1-1-2002

A comprehensive analysis of recently integrated human Ta L1 elements

Jeremy S. Myers
Louisiana State University

Bethaney J. Vincent
Louisiana State University

Hunt Udall
LSUHSC School of Medicine

W. Scott Watkins
University of Utah Health Sciences

Tammy A. Morrish
University of Michigan Medical School

See next page for additional authors

Recommended Citation

Myers, J., Vincent, B., Udall, H., Watkins, W., Morrish, T., Kilroy, G., Swergold, G., Henke, J., Henke, L., Moran, J., Jorde, L., & Batzer, M. (2002). A comprehensive analysis of recently integrated human Ta L1 elements. *American Journal of Human Genetics, 71* (2), 312-326. https://doi.org/10.1086/341718

Am. J. Hum. Genet. 71:312–326, 2002

A Comprehensive Analysis of Recently Integrated Human Ta L1 Elements

Jeremy S. Myers,[1,2,*] Bethaney J. Vincent,[1,2,*] Hunt Udall,[2] W. Scott Watkins,[3] Tammy A. Morrish,[4] Gail E. Kilroy,[1] Gary D. Swergold,[5] Jurgen Henke,[6] Lotte Henke,[6] John V. Moran,[4] Lynn B. Jorde,[3] and Mark A. Batzer[1,2]

[1]Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, Baton Rouge; [2]Departments of Pathology, Genetics, Biochemistry, and Molecular Biology, Stanley S. Scott Cancer Center, Neuroscience Center of Excellence, Louisiana State University Health Sciences Center, New Orleans; [3]Department of Human Genetics, University of Utah Health Sciences Center, Salt Lake City; [4]Departments of Human Genetics and Internal Medicine, University of Michigan Medical School, Ann Arbor; [5]Division of Molecular Medicine, Department of Medicine, Columbia University, New York; and [6]Institut für Blutgruppenforschung, Cologne

The Ta (transcribed, subset a) subfamily of L1 LINEs (long interspersed elements) is characterized by a 3-bp ACA sequence in the 3′ untranslated region and contains ~520 members in the human genome. Here, we have extracted 468 Ta L1Hs (L1 human specific) elements from the draft human genomic sequence and screened individual elements using polymerase-chain-reaction (PCR) assays to determine their phylogenetic origin and levels of human genomic diversity. One hundred twenty-four of the elements amenable to complete sequence analysis were full length (~6 kb) and have apparently escaped any 5′ truncation. Forty-four of these full-length elements have two intact open reading frames and may be capable of retrotransposition. Sequence analysis of the Ta L1 elements showed a low level of nucleotide divergence with an estimated age of 1.99 million years, suggesting that expansion of the L1 Ta subfamily occurred after the divergence of humans and African apes. A total of 262 Ta L1 elements were screened with PCR-based assays to determine their phylogenetic origin and the level of human genomic variation associated with each element. All of the Ta L1 elements analyzed by PCR were absent from the orthologous positions in nonhuman primate genomes, except for a single element (L1HS72) that was also present in the common (*Pan troglodytes*) and pygmy (*P. paniscus*) chimpanzee genomes. Sequence analysis revealed that this single exception is the product of a gene conversion event involving an older preexisting L1 element. One hundred fifteen (45%) of the Ta L1 elements were polymorphic with respect to insertion presence or absence and will serve as identical-by-descent markers for the study of human evolution.

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 ~500,000 L1 LINEs (long interspersed elements) (Smit 1999; Prak and Kazazian 2000; Lander et al. 2001). L1 elements are restricted to mammals, having expanded as a repeated DNA sequence family over the past 100–150 million years (Smit et al. 1995). Full-length L1 elements are ~6 kb long and amplify via an RNA intermediate in a process known as "retrotransposition." L1 integration likely occurs by a mechanism termed "target-primed reverse transcription" (Luan et al. 1993; Kazazian and Moran 1998). This mechanism of mobilization provides two useful landmarks for the identification of L1Hs (L1 human specific) inserts: an endonuclease-related cleavage site (Jurka 1997; Cost and Boeke 1998; Cost et al. 2001) and direct repeats or target site duplications flanking newly integrated elements (Fanning and Singer 1987; Kazazian 2000).

L1 retrotransposons have had a significant impact on the human genome, through recombination (Fitch et al. 1991), alteration of gene expression (Yang et al. 1998; Rothbarth et al. 2001), and de novo insertions that disrupt ORFs and splice sites resulting in human disease (Kazazian et al. 1988; Kazazian 1998; Kazazian and Moran 1998). L1 elements are also able to transduce adjacent genomic sequences at their 3′ end, facilitating exon shuffling (Boeke and Pickeral 1999; Moran et al. 1999; Goodier et al. 2000). In addition, individual mobile elements may undergo post-integration gene conversion events in which short DNA sequences are exchanged by an undefined mechanism, thereby altering the levels of SNP associated with the individual L1 elements (Hardies et al. 1986). Thus, LINEs have exerted a significant influence on the architecture of the human genome.

Even though there are ~500,000 L1 elements in the

Received April 2, 2002; accepted for publication May 9, 2002; electronically published June 17, 2002.

Address for correspondence and reprints: Dr. Mark A. Batzer, Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803. E-mail: mbatzer@lsu.edu

* The first two authors contributed equally to this work.

0002-9297/2002/7102-0011$15.00

human genome, only a limited subset of L1 elements appear to be capable of retrotransposition (Moran et al. 1996; Sassaman et al. 1997). As a result of the limited amplification potential of this diverse gene family, a series of discrete subfamilies of L1 elements exists within the human genome (Deininger et al. 1992; Smit et al. 1995). Each of the L1 subfamilies appears to have amplified within the human genome at different times in primate evolution, making them different genetic ages (Deininger et al. 1992; Smit et al. 1995). The most recently integrated L1 elements within the human genome share a common 3-bp diagnostic sequence within the 3′ UTR, and they comprise almost all of the de novo disease-associated L1 elements within the human genome, as well as several elements that have been shown to be capable of retrotransposition in cell culture (Kazazian and Moran 1998; Boissinot et al. 2000; Sheen et al. 2000). This subfamily was first identified in human teratocarcinoma cells and has been collectively termed "Ta" (for transcribed, subset a) (Skowronski et al. 1988). Some members of the L1 Ta subfamily have inserted in the human genome so recently that they are polymorphic with respect to insertion presence/absence (Boissinot et al. 2000; Sheen et al. 2000). The L1 insertion polymorphisms are a useful source of identical-by-descent variation for the study of human population genetics (Boissinot et al. 2000; Santos et al. 2000; Sheen et al. 2000). Here, we report the analysis of the Ta subfamily of L1 elements from the draft sequence of the human genome.

Material and Methods

Cell Lines and DNA Samples

The cell lines used to isolate primate DNA samples were as follows: human (*Homo sapiens*) HeLa (American Type Culture Collection [ATCC] number CCL2), common chimpanzee (*Pan troglodytes*) Wes (ATCC number CRL1609), pygmy chimpanzee (*P. paniscus*) (Coriell Cell Repository number AG05253), gorilla (*Gorilla gorilla*) Lowland Gorilla (Coriell Cell Repository number AG05251B), green monkey (*Cercopithecus aethiops*) (ATCC number CCL70), and owl monkey (*Aotus trivirgatus*) (ATCC number CRL1556). 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, Asian or Alaskan native, and Egyptian population groups were isolated from peripheral blood lymphocytes (Ausabel et al. 1987), as described elsewhere (Stoneking et al. 1997).

Computational Analyses

The draft sequence of the human genome was screened using the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990), available at the National Center for Biotechnology Information genomic BLAST Web site. A 19-bp oligonucleotide (5′-CCTAATGCTAGATGACACA-3′) that is diagnostic for the L1Hs Ta subfamily was used to query the human genome database with the following optional parameters: filter none and advanced options −e 0.01, −v 600, and −b 600. Copy-number estimates were determined from BLAST search results. Sequences that contained exact matches were subjected to additional analysis as outlined below.

A sequence region of 9,000–10,000 bp, including the match and 1,000–2,000 bp of flanking unique sequence, was annotated using RepeatMasker (version 7/16/00), from the University of Washington Genome Center, or Censor, from the Genetic Information Research Institute (Jurka et al. 1996). These programs annotate repeat-sequence content and were used to confirm the presence of L1Hs elements and regions of unique sequence flanking the elements. PCR primers flanking each L1 element were designed using Primer3 software, available from the Whitehead Institute for Biomedical Research, and were complementary to the unique sequence regions flanking each L1 element. The resultant primers were screened, by standard nucleotide-nucleotide BLAST (blastn), against the nonredundant (nr) and high-throughput (htgs) sequence databases, to ensure that they resided in unique DNA sequences. Primers that resided in repetitive sequence regions were discarded, and, if possible, new primers were then designed. A complete list of all the L1 elements that were identified using this approach and supplemental material from this manuscript are available from the Batzer Lab Web site, in the "Publications" section. Individual L1 DNA sequences were aligned using MegAlign, with the Clustal V algorithm and the default settings (DNAstar, version 5.0 for Windows), followed by manual refinement.

