Cell

LINE-1 Retrotransposition Activity in Human Genomes

Christine R. Beck,[1,*] Pamela Collier,[4] Catriona Macfarlane,[4] Maika Malig,[5] Jeffrey M. Kidd,[5] Evan E. Eichler,[5,6] Richard M. Badge,[4] and John V. Moran[1,2,3,*]
[1]Department of Human Genetics
[2]Department of Internal Medicine
[3]Howard Hughes Medical Institute
University of Michigan Medical School, Ann Arbor, MI 48109, USA
[4]Department of Genetics, University of Leicester, Leicester LE1 7RH, UK
[5]Department of Genome Sciences
[6]Howard Hughes Medical Institute
University of Washington, Seattle, WA 98195, USA
*Correspondence: cregina@umich.edu (C.R.B.), moranj@umich.edu (J.V.M.)
DOI 10.1016/j.cell.2010.05.021

SUMMARY

Highly active (i.e., "hot") long interspersed element-1 (LINE-1 or L1) sequences comprise the bulk of retrotransposition activity in the human genome; however, the abundance of hot L1s in the human population remains largely unexplored. Here, we used a fosmid-based, paired-end DNA sequencing strategy to identify 68 full-length L1s that are differentially present among individuals but are absent from the human genome reference sequence. The majority of these L1s were highly active in a cultured cell retrotransposition assay. Genotyping 26 elements revealed that two L1s are only found in Africa and that two more are absent from the H952 subset of the Human Genome Diversity Panel. Therefore, these results suggest that hot L1s are more abundant in the human population than previously appreciated, and that ongoing L1 retrotransposition continues to be a major source of interindividual genetic variation.

INTRODUCTION

L1s comprise ~17% of human DNA and have been an instrumental force in shaping genome architecture (Lander et al., 2001). Most L1s are molecular fossils that cannot move (retrotranspose) to new genomic locations (Grimaldi and Singer, 1983; Lander et al., 2001). However, a small number of human-specific L1 (L1Hs) elements remain retrotransposition competent (Badge et al., 2003; Brouha et al., 2003; Sassaman et al., 1997). On occasion, their retrotransposition has resulted in sporadic cases of human disease (reviewed in Babushok and Kazazian, 2007; Kazazian et al., 1988).

During the past 15 years, computational, molecular biological, and genomic approaches have been used to identify and characterize L1Hs elements (Badge et al., 2003; Bennett et al., 2004; Boissinot et al., 2000; Boissinot et al., 2004; Brouha et al., 2003; Lander et al., 2001; Moran et al., 1996; Myers et al., 2002; Ovchinnikov et al., 2001; Sheen et al., 2000; Xing et al., 2009). Several themes have emerged from these studies. First, L1Hs elements can be stratified into several subfamilies (pre-Ta, Ta-0, Ta-1, Ta1-d, Ta1-nd) based upon the presence of diagnostic sequence variants contained within their 5′ and/or 3′ untranslated regions (UTRs) (Boissinot et al., 2000; Skowronski et al., 1988; Smit et al., 1995). Second, many L1Hs elements are dimorphic in that they are differentially present in individual genomes and/or are present in an individual but absent from the haploid Human Genome Reference sequence (HGR) (Badge et al., 2003; Bennett et al., 2004; Boissinot et al., 2004; Brouha et al., 2003; Lander et al., 2001; Myers et al., 2002; Xing et al., 2009). Third, it has been estimated that the average human genome contains ~80–100 active (retrotransposition-competent) L1Hs elements, and that only a small number of highly active L1Hs elements ("hot" L1s) account for the bulk of retrotransposition activity in the HGR (Brouha et al., 2003). Those studies, as well as recent efforts to identify insertion, deletion, and inversion polymorphisms (structural variants) in humans (Kidd et al., 2008; Korbel et al., 2007; Tuzun et al., 2005; Xing et al., 2009), indicate that ongoing L1 retrotransposition contributes to interindividual genetic variation.

Here, we employed a fosmid-based, paired-end DNA resource to identify full-length L1Hs elements in the genomes of six individuals of diverse geographic origin. Over half (37/68) of the newly identified L1s were hot for retrotransposition when examined in a cultured cell assay (Moran et al., 1996). Genotyping a subset of these L1s further revealed that some are likely restricted to Africans, whereas others are absent from the Human Genome Diversity Panel (HGDP) (Cann et al., 2002), suggesting that they are present at very low allele frequencies.

