the other L1 subfamilies. As the name suggests, the preTa L1 subfamily is believed to predate the amplification of the L1Hs Ta subfamily in the human lineage. However, the phylogenetic origin and level of human genetic diversity associated with preTa L1 elements remains largely undefined. Here, we report a comprehensive analysis of the preTa L1 subfamily from the draft sequence of the human genome.

Results

L1 preTa subfamily copy number

To identify recently integrated preTa subfamily L1 elements from the human genome, we searched the draft sequence of the human genome (database version: BLASTN 2.2.1 (Apr-13-2001)) using the Basic Local Alignment Search Tool (BLAST)[30] with an oligonucleotide sequence that is complementary to a highly conserved motif in the 3′ untranslated region (UTR) of preTa L1 elements. This 19 base-pair (bp) query sequence (CCTAATGCTAGATGA CACG) includes the preTa subfamily-specific diagnostic mutation "ACG" at its 3′ end (position 5930–5932 relative to LRE-1).[31] We identified 362 unique preTa L1 elements from 2.868×10^9 bp of available human draft sequence. Extrapolating this number to the actual size of the human genome (3.162×10^9 bp), we estimate that this subfamily contains about 400 elements. Taken with the estimate from the L1Hs Ta data,[24] we estimate that there are over 900 human specific LINE-1 elements in the human genome. Of the 362 preTa L1 elements retrieved, six resided at the end of sequence contigs and were not amenable to additional analysis. Of the 356 (362 − 6) remaining elements, 105 (26%) were essentially full length, and 251 were truncated to variable lengths. Alignment and sequence analysis of the full-length elements revealed that 29 contained two intact open reading frames and therefore may be capable of retrotransposition. The complete data set is available on our web site† (under publications).

Estimated subfamily age

The average ages of L1 elements can be determined by the level of sequence divergence from the subfamily consensus sequence using a neutral mutation rate for primate non-coding sequence of 0.15% per million years.[32] The mutation rate is known to be about ten times greater for CpG bases as compared to non-CpG bases as a result of the spontaneous deamination of 5-methyl cytosine.[33] Thus, two age estimates based upon CpG and non-CpG mutations can be calculated for the preTa subfamily of L1 elements. A total of 74,048 bases from the 3′ UTR of 356 preTa L1 elements were analyzed. In all, 361 nucleotide substitutions were observed. Of these, 303 were classified as non-CpG mutations against the backdrop of 71,912 total non-CpG bases, producing a non-CpG mutation density of 0.004213 (303/71,912). Based upon the non-CpG mutation density and a neutral rate of evolution (0.004213/0.0015), the average age of the L1 preTa LINE-1 elements was 2.81 million years old. A total of 58 CpG mutations out of 2136 total CpG nucleotides were found across the same 356 LINE elements, yielding a CpG-based mutation density of 0.027154 (58/2,136). With the expectation that the CpG mutation rate is about tenfold higher than the non-CpG mutation rate, the approximate age of the L1 preTa subfamily using the CpG mutation density is 1.86 million years old. These estimates are in good agreement with one another and taken together, these estimates produce an average age of 2.34 million years old, which is in good agreement with the idea that the preTa L1 subfamily is evolutionarily older than the L1Hs Ta subfamily (estimated average age 1.99 million years).[24,27] In

† http://batzerlab.lsu.edu

addition the average age estimates reported here provide a relative time frame by which to compare L1 retrotransposition activity, and should not be confused with the age of origin.

Similar to the L1Hs Ta subfamily, the preTa L1 subfamily can also be grouped into two subgroups, ACG/A and ACG/G, based on an "A" or "G" base at position 6015 relative to L1.2 (accession number M80343). In order to determine the relative ages of each subgroup, we analyzed the level of sequence divergence in each subgroup. The ACG/A subgroup contained 127 total nucleotide substitutions, with 98 of these classified as non-CpG mutations against the backdrop of 20,402 total non-CpG bases. This yields a non-CpG mutation density of 0.004803 (98/20,402) and produces an estimated age of 3.20 million years old. Of 127 total mutations, 29 were classified as CpG mutations against a backdrop of 606 CpG total bases, which yields a CpG mutation density of 0.047855 (29/606) producing an estimated age of 3.28 million years. The ACG/G subgroup contained 221 total nucleotide substitutions with 191 of these classified as non-CpG mutations against the backdrop of 51,106 total non-CpG bases, which yields a non-CpG mutation density of 0.003737 (191/51,106), producing an estimated age of 2.49 million years old. Of 121 total mutations, 30 were classified as CpG mutations against a backdrop of 1518 CpG total bases, which yields a CpG mutation density of 0.019763 (30/1518) producing an estimated age of 1.35 million years. We calculated the average age of each subgroup as 1.92 and 3.24 million years for the ACG/G and ACG/A, respectively. Although it is likely that the L1Hs Ta subfamily is derived from one of the preTa L1 subsets based on the estimated ages of these L1 subfamilies, the transition intermediates between preTa and Ta subfamilies are not clear.

Features of L1 preTa integration sites

One hallmark of L1 integration is the generation of target site duplications flanking newly integrated elements. Two thousand bases of flanking sequence on each side of the element were searched for target site duplications. Clear target site duplications are considered to be target site duplications at least ten bases in length. Of the 356 elements analyzed, we were able to identify clear target site duplications for 252 elements. We then determined the integration sites for these 252 preTa L1 insertions with clear target duplications. A complete list of L1 integration sites is shown in Table 1, and further supports the notion that some integration sites are more common than others.[6,7,34]

A large number of preTa L1 elements had no observable target duplication sites. One possible explanation for this observation is that these elements have relatively short target site duplications. Alternatively, these elements may represent forward gene conversion events of older pre-existing L1 elements that by mutation, have rendered their target site duplications unrecognizable. Some of these events may also represent integrations that have occurred independent of endonuclease cleavage, that has previously been proposed as a mechanism for the repair of doubled-stranded breaks in DNA.[35–37]

Table 1. PreTa L1 integration sites

PreTa L1 integration sites	Number
TTTT/A	60
TCTT/A	37
CTTT/A	20
TTTA/A	18
TTTC/A	18
TTTT/G	16
TTCT/A	14
TCTT/G	7
CTTT/G	5
ATTT/A	5
CTTT/C	5
TTTT/C	4
TGTT/A	3
TATT/A	3
TATT/G	3
TCTT/C	2
TTTC/C	2
TCTC/A	2
GTTT/A	2
ATTT/C	2
GCTT/T,TTTT/T,TTTG/A,TTTC/T,TTTC/T,TTGT/G, TTAT/A,TGAT/G,TCTT/T,TCAT/A,TATC/A,TATA/T, TAAA/C,GCTT/A,CCTT/A,CATT/G,CATT/A,ACTT/G, ACTT/A,ACTA/C,ACCT/A,ACAC/T,ACAA/A,AAAA/A	1 each

To further characterize the preTa L1 insertions, we determined the DNA base content for sequence blocks 1 kb and 2 kb flanking all preTa L1 insertion sites with target site duplications of at least 10 bp. Flanking sequence was then grouped according to GC content with only data for the 1 kb sequence blocks shown in Figure 1. Our data suggest that preTa L1 elements integrate preferentially in genomic regions with GC content less than 36%, but are present in genomic regions with GC content as low as 26% and as high as 52%. A similar insertion site preference was observed for 2 kb sequence blocks as well as for the previously reported L1 Ta subfamily[24] and other L1 subfamilies.[38] In addition, we also analyzed preTa L1 elements inserted in repetitive sequences and

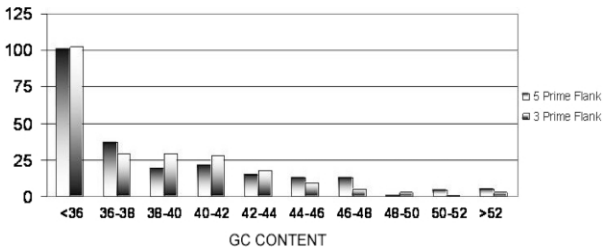

Figure 1. Analysis of preTa L1 pre-integration sites. GC content was calculated for L1 insertion flanking sequences of 1 kb and 2 kb. The 1 kb results are shown here.

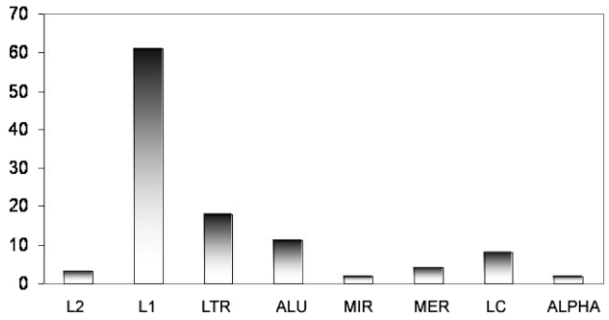

Figure 2. PreTa L1 integrations within other repetitive elements. PreTa L1 insertions within mobile elements were grouped according to the element in which they inserted. Mobile element categories include LINE-2 (L2), LINE-1 (L1), long terminal repeats (LTR), *Alu* (ALU), mammalian-wide interspersed repeats (MIR), medium reiteration frequency sequences (MER), low complexity sequence (LC), Alphoid satellite repeats (ALPHA).

grouped them according to the repeat family in which they reside (Figure 2). This analysis showed that preTa L1 elements insert most frequently in other L1 elements, which is expected both because L1 sequences occupy a large percentage of the human genome and because L1 elements are less GC-rich relative to other mobile element families, such as *Alu* elements, making them more susceptible to subsequent L1 integrations. Lastly, preTa L1 containing regions were analyzed to determine the distance from the integration to the nearest gene. A total of 12 preTa L1 elements reside within 25 kb of novel or known genes as denoted by GenBank annotation, including one full length preTa element, L1AD242, which inserted into intron 23-24 of the retinoblastoma susceptibility protein 1 gene and accounts for 6072 bp of the 7988 bp intron.

Sequence diversity

PreTa L1 sequence diversity is also created by variable 5′ truncation with some of the elements in the human genome only a few hundred base-pairs in length, whereas some full-length elements are over 6000 base-pairs. This phenomenon is classically attributed to the lack of processivity of the reverse transcriptase enzyme in the creation of the L1 cDNA. The point of truncation is traditionally believed to occur as a function of length, where shorter inserts are more likely to occur in the human genome than longer elements.[39] Our data show that there is an enrichment of full-length elements in the human genome, and like the Ta L1 elements many preTa L1 elements have been faithfully replicated in their entirety and inserted into new genomic locations. Of the 356 elements examined (362 total minus six elements located at the end of sequencing contigs), 97 were over 6000 base-pairs long, representing a much larger preTa L1 size class than any other size class (Figure 3). By contrast, very few elements were found in the size ranges between 4000 and 5500 bases, with only 14 of the 356 elements truncated to this particular size range. A bimodal distribution in the size of the elements is created, since there are a significant number of preTa L1 elements that are severely 5 prime truncated and those that are full-length with the average preTa element length of roughly 2700 bp and the median preTa element

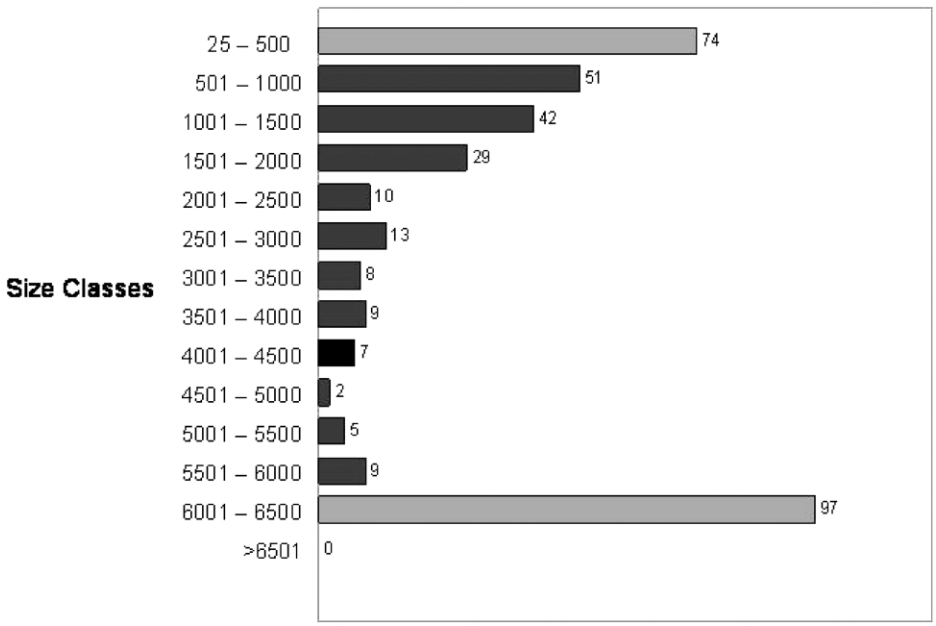

Figure 3. PreTa L1 element genomic size distribution. The following schematic shows the size distribution of preTa L1 elements. Elements are grouped in 500 bp intervals ranging from 25 bp in length to >6501 bp in length. The two most common size intervals are denoted in gray.