PCR Amplification

PCR amplification of 262 individual L1 elements was performed in 25-μl reactions that contained 50–100 ng of template DNA; 40 pmol of each oligonucleotide primer (see table A1, available online only); 200 μM of deoxyribonucleoside triphosphates, in 50 mM KCl and 10 mM Tris-HCl (pH 8.4); 1.5 mM $MgCl_2$; and 1.25 U of *Taq* DNA polymerase. Each sample was subjected to the following amplification conditions for 32 cycles: an initial denaturation at 94°C for 150 s, 1 min denaturation at 94°C, and 1 min at the annealing temperature (specific for each locus, as shown in table 1 and appendix A, available online only), followed by extension at 72°C for 10 min. For analysis, 20 μl of each sample was fractionated on a 2% agarose gel with 0.05 μg/ml ethidium bromide. PCR products were directly visualized using UV fluorescence. The human genomic diversity asso-

Table 1

Summary of Ta L1 Element Computational and PCR Analysis

Classification	No. of Elements
Successful PCR analysis	262
L1 elements inserted in other repeats	137
L1 elements located at the end of sequencing contigs	69
Total Ta L1 elements analyzed	468

NOTE.—A full summary of GenBank accession numbers, PCR primers and conditions, and PCR amplicon sizes for these loci is shown in table A1, available online only, and is also available at the Batzer Lab Web site.

ciated with each Ta L1 element was determined by the amplification of 20 individuals from each of four geographically distinct populations (African American, Asian or Alaskan native, European German, and Egyptian).

Cloning and Sequence Analysis

L1 element–related PCR products were cloned using the Invitrogen TOPO TA Cloning Kit, according to the manufacturer's instructions, and were sequenced using an Applied Biosystems 3100 automated DNA sequencer, by the chain-termination method (Sanger et al. 1977). The DNA sequence for the common and pygmy chimpanzee orthologs of L1HS72 were assigned GenBank accession numbers AF489459 and AF489460, respectively. Additional diverse human sequences from L1HS72 were assigned GenBank accession numbers AF489450–AF489458. DNA sequences derived from L1 pre-integration sites were assigned GenBank accession numbers AF461364, AF461365, AF461368–AF461383, AF461386, and AF461387.

Results

L1 Ta Subfamily Copy Number and Age

To identify recently integrated Ta L1 elements from the human genome, we searched the draft sequence of the human genome (BLASTN database, version 2.2.1), using BLAST (Altschul et al. 1990) with an oligonucleotide that is complementary to a highly conserved sequence in the 3′ UTR of Ta L1 elements. This 19-bp query sequence (CCTAATGCTAGATGACACA) includes the Ta subfamily–specific diagnostic mutation ACA at its 3′ end at positions 5930–5932 relative to L1 retrotransposable element–1 (Dombroski et al. 1991). We identified 468 unique Ta 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 ~520 elements. Of the 468 elements retrieved, 69 resided at the end of sequence contigs and were not amenable to additional in vitro wet-bench analysis. Of the 399 remaining elements, 124 (31%) of the elements were essentially full length, and the remaining 275 were truncated to variable lengths. Alignment and sequence analysis of the full-length elements revealed that 44 contained two intact ORFs and therefore may be capable of retrotransposition. This estimate of putative retrotransposition-competent L1 elements is in good agreement with the initial analysis of the draft sequence of the human genome (Lander et al. 2001).

The ages of L1 elements can be determined by the level of sequence divergence from the subfamily consensus sequence by use of a neutral mutation rate for primate noncoding sequence of 0.15% per million years (Miyamoto et al. 1987). The mutation rate is known to be ~10 times greater for CpG bases as compared to non-CpG bases, as a result of the spontaneous deamination of 5-methyl cytosine (Bird 1980). Thus, two age estimates that are based on CpG and non-CpG mutations can be calculated for the Ta subfamily of L1 elements. A total of 89,929 bp from the 3′ UTR of 459 Ta L1Hs elements were analyzed, and L1 elements characterized elsewhere were excluded from this analysis—along with nine elements that, according to the nucleotide present at position 6015 in the 3′ UTR of the elements, do not technically belong to the Ta subfamily (Ovchinnikov et al. 2001). Three hundred thirty-one total nucleotide substitutions were observed. Of these, 263 were classified as non-CpG mutations against the backdrop of 88,141 total non-CpG bases, thereby producing a non-CpG mutation density of 0.002984. Based on the non-CpG mutation density and a neutral rate of evolution (0.002984/0.0015), the average age of the Ta L1 elements was 1.99 million years. A total of 68 CpG mutations were found across these 459 L1 elements from 1,788 total CpG nucleotides, thereby yielding a CpG-mutation rate of 0.038031. With the expectation that the CpG mutation rate is ~10-fold higher than the non-CpG mutation rate, the approximate age (obtained using the CpG mutation density) of the L1Hs Ta subfamily is 2.54 million years. These estimates are in good agreement with one another, as well as with previous estimates derived from an analysis of a small number of Ta L1 elements (Boissinot et al. 2000).

Nine of the 468 elements analyzed do not technically belong to the Ta subfamily of L1 elements, on the basis of a single-nucleotide substitution (L1HS19, -72, -274, -309, -318, -325, -390, -399, and -493) that is also considered diagnostic for the L1 Ta subfamily. Although they all have the 19-bp query sequence ending in ACA in the 3′ UTR at positions 5930–5932, they lack a G at position 6015 (Ovchinnikov et al. 2001) and instead contain an A at that position, which is a diagnostic feature found in older primate-specific L1PA10–L1PA2 subfamilies (Smit et al. 1995). Thus, these elements may be Ta L1

elements that have undergone fortuitous single-base substitutions of the ancestral nucleotide, may be Ta L1 elements that have undergone backward gene-conversion events, or may simply be older, "pre-Ta" L1 elements that were generated by a source gene (or source genes) that did not contain this diagnostic base. To determine the effect that the Ta versus non-Ta designation has on the calculated age estimate, we examined a total of 1,807 bp from the 3′ UTRs of these nine elements. There were 27 non-CpG mutations from a total of 1,771 non-CpG bases, thereby yielding a mutation density of 27/1,771, or 0.015246. Dividing by the neutral rate of evolution for primate noncoding sequence (0.015246/0.0015), we arrive at an estimated age of 10.16 million years. This is significantly older than the average age of 2.26 million years that was calculated from the larger data set (i.e., the data set of Ta L1 elements only). The CpG mutation density in the elements was also calculated. There were 2 CpG mutations from 36 CpG bases, thereby producing a CpG mutation density of 2/36, or 0.056. We divide this figure by the projected CpG mutation rate (0.056/0.015), arriving at an estimated age of 3.73 million years. This figure is lower than the non-CpG mutation rate, but it still suggests that these elements are at least twice as old as their true Ta counterparts. In addition, all but one of these Ta L1 elements (L1HS493) were monomorphic for the presence of the L1 element in the human population. Thus, the higher levels of nucleotide diversity and the absence of associated insertion polymorphism of eight of these L1 elements are consistent with their being older members of the L1 Ta subfamily, whereas L1HS493 may be the product of a gene-conversion event.

The nucleotide-sequence substitution patterns were further examined with respect to the levels of presence/absence of insertion polymorphism associated with each of the L1 elements (as outlined in detail below, in the "L1 Element–Associated Human Genomic Diversity" subsection). The 3′ UTRs of 139 fixed-present elements were analyzed for both CpG and non-CpG mutations and had an estimated average age of 2.45 million years. This calculation yields an age that is somewhat older than the average age that was predicted for the subfamily as a whole—a finding that was expected, since these elements are thought to have inserted during the early stages of L1Hs Ta expansion in the human genome, such that they have become fixed across diverse human populations. Similar calculations were repeated for the high-frequency, intermediate-frequency, and low-frequency L1 Ta insertion polymorphisms, with average ages of 2.24, 2.06, and 1.69 million years, respectively. Although the age differences across different insertion frequencies are not significantly different (P values >.05) when tested with a one-tailed t test, they do suggest a progressive decrease in the calculated age of each group, with corresponding decreases in insertion frequency. This is exactly what would be expected under a model in which newer elements arose more recently and have lower allele frequencies in the human population.