RESULTS

An Experimental Strategy to Identify Full-Length Human-Specific L1s

To identify novel, full-length L1s in the genomes of geographically diverse individuals, we exploited a fosmid-based, paired-end DNA sequencing strategy that previously was used to

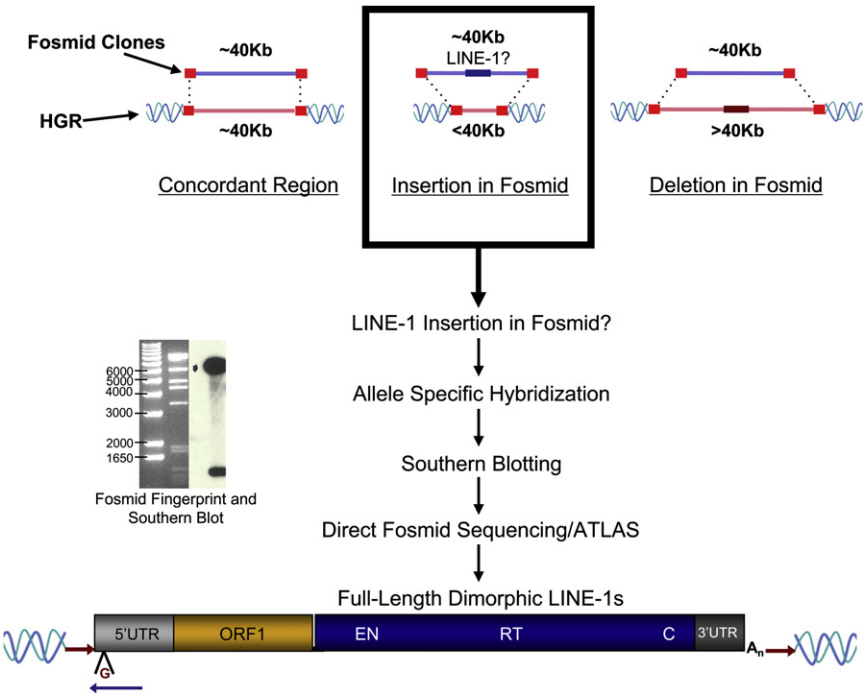

Figure 1. A Strategy for Identifying Dimorphic L1Hs Elements in Individual Human Genomes

In silico comparison of the fosmid end sequences (red squares) from individual genomic libraries (blue horizontal line) and the HGR (pink horizontal line) enables the detection of fosmids that may contain insertions or deletions with respect to the HGR (see dashed lines). Insertion fosmids were screened by allele-specific oligonucleotide hybridization to detect characters that are present in the 5′ UTR of newer L1 elements (one discriminating character utilized, a deletion of the G residue at bp 74 in recent L1s, is indicated in maroon). Putative L1Hs-containing fosmids were analyzed by Southern blotting with a 5′ UTR probe (blue arrow). A representative digest and Southern blot is shown. The ~6 kb band is diagnostic for the full-length L1. The additional hybridizing band (~1.3 kb band liberated from the L1 5′ flank in this Southern blot example) serves to distinguish individual fosmids. ATLAS and/or DNA sequencing confirmed the presence of a dimorphic, full-length L1Hs insertion. The endonuclease (EN), reverse transcriptase (RT), and cysteine-rich (C) domains of ORF2 (blue rectangle) are indicated.

identify structural variants in human DNA (Kidd et al., 2008; Tuzun et al., 2005). Fragments of genomic DNA approximately 40 kb in size were individually cloned using fosmid vectors (see Experimental Procedures). Sequence reads were obtained from both ends of each insert (paired-end sequences) and compared to the HGR. End sequences from genomic fragments that do not differ significantly in size from the HGR will map ~40 kb away from each other. In contrast, paired-end sequences derived from genomic fragments containing a full-length, dimorphic ~6 kb L1Hs element will be separated by ~34 kb when mapped to the HGR (Figure 1) (Tuzun et al., 2005). In general, the predicted variants were required to be supported by two fosmid clones containing putative insertions from the same individual. The size cutoffs used in our screening protocols are biased to allow the identification of full-length or near full-length L1 insertion polymorphisms, but not severely 5′ truncated L1 sequences, which are replication deficient (Table 1). Through this scheme, we should be able to identify the bulk of full-length L1s in an individual genome that are dimorphic when compared to the HGR.