Table 2. Summary of preTa L1 analysis

Loci analyzed by PCR	254
Fixed present	200
High frequency insertion polymorphisms	11
Intermediate frequency insertion polymorphisms.	22
Low frequency insertion polymorphisms	0
Total preTa insertion polymorphisms	33
Inserted in paralogous sequences	3
No pre-integration site amplified in primates	9
No PCR results	9
Loci not analyzed by PCR	
L1 elements inserted in other repeats	102
End of contig	6
Total preTa L1 elements analyzed	362

length of roughly 1600 bp. A total of 196 elements were small, with sizes less than 2000 bp, with 125 of these only 50–1000 bases in length. In addition 28% (100/356) of the preTa L1 elements examined were inverted at their 5 prime end, which is believed to occur by an event known as twin priming where target-primed reverse transcription is interrupted by a second internal priming event, resulting in an inversion of the 5 prime end of the newly integrated LINE element.[40] Although L1 truncation is most likely the result of the relatively low processivity of the L1 reverse transcriptase, processes that form secondary structures in the RNA or DNA strands present at the integration site, like twin priming, may also be associated with L1 truncation. One expectation of this model is that a common truncation point should exist for L1 preTa elements. However, from our data we were not able to identify any common truncation points.

Similar to other L1 elements, preTa L1 elements exhibit a significant amount of sequence diversity in the 3 prime tails. In general, the 3 prime tails found in this L1 subfamily range in size from 4 bp to over 1600 bp in length. In all 64% contain AT-rich low complexity sequence, 13% have homopolymeric A tails with an average tail length of 15 bp, 6% have simple sequence repeats with the most common repeat family $TAAA_n$, and 17% contain complex sequence likely resulting from 3 prime transduction events. Three-prime transduction by L1 elements is a unique duplication event that occurs when an L1 sequence is transcribed

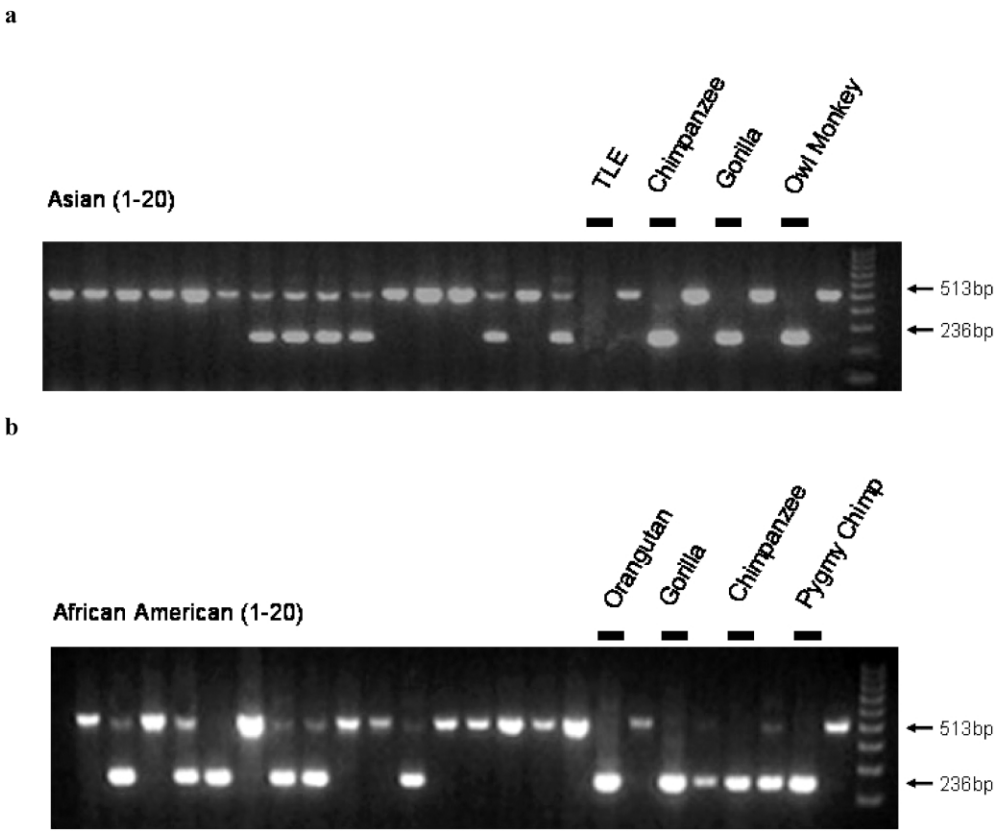

Figure 4. PreTa L1 insertion polymorphisms. This Figure is an agarose gel chromatograph of the PCR products from a survey of the human genomic variation associated with L1AD125. Amplification of the pre-integration site of this locus generates a 236 bp PCR product. Amplification of a filled site generates a 513 bp product (using flanking unique sequence primers). In this survey of human genomic variation 20 individuals from each of four diverse populations were assayed for the presence or absence of the L1 element, with Asian samples shown in (a) and African Americans shown in (b). The control samples are denoted by the black lines and were TLE buffer (10 mM Tris–HCl, 0.1 mM EDTA), common chimpanzee, pygmy chimpanzee, gorilla, orangutan and owl monkey DNA templates. In addition, this particular L1 element was absent from the genomes of non-human primates.

Table 3. PreTa L1 primers, PCR conditions, and associated human genomic diversity

Name	Accession	Chrm. loc.[b]	Forward primer	Reverse primer	Human diversity[c]	AT (F,R)[d]	AT (ACG)[d]	PCR product sizes[a]		
								Filled	Empty	Subfamily specific
L1AD1	AC080166.6	2	AATTCGCTGCATAATTTCTT	AAACATATGGCCATCTTGAC	FP	55	60	6835	249	578
L1AD2	AC090955.2	3	TTTTCTCCATGACTTGAGATGGT	TGCAATCATGAAAACCAGTG	FP	60	60	6308	245	265
L1AD3	AC018878.8	2	TGCACATGGATGTGTAAGAATAC	TTCTTCCCATAAGCATTGGT	FP	60	60	6448	339	245
L1AD4	AC053545.5	4	TTGATGCATTTCTGCATAAGG	CCAAGATTTTGGCTAGCATTT	FP	55	60	4528	295	188
L1AD5	AC079801.2	16	TCATCTCACAGAGCTCACAG	CTAGGAATCCTTCTGTCTGG	NP		60	749	326	150
L1AD6	AC073647.9	7	GCAAACACTGGTTCAAGAAG	TGGAGATAGTGTAGGCACAG	FP	55	60	1741	87	233
L1AD7	AC093607.3	4	Inserted in repeats		R					
L1AD8	AC079926.7	4	GCCTCTTTCTTAGTCAAGCA	AGGTCACAAGGGACATTTCT	NP		60	857	417	208
L1AD9	AC012593.8	2	CAGGTAGGGGAAAGGAGGAG	TGGGCTTATTATCCCCTTGA	FP	55	60	1034	392	342
L1AD10	AC016906.7	2	TGTATTTACCGGGGATGAGG	GCTGTCCCAAATTTCCAGAG	IF	60	60	3602	172	229
L1AD11	AC018465.8	2	GCACCTTGCTATTTGTTTTCT	CCCTAGAGCAATCACCAAAGA	FP	60	60	6515	458	185
L1AD12	AC083950.4	2	GGATAGGCAATGTGTTAGGT	TGCAGAGGCAGTTGTAACAT	FP	55	60	1106	603	303
L1AD13	AC097484.3	4	AAACCTATACATAGAAAATTGCTG	ACCCAGAACAAATGAACACT	FP	60	55	1368	473	424
L1AD14	AC012665.8	2	TTCTGCAACTATAGCCGTAA	ACAACAGACACAGAAGCAAA	IF	60	55	6187	136	173
L1AD15	AC093584.3	4	Inserted in repeats		R					
L1AD16	NG_000004.1	UNK	GGTTGAGAACCACTGTCATAA	GCCAGTGCTTAGATTTACCA	FP	60	60	6213	145	260
L1AD17	AC105459.1	7	ATTCCCCATTTTACGATTTT	GCTACTGCCGTGTTTTACA	FP	55	60	440	276	309
L1AD18	AC096764.3	2	AGATGCCCGGTCTACTACTT	AGCACTTTAAAGGCATCAAC	FP	55	60	3467	151	249
L1AD19	AC009156.9	16	ATATTGGCCAAAGCCTCTTA	TGGCAAGTCCTGAATGATAA	IF	55	55	3974	88	191
L1AD20	AC009156.9	16	CATTAGCAAGCTGATTCAAA	CTTTTGCCATGATTAGTGGT	HF	55	60	474	147	205
L1AD21	AC097522.4	4	CAGAAAGTCATCTCATCTTCC	TAAAGCATTCGTTGTTGTTG	FP	55	60	6528	353	587
L1AD22	AC092570.3	2	CCTCCTCACCTCCTTTTAAT	ATGAAGGGAAACGAGAAAAG	FP	55	60	562	63	220
L1AD23	AC018673.4	12	Inserted in repeats		R					
L1AD24	AC097451.2	4	TCGTTCCTCATCTCTTTGTT	AGCAAAAGCAGTCACTTTTC	FP	55	55	3467	382	396
L1AD25	AC023154.5	4	Inserted in repeats		R					
L1AD26	AC096769.3	4	TTGAGTTTTCCCTCCATGAAA	TCTGATGAATTGTGCCTGACA	FP	60	60	381	157	263
L1AD27	AC093877.3	4	AATATTTAACATGGCCCATAA	GGCATTGGTGTCAATGAGAA	FP	60	60	1171	110	834
L1AD28	AC096749.2	4	GAAGGCTTTATACTCCTTCTTGGA	TCATGGGAGATTTTTCAACTTTC	FP	55	60	6459	419	330
L1AD29	AC105150.2	8	GGACAGAAATACTGGCATCT	CACAATCTTATCTCAAGGGAAT	FP	60	60	6398	318	354
L1AD30	AC055820.7	18	CTTGATGGCAATACAGCCTAA	CCATTAATGTGGGCTCATAATCT	FP	60	60	1855	78	208
L1AD31	AC018626.8	18	GGGAAACGACAGAAGATGGA	GAATTTTGATTTGTGGGCATA	FP	60	60	1143	209	204
L1AD32	AC091613.3	1	End of contig		EC					
L1AD33	AC092798.3	3	Inserted in repeats		R					
L1AD34	AC012642.5	5	GGCTTGTGCTACACAGAGTT	CCAACCAGGAACAATAAAAG	FP	55	55	2816	519	247
L1AD35	AC021538.8	UNK	AAATGCCCACAAAATTCCTG	CCATGGGAGCTACTGGAAAA	FP	55	60	984	386	479
L1AD36	XM_037013.1	UNK	End of contig		EC					
L1AD37	AC099515.2	5	Inserted in repeats		R					
L1AD38	AC026703.4	5	CCCAGTTCTCCAAAATATCA	CACTTGCCTATGGTTCATTT	FP	55	55	5984	240	468
L1AD39	AC078857.12	3	Inserted in repeats		R					
L1AD40	AC078857.12	3	TCGTGACCTTATTAGCCACT	CCTCCATTTGCTACCTAGAG	FP	60	60	1680	512	633
L1AD41	AC078857.12	3	TGTTATTTCAGCTTTAACCATCAA	TTTAAAAATCAAGTATGGGAAAAA	FP	55	55	1202	141	242
L1AD42	AC093515.3	16	Inserted in repeats		R					
L1AD43	AC011597.27	3	Inserted in repeats		R					