L1 Element–Associated Human Genomic Diversity

Of the 468 Ta L1Hs elements isolated *in silico,* 262 were further analyzed using a PCR-based assay and flanking unique sequence primers as described elsewhere (Sheen et al. 2000) (table 1; also see appendix A, available online only). The remaining elements were not suitable for further analysis, for various reasons. Some (137) of the L1 elements were inserted into other repetitive regions of the genome such that flanking unique sequence PCR primers could not be designed. Sixty-nine additional elements resided at the end of sequencing contigs in GenBank, so the lack of flanking unique sequence information made PCR-primer design in this region impossible. Three elements—L1HS17, L1HS47, and L1HS63—produced inconclusive PCR results because of the amplification of paralogous genomic sequences as described elsewhere (Batzer et al. 1991). Another five elements produced nonspecific PCR results, and they were excluded from further analysis. Thirty-six of the Ta L1 elements mapped to chromosome X, and 10 mapped to chromosome Y (table 1; also see appendix A, available online only). All of the Ta L1 elements from chromosomes X and Y were tested using human DNA samples in which the gender had been determined using a PCR-based assay that was described elsewhere (Eng et al. 1994). The human genomic diversity associated with the autosomal and sex-linked Ta L1 elements is summarized in table 2 and appendix A, available online only.

A high degree (45%) of insertion polymorphism was found in the 254 (i.e., 262 − 8) remaining elements that were subjected to the two-step PCR-based assay across 80 individuals from four geographically diverse human populations (table 2; also see appendix A, available online only). One hundred thirty-nine of the Ta L1 elements were fixed present, meaning that every individual tested was homozygous (i.e., +/+) for the presence of the L1 repeat. These elements are likely to be slightly older than their polymorphic counterparts, having inserted into the human genome prior to the migration of humans from Africa. By contrast, 115 of the elements assayed by PCR were polymorphic, to some degree, in the populations that were surveyed. A survey of human genomic diversity associated with a severely truncated L1 element is shown in figure 1. A sample of the human genomic diversity associated with relatively long L1 insertion polymorphism is shown in figure 2. Thirty-seven of the Ta L1 elements were high-frequency insertion polymorphisms with an L1 allele frequency that was >0.67, so that most of the individuals were homozygous for the presence of the L1

Table 2

Summary of Ta L1 Element–Associated Human Genomic Diversity

Classification	No. of Elements
Autosomal Ta L1 elements:	
HF	36
IF	55
LF	15
VLF/fixed absent	3
Fixed present	129
X-linked Ta L1 elements:	
HF	1
IF	1
LF	4
VLF/fixed absent	0
Fixed present	8
Y-linked Ta L1 elements:	
Polymorphic	0
Fixed present	2

NOTE.—The L1 Ta insertion polymorphisms are classified according to allele frequency as high-frequency (HF) (present in more than 2/3 but not in all chromosomes tested), intermediate-frequency (IF) (present in more than 1/3 of chromosomes tested but in no more than 2/3 of the chromosomes), low-frequency (LF) (present in no more than 1/3 of the chromosomes tested), or very-low-frequency (VLF) (or "private") insertion polymorphisms. A full summary of the genotypes for each locus, L1 allele-frequency data, and heterozygosity values is shown in tables A2 and A3, available online only, and is also available at the Batzer Lab Web site.

element. Fifty-six of the polymorphic elements were intermediate frequency, with an L1 allele frequency >0.33 but <0.67 across the diverse human populations sampled. Nineteen of the 254 elements had insertion allele frequencies <0.33, and these were termed "low-frequency insertion polymorphisms." These elements include some of the youngest members of the subfamily, having inserted into the human genome so recently that the element appears in the genomes of only a handful of individuals who were screened in our assay. Three Ta L1 elements—L1HS44, L1HS287, and L1HS373—appeared to be absent from the genomes of all the individuals tested, and one of these (L1HS373) is full length and has two functional ORFs, suggesting that it may be retrotransposition competent. Previous experiments with *Alu* elements have shown not only that these types of elements are indeed present within the genomic clone that was sequenced as part of the human genome project but also that they represent relatively rare, "private" mobile-element insertion polymorphisms (Carroll et al. 2001).

Overall, the unbiased heterozygosity values across all of the L1 elements subjected to PCR analysis were similar across the four populations, with values of 0.265 in African Americans, 0.233 in Asians, 0.252 in European Germans (i.e., white Germans of European descent), and 0.250 in Egyptians (table 2; also see appendix A, available online only). However, several of the polymorphic elements individually exhibited unbiased heterozygosity values that approached 0.5, the theoretical maximum for biallelic loci. A subset of 31 of the 115 L1 insertion polymorphisms are, to some degree, population specific, meaning that insertion frequencies differ by ⩾25% in one of the tester populations, relative to the other three populations that were surveyed. Detailed analysis of the human genomic variation associated with the polymorphic L1 elements will prove useful for the study of human population genetics.

To determine if the L1 insertion polymorphisms were in Hardy-Weinberg equilibrium (HWE), we performed a total of 460 χ^2 tests for goodness of fit. A total of 77 deviations from Hardy-Weinberg expectations were observed in the comparisons. However, 73 of the deviations were the result of low expected numbers. The remaining four tests that deviated from HWE did not cluster by locus or population. A total of 23 deviations from HWE would be expected by chance alone at the 0.5% significance interval. In addition, we applied Fisher's exact test to the data, using the Genetic Data Analysis program. The test yielded only 22 of 436 significant comparisons, which is approximately what would be expected on the basis of chance alone. By Fisher's exact test, only 6 of the 436 comparisons were significant at the .01 level, and they did not cluster across all populations at any locus tested. Therefore, we conclude that these L1 insertion polymorphisms do not significantly depart from HWE.

Phylogenetic Origin

Almost all of the Ta L1 elements analyzed using PCR were located in the human genome and were absent from the orthologous positions within nonhuman primate genomes. Only a single truncated L1 element (L1HS72) produced unexpected results when subjected to the initial PCR by use of external flanking primers and nonhuman primate DNA as a template. The 825-bp amplicon that corresponded to the L1HS72 insertion was found in loci in all 80 human individuals tested, as well as in the orthologous loci from the common chimpanzee and pygmy chimpanzee genomes (fig. 3*A*). However, the gorilla, green monkey, and owl monkey only amplified the small PCR product corresponding to the empty allele or pre-integration site (fig. 3*A*). Subsequent PCRs by use of the internal subfamily-specific ACA primer and the 3′ flanking primer across the same DNA templates produced a characteristic L1 filled-site amplicon only in the human individuals and not in any of the nonhuman primate genomes (chimpanzee, gorilla, green monkey, and owl monkey). It appeared that we had potentially isolated a Ta L1 element that inserted into the genome before the divergence of humans from African apes, but the second PCR by use of the internal subfamily-specific

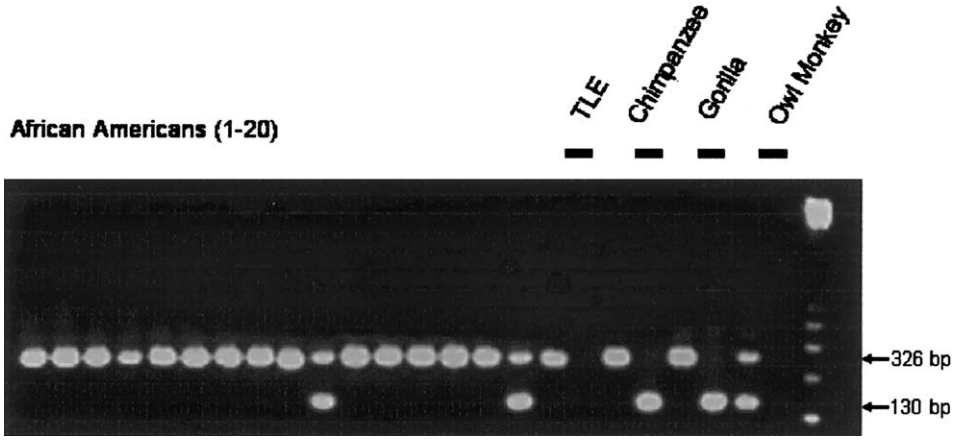


Figure 1 Human diversity associated with a truncated Ta L1Hs element, as shown by an agarose gel chromatograph of the PCR products from a survey of the human genomic variation associated with L1HS7. Amplification of the pre-integration site of this locus generates a 130-bp PCR product; amplification of a filled site generates a 326-bp product (by use of 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 only the African American samples shown here; the control samples (*gray lines*) were TLE buffer (i.e., 10 mM Tris-HCl:0.1 mM EDTA), common chimpanzee, gorilla, and owl monkey DNA templates. Most of the individuals surveyed were homozygous for the presence of the L1 element; in addition, this particular L1 element was absent from the genomes of nonhuman primates.

ACA primer and the 3′ flanking primer again produced the expected product that corresponded to the presence of this Ta L1 element only in humans. These data suggest that there is a difference in the sequence structure of this L1 element in the human genome, as compared to the common and pygmy chimpanzee genomes, which contained putative Ta L1 filled alleles.