Fosmids fulfilling the above mapping criterion were subjected to a series of screens (Figure 1). First, allele-specific oligonucleotide hybridization using probes directed against diagnostic sequences in the L1Hs 5′ UTR identified insertion fosmids that contained putative dimorphic L1Hs elements (Boissinot et al., 2000; Tuzun et al., 2005). Second, Southern blotting with a probe directed against the 5′ UTR of L1.3 (accession# L19088) enabled the identification of fosmids that contained putative full-length L1Hs elements (Dombroski et al., 1993; Sassaman et al., 1997). Third, a suppression PCR-based method (ATLAS) (Badge et al., 2003) and/or direct sequencing was used to verify the presence of a full-length (or near full-length) L1Hs element in the fosmid. Finally, genomic sequences flanking the 5′ and 3′ ends of the newly identified L1Hs elements were used as probes in BLAT searches (http://genome.ucsc.edu/cgi-bin/hgBlat?command = start) (Kent, 2002) to confirm that the L1 was absent from the HGR (NCBI build 36.1/hg18). Flanking sequences also were used to determine whether any of the L1Hs elements were present in a database of known polymorphic retrotransposon insertions (dbRIP; http://dbrip.brocku.ca/) (Wang et al., 2006). Two additional L1Hs elements were identified through direct sequencing of the fosmids (#1-(2-1) and 3-(2-1)).

Identification of Full-Length L1Hs Elements from Geographically Diverse Individuals

We first conducted a pilot study to examine a fosmid library from a female individual (G248; NA15510) for full-length L1Hs insertions (Table 1) (Tuzun et al., 2005). Despite the fact that this library was optimized for identifying ~8 kb insertion polymorphisms as part of the Human Genome Structural Variation project (HGSV) (Kidd et al., 2008; Tuzun et al., 2005), we were able to identify five novel L1Hs elements using our screening protocol (Table 1).

The above data provided "proof of principle" that our strategy was effective for identifying full-length, dimorphic L1Hs elements. Thus, we next screened fosmid libraries from five females representing four distinct geographic populations that were studied as part of the HapMap project (one Japanese [NA18956], one Chinese [NA18555], one Western European CEPH [NA12878], and two Yoruban individuals [NA19240, NA19129]) (International HapMap Consortium, 2005; Kidd et al., 2008). Size cutoffs allowed detection of insertion polymorphisms as small as ~4.2–5.5 kb and enabled the identification of an additional 64 L1Hs elements (Table 1) (Kidd et al., 2008). As our strategy is biased

Table 1. Summary of Data for the Six Libraries

Individual/Library Data						LINE-1 Data				
Library ID	Coriell ID	Population	Library Mean In Silico Insert Size	SD (kb)	Detection Limit (kb)	Dimorphic Elements	Novel (Not in dbRIP)	Active	Hot	HGR "Hot" Elements
G248	NA15510	N/A	39.89	2.75	8.25	5	5	4	4	2
ABC9	NA18956	Japan	39.51	2.26[b]	4.52[b]	16	16	9	8	2
ABC10	NA19240	Yoruba[a]	41	1.84	5.52	20	18	11	9	2
ABC11	NA18555	China	40.03	1.77	5.31	13	12	9	8	2
ABC12	NA12878	CEPH[a]	39.75	1.4	4.2	8	7	4	3	2
ABC13	NA19129	Yoruba[a]	39.29	1.77	5.31	7	7	6	5	2
Total						69/68[c]	65	43	37	

Column 1: library identifiers. Column 2: Coriell identifier of individuals analyzed. Column 3: population of origin for individuals in the HapMap study. Column 4: the average insert size of each individual library (in kb). Column 5: the standard deviation in insert size of each individual library. Column 6: the detection limit for the size of insertions in each library. For ABC9 a more reduced threshold was applied than that used previously (Kidd et al., 2008). Column 7: the number of elements found in each library that are absent from the HGR. Column 8: the number of elements from column 7 that are not completely annotated in dbRIP (Wang et al., 2006). Column 9: the number of elements from column 7 that were active in retrotransposition assays. Column 10: elements from column 9 that retrotransposed at levels > 10% of L1.3, a known active element. Column 11: the number of the HGR hot elements that were present in each individual (Brouha et al., 2003).