(*continued*)

Table 3 Continued

Name	Accession	Chrm. loc.[b]	Forward primer	Reverse primer	Human diversity[c]	AT (F,R)[d]	AT (ACG)[d]	PCR product sizes[a]		
								Filled	Empty	Subfamily specific
L1AD44	AC079943.18	3	ATGCCATCCCCTGGATTT	TGGTTGCTCCAAAGGAACTT	NP		60	6591	530	316
L1AD45	AC061710.16	3	GAGCAAATTTGTCAGACAGAACA	TGGGATGGTTGAAATCAAATG	FP	60	60	3854	147	199
L1AD46	AC072051.8	UNK	CCCTATTTTCCCCATCATCA	AAGCAGGCAGATGGTCACTT	FP	55	60	3706	69	166
L1AD47	AC008006.10	18	CGTCACACACATAACCAGAG	GATCAGGAATATGGCAAAGA	FP	55	55	471	212	284
L1AD48	AC027553.6	UNK	TGCATGAAGCACTACTCAAAGA	TGCAAGATGTGTCAGTATTTAGC	FP	60	60	6181	106	226
L1AD49	AC018991.10	UNK	Inserted in repeats		R					
L1AD50	AC008948.8	5	Inserted in repeats		R					
L1AD51	AC008728.7	5	Inserted in repeats		R					
L1AD52	AC093566.3	8	Inserted in repeats		R					
L1AD53	AC020783.8	8	Inserted in repeats							
L1AD54	AC068062.5	10	CCTTTGTTTCTTGGGTGTGG	CCCACATCACCAAACCATTT	FP	60	60	357	128	212
L1AD55	AC064875.5	2	GCCACACTCCTTTGTTTGCT	CAAGCACAAAAGCAGGAACA	FP	60	60	724	193	273
L1AD56	AC073275.8	7	Inserted in repeats		R					
L1AD57	AC010747.10	2	CGGAAAATTTGTTACTTGCT	AGGTATGCTGCATTTCTTTC	FP	55	55	3903	97	272
L1AD58	AC012509.13	2	CCCTGGATGCTGAGTTTCTT	TCCATCTGGCATTGACTCAG	FP	60	60	1062	139	213
L1AD59	AC009964.11	2	TGGGACATTGACTCCTACTC	GGCATAGGTTTCTGGAAGTA	NP		60	760	340	282
L1AD60	AC009961.11	2	Inserted in repeats		R					
L1AD61	AC078851.4	2	TTTATGCTGATCACTGTTCTTC	AACTAGTTGCATCGTGATCATA	FP	60	55	2090	80	208
L1AD62	AC016720.9	2	CTTTCGCATCATCGTAAAGT	ATTGCCAACTGGTTACAAAG	FP	55	55	2886	114	261
L1AD63	AC012492.9	2	AAAAACCCTTTAAGCTCAGT	TGGAAGCATACAAAATGAAA	FP	55	55	6402	342	180
L1AD64	AC069285.8	7	GCCACTGCTAATCAATTCAC	CCAAAGCAGACACAATTTCT	PARALOG	55	60	6131	77	172
L1AD65	AC026029.8	4	TTTCCTCAAAGTTGATGCTC	CCTGGAAGGCATAACTGATA	NP		55	6787	271	575
L1AD66	AC025223.6	2	TATCCAAATATCCCTTGCAG	TTGTAGTTTGTGGAACTGGA	PARALOG	55	55	717	201	197
L1AD67	AC095347.6	12	Inserted in repeats		R					
L1AD68	AC069242.13	3	CCTATGGATGAAAAATGGAC	TCTGAAAATGTTGCCATTG	FP	55	55	294	111	176
L1AD69	AC092325.2	16	Inserted in repeats		R					
L1AD70	AC079841.10	3	TCCAAGAGCAGGCAGTATTA	TTCCTGACTACTCCAGTTCAG	FP					
L1AD72	AC092468.9	3	GTGCAGGTGTAAGGAAGAAA	GTCTTCAAACCAGACTGCAT	FP	55	55	546	93	218
L1AD73	AC097657.3	4	TGATTTGCAGTATTTTTCCTT	GCATGACCCAGATTAGAAAA	FP	55	55	1148	126	168
L1AD74	AC097463.2	2	No results		NR					
L1AD75	AC092018.2	1	TTTCTCTCCCTCAAGCCTTTT	CCAAAATTCATGCTGGGAAC	IF	60	60	1636	129	124
L1AD76	AC027345.5	4	AAACCTCCCTTTAGTCTCCA	CACCAGACCCAATTTTTAGA	FP	55	60	4500	221	173
L1AD77	AC097110.1	4	TCAAGGAAGGGAGTTAAAAA	ACTTCTTTCATGCCCCTTAT	HF	55	55	991	729	237
L1AD78	AC026439.4	5	TCTTGAGGCTTGCAAATACT	ATGAGCAACAAGAAATCACC	FP	55	60	1559	295	306
L1AD79	AC016620.7	5	Inserted in repeats		R					
L1AD80	AC092185.3	3	AAGCAGTATGTCTGGCACA	ACAAACTGACACTCCAAACC	FP	60	55	6148	72	197
L1AD82	AC022165.8	16	GGTGTCTCCACAGTTGATTC	CCACCGCCAGATTTTACTA	HF	55	55	2876	117	196
L1AD84	AC090525.8	12	TTCCCTGGGTCACTTTTCTC	TGCCAAATTGCTTTGCATAC	FP	55	55	2068	255	333
L1AD85	AC026120.33	12	Inserted in repeats		R					
L1AD86	AC093865.2	2	ACATGATGTCCCATCTTCCA	AAGAGCCATATGAGAGCTTCC	FP	60	60	1046	271	304
L1AD87	AC022446.6	5	AATTTTTCCCCACATGTTC	ACAGAATGGATTTAGCTTGC	FP	60	60	3761	118	248
L1AD88	AC090519.3	15	Inserted in repeats		R					
L1AD89	AC084819.17	12	Inserted in repeats		R					

(*continued*)

Table 3 Continued

Name	Accession	Chrm. loc.[b]	Forward primer	Reverse primer	Human diversity[c]	AT (F,R)[d]	AT (ACG)[d]	PCR product sizes[a]		
								Filled	Empty	Subfamily specific
L1AD90	AC092601.3	2	Inserted in repeats		R					
L1AD91	AC008571.6	5	TGCTAAACAGAAGGCACATA	ATAGATCCATCTGCCAAATC	FP	55	60	6266	170	314
L1AD92	AC092638.2	2	TTATCCAAAGAAGGGGAAAGG	TTTGCCTTATAAGCATTGTGAAAA	FP	55	55	6224	181	195
L1AD93	AC096653.1	4	CAACACTCATTACAACCTGTG	CAGAGTTTATCAGCCAGACC	FP	60	60	2336	382	399
L1AD94	AC092581.2	4	CTCCACGTTAACAGATAGGG	TGAGCTTCACTTAACCACTG	FP	60	60	507	341	239
L1AD95	AC096569.1	2	CCAGCACTGATTTCATAGATGC	TTCAGACAACTGAAGTGCCTTT	FP	55	55	6161	89	224
L1AD96	AC092631.1	4	TAATTAGGTAACGCCTGTGG	CAGGAAGCCTAAACTGCTT	IF	60	60	932	98	245
L1AD97	AC008709.6	5	CCCCAGGCTTTTGAAAATTA	ATTCTCGGGGTCCCAATTAC	FP	60	55	6164	111	214
L1AD98	AC060796.7	17	ATGGAAAGGGGAAGATTTTA	GGCTATACTACAACATCCCTCA	FP	55	55	6164	126	203
L1AD99	AC090791.6	11	GTGACACAAAAAGCACAATTAC	CAATGATTCATGAGTTGGAA	FP	55	55	2737	292	303
L1AD100	AC026729.5	5	CCTGGGTCACAATATGAAGA	TCTGATAACCAGAAGATGAAGA	HF	55	60	6324	258	352
L1AD101	AC025467.5	5	AGTCTCCCTTTCAGAAGCA	AATGCTGGGAATCTTACCTC	IF	55	55	6091	66	163
L1AD102	AC025467.5	5	GAATGGGGTGTGCTGTAA	TTTTAACAAGATCCCAGACC	IF	60	55	3721	78	164
L1AD105	AC010275.6	5	ATTCTCGGGGTCCCAATTAC	CCCCAGGCTTTTGAAAATTA	FP	55	55	6164	111	214
L1AD108	AC008550.5	5	CACAATCATACCTTCCCAACTG	CAGATGAGACTTTGGACGTGA	FP	60	60	6154	84	187
L1AD110	AC092721.2	16	ATTTTGTGGTTCAGCATTTT	CATAGAAAAGGGAACAAATGA	FP	60	60	1590	82	226
L1AD111	AC092357.2	16	AAAAGTTGTTTTCCTGATTTTT	AGTTTTCTCTGCAGCTCATC	FP	56	55	6252	188	184
L1AD112	AC034219.5	5	TTTCCAAAAACAGCTAGGAG	CGTTTTTCTAGCTTAGCAATG	FP	55	55	406	106	209
L1AD113	AC005406.2	UNK	ACCTTGATTGCAAATTGTTT	GGTTTCTTGGCCTCTTTACT	FP	60	60	2881	80	189
L1AD114	AC020651.19	3	Inserted in repeats		R					
L1AD115	AC084032.23	12	AACTGCCATGAAAACTTACC	AAAGATTGTCCACATCAAGG	FP	55	60	253	100	190
L1AD116	AC025176.5	5	End of contig		EC					
L1AD117	AC022024.6	10	CAGCAACCATAGGTTGATAAG	GGATTACTGCCCAAAGAAAC	FP	60	60	852	487	310
L1AD118	AC026113.25	12	GACTGCTGGATCAAATGTTAG	ACCACCTTACTCCTGCTACA	R	55	60	6231	188	272
L1AD119	AC024941.30	12	CTTTATTCATGGCAGAAAGC	CTCATGAGATCTGGTTGTTT	R	55	60	1347	112	249
L1AD120	AC066613.7	UNK	Inserted in repeats		R					
L1AD121	AC010857.8	4	Inserted in repeats		R					
L1AD122	AC011712.6	18	CCCAGGGGAATATATGGAAATTA	AATTGAATGCAGATGGTTTTACC	FP	60	55	6631	139	608
L1AD123	AC010928.7	18	CCAGGAGTCAGAGGATTACA	TCTGTTGTGAGAAGCAAATG	FP	60	60	410	98	172
L1AD124	AC013759.6	18	Inserted in repeats		R					
L1AD125	AC013759.6	18	AAACGGTGAAGGAAATGTTG	GACATGAGCAACCATCAGGA	IF	60	60	513	236	309
L1AD126	AC021082.4	5	Inserted in repeats		R					
L1AD127	AC012323.7	16	Inserted in repeats		R					
L1AD128	AC025097.41	UNK	Inserted in repeats		R					
L1AD130	AC039057.8	UNK	Inserted in repeats		R					
L1AD131	AC073258.9	7	Inserted in repeats		R					
L1AD132	AC017014.4	2	GGGAAGTGAAGGCTAACATA	ACCATGGAGCTCAATTTACA	FP	60	60	469	84	187
L1AD133	AC069294.5	7	GGTTGAGAACCACTGTCATAA	GCCAGTGCTTAGATTTACCA	FP	60	60	6212	145	259
L1AD134	AC084732.1	4	CTACCCAGAACAAATGAACAC	AACCTATACGTAGAAAATTGCTG	FP	60	60	1368	475	422
L1AD135	AC008276.4	2	CTCAAGGGTTCTCATCACTAA	GGAAAGGATACCACAATCAA	HF	60	60	1871	87	191
L1AD136	AC017015.4	3	TGGCTGACAAATTGGTGATT	CCCATGTGAACTGCATTGAA	FP	60	60	712	293	217
L1AD137	AC010970.3	Y	Inserted in repeats		R					
L1AD138	AC012284.5	15	GAGCTGAAGAAACAAAGGAA	ACCTCAAATTCATTTTGGAA	FP	55	60	780	75	200