Gene Conversion

To precisely define the sequence structure of the L1HS72 locus, we cloned and sequenced, for further analysis, the PCR amplicons from several human genomes, as well as those from the common chimpanzee and the pygmy chimpanzee (fig. 3*B*). Sequence analysis of the orthologous sites from the common and pygmy chimpanzee genomes revealed the presence of an older, primate-specific L1 element that had the greatest sequence identity to the L1PA3 subfamily (fig. 3*B*). Interestingly, this L1 element shared identical target-site duplications with that of the Ta L1 element that was present in the human samples that we studied. Both the human sequence and the chimpanzee sequence also contained many of the diagnostic mutations characteristic of an L1PA3 element. However, only the human L1 sequences contained the Ta diagnostic ACA mutation at positions 5930–5932 in the 3′ UTR. The common and pygmy chimpanzee sequences contained GAT at this position and an additional A mutation at diagnostic position 6015, both of which are characteristic of older L1PA elements (L1PA6–L1PA2). The most likely explanation for the presence of the L1Hs Ta ACA sequence in the human L1 element is a forward gene-conversion event that affected a preexisting older L1 element at this locus. To further investigate the putative gene conversion at this locus, we cloned and sequenced alleles derived from African American, Asian, European German, and Egyptian genomes. Although there was a limited sample size, all nine individuals who were sequenced contained the ACA sequence, and at least four samples (European Germans 1 and 2 and Egyptians 2 and 3) contained SNPs, three of which occur at a specific CpG dinucleotide (fig. 3*B*). Therefore, we conclude that gene-conversion events have altered the L1 Ta subfamily–specific diagnostic nucleotide positions at this locus within the human lineage.

To begin to examine the level of gene conversion across the entire Ta subfamily, we examined multiple-sequence alignments of the 459 Ta L1Hs elements. Close inspection of the multiple-sequence alignment revealed some highly variable sequence features that were unexpected among such a young L1 subfamily, in which we would expect low levels of nucleotide divergence. It appears that many of the single-base substitutions in Ta L1 elements are not completely random mutation events. In fact, it became clear that a substantial number of the elements possessed specific mutations that are diagnostic for older L1PA primate-specific elements in addition to the younger diagnostic mutations. These mosaic elements all possessed the 19-bp Ta L1 consensus sequence, but they also contained short tracts of sequence diagnostic for other L1 subfamilies.

There are two possible explanations for the presence

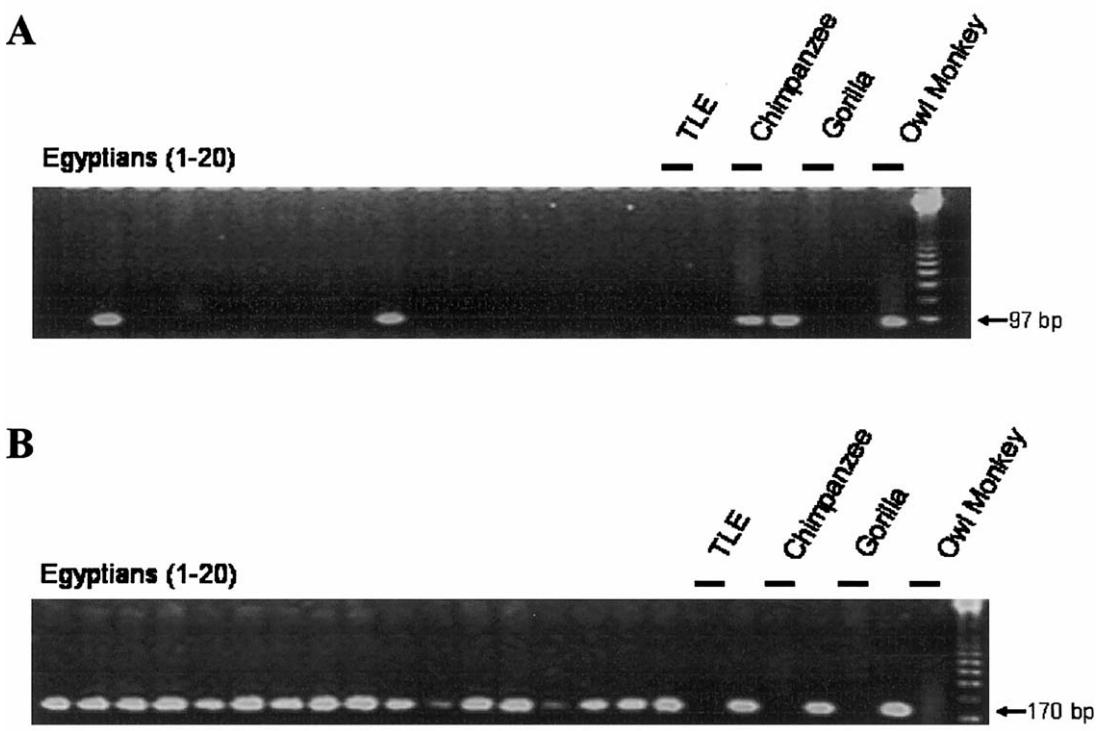
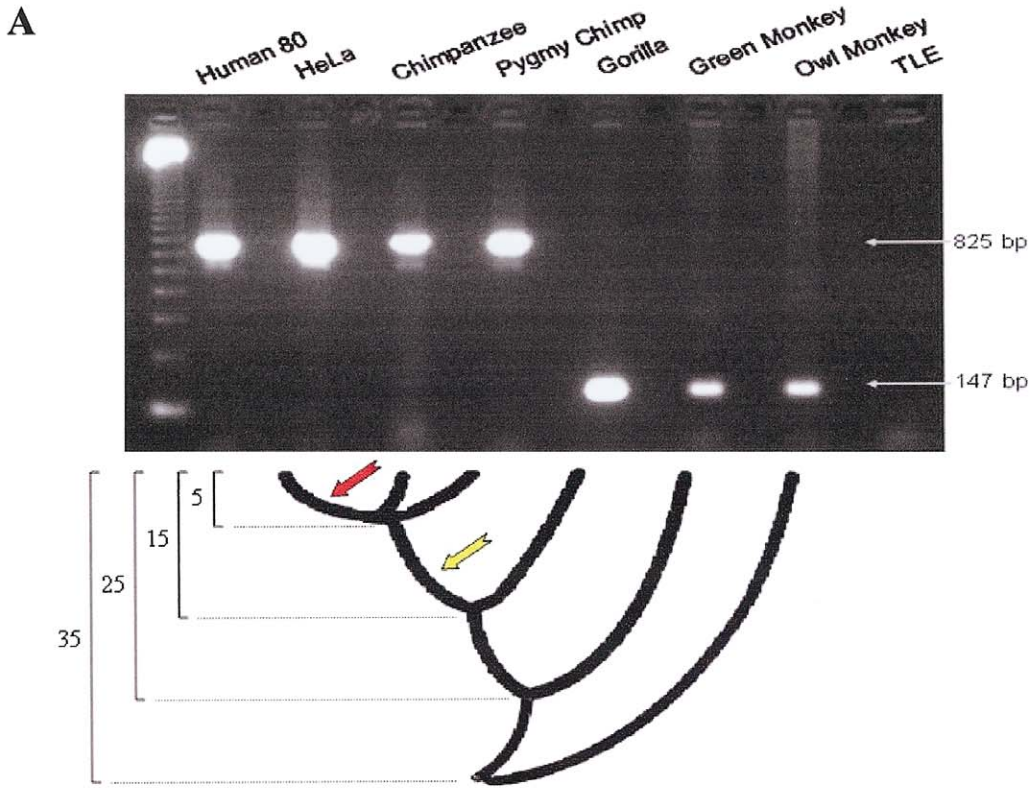


Figure 2 Human diversity associated with a long L1Hs Ta insertion polymorphism, as shown by an agarose gel chromatograph of the PCR products from a survey of the human genomic variation associated with L1HS364. Because of the size (~6,000 bp) of this L1 element, two separate PCRs are performed to genotype individual samples. In the first reaction, flanking unique sequence primers were used to genotype the empty alleles (*A*); amplification of empty alleles from this locus generates a 97-bp PCR product. In the second reaction, a Ta subfamily–specific internal primer termed "ACA" and the 3′ flanking unique sequence primer were used to genotype filled sites (*B*); the amplification of filled sites generates a 170-bp product. 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 only the Egyptian samples shown here; the control samples (*black lines*) were TLE buffer, common chimpanzee, gorilla, and owl monkey DNA templates. This particular L1 insertion polymorphism is a high-frequency insertion polymorphism, and most of the individuals surveyed have L1 filled chromosomes.

of these mosaic elements. The first theory is that L1Hs Ta source genes, while acquiring the young diagnostic mutations of the L1Hs Ta subfamily, also retained many of the other diagnostic mutations of their older L1 subfamily progenitors. Over time, this gave rise to elements with combinations of young and old mutations, as proposed in the master-gene theory of LINE and short-interspersed-element (SINE) amplification (Deininger et al. 1992). The second theory is that some of these mosaic elements are products of gene-conversion events—that is, a nonreciprocal transfer of sequence between a pair of nonallelic genomic DNA sequences, such as interspersed repeats. The donor sequence is unchanged, and the recipient sequence gains some of the donor sequence; alternatively, a nonintegrated LINE cDNA may also serve as the donor sequence for the gene conversion. Gene conversion between SINEs and LINEs is a significant influence on the genomic landscape of young *Alu* elements, creating hybrid sequence mosaics of the various mobile-element subfamilies (Batzer et al. 1995; Kass et al. 1995; Roy et al. 2000; Roy-Engel et al. 2001, 2002). Gene conversion may contribute to as much as 10%–20% of the sequence variation between recently integrated *Alu* elements (Roy et al. 2000). It is likely that the same process may also alter the sequence diversity of L1 elements, since they are also part of a large, nearly identical multigene family and since they have previously been shown to have undergone limited gene conversion (Hardies et al. 1986; Burton et al. 1991). Unfortunately, the vast majority of primate L1 subfamily structure has only been deduced computationally and has not been verified at the wet bench, to precisely define the expansion of L1 elements in a phylogenetic context. Therefore, it is currently not possible to accurately estimate the level of gene conversion between L1 elements within the genome.