[a] Daughters of Hap Map trios.

[b] Differs from Kidd et al. (2008).

[c] One element recurred in ABC11 & 12- #4-1 and #5-77. Neither allele is active, and the element is in dbRIP.

toward finding novel, full-length L1s, we generally observed a decrease in the number of L1Hs elements identified in each successive library screen (e.g., ABC13 was the last library analyzed and contained relatively few novel L1Hs elements). In total, we identified 69 L1Hs elements that were absent from the HGR, one of which was identified in two different individuals (#4-1 and 5-77). This element also was completely annotated in dbRIP, unlike 65 of the distinct 68 L1s identified in this study (Table 1). The number of elements discovered at each stage of the analysis is detailed in the Extended Experimental Procedures.

Many of the Newly Identified L1Hs Elements Are Hot for Retrotransposition

We next tested if the L1Hs elements identified in our screens were active for retrotransposition in cultured cells. Sixty-seven elements were cloned into either a pBluescript and/or pCEP4 L1 expression vector that contained an *mneoI* retrotransposition indicator cassette in its 3′ UTR (#2-42 was refractory to cloning; details in Experimental Procedures) (Freeman et al., 1994; Moran et al., 1996). The pBluescript-based L1 constructs lack an exogenous promoter; thus, L1 expression is driven from its native 5′ UTR. Elements isolated from libraries ABC11–13 were assayed in this context. L1s isolated from the G248, ABC9, and ABC10 libraries were assayed in pCEP4 (CMV+/5′UTR+) and/or pBluescript (5′UTR+) based contexts. The resultant plasmids were transfected into HeLa cells and successful retrotransposition events were detected as G418-resistant foci (Figure 2A) (Moran et al., 1996). Retrotransposition activities are reported relative to L1.3, and hot refers to an L1 that jumps at >10% of L1.3 (see Table S1 available online). Notably, 22 elements yielded similar retrotransposition efficiencies relative to L1.3 when tested in either a CMV+/5′UTR+ or a 5′UTR+ context (data not shown). Since the subcloning procedure does not involve PCR, we truly are testing the retrotransposition capability of each of the identified L1Hs elements in our screen.

Each individual contained between three and nine hot L1s in their genome and 55% (37/67) of the L1Hs elements tested were hot for retrotransposition (Figures 2A and 2B; Table 1). These 37 hot L1Hs elements represent an approximately 4-fold increase in the number of hot L1s identified in previous studies (Badge et al., 2003; Brouha et al., 2002, 2003; Kimberland et al., 1999; Lander et al., 2001; Sassaman et al., 1997). Examination of the 3′ UTR sequences of the 68 L1s uncovered six elements that contain an ACG in place of the Ta subfamily diagnostic ACA characters. These elements are termed "pre-Ta" and represent an older L1 subfamily (Boissinot et al., 2000; Brouha et al., 2003; Kazazian et al., 1988; Lander et al., 2001; Myers et al., 2002; Skowronski et al., 1988). Two pre-Ta L1s (#3-5 and 5-55) were hot for retrotransposition (Figure 2B; Table S1). These data agree with previous studies, which showed that a de novo insertion of a pre-Ta L1 into the *Factor VIII* gene resulted in a sporadic case of hemophilia A (Kazazian et al., 1988).

Hallmarks and Insertion Locations of L1s Identified in This Study

We next sequenced each L1Hs element in its entirety and compared these data to fosmid sequences previously deposited in GenBank (Kidd et al., 2008). We annotated each L1 for hallmarks of retrotransposition as well as their chromosomal environments (Table S2). In general, the L1Hs elements were flanked by target-site duplications that ranged from 6 to 20 bp, inserted into an L1 endonuclease consensus cleavage sequence (Cost and Boeke, 1998; Feng et al., 1996; Morrish et al., 2002), and their 3′ ends had either homopolymeric poly(A) tails that ranged from ~8–41bp in size or interrupted poly(A) tails/3′ transductions ranging from ~18 bp to 1105 bp in length (Table S2) (Goodier et al., 2000; Holmes et al., 1994; Moran et al., 1999; Pickeral et al., 2000).