(*continued*)

Table 3 Continued

Name	Accession	Chrm. loc.[b]	Forward primer	Reverse primer	Human diversity[c]	AT (F,R)[d]	AT (ACG)[d]	PCR product sizes[a]		
								Filled	Empty	Subfamily specific
L1AD139	AC009479.4	Y	Inserted in repeats		R					
L1AD140	AC010722.2	Y	TTCAGGAACATTGCTATGAGGAT	TAGGCATTTATCATGTGCTC	FP	55	55	1643	218	283
L1AD143	AC079175.24	X	CAGTAAACTGGGCTGCTATC	GAGAGTCAAGCAGTGGGTAA	FP	55	60	5078	80	208
L1AD144	AC023842.5	8	CACAAGATTCAATACCTGAGTGACA	TGGGCATTACTAGTTGAACCTAAAG	FP			1641	141	261
L1AD145	AC087883.12	12	GAAGGAAGCCCCCATATGAT	GAGGTGAAAGGCCATTAAAGAA	FP	60	55	473	147	243
L1AD146	AF280107.1	UNK	End of contig		EC					
L1AD147	AC063951.22	12	End of contig		EC					
L1AD148	AC024060.5	3	AACTTCCTTAGGACCTCATTT	TGTGTTTAACGTTCTAAACCTG	FP	60	60	1361	65	229
L1AD149	AC087433.4	15	CCGAAACACAGATAAGCACT	AGTGTAAAAATCTGCATAGCC	FP	55	55	2160	508	274
L1AD150	AC073572.19	12	ATTCCCCCAATTCTCCAAAA	GCAAGGGCCAACTATGCTAA	FP	55	55	1195	124	187
L1AD151	AC023795.18	12	Inserted in repeats		R					
L1AD152	AC079865.14	12	GGGAGATCCAGACATACAAC	TGTGTAACTCTTTTGCGATG	FP	60	60	569	369	341
L1AD153	AC058784.17	13	Inserted in repeats		R					
L1AD154	AC023812.7	3	ACCTCTACCTTACCACACCA	CCTAACTCAGGTCATTCTGC	FP	60	60	1475	175	260
L1AD155	AC018923.21	3	Inserted in repeats		R					
L1AD156	AC008436.5	5	Inserted in repeats		R					
L1AD159	AC008496.5	5	Inserted in repeats		R					
L1AD160	AC034194.4	3	AGAGCTACATGGCTAAATGC	TCTGCAGTTTTAACACCTCTT	IF	60	55	543	238	261
L1AD161	AC011546.6	19	Inserted in repeats		R					
L1AD162	AC020717.3	X	TTCCTATAGGCTTGAATGGA	TTTTGGTGCCCAATAGTATC	FP	55	60	2923	198	219
L1AD163	AC007132.3	2	CCCAGTATGTCCCTCACTCAG	TAGGCAAACCCCAATTGAAA	FP			6359	315	351
L1AD164	AC006968.2	X	TTCCCTGTCCAATGTAAAGAA	AAAGTGCATATTGCACAGGA	FP	55	55	836	107	158
L1AD165	AC010685.3	Y	Inserted in repeats		R					
L1AD166	AC010889.3	Y	CCCTAACATTTCAAAATGCACTG	ATTTTTCCAACTACTGGCACTCA	FP	60	60	1256	162	214
L1AD167	AC006334.3	7	Inserted in repeats		R					
L1AD168	AC009489.3	Y	TGCCTTTATAATATGGAAATGCAG	TGCTCATGGAGTCAGAATATGAA	FP	55	55	1080	196	183
L1AD169	AC011745.4	Y	TCCCATTGCATTTAGCAGATT	AGGCCTGTATTTCAATTGTGCTT	FP	60	55	3676	95	265
L1AD170	AC007278.3	2	GTCTATTAATCCCCCTCCAC	CAACGTTGAAAAGATGTAGAGA	FP	60	52	6149	87	174
L1AD171	AC006992.2	7	TGGAACTATTTCAGGAAATTAAA	AACAAGGGGGAAGAGAATAA	FP	55	55	6278	197	234
L1AD172	AC006362.2	7	Inserted in repeats		R					
L1AD173	AC015542.17	3	TTCCAATATACTTTGCCCTTA	AGTAGGCATCAGCAACAGTC	FP	55	55	546	393	322
L1AD174	AC022013.3	3	TTTGGGGAGAACTATCTGTG	GCTTGGACATTGGAATTTT	FP	60	54	399	118	188
L1AD176	AC026204.4	3	GCACTCTCATTTACTGCTGA	CCACCTTTTACTATTTTGGTG	IF	60	55	838	494	195
L1AD177	AC018514.7	14	ACCAGATGGAAGCTAGATGA	AAGTTTCCAAGGGAAATCAG	FP	60	55	6370	256	373
L1AD178	AC058791.3	7	ATTGTTTAGGGGAAAAGGAC	CCAAAAGCAGGTTAATTCTC	FP	55	55	629	203	322
L1AD179	AC013738.4	10	ACTCCACTTTAATTCGCAAG	GAAGGCGAGAAACTGTAGAA	FP	55	60	1056	113	289
L1AD180	AL627250.8	X	Inserted in repeats		R					
L1AD181	AL449304.19	9	TTCCATAGCCATTGATTACA	AATTTTCAGGCACGTTTTTA	FP	55	55	652	286	446
L1AD182	AL137787.11	X	Inserted in repeats		R					
L1AD183	AL445312.5	X	GTCCAGAAGTCTCTCCTGTT	CGATTGCAGGCTTTCTAATA	FP	55	60	2873	105	413
L1AD184	AL360020.15	9	Inserted in repeats		R					
L1AD185	AL391260.13	10	TTCTGTAGGGCTCCTGACTA	ATTCACAGTTCCCCGTAGTA	FP	60	55	7905	185	1829
L1AD186	AC016951.9	3	ACTTGAAATTGGGGTAGATG	ATTTTCTAGAGGGCTCCTTG	IF	60	59	843	190	206

(*continued*)

Table 3 Continued

Name	Accession	Chrm. loc.[b]	Forward primer	Reverse primer	Human diversity[c]	AT (F,R)[d]	AT (ACG)[d]	PCR product sizes[a]		
								Filled	Empty	Subfamily specific
L1AD187	AL365258.24	1	Inserted in repeats		R					
L1AD188	AL603765.6	1	Inserted in repeats		R					
L1AD189	AL596326.5	1	TGTTTCATGGAGTGTATTTCA	TGCAATGTTAGAAGAAGTGG	HF	55	55	456	198	289
L1AD190	AL606752.11	1	GCTTGACACATAGTGCTTGA	AAATGTGGCATTATTTTCACT	FP	60	60	462	250	193
L1AD191	AL589877.13	X	ACCCAGAAACGCATATACAC	GCAAATTGCAACAAGATAAA	FP	55	55	1926	591	352
L1AD192	AL513493.11	1	TGTCCAATTAAAAGGCACAT	TGGAATATCTTTTTCTGCCTA	FP	55	60	941	134	322
L1AD193	AL359733.15	1	TCTTTTACTCCCAAAAGGAA	TTGGGTAGATGAAGATGACC	NP		55	1900	260	292
L1AD194	AL357873.17	1	GCCCTGGATGTAGTGTATGT	CTCTCTCTTCATCCGTTCAG	FP	55	55	974	144	256
L1AD195	AL592494.4	1	No results		NR	55	55			
L1AD196	Z82209.2	X	TTCTCTCCTAACCCTCTTGG	TTTAGGGTATGCGGTAGAAG	FP	60	55	6581	349	385
L1AD197	AL354949.10	1	GAAACTGAGATTCACGGAAG	AGTTTCTCATCCCACCTTCT	FP	60	60	6437	360	467
L1AD198	AL138785.8	1	GCTTCACCTCACTAGCCTTA	CTCACAAAGCAGCATTTACA	FP	60	60	456	87	163
L1AD199	AL445197.4	1	TTCAGCATATCTGCAAAGTG	GAAAGGATTCTCATTTCCTG	FP	55	60	626	216	341
L1AD200	AL136224.24	6	CAGTCTATCAATTCCTGTTGG	TGATCATCCAGCTCAATTACT	FP	60	55	2353	472	440
L1AD201	AL607144.5	13	CAGACTTGGGCATCTTTTAG	AAAACATCAGGGCCAAATA	FP	55	57	1328	148	178
L1AD202	AL513324.8	10	Inserted in repeats		R					
L1AD203	AL390834.24	10	Inserted in repeats		R					
L1AD204	AF245226.1	21	Inserted in repeats		R					
L1AD205	AL596342.3	1	GACTCTTCCCCTTGAGAATC	GCATGCCTACGATCTCTTAT	FP	55	55	381	222	253
L1AD206	AL603902.4	6	Inserted in repeats		R					
L1AD207	AL592067.4	13	ATTTAGGTATGCGTTTCAGC	ACATCTCTTCATGCCTTCAG	FP	55	55	999	422	238
L1AD208	AL353743.22	9	ATCTCCTATCCCCTTAGCTG	AACCCAAGAGTCACAGTTGA	HF	60	60	1978	530	280
L1AD209	AL139282.10	1	TTGAGTCAAGGAAAAATAATGA	AAAGCAAGGCAGGTATGTTA	FP	60	60	1667	214	245
L1AD210	AL512504.9	X	Inserted in repeats		R					
L1AD211	AL590439.12	10	ATATTGATTTGGCATCCTGA	GTAAACGTTCTAGCCAAAGC	FP	60	60	6207	155	169
L1AD212	AC007347.3	16	CACGGGAGAAGATTTATGTC	TTGTACCTACTCCACCCAAG	FP	54	55	6400	210	310
L1AD213	AC007262.4	14	GCCATAAACAGAAAACCATT	GTTGCAGAAATAACAGCACA	IF	60	60	494	182	294
L1AD214	AC007221.2	16	GCAGTCAACATCTTCCAGTA	TGAGCTAGAATCCCAAAGAT	FP	55	60	6267	135	324
L1AD215	AC007115.1	12	TGAAGAACCTTCACGTAAGAA	AAATATGATGCTTTGCTTCC	FP	60	55	556	176	362
L1AD216	AC006143.1	X	GAGGCTTACTGGAAGCATAG	CTCACGGTTGATGTCACTTT	FP	60	60	1494	430	520
L1AD217	AC011594.8	12	CTGGCCAAAGAGGTAGTTT	CAAAAGAGCATGGTACTGGT	NP		55	7620	479	537
L1AD218	AC004141.1	7	TCCTTAACCTAGTTGCTCCA	AGGGTACATTGAAGTTGAGG	NR	60	60	624	340	458
L1AD219	AC002076.1	7	AGGGAATATTTGGGACATCT	CCCCACCACACTAGAAACTA	IF	60	60	6418	354	391
L1AD220	AC003085.1	7	CCAGGGAACTTGATTTTAGA	CAATTGGATAAGAGGGACTG	FP	60	55	6500	303	199
L1AD221	AC004161.1	UNK	Inserted in repeats		R					
L1AD222	AC006204.1	7	TTTGGAAGCTTCACTTTAGC	TGGCCTTAATATTTTAGCAAC	FP	60	60	590	167	246
L1AD224	AL356096.11	13	Inserted in repeats		R					
L1AD225	AL513355.16	10	CGGTTCTAAACACCATTTGT	TTATGGCCCTTAATTTCATC	FP	60	60	1739	177	192
L1AD226	AL358873.25	6	GCATCTTTGAATCAACAAGTC	TGTATCTAACTATTCCCAGTGATT	FP	60	55	986	238	751
L1AD227	AC004822.1	X	TTGAGAGCATCCATATTTCC	CCAACCTCAGATTACCAAGA	FP	55	60	768	115	202
L1AD228	AC005053.1	7	Inserted in repeats		R					
L1AD229	AL450312.10	9	Inserted in repeats		R					
L1AD230*	AL583806.7	6	GCAATCCATAGACAACCAAT	AGGAGGAATATGCAAACTGA	HF	55	55	2249	599	338