Sequence Diversity

One hallmark of L1 integration is the generation of target-site duplications flanking newly integrated ele-



B

Position information	ORF 2 (5536,5539)	(5591)	(5713)	(5745)	(5767)	(5791)	3′UTR (5828)	(5851)	(5864)	(5871)	(5879)	(5930-5932)	(5946)	(5990)	(6015)	
	[AAGATTCTA]															[AAGATTCTA]
L1HS-Ta consensus	5′	(T..G)	*	C	A	C	G	C	..	T	T	G	G	C	ACA	CpG	A	G	3′
L1HS-Ta0 consensus	5′	(G..C)	*	C	A	C	G	C	..	T	T	G	G	C	ACA	CpG	A	G	3′
L1HS-Ta1 consensus	5′	(T..G)	*	C	A	C	G	C	..	T	T	G	G	C	ACA	CpG	A	G	3′
L1HS72 GenBank	5′	TSD	(G..C)	*	T	G	T	A	T	..	A	C	A	T	G	T	TSD		3′
African American 1	5′	TSD	(G..C)	*	T	G	T	A	T	..	A	C	A	T	G	T	TSD		3′
African American 2	5′	TSD	(G..C)	*	T	G	T	A	T	..	A	C	A	T	G	T	TSD		3′
Asian 1	5′	TSD	(G..C)	*	T	G	T	A	T	..	A	C	A	T	G	T	TSD		3′
Asian 2	5′	TSD	(G..C)	*	T	G	T	A	T	..	A	C	A	T	G	T	TSD		3′
European German 1	5′	TSD	(G..C)	*	T	G	T	A	T	..	A	C	A	T	G	T	TSD		3′
European German 2	5′	TSD	(G..C)	*	T	G	T	A	T	..	A	C	A	T	G	T	TSD		3′
Egyptian 1	5′	TSD	(G..C)	*	T	G	T	A	T	..	A	C	A	T	G	T	TSD		3′
Egyptian 2	5′	TSD	(G..C)	*	T	G	T	A	T	..	A	C	G	T	G	T	TSD		3′
Egyptian 3	5′	TSD	(G..C)	*	T	G	T	A	T	..	A	C	A	T	G	T	TSD		3′
Chimpanzee	5′	TSD	(G..C)	*	T	G	T	T	..	A	C	T	G	GAT		T	TSD		3′
Pygmy Chimpanzee	5′	TSD	(G..C)	*	T	G	T	T	..	A	C	T	G	GAT		T	TSD		3′

Figure 3 L1HS72 gene conversion. *A,* Agarose gel chromatograph of the PCR products derived from the amplification of L1HS72 in a series of human and nonhuman primate genomes, with a schematic of the primate evolutionary tree over the past 35 million years shown below. The yellow notched arrow represents the approximate time period when the L1HS72 element first integrated, and the red notched arrow represents the approximate time period of the gene conversion event of the preexisting L1 element. The fragment-length marker is a 123-bp ladder. *B,* Sequence alignment generated by sequencing the L1HS72 amplicons from nine diverse humans. Sequences are compared relative to L1Hs Ta consensus sequence and the L1HS72 sequence obtained from GenBank with only the diagnostic bases shown and positions reported relative to L1 retrotransposable element–1 (Dombroski et al. 1991). The G and C at positions 5536 and 5539 are indicative of the Ta-0 subset, whereas the Ta-1 subset has T and G at these nucleotides (Boissinot et al. 2000). The G at position 6015 (in addition to the ACA at positions 5930–5932) is diagnostic for the L1Hs Ta subfamily (Ovchinnikov et al. 2001). The target-site duplication sequence (TSD) is shown in brackets. The mosaic elements seen in the human samples are believed to be the result of at least one gene conversion, some time after the divergence of humans from the great apes (approximately five million years ago), of a preexisting L1 element with a younger L1Hs element. In the representation of nucleotides, different colors are used to denote conserved sequences and sequence variations between samples: green denotes bases unique to the common and pygmy chimpanzee genomes; blue denotes nucleotides unique to the human samples; orange denotes shared bases conserved between the common chimpanzee, pygmy chimpanzee, and human samples; and red denotes SNPs, within L1HS72, in the human population.

ments. Two thousand base pairs of flanking sequence on each side of the element were searched for target-site duplications. Direct repeats >10 bp long are considered to be clear target-site duplications. Of the 399 elements (i.e., a total of 468 elements minus the 69 elements located at the end of sequencing contigs), we were able to identify clear target-site duplications for 272 elements. All elements with clear target-site duplications had endonuclease sites that matched those described elsewhere (Feng et al. 1996; Jurka 1997; Cost and Boeke 1998). A total of 13 elements (L1HS45, -70, -172, -178, -284, -372, -415, -416, -442, -443, -448, -513, and -558) apparently lacked target-site duplications or contained short target-site duplications. To further investigate these elements, PCRs specific for the pre-integration sites for those elements listed were performed on the common chimpanzee, pygmy chimpanzee, and, when possible, human samples. The resulting amplicons were cloned and sequenced, to unambiguously define the pre-integration site for each element. The resulting pre-integration sites were then compared with the original GenBank sequence for each locus.

All 13 of the L1Hs elements lacked obvious target-site duplications when compared with the common and pygmy chimpanzee pre-integration-site sequences. In addition, L1HS178 and L1HS284 had no observable target-site duplications and atypical endonuclease-cleavage sites. One possible explanation for this observation is that these elements have integrated independent of endonuclease cleavage of target sequence, which has elsewhere been proposed as a mechanism for the repair of double-stranded breaks in DNA (Moore and Haber 1996; Teng et al. 1996; Morrish et al. 2002). Alternatively, these elements may represent forward gene-conversion events of preexisting L1 elements that, by mutation, have rendered their target-site duplications unrecognizable. However, because little is known about the rates of these events in mammalian cells, further studies are required in order to resolve the mechanism underlying these integration events.

Another aspect of L1Hs Ta sequence diversity is created by variable 5′ truncation such that some of the elements in the human genome are only a few hundred base pairs long, whereas some full-length elements are >6,000 bp long. This phenomenon is classically attributed to the lack of processivity of the reverse-transcriptase enzyme in the creation of the L1 cDNA copy. 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 are longer elements (Grimaldi et al. 1984). Our data show that there is an enrichment of full-length elements in the human genome and that many Ta elements have been faithfully replicated in their entirety and inserted into new genomic locations. Of the 399 elements examined, 119 were >6,000-bp long, representing an L1 Ta size class much larger than any other (fig. 4). By contrast, very few elements were found in the size class ranging between 3,500 and 5,500 bp, with only 22 of the 399 elements truncated to this particular size class. A bimodal distribution of the size of the elements is created, since there are a significant number of Ta L1 elements that are severely 5′ truncated and that are full length. One hundred ninety-eight elements were extremely small, having sizes <2,000 bp, and 118 of these elements were between 25 and 1,000 bp long. The distribution is noteworthy, although the mechanism by which these are enriched in the human genome remains to be determined. In addition, 20% (79/399) of the L1Hs elements examined are inverted at their 5′ end—which is an occurrence that is believed to be due to an event known as "twin priming" (Ostertag and Kazazian 2001), in which target-primed reverse transcription is interrupted by a second internal priming event, resulting in an inversion of the 5′ end of the newly integrated LINE. Although L1 truncation is most likely the result of the relatively low processivity of the L1 reverse transcriptase, processes, like twin priming, that form secondary structures in the RNA or DNA strands present at the integration site may also be associated with L1 truncation.