A subset of the elements (~32/68) contained an additional 1–14 bp of untemplated nucleotides at their 5′ ends, termed 5′

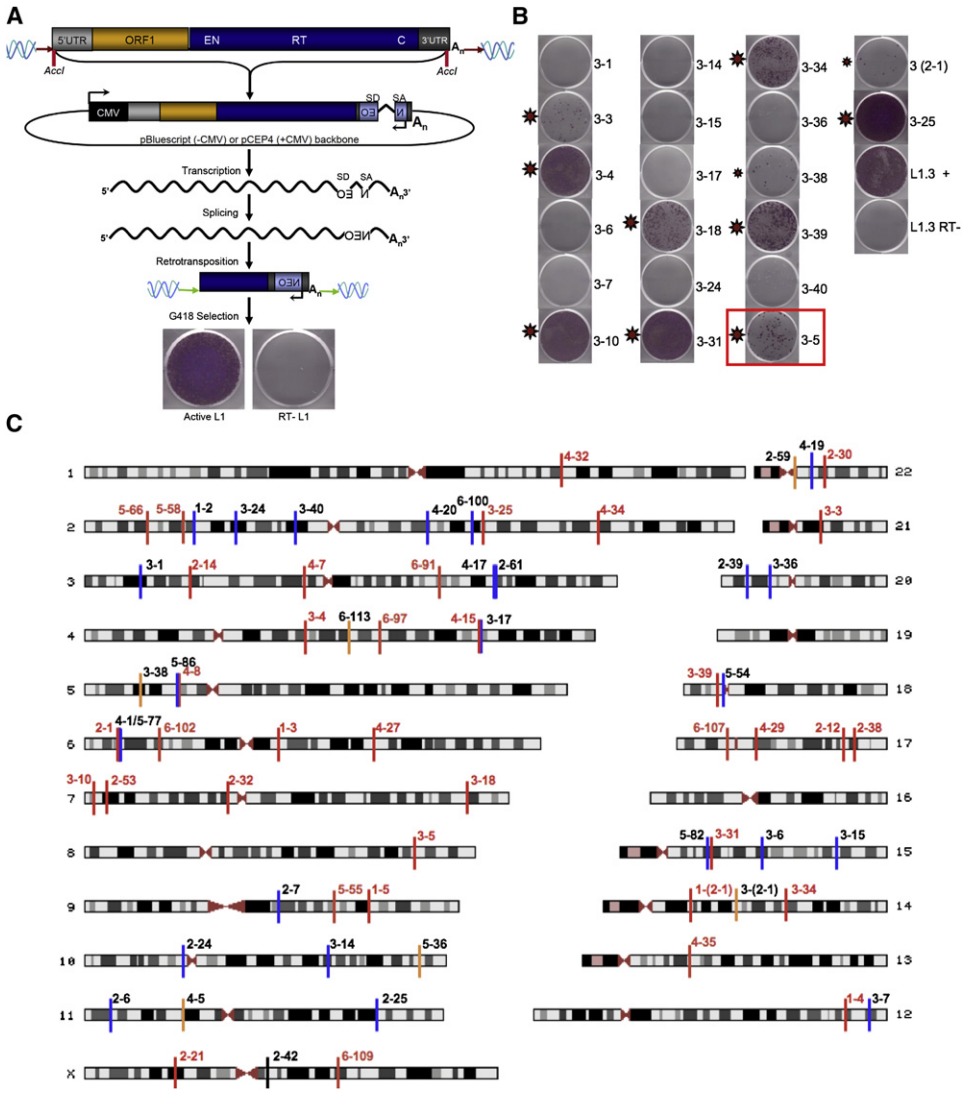

Figure 2. L1Hs Activity in Six Human Genomes

(A) Cloning strategy: All but one L1Hs element were cloned directly from fosmids using AccI sites in their 5′ UTRs and 3′ UTRs, respectively (red vertical lines; see Experimental Procedures). The L1s then were ligated into vectors that either contain or lack a CMV promoter (black rectangle). Both vectors contain the *mneoI* retrotransposition indicator cassette (light blue) in the L1 3′ UTR. This cassette allows for detection of retrotransposition events in a cell culture retrotransposition assay. SD = splice donor. SA = splice acceptor. Active elements confer G418 resistance to HeLa cells, whereas defective elements, as illustrated by the RT mutant control (RT- L1), do not.