(*continued*)

Table 3 Continued

Name	Accession	Chrm. loc.[b]	Forward primer	Reverse primer	Human diversity[c]	AT (F,R)[d]	AT (ACG)[d]	PCR product sizes[a]		
								Filled	Empty	Subfamily specific
L1AD232	AL583825.8	1	TCCCAGAACTACCTCATAACA	GAGGAAGACAGTGTCACAGA	IF	60	60	1162	219	329
L1AD233	AF207955.1	21	AGGGGTAGATTTTGTTCAGA	AGGACCATTTGCAATGTTAG	FP	60	60	1283	747	667
L1AD234	AL391992.8	10	TGGCTAGTCACCCTAAAAGA	GTTTTATAGGCTTGCATTGG	FP	60	55	6487	388	360
L1AD235	AL160234.3	14	GGAGCTATTAAGCCACAAAA	GAGAGGGTATCCTCGTCTTA	FP	55	60	6771	694	326
L1AD236	AL079307.7	14	GAATGGGGAATTATACGTGA	GTAAGGCACTTGGAAATGTG	FP	60	60	6260	196	295
L1AD237	AL162431.17	1	AAGTGAATGTGGATTTACCC	TCTCAAGGAAATCAGCTCTT	FP	60	60	6499	435	324
L1AD238	AL389895.3	14	ACTTTTATGCCTGAAACCTG	ATCCTTTCTCAGAGGGATCT	FP	60	60	6370	325	278
L1AD239	AL357045.10	1	Inserted in repeats		R					
L1AD240	AL591770.1	14	GTCTCAGACACACAAGCTCA	TTGGCCACTCATCTATCTTT	HF	60	60	540	222	258
L1AD241	AL512310.3	14	Inserted in repeats		R					
L1AD242	AL136960.4	13	CCCCTGAAGAGTCCATATAA	CCTAACAGTCAGGAAAGCTG	FP	55	55	6347	288	197
L1AD243	AL445466.9	1	CTGCTTGTCTTTGGTCTGAT	GTGATCCTGTAGGCCTTCTT	FP	60	60	2933	410	1229
L1AD244	AL512790.1	14	GCATCCGTTTCTCTGATG	TGCAGATTGTACAGAAAAGC	FP	60	60	1394	166	296
L1AD245	AL136295.3	14	ACTTTAGGATTCCGTGGTTT	AATGCTGTTAGAGGAGGATTC	FP	55	60	2193	482	222
L1AD246	AL391838.9	13	Inserted in repeats		R					
L1AD247	AL512662.8	UNK	Inserted in repeats		R					
L1AD248	AL138694.18	UNK	Inserted in repeats		R					
L1AD249	AL133241.3	14	Inserted in repeats		R					
L1AD250	AL121852.3	14	CCCTCAAGAACGATTTTATG	TGTCTAGAATGTTCCCCTTT	FP	60	60	6397	280	237
L1AD251	AL117191.6	14	CTGTGGAGGAAACATTGAAG	TCACACTCAAAGACTCCTTTC	IF	60	60	1995	172	288
L1AD252	AL590370.2	6	GTGAAGGGCACTGGTTATTA	TAATGAAATCGGACCTGTCT	FP	60	60	6498	408	202
L1AD253	AL163613.2	14	TTGCCTAGCTTTTCTACCAG	TTCAAGCTACCTTCTCAAGC	IF	60	60	1369	726	180
L1AD254	AL118557.5	14	ACCTTGACATTCCTGCAA	AATCCACCTGCAGACATTAC	FP	60	60	1000	143	514
L1AD255	AL117693.5	14	TCATTGTTCTATCCATGCCTTTT	GTAGGTTTGGGGCTGGAAAT	IF	55	60	961	197	228
L1AD256	AL161804.4	14	Inserted in repeats		R					
L1AD257	AL359545.12	10	Inserted in repeats		R					
L1AD258	AL358293.4	14	GGTTCAATTGAGCGTTACTT	TGCTGATATAGCACCTAGCA	FP	60	60	6800	735	300
L1AD259	AL158111.5	14	Inserted in repeats		R					
L1AD260	AL133238.3	14	GGTGGATGTATCCATTGTTT	TTTATGCATGCAAGAAATGA	FP	55	55	627	436	464
L1AD261	AL049838.3	14	CTATGGACCCATCTGACTGT	AGTTATTAAACCGGCCACTA	FP	60	60	6269	222	245
L1AD262	AC006568.7	4	ACACGGAGACACTTCAAATC	ACCCGTTATTGTGTTCAGAC	FP	60	60	6424	363	407
L1AD263	AL355481.12	13	GGCTACTTTGGCTTCTGTAA	ATTTGCTCAAACATTTCTGG	FP	55	55	5616	511	531
L1AD264	AL031681.16	20	GGGGAAGTTCCTCCTATATT	AAATGGTAGGTTGGTTTATCA	IF	60	60	1699	501	350
L1AD265	AL589693.3	6	ATAAATTTTCAGGCCTTTCC	GAACAAATTAGACACCATAAGGA	FP	60	60	6218	172	189
L1AD266	AL365508.19	6	Inserted in repeats		R					
L1AD267	AL445258.4	X	Inserted in repeats		R					
L1AD268	AL034425.9	20	GTTTAACCCAGCTGTCCAT	TCCTGTCTCATTTGCTTACC	FP	60	60	2022	361	395
L1AD269	AL136090.12	20	TGACATGGGAGCAATAATAGT	CAGGTGAAATGTATTGAAGGA	FP	55	55	1933	315	371
L1AD270	AL135936.11	20	Inserted in repeats		R					
L1AD271	AL390057.12	6	Inserted in repeats		NR					
L1AD272	AL161901.18	13	Inserted in repeats		R					
L1AD273	AC006947.2	17	GCCTGCTACATGTTCAGAT	CCATCCTTTCTGGAGTGAT	FP	60	60	6252	214	243
L1AD274	AL161938.6	20	Inserted in repeats		R					

(*continued*)

Table 3 Continued

Name	Accession	Chrm. loc.[b]	Forward primer	Reverse primer	Human diversity[c]	AT (F,R)[d]	AT (ACG)[d]	PCR product sizes[a]		
								Filled	Empty	Subfamily specific
L1AD275	AL157380.15	X	Inserted in repeats		R					
L1AD276	AL031679.1	20	ATTCTTCCTGCCACCTTATG	TTAATAGCTGAGCATCATGG	FP	60	60	993	492	372
L1AD277	AC006265.1	17	GTACAAACCATGGACCAGTT	ATGCAAGTATTTGGCATCTT	FP	55	60	6451	386	239
L1AD278	AL121757.7	UNK	Inserted in repeats		R					
L1AD279	AL157881.14	UNK	Inserted in repeats		R					
L1AD280	AC006131.1	UNK	Inserted in repeats		R					
L1AD281	AF036938.1	X	CAGAGTGAAGTGCTTGGTTT	CTTAATATTTGGGCCATGC	NR	60	55	1342	494	590
L1AD282	AL450303.10	1	No results		NR					
L1AD283	AL358434.6	UNK	Inserted in repeats		R					
L1AD284	AL357141.8	6	No results		NR					
L1AD285	AL359252.17	6	ATCCAATCACCATCATCAGT	ACCTGTGTTCCTATCTTTGC	FP	55	55	823	423	272
L1AD286	AL354937.12	9	TTTAACAACGCACACTTAGC	ATTAAGCAATGGCAGGAAT	FP	60	60	1385	337	444
L1AD287	AL356430.19	13	TTGAAATCAATAATGAGGGATA	AACATCAGTCAGCTAAAGCA	FP	55	55	518	277	256
L1AD288	AL121574.19	UNK	Inserted in repeats		R					
L1AD289	AL390039.10	UNK	Inserted in repeats		R					
L1AD290	AL158167.15	10	CCATGCCTCAACATCTCA	ACCTTCCTTATCTTCCCTTG	IF	60	60	750	175	237
L1AD291	AL157398.6	10	TGGAAAAATATCCCATAATGA	TTTCAGATGGTTTTTCAACA	FP	55	55	6277	180	311
L1AD292	AL136970.8	6	GGCAAATTGAGTCAAAGATG	AACTCATTCACAGTAGCAACAA	FP	60	60	6281	206	200
L1AD293	AL136117.12	6	TGGGAATCAGGAAATTTAAC	CCTATTTCTTGGGTTTTCTG	FP	60	60	2300	199	429
L1AD294	AL356286.8	X	Inserted in repeats		R					
L1AD295	AL158201.19	X	AAAGAAAGAAAACACCCACA	CTCACGTATTATTCCGATTTG	NP		60	2579	245	699
L1AD296	AL136441.16	13	AACCAAGGACTTACACATGC	ACTACCACTCATCCAGCAAA	FP	60	60	6518	461	261
L1AD297	AL357499.10	UNK	Inserted in repeats		R					
L1AD298	AL136455.6	1	TGCCACATCTGTTCAGTAAA	GAAATAGGCTCGTTTTCTCT	FP	60	60	1906	399	351
L1AD299	AL359502.14	13	TTAATGCAAGCAGAGTTTCC	TAAGAACCCATGGTCCAGTA	FP	60	55	6269	180	291
L1AD301	AL139334.10	6	AGTTGTCTGAGGAAACACCA	TACGCAGCATCAAGTAAAGA	FP	60	60	1823	700	288
L1AD303	AL139092.12	6	Inserted in repeats		R					
L1AD304	AC005358.1	17	ATCAGTGGTTCTTTGTCCTG	AGCAGTTCACAGTCCTTAGC	FP	55	55	1230	226	248
L1AD305	AC004768.1	5	GCCAGGAGATAATTTGTAGC	TACCTTGCCAGTAACCTTCT	FP	60	60	2726	386	330
L1AD306	AC004389.1	X	End of contig		EC					
L1AD307	AC004074.1	X	Inserted in repeats		R					
L1AD308	AC004523.1	UNK	Inserted in repeats		R					
L1AD309	AL138702.8	13	GCATTGCAGAAGAAAGCTA	TACCTCCAAGGCAAAACTTA	FP	60	60	1547	273	293
L1AD310	AL121946.20	6	CAACACACGTACAGGTATGC	TTAGCCTCTGTCTTTTGTGC	IF	60	55	6557	519	372
L1AD311	AL135932.7	11	TGACCTGTTCTGATGATTGA	CTTCTCAGGGTATCTGTCCA	FP	55	55	2281	271	327
L1AD312	AL136086.8	1	TTGGGGATAACTTTAACTGC	CCTTTTCATCCTCATGTTTT	IF	55	60	6284	228	209
L1AD313	AL137026.21	10	GCAGGAGAGAGTAAAGGGTTA	TGACAACCACTGCTATCAAG	FP	60	60	1382	86	165
L1AD314	AL121938.10	6	GGCTCAGGGAGATTTGATA	TCTGTTGTACTCTTTCAGGAACT	FP	60	55	3462	311	322
L1AD315	AL121933.15	6	GGTAACTAAAGCCATTGCAG	TATCTTTGGATGCTGCATAA	FP	55	55	2636	429	316
L1AD316	AL133547.16	9	Inserted in repeats		R					
L1AD317	AL157378.8	6	Inserted in repeats		R					
L1AD318	AL355871.5	1	TGTGGCTAATTCTGAGACCT	ACATGAGTTATCGTGGCATC	IF	60	60	631	176	175
L1AD319	AL157361.6	13	CCCAATGAACCTGTTGTAGT	GGATTTACATGCCACTTAGG	FP	55	60	392	188	241

(*continued*)