We also observed a significant amount of sequence diversity in the 3′ tails of members of the L1Hs Ta subfamily. The 3′ tails within this L1 subfamily range in size from 3 to >1,000 bp. Thirty-six percent contain AT-rich

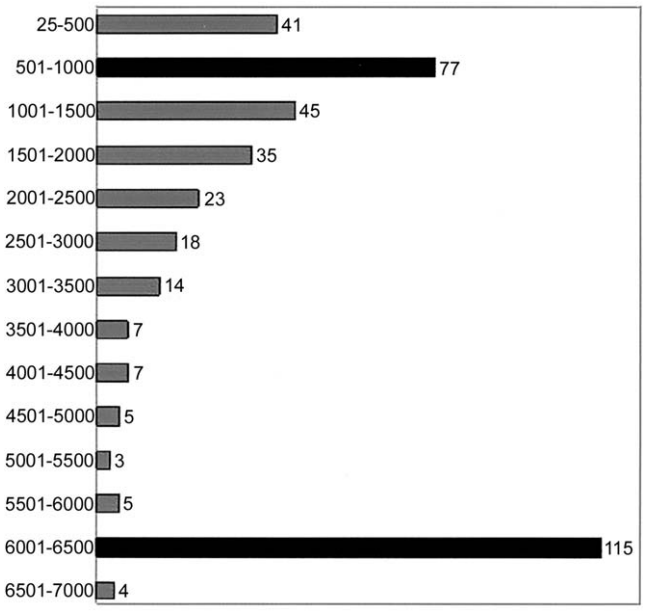


Figure 4 Ta L1 element size classes (in bp), showing the size distribution of Ta L1Hs elements. Elements are grouped in 500-bp intervals ranging from <500 bp to 7,000 bp long. The two most common size intervals are shown in black.

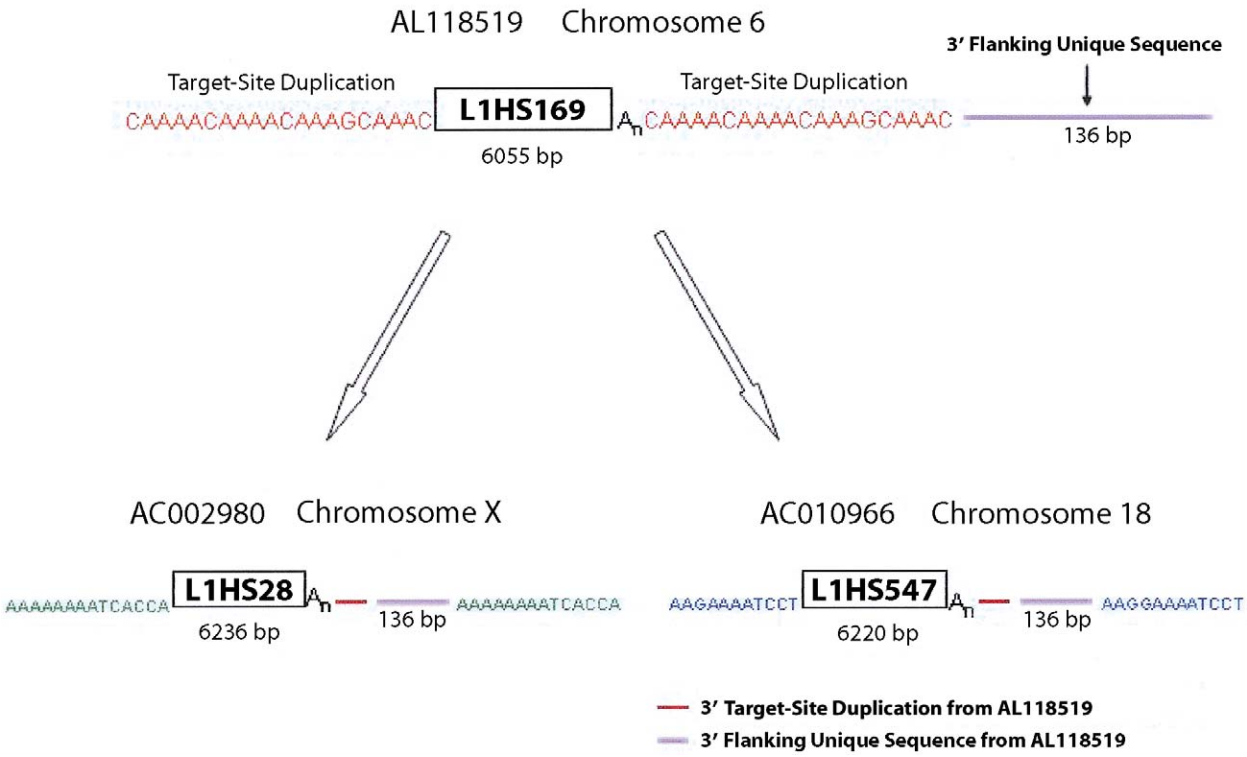


Figure 5 L1HS169-mediated transduction, showing an L1Hs transduction event. L1HS169 marked by clear target-site duplications is the putative source gene for L1HS28. The L1HS28 insertion contains 3′ flanking sequences identical to that of L1HS169 and unique target-site duplications flanking this entire sequence—suggesting that L1HS28 was created from a read-through transcript of L1HS169 that, to give rise to L1HS28, integrated into a new location on chromosome X. In addition, a second transduction event—L1HS547, from chromosome 18—is also flanked by unique target-site duplications and was also derived from L1HS169.

low-complexity sequence, 31% have homopolymeric A tails, 5% have simple sequence repeats with the most common repeat family TAAA, and 26% contain complex sequence that likely results from 3′ transduction events. The diversity in the tails of the L1 elements is not surprising, since previous studies have shown an association, as well as direct evidence that mobile-element–related simple-sequence-repeat motifs mutate to form nuclei for the generation of simple sequence repeats (Economou et al. 1990; Arcot et al. 1995; Ovchinnikov et al. 2001). Three-prime transduction by L1 elements is a unique duplication event that involves retrotransposons and that has elsewhere been described, in detail, in L1 elements (Boeke and Pickeral 1999; Moran et al. 1999; Goodier et al. 2000). We have identified a number of 3′ transduction events that are mediated by Ta L1Hs elements and believe that these elements have transduced a total of ~8,500 bp of sequence. We have also taken advantage of the L1 element–mediated transduction to computationally identify a putative retrotransposition-competent L1 Ta source gene. L1HS169 has a 136-bp fragment that is located outside its direct repeats and that is adjacent to its 3′ tail; this fragment is also found adjacent to the 3′ tail of L1HS28 but inside its direct repeat (fig. 5). This suggests that L1HS28 is a daughter copy, or the progeny, of the full-length element L1HS169. In addition, AC010966 from chromosome 18 appears to be a transduction event that was also generated from an L1HS169 read-through transcript. Therefore, we conclude that L1HS169 is responsible for multiple transduction events in the human genome and has produced two independent L1 integrations located on chromosomes X and 18.

Discussion

Here we report a comprehensive analysis of the dispersion and insertion polymorphism of the youngest known L1 subfamily (i.e., Ta) within the human genome. The computational approach described herein provides an efficient and high-throughput method for the recovery, from the

human genome, of Ta L1Hs elements, many of which will be polymorphic for insertion presence/absence in individual human genomes. Individual L1 insertion polymorphisms that were identified are the products of unique insertion events within the human genome. Because each L1 element integrates into the human genome only once, individuals that share L1 insertions (and insertion polymorphisms) inherited them from a common ancestor, thereby making the L1 filled sites identical by descent. This distinguishes L1 insertion polymorphisms and other mobile-element insertion polymorphisms from other types of genetic variation—including microsatellites (Nakamura et al. 1987) and RFLPs (Botstein et al. 1980)—that are not necessarily homoplasy free. In addition, the ancestral state of an L1 insertion is known to be the absence of the L1 element. Knowledge about the ancestral state of L1 insertions facilitates the rooting of trees of population relationships by use of minimal assumptions. Therefore, the 115 new L1 insertion polymorphisms reported herein appear to have genetic properties that are similar to those of *Alu* insertion polymorphisms (Batzer et al. 1991, 1994; Perna et al. 1992; Hammer 1994; Stoneking et al. 1997; Jorde et al. 2000), and they will serve as an additional source of identical-by-descent genomic variability for the study of human population relationships.

It is noteworthy that the computational identification of L1 insertion polymorphisms introduces a selection for only those elements present in the draft-sequence database. As a result, elements that are not present in the database cannot be identified. This has important consequences with respect to the frequency spectrum of the elements identified. By use of this type of approach, a number of different types of L1 insertion polymorphisms are identified that vary in the frequency of the L1 insertion allele. By contrast, PCR-based display approaches provide an alternative method for the ascertainment of mobile-element insertion polymorphisms from the human genome (Roy et al. 1999; Sheen et al. 2000; Ovchinnikov et al. 2001). In these approaches, polymorphic mobile elements are directly identified; however, elements that are polymorphic but have higher allele frequencies (i.e., high-frequency insertion polymorphisms) are lost in the process, since most genomes will contain at least one filled allele that contains the mobile element and would not be scored as an insertion polymorphism. Therefore, more population-specific or private mobile-element insertion polymorphisms will be identified using PCR-based displays or other types of direct selection (Roy et al. 1999; Sheen et al. 2000; Ovchinnikov et al. 2001). Using our computational approach, we recovered only 14 of 49 Ta L1 elements that were elsewhere identified using PCR-based displays (Sheen et al. 2000; Ovchinnikov et al. 2001) and that had sufficient flanking unique DNA sequences for comparison to the data set that we studied. Thus, computational and experimental ascertainment of mobile-element insertion polymorphisms are quite complementary approaches for the identification of new mobile-element insertion polymorphisms.