Table 3 Continued

Name	Accession	Chrm. loc.[b]	Forward primer	Reverse primer	Human diversity[c]	AT (F,R)[d]	AT (ACG)[d]	PCR product sizes[a]		
								Filled	Empty	Subfamily specific
L1AD320	AL157360.8	UNK	TCCAATGTTCTCTTAGAGGAGT	TCAACATGCAAAAGACTGAA	FP	60	55	489	114	248
L1AD321	AL139115.5	9	CTTGTCCATTTTCTCCACTG	CAACCCAGTAACTCCACTTC	FP	60	60	1193	80	200
L1AD322	AL049796.28	1	TTCTTCCTGGAAAATTGCTA	TTCCTATGAATCCAGTAGTGC	FP	55	60	6512	434	251
L1AD323	AL117345.21	6	GATGGCTTCAAATCCTTCTT	CACTTCAGATAGAACAAGAGCA	FP	60	55	3744	395	379
L1AD324	AL109920.15	6	TATCATTCCTTCAGGCCATA	GGTGAATGCTTTGGACTTTA	FP	55	60	1568	249	280
L1AD325	Z98950.1	X	TCGGCAGCACATATACTAAA	TCCATAGCCAAGTGAGTTTT	FP	60	55	1001	207	283
L1AD326	AL050309.4	X	Inserted in repeats		R					
L1AD327	AL030998.1	X	AAAACATATTTGGAGGAGCA	GTGACCTGGTGTTTTTGTCT	FP	55	55	6315	202	314
L1AD328	AL133353.6	10	TGCTAATAAAAGCACTCTGAAA	AAGATGGTGAATGTTGTAGGA	FP	55	60	2610	155	284
L1AD329	AL136169.6	UNK	Inserted in repeats		R					
L1AD330	AL133404.8	6	Inserted in repeats		R					
L1AD331	AL136363.4	X	ATTTCTTCTGCAGCTCTGAC	CATGATAACTTTGGTTTGTCAC	FP	60	60	6213	188	279
L1AD332	AL133247.1	2	TGACTGACCACTGTATGGAA	GTGGCTGTTTGGATTCTTTA	FP	60	60	1399	204	247
L1AD333	AL078604.10	6	Inserted in repeats		R					
L1AD334	AL021877.1	22	TTGACTTGTTTAGAAAGGGATT	GGATAAAGCTGAAAGCTCAA	FP	55	60	6322	233	215
L1AD335	Z70758.1	X	TCATCCAGCATTGAATCAG	TTGGTAGAAAGTGAAGTGGAG	FP	60	60	571	199	238
L1AD336	AL096706.10	UNK	Inserted in repeats		R					
L1AD337	AL049589.15	X	Inserted in repeats		R					
L1AD338	AL021069.1	1	AAGAATCCAATTTGCAACAG	TTTGATTCGGATTACACTGA	FP	60	60	6248	173	233
L1AD339	Z97181.1	X	GTTAAAATGCCAGGCTGAT	TGAGAAATGTGTTCTCCAAA	FP	55	55	1169	136	349
L1AD340	AL031117.1	X	Inserted in repeats		R					
L1AD341	AL034348.5	6	TGACTTCCATTTCAGGTACTC	CCACATTAGAGGTTTTCCAA	FP	55	60	4229	143	293
L1AD342	AL022399.2	1	TATGCATTTCCATGACTTGA	GTGGTAGGAGTAGGGGAAAG	FP	60	60	6795	342	708
L1AD343	AL033530.1	1	Inserted in repeats		R					
L1AD344	AL031313.1	X	Inserted in repeats		R					
L1AD345	AL023806.1	6	AGTACCAATGAAGTGCCATT	CAGGAGCATAAATAGGACCA	FP	60	60	1770	379	500
L1AD346	Z80232.1	X	CGGAAAATCCTCAGTCATC	ATGCCACAGCTTAAAAGTTC	FP	60	60	1065	261	309
L1AD347	Z84720.1	X	Inserted in repeats		R					
L1AD348	Z93018.1	X	No results		NR					
L1AD349	Z99128.1	6	AGCACTCCTTTTATGAAGTCAACC	AGAGGAGAGAGTGGTTGATATTGG	FP	55	55	2851	1223	565
L1AD350	Z82170.1	UNK	GGCAGACCAAATGGATTAT	GATCCAAATATCAGACAAAATGT	FP	55	60	6342	288	184
L1AD351	Z95126.1	X	TGACATGCTTCCCTAAGTTT	TATAGAAAGTGAGGCCCAGA	FP	60	60	537	363	313
L1AD352	Z95325.2	X	CTTGCTGAATTAATCCCTTT	GGAAGAAATGATCCATAAGAAA	FP	55	55	3497	355	346
L1AD353	AL022308.1	X	CAAGGGGAAATCTCACAATA	GGACTTTGGGACTTACATCA	PARALOG	55	60	6238	174	263
L1AD354	AL023095.1	X	TCATCTTGCTCCCAAATATC	TCCTTAACACAGTCAAGTGAAC	FP	60	60	4839	170	338
L1AD355	Z98948.1	X	No results		NR					
L1AD356	AC000111.1	7	TGTGGCTATGTGAGATGAGA	CCTTAATTTGAGGGGTTTTT	FP	55	55	4633	326	385
L1AD357	AP004241.2	11	CATAGGACGTTCAAGTGTGA	ATTGTCTATGGCTGCTTTCT	FP	60	55	765	387	593
L1AD358	AP002803.3	11	AGGTTTTGAGGTTTGCTGTA	TCCCAATAATCACTTTCCAC	FP	55	55	6274	205	264
L1AD359	AP002002.4	11	AAGGGCATATAAAACTGGTG	GCACCCATTAACTCATCATT	FP	55	60	6460	356	328
L1AD360	AP000764.4	11	CCATGCTTTCCACTCTTTAT	GCAGAAAAGGGTGTTCATA	FP	60	60	379	179	240
L1AD361	AP002784.3	11	GGAAAAATGACAGTCAGGAG	GCCTACCCAATGAATATCCT	HF	60	60	1031	149	258
L1AD362	AP003719.3	11	Inserted in repeats		R					

(*continued*)

Table 3 Continued

Name	Accession	Chrm. loc.[b]	Forward primer	Reverse primer	Human diversity[c]	AT (F,R)[d]	AT (ACG)[d]	PCR product sizes[a]		
								Filled	Empty	Subfamily specific
L1AD363	AP000811.4	11	CCATTACTTGAAGCAGAACC	CTGTGGGTCTCAGATCATTT	FP	55	55	6419	367	175
L1AD364	AP001977.4	11	TAAACTGGGGCTAGAAGTCA	CCAATTGAGAACCATCTTGT	FP	55	55	6335	383	344
L1AD365	AP002982.2	8	ACAGAGATTTCCTGGGCACT	TCAAACTGCATGCAAAATCC	FP	55	55	811	109	208
L1AD367	AP000789.4	11	CCAACAGGGATCAAAGGTTC	GCCACCTTGAGTTGGTGAAG	FP	55	55	378	147	175
L1AD368	AP002006.5	11	TTTCTTTTCCTACTCTCCCTCTC	GAGAAATAAAGGCAATTGCTCAC	NP		55	4593	186	922
L1AD369	AP001485.4	11	AAAACATATAAGCGGCCAAC	CAGCACCTGTTATGGTTTGA	FP	60	55	2437	466	187
L1AD370	AP000462.2	11	TAAGAAGAGGGGAGGAGACT	GCCTCTATGAAGCAGGTATG	FP	55	60	793	178	237
L1AD371	AP001709.1	21	CTAAATTGCTCCATTCCTTG	ATCACTGTAGGGTGATCCAG	HF	55	55	2525	581	562
L1AD372	AP001678.1	21	CTTACGCCTCAATTATCTGG	TGCAATTGATCTTACAAGGA	FP	55	55	2325	280	269
L1AD373	AP001674.1	21	CAAATAGCCAGCACAATATG	TTGTCATTGGTCTTTTGTCA	FP	55	60	823	165	226
L1AD374	AP001669.1	21	Inserted in repeats		R					
L1AD375	AB009801.1	14	AATCCACCTGCAGACATTAC	AGAACATCCCCTATCCAAAC	FP	55	55	688	87	202
L1AD382	Z95325.2	X	Inserted in repeats		R					
L1AD383	AC090791.6	11	TGGTGGTCTCAGAGTAAACA	ACCCAAAACATCATTAGTGC	FP	60	60	1642	117	1026
L1AD384	AL136441.16	13	Inserted in repeats		R					
L1AD385	AP003123.2	11	GCACAGGTTTATCTCCTTGA	ATTGAAGACCTGCAATTTGT	FP	55	55	6379	284	287
L1AD386	AC114975.2	5	Inserted in repeats		R					
L1ADY8	AC010970.3	Y	TCACACGTATCCCTTTGCAG	TTTTCTGTGAAACATCTTGGAGA	FP	55	55	1813	115	204

∗ Indicates L1 preTa element identified by Ovchinnikov 2002 (Ref. 28).

[a] PCR product sizes: empty product size is calculated computationally by removing the L1 preTa elements and one direct repeat from identified filled site. Subfamily-specific product size is calculated from internal subfamily-specific primer located in the 3′ UTR to the proximal 3′ primer. In cases where target site duplication sequences were not found flanking the element PCR product sizes may vary from those reported.

[b] Chromosomal location was determined from accession information or by PCR analysis of NIGMS monochromosomal hybrid cell line DNA samples. L1 elements with unknown locations are denoted UNK.

[c] Elements at the end of sequencing contigs are denoted (EC), those residing in other repeats (R), those having paralogs (PARALOG), and elements with inconclusive PCR results (NR). Elements represented here are classified according to allele frequency as: high frequency (HF), intermediate (IF), no pre-integration site in primate samples tested (NP), or as fixed present (FP) insertions. Fixed present: every individual tested had the LINE element in both chromosomes. Intermediate frequency insertion polymorphism: the element is present in more than 30% of alleles tested and no more than 70% of the alleles. High frequency insertion polymorphism: the element is present in more than 70% but not all alleles tested. Indeterminable data is denoted (−).

[d] Amplification of each locus required 2:30 minutes at 94 °C initial denaturing, and 32 cycles for one minute at 94 °C, one minute at annealing temperature (AT), and one minute elongation at 72 °C. A final extension time of ten minutes at 72 °C was also used.

Table 4. Autosomal preTa L1 allele frequency and heterozyosity

	African American genotypes					Asian genotypes					European genotypes					South American genotypes					
Element	+/+	+/−	−/−	f[a]	Het[b]	+/+	+/−	−/−	f[a]	Het[b]	+/+	+/−	−/−	f[a]	Het[b]	+/+	+/−	−/−	f[a]	Het[b]	Avg Het [c]
L1AD10	0	5	14	0.13	0.23	0	8	12	0.20	0.33	3	7	7	0.38	0.49	3	7	10	0.33	0.45	0.37
L1AD14	9	10	1	0.70	0.43	4	8	7	0.42	0.50	16	4	0	0.90	0.18	17	2	1	0.90	0.18	0.33
L1AD19	13	7	0	0.83	0.30	15	2	0	0.94	0.11	14	6	0	0.85	0.26	14	6	0	0.85	0.26	0.23
L1AD20	18	2	0	0.95	0.10	19	1	0	0.98	0.05	16	0	0	1.00	0.00	19	0	0	1.00	0.00	0.04
L1AD75	0	5	15	0.13	0.22	0	1	18	0.03	0.05	1	9	9	0.29	0.42	0	9	11	0.23	0.36	0.26
L1AD77	1	5	11	0.21	0.34	0	1	19	0.03	0.05	0	3	13	0.09	0.18	0	6	12	0.17	0.29	0.21
L1AD82	19	1	0	0.98	0.05	17	0	0	1.00	0.00	20	0	0	1.00	0.00	19	1	0	0.98	0.05	0.03
L1AD96	13	5	2	0.78	0.36	15	1	0	0.97	0.06	5	10	5	0.50	0.51	11	7	2	0.73	0.41	0.34
L1AD100	19	0	0	1.00	0.00	19	0	1	0.95	0.10	20	0	0	1.00	0.00	20	0	0	1.00	0.00	0.02
L1AD101	16	4	0	0.90	0.18	10	5	0	0.83	0.29	13	6	1	0.80	0.33	11	9	2	0.70	0.43	0.31
L1AD102	14	0	0	1.00	0.00	14	1	0	0.97	0.07	12	1	2	0.83	0.29	0	4	16	0.10	0.18	0.13
L1AD125	12	7	1	0.78	0.36	14	6	0	0.85	0.26	20	0	0	1.00	0.00	19	1	0	0.98	0.05	0.17
L1AD135	19	1	0	0.98	0.05	20	0	0	1.00	0.00	20	0	0	1.00	0.00	20	0	0	1.00	0.00	0.01
L1AD160	11	5	4	0.68	0.45	5	11	4	0.53	0.51	4	12	1	0.59	0.50	4	8	3	0.53	0.51	0.49
L1AD176	7	3	2	0.71	0.43	2	9	5	0.41	0.50	0	1	15	0.03	0.06	1	0	11	0.08	0.16	0.29
L1AD186	4	7	8	0.39	0.49	14	5	1	0.83	0.30	5	10	2	0.59	0.50	4	11	5	0.48	0.51	0.45
L1AD189	14	5	0	0.87	0.23	19	0	0	1.00	0.00	20	0	0	1.00	0.00	19	1	0	0.98	0.05	0.07
L1AD208	14	6	0	0.85	0.26	19	0	0	1.00	0.00	14	0	0	1.00	0.00	14	0	0	1.00	0.00	0.07
L1AD213	7	9	3	0.61	0.49	2	12	5	0.42	0.50	2	2	5	0.33	0.47	8	5	7	0.53	0.51	0.49
L1AD219	3	14	3	0.50	0.51	0	10	10	0.25	0.38	1	5	14	0.18	0.30	2	11	7	0.38	0.48	0.42
L1AD230	14	6	0	0.85	0.26	19	0	0	1.00	0.00	20	0	0	1.00	0.00	20	0	0	1.00	0.00	0.07
L1AD232	13	7	0	0.83	0.30	8	7	3	0.64	0.47	12	2	0	0.93	0.14	13	4	1	0.83	0.29	0.30
L1AD240	13	3	0	0.91	0.18	20	0	0	1.00	0.00	13	0	0	1.00	0.00	20	0	0	1.00	0.00	0.04
L1AD251	3	9	7	0.39	0.49	10	8	2	0.70	0.43	14	4	0	0.89	0.20	8	11	1	0.68	0.45	0.39
L1AD253	11	6	3	0.70	0.43	0	14	5	0.37	0.48	4	8	7	0.42	0.50	0	6	14	0.15	0.26	0.42
L1AD255	1	8	10	0.26	0.40	1	9	10	0.28	0.41	6	7	7	0.48	0.51	3	14	3	0.50	0.51	0.46
L1AD264	4	10	6	0.45	0.51	2	9	8	0.34	0.46	2	7	7	0.34	0.47	3	11	6	0.43	0.50	0.48
L1AD290	7	12	1	0.65	0.47	4	8	7	0.42	0.50	3	13	0	0.59	0.50	6	9	5	0.53	0.51	0.49
L1AD310	5	6	7	0.44	0.51	0	5	15	0.13	0.22	5	2	5	0.50	0.52	6	5	7	0.47	0.51	0.44
L1AD312	0	4	16	0.10	0.18	11	6	2	0.74	0.40	2	9	5	0.41	0.50	2	7	9	0.31	0.44	0.38
L1AD318	4	12	4	0.50	0.51	2	12	6	0.40	0.49	4	8	8	0.40	0.49	3	11	6	0.43	0.50	0.50
L1AD361	17	3	0	0.93	0.14	19	0	0	1.00	0.00	20	0	0	1.00	0.00	20	0	0	1.00	0.00	0.04
L1AD371	15	5	0	0.88	0.22	18	2	0	0.95	0.10	20	0	0	1.00	0.00	20	0	0	1.00	0.00	0.08

[a] f represents the frequency of the element.
[b] This is unbiased heterozygosity.
[c] Average heterozygosity is the average heterozygosity for all populations.

along with genomic sequence at its 3 prime end. This sequence then integrates at a different genomic location, resulting in duplication of the source L1 sequence and the 3 prime genomic sequence flanked by target site duplications.[16–18] We have identified 50 3 prime transduction events mediated by preTa L1 elements and believe that these elements have transduced approximately 10,400 total bases of sequence with one transduction event responsible for duplicating a region over 1600 bp. The diversity observed in the tails of the L1 elements is not surprising, since previous studies have shown an association as well as direct evidence that simple sequence repeat motifs present in the 3 prime tail of mobile elements can mutate, serving as nuclei for the generation of simple sequence repeats.[41–43] A complete list of the preTa elements involved in transduction events is located at our web site†.

L1 associated human genomic diversity

Of the 362 preTa L1 elements isolated *in silico*, 102 of the elements were inserted into other repetitive regions of the genome such that flanking unique sequence PCR primers could not be designed. Six additional elements resided at the end of sequencing contigs in GenBank and lacked unique flanking sequence information, making PCR primer design in this region impossible. The remaining 254 were analyzed using a subfamily-specific PCR assay and flanking unique sequence primers as previously described[28] (summarized in Table 2). Three elements out of 254, produced inconclusive PCR results because of the amplification of paralogous genomic sequences as described previously.[44] Nine elements produced non-specific PCR results, and were excluded from further analysis. Another nine elements produced subfamily-specific PCR products in all human samples tested, but did not produce pre-integration sites in both human and non-human primate genomes. This may be the result of some type of large deletion event that occurred in the human genome and not in the genome of non-human primates, making the non-human primate pre-integration site much larger than expected and not detectable by our assay as reported previously.[19] Alternatively this could also be the result of mutations in the oligonucleotide hybridization sites rendering them ineffective for PCR. In addition, we identified 36 preTa L1 elements that mapped to the X chromosome and eight that mapped to the Y chromosome, all of which were fixed present in the individuals tested (Table 3). The human genomic diversity associated with the autosomal preTa L1 elements is shown in Tables 3 and 4.

A total of 293 (254-9-9-3) preTa L1 elements produced unambiguous results when analyzed by a two-step PCR assay across 80 individuals from four geographically diverse human populations with 33 (14%) being polymorphic with respect to insertion presence/absence (Tables 3 and 4). Examples of human genomic diversity associated with preTa L1 insertion polymorphisms are shown in Figure 4(a) and (b). Of the preTa L1 elements, 11 were high frequency insertion polymorphisms with L1 element allele frequencies greater than 0.70, so that most of the individuals were homozygous (+/+) for the presence of the LINE element. Of the polymorphic elements, 22 were intermediate frequency, with a LINE element allele frequency greater than 0.30 but less than 0.70 across the diverse human populations sampled. None of the L1 preTa elements tested had insertion allele frequencies less than 0.30. One possible explanation for the absence of low frequency preTa insertion polymorphisms would be that the preTa subfamily has largely undergone retrotranspositional quiescence and is no longer generating new copies. As a result, the number of low frequency preTa insertion polymorphisms in the human genome would be limited. It is also possible that the newly integrated preTa L1 elements are removed from the human genome as a result of negative selection. However, we consider the former explanation more likely based upon the threefold higher levels of insertion polymorphism in the Ta subfamily as compared to the preTa subfamily (45% *versus* 15%) as well as the previously reported frequency distribution of Ta L1 insertion polymorphisms in the human genome.[24]

A total of 200 preTa L1 elements were fixed present in the human genome. These elements are likely to be slightly older than their polymorphic counterparts, having inserted into the human genome prior to the radiation of humans from Africa. Overall, the unbiased heterozygosity values across all of the L1 elements subjected to PCR analysis were similar across the four populations, with values of 0.306 in African Americans, 0.243 in Asians, 0.252 in Europeans, and 0.269 in South Americans with the African American population being the most diverse with respect to preTa L1 alleles (Table 4). However, several of the polymorphic elements individually exhibited unbiased heterozygosity values that approached 0.5, the theoretical maximum for bi-allelic loci.

In order to determine whether the LINE insertion polymorphisms were in Hardy–Weinberg Equilibrium (HWE) we compared expected genotype frequencies with observed genotype frequency using chi-square tests for goodness of fit. A total of 132 chi-square tests for goodness of fit are theoretically possible. However, 28 of the comparisons involved populations that were monomorphic for the presence of the L1 insertion leaving 104 possible tests. A total of 23 deviations from Hardy–Weinberg expectations were observed in the comparisons. A total of 18 of the deviations were the result of low expected genotype frequencies. Of the remaining five tests that deviated from HWE, none clustered by population or locus. This deviation is not surprising, since a total of 5.15 deviations from HWE would be expected by

chance alone at the 5% significance level. One shortcoming of this method is its inability to deal with low expected genotype frequencies. To further test these polymorphisms for HWE, we performed an exact test for Hardy–Weinberg proportions using the Markov chain test available in the Arlequin program,[45] which is not hindered by low expected frequencies. The exact test showed that none of the 104 comparisons deviated from HWE proportions at the 1% level. Therefore we conclude that the newly identified L1 insertion polymorphisms do not significantly depart from HWE.

Discussion

Here, we report a comprehensive analysis of the dispersion and insertion polymorphism associated with the preTa L1 subfamily within the human genome. We estimate that there are approximately 900 lineage-specific L1 elements present in the entire human genome. In addition, given the median size for preTa and Ta L1 elements (~1600 bp) and a conservative copy number estimate of 900 elements, we estimate that human lineage-specific L1 retrotransposition has been responsible for increasing the size of the human genome by roughly 1.4 million bases.

The level of sequence diversity, estimated age, and the reduction of human genomic variation associated with this L1 subfamily relative to the Ta L1 subfamily provide strong evidence suggesting that the expansion of preTa L1 elements began prior to the expansion of the Ta L1 subfamily that has been analyzed in detail previously.[24,27] However, the expansion of preTa L1 elements also appears to have occurred over a time frame that predated the radiation of humans from Africa and continued until very recently, in fact it may still be occurring at a very low level within the human lineage. Thus, we conclude that the expansion of preTa and Ta L1 elements occurred in an overlapping time frame in the human lineage. The reason(s) for the relative retrotranspositional quiescence of preTa elements remain unknown. However, they may relate to alterations in the ORF2 protein of the preTa elements, decreased transcription from the preTa "source" elements or a decrease in the ability of the elements to undergo target-primed reverse transcription.[46] Further studies using *in vitro* systems to measure retrotransposition[25] will be required to definitively address this question.

Sequence analysis of the preTa L1 insertions suggests that they have a slight preference for integrating into regions of the genome with low GC content. This observation is contradictory to that previously reported,[43] but is in agreement with results obtained by The International Human Genome Sequencing Consortium.[3] The reason for this integration site preference is unclear, but may result from a subtle sequence preference of the preTa-encoded endonuclease. Alternatively, this observation may reflect limitations on L1 preTa insertion events imposed by chromatin organization. However, it is likely that both factors, as well as others not mentioned here, are important in determining where in the human genome young L1 elements will integrate. It is also interesting to note that some preTa L1 insertions have occurred adjacent to known genes. The persistence of these newly integrated preTa L1 elements in these regions of the human genome is most likely indicative that they have had no negative effects with respect to the function of these genes.

Of the essentially 105 full length L1 preTa elements identified, 29 have both open reading frames intact and are putatively retrotransposition-competent elements. The data collected from the L1 preTa subfamily along with the L1Hs Ta subfamily (44 elements) yield a computational estimate of 73 active L1 elements within the genome that is comparable to previous estimates of the number of potentially active L1 elements in the human genome.[26] Collectively, these data suggest that L1 elements from multiple subfamilies may still be capable of retrotransposition within the human lineage. In addition, it is also important to mention that those full-length elements that no longer have intact open reading frames could have previously served as active source or driver genes for the expansion of pre Ta L1 elements, but have accumulated mutations over time that subsequently inactivated them.