The L1 Ta subfamily can be further subdivided—according to the nucleotides that are present, within ORF 2, at positions 5536 and 5539—into Ta-0 and Ta-1 (Boissinot et al. 2000). Ta-0 L1 elements are believed to be evolutionarily older, and they possess a G at position 5536 and a C at position 5539. Ta-1 L1 elements, however, have a T at position 5536 and a G at nucleotide 5539. Ta-1 L1 elements are considered to be younger, and it is believed that all actively transposing elements in humans belong to the Ta-1 subset of L1 elements (Boissinot et al. 2000). One hundred ninety-two of the 459 Ta elements identified from the draft human genomic sequence belong to the younger Ta-1 subset, and 137 belong to the Ta-0 subset. Another 105 of the elements either are 5′ truncated such that they terminated before these positions at 5536 and 5539 or are inverted or rearranged in the region in question. An additional 25 elements are sequence intermediates between Ta-1 and Ta-0.

Inspection of the insertion polymorphism data for each of these Ta subsets showed that only 35% of the Ta-0 L1 elements analyzed by PCR were polymorphic, with the remaining 65% being fixed present in the human populations screened. Consistent with the idea that Ta-0 L1 elements are older, 9 of the polymorphic elements were high-frequency insertion polymorphisms, 10 were intermediate-frequency insertion polymorphisms, and only 5 were low-frequency insertion polymorphisms. None of the Ta-0 L1 elements were fixed absent or very low frequency in the populations that were analyzed. By contrast, 56% of the Ta-1 L1 elements were polymorphic with respect to presence—with 18 high-frequency, 27 intermediate-frequency, and 11 low-frequency insertion polymorphisms. In addition, we can use the non-CpG mutation density in Ta-0 and Ta-1 L1 elements to calculate the estimated age of each of the Ta-derivative subfamilies. The non-CpG mutation density for the Ta-0 and Ta-1 L1 elements was 0.003103 and 0.002560, respectively. Using a neutral rate of evolution of 0.15% per million years (Miyamoto et al. 1987), we derive estimates of 2.07 (i.e., 0.003103/0.0015) million years and 1.71 (i.e., 0.002560/0.0015) million years from the Ta-0 and Ta-1 subsets, respectively. Although these estimates are not significantly different from each other, they do support the notion that the Ta-0 L1 elements are slightly older than the Ta-1 L1 elements, as do the differences in insertion polymorphism. In addition, they provide direct evidence that the Ta-0 and Ta-1 subsets have simultaneously amplified within the human genome.

Forty-four of the 124 full-length Ta L1Hs elements

that were identified have both ORFs intact and are presumably retrotransposition-competent elements. This compares favorably with previous estimates of the number of potentially active L1 elements in the human genome (Sassaman et al. 1997). In addition, it is also important that those full-length elements that no longer have intact ORFs might have previously acted as active "source," or driver, genes for the expansion of Ta L1 elements but might have accumulated mutations over time that inactivated them. These data, as well as data from the previous studies involving the isolation and amplification of some of these full-length Ta L1 elements within tissue-culture systems, demonstrate that multiple L1 elements have expanded within the human genome in an overlapping time frame. It is interesting to compare the amplification of the L1 elements to that of the *Alu* SINEs within the human genome. In the case of the L1 elements, one major family (Ta) with two subdivisions (Ta-0 and Ta-1) has expanded to a copy number of ~500 elements in the past four to six million years since the divergence of humans and African apes. By contrast, the expansion of *Alu* elements is characterized by the amplification of at least three major lineages, or subfamilies of elements, that have collectively generated ~5,000 copies (Batzer and Deininger 2002). On the basis of these copy numbers alone, it would appear that *Alu* elements have been 10 times more successful than L1 elements have been with respect to duplicating themselves, within primate genomes, over the past four to six million years. However, if we make the estimate relative to the total family size of 500,000 L1 elements or 1.1 million *Alu* elements (Lander et al. 2001), then the relative difference is merely fivefold. This difference in amplification is also apparent across the entire expansion of these repeated DNA sequence families, since the L1 elements have expanded to only 500,000 copies in 150 million years, whereas the *Alu* elements have expanded to 1.1 million copies in only 65 million years.

Since *Alu* and L1 elements are thought to utilize the same enzymatic machinery for their mobilization, the differential amplification of both young and old *Alu* and L1 elements within primate genomes is quite interesting (Boeke 1997). The two different classes of repeats putatively compete for access to the same reverse transcriptase and endonuclease; thus, it is possible that *Alu* elements are currently more effective than the L1 elements at attracting the replication machinery within the human genome. If this competition between interspersed elements is important, then we may expect to see differential rates of L1 and *Alu* expansion in different nonhuman primate genomes as the elements compete for the common components involved in mobilization. Differential mobilization of SINEs and LINEs has been elsewhere reported in rodent genomes (Kim and Deininger 1996; Ostertag et al. 2000). Therefore, it would not be surprising to see something similar in nonhuman primate genomes. Alternatively, the differential amplification may reflect differences in selection against new L1 and *Alu* insertions within the human genome (Lander et al. 2001). Since L1 elements are typically much larger than *Alu* repeats, it is easy to envision that the larger insertions would be much more disruptive to the genome than the shorter *Alu* insertions are. This type of selection has been suggested as one potential explanation for the differential distributions of L1 elements (Boissinot et al. 2001) and of *Alu* and L1 elements (Lander et al. 2001; Ovchinnikov et al. 2001) throughout the human genome. However, the argument that selection is responsible for the differential distribution of *Alu* sequences has recently been questioned (Brookfield 2001). Further studies of the expansion of interspersed elements within the genomes of nonhuman primates will be required in order to definitively address these questions.

Our analysis of mosaic Ta L1Hs elements suggests that gene conversion alters the sequence diversity within these elements. This is not surprising, since previous studies have indicated that gene conversion plays a role in the generation of sequence diversity in *Alu* repeats (Maeda et al. 1988; Batzer et al. 1995; Kass et al. 1995; Roy et al. 2000; Carroll et al. 2001; Roy-Engel et al. 2002), as well as the generation of sequence diversity in L1 elements, within the genome (Hardies et al. 1986; Burton et al. 1991; Tremblay et al. 2000). Unfortunately, an accurate estimate of L1-based gene conversion is not yet possible, because primate L1 subfamily structure is not yet clearly defined. However, gene conversion appears to play a significant role in the sculpting of human genomic diversity (Ardlie et al. 2001; Frisse et al. 2001). Because of the hierarchical subfamily structure of *Alu* and LINEs and because of the defined pattern of ancestral mutations, these elements provide a unique opportunity for the estimation of gene conversion throughout the genome. It is also important to consider that the gene conversion between large multigene families, such as SINEs and LINEs, may occur by a mechanism that is completely different from that which occurs at other unique and low-repetition sequences within the human genome. Nevertheless, large-scale studies of orthologous sequences from the same L1 element in different human genomes will begin to quantitatively address this issue and also will provide insight into the molecular mechanism that drives the process. In addition, detailed pedigree analyses or studies of germ cell–derived L1 diversity will provide insight into the germ line rate of gene conversion between L1 elements. Clearly, L1 elements continue to have a significant impact on human genetic diversity—through recombination, insertional mutagenesis, gene conversion, sequence transduction, and the generation of other

simple-sequence-repeat motifs (Kazazian and Moran 1998; Goodier et al. 2000; Ovchinnikov et al. 2001).

Acknowledgments

This research was supported by National Institutes of Health grants R01 GM59290 (to L.B.J. and M.A.B.), R21 CA87356-02 (to G.D.S.), and R01 GM60518 (to J.V.M.); by support from the W. M. Keck Foundation (to J.V.M.); by Louisiana Board of Regents Millennium Trust Health Excellence Fund grants (2000-05)-05, (2000-05)-01, and (2001-06)-02 (to M.A.B.); and, through award 2001-IJ-CX-K004 (to M.A.B.), by the Office of Justice Programs, National Institute of Justice, U.S. Department of Justice. Points of view expressed in this article are those of the authors and do not necessarily represent the official position of the U.S. Department of Justice.

Electronic-Database Information

Accession numbers and URLs for data presented herein are as follows:

Batzer Lab, http://batzerlab.lsu.edu/

BLAST, http://www.ncbi.nlm.nih.gov/blast/

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for the DNA sequences from the common and pygmy chimpanzee orthologs of L1HS72 [accession numbers AF489459 and AF489460]; diverse DNA sequences from L1HS72 [accession numbers AF489450–AF489458]; and Ta L1 element preintegration site sequences, namely, L1HS45 [accession numbers AF461364 and AF461365], L1HS172 [accession numbers AF461368 and AF461369], L1HS178 [accession numbers AF461370 and AF461371], L1HS284 [accession numbers AF461372 and AF461373], L1HS372 [accession numbers AF461374 and AF461375], L1HS416 [accession numbers AF461376 and AF461377], L1HS442 [accession numbers AF461378 and AF461379], L1HS443 [accession numbers AF461386 and AF461387], L1HS513 [accession numbers AF461380–AF461382], and L1HS558 [accession number AF461383])

Genetic Information Research Institute Censor Server, http://www.girinst.org/Censor_Server-Data_Entry_Forms.html

Primer3, http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi

RepeatMasker Web Server, http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA (1995) Alu repeats: a source for the genesis of primate microsatellites. Genomics 29:136–144

Ardlie K, Liu-Cordero SN, Eberle MA, Daly M, Barrett J, Winchester E, Lander ES, Kruglyak L (2001) Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. Am J Hum Genet 69:582–589

Ausabel FM, Brent R, Kingston ME, Moore DD, Seidman JG (1987) Current protocols in molecular biology. John Wiley & Sons, New York

Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. Nat Rev Genet 3:370–379

Batzer MA, Gudi VA, Mena JC, Foltz DW, Herrera RJ, Deininger PL (1991) Amplification dynamics of human-specific (HS) Alu family members. Nucleic Acids Res 19:3619–3623

Batzer MA, Rubin CM, Hellmann-Blumberg U, Alegria-Hartman M, Leeflang EP, Stern JD, Bazan HA, Shaikh TH, Deininger PL, Schmid CW (1995) Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. J Mol Biol 247:418–427

Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, Novick GE, Ioannou PA, Scheer WD, Herrera RJ, Deininger PL (1994) African origin of human-specific polymorphic Alu insertions. Proc Natl Acad Sci USA 91:12288–12292

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499–1504

Boeke JD (1997) LINEs and Alus—the polyA connection. Nat Genet 16:6–7

Boeke JD, Pickeral OK (1999) Retroshuffling the genomic deck. Nature 398:108–109

Boissinot S, Chevret P, Furano AV (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol 17:915–928

Boissinot S, Entezam A, Furano AV (2001) Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol 18:926–935

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Brookfield JF (2001) Selection on Alu sequences? Curr Biol 11:R900–R901

Burton FH, Loeb DD, Edgell MH, Hutchison CA 3d (1991) L1 gene conversion or same-site transposition. Mol Biol Evol 8:609–619

Carroll ML, Roy-Engel AM, Nguyen SV, Salem AH, Vogel E, Vincent B, Myers J, Ahmad Z, Nguyen L, Sammarco M, Watkins WS, Henke J, Makalowski W, Jorde LB, Deininger PL, Batzer MA (2001) Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J Mol Biol 311:17–40

Cost GJ, Boeke JD (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

Cost GJ, Golding A, Schlissel MS, Boeke JD (2001) Target DNA chromatinization modulates nicking by L1 endonuclease. Nucleic Acids Res 29:573–577

Deininger PL, Batzer MA, Hutchison CA 3d, Edgell MH (1992) Master genes in mammalian repetitive DNA amplification. Trends Genet 8:307–311

Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian HH Jr (1991) Isolation of an active human transposable element. Science 254:1805–1808

Economou EP, Bergen AW, Warren AC, Antonarakis SE (1990) The polydeoxyadenylate tract of Alu repetitive elements is

polymorphic in the human genome. Proc Natl Acad Sci USA 87:2951–2954
Eng B, Ainsworth P, Waye JS (1994) Anomalous migration of PCR products using nondenaturing polyacrylamide gel electrophoresis: the amelogenin sex-typing system. J Forensic Sci 39:1356–1359
Fanning TG, Singer MF (1987) LINE-1: a mammalian transposable element. Biochim Biophys Acta 910:203–212
Feng Q, Moran JV, Kazazian HH Jr, Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87:905–916
Fitch DH, Bailey WJ, Tagle DA, Goodman M, Sieu L, Slightom JL (1991) Duplication of the γ-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. Proc Natl Acad Sci USA 88:7396–7400
Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet 69:831–843
Goodier JL, Ostertag EM, Kazazian HH Jr (2000) Transduction of 3′-flanking sequences is common in L1 retrotransposition. Hum Mol Genet 9:653–657
Grimaldi G, Skowronski J, Singer MF (1984) Defining the beginning and end of KpnI family segments. EMBO J 3: 1753–1759
Hammer MF (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
Hardies SC, Martin SL, Voliva CF, Hutchison CA 3d, Edgell MH (1986) An analysis of replacement and synonymous changes in the rodent L1 repeat family. Mol Biol Evol 3: 109–125
Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA (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 (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci USA 94:1872–1877
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
Kass DH, Batzer MA, Deininger PL (1995) Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. Mol Cell Biol 15:19–25
Kazazian HH Jr (1998) Mobile elements and disease. Curr Opin Genet Dev 8:343–350
——— (2000) L1 retrotransposons shape the mammalian genome. Science 289:1152–1153
Kazazian HH Jr, Moran JV (1998) The impact of L1 retrotransposons on the human genome. Nat Genet 19:19–24
Kazazian HH Jr, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332:164–166
Kim J, Deininger PL (1996) Recent amplification of rat ID sequences. J Mol Biol 261:322–327
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
Luan DD, Korman MH, Jakubczak JL, Eickbush TH (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
Maeda N, Wu CI, Bliska J, Reneke J (1988) Molecular evolution of intergenic DNA in higher primates: pattern of DNA changes, molecular clock, and evolution of repetitive sequences. Mol Biol Evol 5:1–20
Miyamoto MM, Slightom JL, Goodman M (1987) Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. Science 238:369–373
Moore JK, Haber JE (1996) Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. Nature 383:644–646
Moran JV, DeBerardinis RJ, Kazazian HH Jr (1999) Exon shuffling by L1 retrotransposition. Science 283:1530–1534
Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr (1996) High frequency retrotransposition in cultured mammalian cells. Cell 87:917–927
Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato T, Taccioli G, Batzer MA, Moran JV (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet 31:159–165
Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, White R (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622
Ostertag EM, Kazazian HH Jr (2001) Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res 11:2059–2065
Ostertag EM, Prak ET, DeBerardinis RJ, Moran JV, Kazazian HH Jr (2000) Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Res 28:1418–1423
Ovchinnikov I, Troxel AB, Swergold GD (2001) Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res 11:2050–2058
Perna NT, Batzer MA, Deininger PL, Stoneking M (1992) Alu insertion polymorphism: a new type of marker for human population studies. Hum Biol 64:641–648
Prak ET, Kazazian HH Jr (2000) Mobile elements and the human genome. Nat Rev Genet 1:134–144
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
Roy AM, Carroll ML, Kass DH, Nguyen SV, Salem AH, Batzer MA, Deininger PL (1999) Recently integrated human Alu repeats: finding needles in the haystack. Genetica 107:149–161
Roy AM, Carroll ML, Nguyen SV, Salem AH, Oldridge M, Wilkie AO, Batzer MA, Deininger PL (2000) Potential gene conversion and source genes for recently integrated Alu elements. Genome Res 10:1485–1495
Roy-Engel AM, Carroll ML, El-Sawy M, Salem AE, Garber RK, Nguyen SV, Deininger PL, Batzer MA (2002) Non-traditional Alu evolution and primate genomic diversity. J Mol Biol 316:1033–1040
Roy-Engel AM, Carroll ML, Vogel E, Garber RK, Nguyen SV, Salem AH, Batzer MA, Deininger PL (2001) Alu insertion

polymorphisms for the study of human genomic diversity. Genetics 159:279–290

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467

Santos FR, Pandya A, Kayser M, Mitchell RJ, Liu A, Singh L, Destro-Bisol G, Novelletto A, Qamar R, Mehdi SQ, Adhikari R, de Knijff P, Tyler-Smith C (2000) A polymorphic L1 retroposon insertion in the centromere of the human Y chromosome. Hum Mol Genet 9:421–430

Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH Jr (1997) Many human L1 elements are capable of retrotransposition. Nat Genet 16:37–43

Sheen FM, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, Batzer MA, Swergold GD (2000) Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res 10:1496–1508

Skowronski J, Fanning TG, Singer MF (1988) Unit-length LINE-1 transcripts in human teratocarcinoma cells. Mol Cell Biol 8:1385–1397

Smit AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9:657–663

Smit AF, Toth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246:401–417

Stoneking M, Fontius JJ, Clifford SL, Soodyall H, Arcot SS, Saha N, Jenkins T, Tahir MA, Deininger PL, Batzer MA (1997) Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. Genome Res 7:1061–1071

Teng SC, Kim B, Gabriel A (1996) Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks. Nature 383:641–644

Tremblay A, Jasin M, Chartrand P (2000) A double-strand break in a chromosomal LINE element can be repaired by gene conversion with various endogenous LINE elements in mouse cells. Mol Cell Biol 20:54–60

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