The computational identification approach described here provides an efficient and high-throughput method for recovering preTa L1 elements from the human genome, some of which are polymorphic for insertion presence/absence in individual human genomes. Individual L1 insertion polymorphisms identified, similar to other mobile element insertion polymorphisms, are the products of unique insertion events within the human genome. Because each L1 element integrates only once into the human genome, individuals that share L1 insertions (and insertion polymorphisms) inherited them from a common ancestor, making the L1 filled sites identical by descent.[24,28] This distinguishes L1 insertion polymorphisms from other types of genetic variation that may not be derived from a single ancestral allele, including microsatellites[47] and restriction fragment length polymorphisms.[47,48] In addition, the ancestral state of an L1 insertion is known to be the absence of the L1 element. Therefore the 33 new L1 insertion polymorphisms reported here appear to have genetic properties similar to the previously identified *Alu*[44,49–53] and L1[24,27,28] insertion polymorphisms and provide a unique form of genetic variation present in the human population that will serve as an additional source of identical by descent genomic variability for the study of human population relationships.

Materials and Methods

Cell lines and DNA samples

The cell lines used to isolate primate DNA samples were as follows: human (*Homo sapiens*) HeLa (ATCC CCL2), common chimpanzee (*Pan troglodytes*) Wes (ATCC CRL1609), pygmy chimpanzee (*Pan paniscus*) Coriell Cell Repository Number AG05253, gorilla (*Gorilla gorilla*) Lowland Gorilla (Coriell Cell Repository Number AG05251B), green monkey (*Cercopithecus aethiops*) ATCC CCL70, owl monkey (*Aotus trivirgatus*) OWK (OWKidney) ATCC CRL 1556, and Orangutan (*Pongo pygmaeus*) (Coriell Primate Panel PRP00001 Cell Repository Number NG12256). Cell lines were maintained as directed by the source and DNA isolations were performed using Wizard genomic DNA purification (Promega). Human DNA samples from the European, African American, and Asian population groups were isolated from peripheral blood lymphocytes[54] available from previous studies.[50] South American Human DNA was obtained from Coriell Human Variation Panels HD17 and HD18.

Computational analyses

The draft sequence of the human genome was screened using the Basic Local Alignment Search Tool (BLAST)[30] available at the National Center of Biotechnology Information Genomic Blast page †. A 19 bp oligonucleotide, 5′-CCTAATGCTAGATGACACG-3′ that is diagnostic for the preTa subfamily was used to query the Human Genome database with the following optional parameters: filter none; advanced options −e 0.1, −v 600, −b 600. Copy number estimates were determined from BLAST search results. Sequences containing exact matches were subjected to additional analysis as outlined below.

A sequence region of 9000–10,000 bases, including the match and 1000–2000 bases of flanking unique sequence were annotated using RepeatMasker version 7/16/00 from the University of Washington Genome Center Server‡ or Censor from the Genetic Information Research Institute§.[55] These programs annotate repeat sequence content and were used to confirm the presence of preTa L1 elements and regions of unique sequence flanking the elements. PCR primers flanking each L1 element were designed using Primer3 software available at the Whitehead Institute for Biomedical Research‖ and were complementary to the unique sequence regions flanking each L1 element. The resultant primers were screened with standard nucleotide-nucleotide BLAST [blastn] against the non-redundant (nr) and high-throughput (htgs) sequence databases to ensure they resided in unique DNA sequences. Primers residing in repetitive sequence regions were discarded and new primers designed if possible. A complete list of all the L1 elements identified using this approach is available from our website‡. Individual L1 DNA sequences were aligned using MegAlign with the ClustalW algorithm and the default settings (DNAstar version 5.0 for Windows) followed by manual refinement.

PCR amplification

PCR amplification of 255 individual L1 elements was carried out in 25 μl reactions containing 20–100 ng of template DNA, 40 pM of each oligonucleotide primer (Table 1), 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). Each sample was subjected to the following amplification for 32 cycles: an initial denaturation of 150 seconds at 94 °C, one minute denaturation at 94 °C, one minute at the annealing temperature (specific for each locus), and an extension at 72 °C for one minute. Following the cycles a final extension was performed at 72 °C for ten minutes. For analysis, 20 μl of each sample was fractionated on a 2% (w/v) agarose gel with 0.05 μg/ml ethidium bromide. PCR products were directly visualized using UV fluorescence. The human genomic diversity associated with each L1 preTa element was determined by the amplification of 20 individuals from each of four geographically distinct populations (African American, Asian, European, and South American) for a total of 160 chromosomes.

† http://www.ncbi.nlm.nih.gov/BLAST/

‡ http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker

§ http://www.girinst.org/Censor_Server-Data_Entry_Forms.html

‖ http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgiand

Acknowledgements

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

References

1. Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**, 657–663.
2. Prak, E. T. & Kazazian, H. H., Jr (2000). Mobile elements and the human genome. *Nature Rev. Genet.* **1**, 134–144.
3. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
4. Smit, A. F., Toth, G., Riggs, A. D. & Jurka, J. (1995). Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* **246**, 401–417.
5. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.
6. Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl Acad. Sci. USA*, **94**, 1872–1877.
7. Cost, G. J. & Boeke, J. D. (1998). Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry*, **37**, 18081–18093.

8. Cost, G. J., Golding, A., Schlissel, M. S. & Boeke, J. D. (2001). Target DNA chromatinization modulates nicking by L1 endonuclease. *Nucl. Acids Res.* **29**, 573–577.
9. Fanning, T. G. & Singer, M. F. (1987). LINE-1: a mammalian transposable element. *Biochim. Biophys. Acta*, **910**, 203–212.
10. Kazazian, H. H., Jr (1998). Mobile elements and disease. *Curr. Opin. Genet. Dev.* **8**, 343–350.
11. Deininger, P. L. & Batzer, M. A. (1999). Alu repeats and human disease. *Mol. Genet. Metab.* **67**, 183–193.
12. Kazazian, H. H., Jr, Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G. *et al.* (1988). Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, **332**, 164–166.
13. Yang, Z., Boffelli, D., Boonmark, N., Schwartz, K. & Lawn, R. (1998). Apolipoprotein(a) gene enhancer resides within a LINE element. *J. Biol. Chem.* **273**, 891–897.
14. Rothbarth, K., Hunziker, A., Stammer, H. & Werner, D. (2001). Promoter of the gene encoding the 16 kDa DNA-binding and apoptosis-inducing C1D protein. *Biochim. Biophys. Acta*, **1518**, 271–275.
15. Fitch, D. H., Bailey, W. J., Tagle, D. A., Goodman, M., Sieu, L. *et al.* (1991). Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. *Proc. Natl Acad. Sci. USA*, **88**, 7396–8400.
16. Boeke, J. D. & Pickeral, O. K. (1999). Retroshuffling the genomic deck. *Nature*, **398**, 108–109.
17. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H., Jr (1999). Exon shuffling by L1 retrotransposition. *Science*, **283**, 1530–1534.
18. Goodier, J. L., Ostertag, E. M. & Kazazian, H. H., Jr (2000). Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657.
19. Gilbert, N., Lutz-Prigge, S. & Moran, J. (2002). Genomic deletions created upon LINE-1 retrotransposition. *Cell*, **110**, 315–325.
20. Kazazian, H. & Goodier, J. (2002). LINE drive. Retrotransposition and genome instability. *Cell*, **110**, 277–280.
21. Symer, D., Connelly, C., Szak, S., Caputo, E., Cost, G. *et al.* (2002). Human l1 retrotransposition is associated with genetic instability *in vivo*. *Cell*, **110**, 327–338.
22. Deininger, P. L., Batzer, M. A., Hutchison, C. A., III & Edgell, M. H. (1992). Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8**, 307–311.
23. Skowronski, J., Fanning, T. G. & Singer, M. F. (1988). Unit-length line-1 transcripts in human teratocarcinoma cells. *Mol. Cell. Biol.* **8**, 1385–1397.
24. Myers, J. S., Vincent, B. J., Udall, H., Watkins, W. S., Morrish, T. A. *et al.* (2002). A comprehensive analysis of recently integrated human Ta L1 elements. *Am. J. Hum. Genet.* **71**, 312–326.
25. Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D. *et al.* (1996). High frequency retrotransposition in cultured mammalian cells. *Cell*, **87**, 917–927.
26. Sassaman, D. M., Dombroski, B. A., Moran, J. V., Kimberland, M. L., Naas, T. P. *et al.* (1997). Many human L1 elements are capable of retrotransposition. *Nature Genet.* **16**, 37–43.
27. Boissinot, S., Chevret, P. & Furano, A. V. (2000). L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol.* **17**, 915–928.
28. Sheen, F. M., Sherry, S. T., Risch, G. M., Robichaux, M., Nasidze, I. *et al.* (2000). Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. *Genome Res.* **10**, 1496–1508.
29. Ovchinnikov, I., Rubin, A. & Swergold, G. D. (2002). Tracing the LINEs of human evolution. *Proc. Natl Acad. Sci. USA*, **99**, 10522–10527.
30. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
31. Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F. & Kazazian, H. H., Jr (1991). Isolation of an active human transposable element. *Science*, **254**, 1805–1808.
32. Miyamoto, M. M., Slightom, J. L. & Goodman, M. (1987). Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. *Science*, **238**, 369–373.
33. Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucl. Acids Res.* **8**, 1499–1504.
34. Feng, Q., Moran, J. V., Kazazian, H. H., Jr & Boeke, J. D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, **87**, 905–916.
35. Morrish, T. A., Gilbert, N., Myers, J. S., Vincent, B. J., Stamato, T. *et al.* (2002). DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nature Genet.* **31**, 159–165.
36. Moore, J. K. & Haber, J. E. (1996). Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. *Nature*, **383**, 644–646.
37. Teng, S. C., Kim, B. & Gabriel, A. (1996). Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks. *Nature*, **383**, 641–644.
38. Szak, S. T., Pickeral, O. K., Makalowski, W., Boguski, M. S., Landsman, D. *et al.* (2002). Molecular archeology of L1 insertions in the human genome. *Genome Biol.* **3** research0052.
39. Grimaldi, G., Skowronski, J. & Singer, M. F. (1984). Defining the beginning and end of KpnI family segments. *EMBO J.* **3**, 1753–1759.
40. Ostertag, E. M. & Kazazian, H. H., Jr (2001). Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* **11**, 2059–2065.
41. Arcot, S. S., Wang, Z., Weber, J. L., Deininger, P. L. & Batzer, M. A. (1995). Alu repeats: a source for the genesis of primate microsatellites. *Genomics*, **29**, 136–144.
42. Economou, E. P., Bergen, A. W., Warren, A. C. & Antonarakis, S. E. (1990). The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. *Proc. Natl Acad. Sci. USA*, **87**, 2951–2954.
43. Ovchinnikov, I., Troxel, A. B. & Swergold, G. D. (2001). Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. *Genome Res.* **11**, 2050–2058.
44. Batzer, M. A., Gudi, V. A., Mena, J. C., Foltz, D. W., Herrera, R. J. *et al.* (1991). Amplification dynamics of human-specific (HS) Alu family members. *Nucl. Acids Res.* **19**, 3619–3623.
45. Guo, S. W. & Thompson, E. A. (1992). Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361–372.

46. Moran, J. V. (1999). Human L1 retrotransposition: insights and peculiarities learned from a cultured cell retrotransposition assay. *Genetica*, **107**, 39–51.
47. Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T. *et al.* (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science*, **235**, 1616–1622.
48. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331.
49. Batzer, M. A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D. H. *et al.* (1994). African origin of human-specific polymorphic Alu insertions. *Proc. Natl Acad. Sci. USA*, **91**, 12288–12292.
50. Stoneking, M., Fontius, J. J., Clifford, S. L., Soodyall, H., Arcot, S. S. *et al.* (1997). Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.* **7**, 1061–1071.
51. Jorde, L. B., Watkins, W. S., Bamshad, M. J., Dixon, M. E., Ricker, C. E. *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.
52. Perna, N. T., Batzer, M. A., Deininger, P. L. & Stoneking, M. (1992). Alu insertion polymorphism: a new type of marker for human population studies. *Hum. Biol.* **64**, 641–648.
53. 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.
54. Ausabel, F. M., Brent, R., Kingston, M. E., Moore, D. D. & Seidman, J. G. (1987). *Current Protocols in Molecular Biology*, Wiley, New York.
55. Jurka, J., Klonowski, P., Dagman, V. & Pelton, P. (1996). CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20**, 119–121.

Edited by J. Karn

(Received 4 November 2002; received in revised form 18 December 2002; accepted 20 December 2002)

Supplementary Material is available on Science Direct