(B) Representative G418-resistant foci for the 20 elements from the Yoruban library, ABC10: Nine of these elements were highly active (large suns to the left of assay image), and two more retained a low level of activity (small suns). One element (#3-5, red box) is a hot pre-Ta L1 (#3-5 was tested in a pBluescript backbone [5′UTR+]; all others were tested in a pCEP4 [CMV+/5′UTR+]) backbone (Extended Experimental Procedures). Table S1 displays retrotransposition efficiencies for each L1 identified in this study. Figure S1 provides details on the EN-deficient element #3-24.

(C) The 68 distinct L1Hs elements identified in this study and their positions in the genome: Red vertical lines and text represent hot or highly active elements. Orange vertical lines with black text represent low-level activity elements. Blue vertical lines with black text represent "dead" or inactive elements. The black line indicates the one untested element (#2-42). Ideograms were adapted from UCSC genome browser: http://genome.ucsc.edu (Kent et al., 2002).

end heterogeneity (Athanikar et al., 2004; Lavie et al., 2004). Five of these L1s have an extra G at their 5′ ends, and one has three extra Gs when compared to a hot L1Hs consensus sequence (Brouha et al., 2003). These extra nucleotides potentially could result either from a terminal transferase activity associated with the L1 reverse transcriptase or from reverse transcription of the 7-methyl-guanosine cap at the 5′ end of L1 RNA (Boeke, 2003; Gilbert et al., 2005; Symer et al., 2002). The majority of elements identified were full-length; however, we also found seven elements (e.g., #1-5 and 2-30) that were truncated within their 5′ UTR. These data, along with the fact that the fosmid libraries provided ~4- to 5-fold coverage of each haplotype from the six individuals (Kidd et al., 2008), indicate that our screening procedure identified the majority of the dimorphic full-length L1s in these genomes.

The 68 L1Hs elements were dispersed throughout the genome. We did not identify L1Hs elements on chromosomes 16 or 19 (Figure 2C); however, this result probably reflects our small sample size rather than a systematic bias against their ability to insert on these chromosomes (Lander et al., 2001). Consistently, we previously were able to detect the insertion of engineered L1s into chromosomes 16 and 19 of HeLa cells (Gilbert et al., 2005).

Approximately 32% (22/68) of L1Hs elements were present in the introns of known RefSeq genes (http://www.ncbi.nlm.nih.gov/RefSeq/), and mutations in several of these genes are implicated in human genetic disorders (Table S3). Thirteen L1 insertions were in the antisense orientation (i.e., were transcribed in the opposite orientation to the gene), whereas nine L1 insertions were in the same transcriptional orientation as the gene. Since ~26%–38% of the genome is spanned by genes (Venter et al., 2001), the data suggest that the L1s have inserted randomly with respect to gene content, which is in agreement with previous studies (Gilbert et al., 2002, 2005; Ovchinnikov et al., 2001; Symer et al., 2002).

Our sequencing studies uncovered several expected trends and some unexpected results. All 37 hot L1 elements and the 6 low-level activity elements had two intact open reading frames (ORFs). A consensus sequence derived from these 37 hot L1s was identical at the amino acid level to a previously derived consensus (Brouha et al., 2003) (data not shown).

Inactive elements generally had frameshift (5/24) or chain-terminating nonsense mutations (9/24) in at least one of the L1 ORFs. However, ten of these low-level activity or inactive elements contained two intact open reading frames. One L1 (#3-24) contained an S228P missense mutation within the endonuclease (EN) domain of ORF2p (Feng et al., 1996; Weichenrieder et al., 2004). Though L1s containing EN mutations are unable to retrotranspose in HeLa cells, they can retrotranspose in Chinese Hamster Ovary (CHO) cells deficient in the nonhomologous end-joining (NHEJ) pathway of DNA repair, presumably by parasitizing a free 3′ OH group to initiate target-primed reverse transcription (TPRT) (Morrish et al., 2002, 2007). Interestingly, although #3-24 is inactive in NHEJ-proficient cell lines, the L1 retrotransposed at roughly 60% the efficiency of the wild-type control, L1.3, in NHEJ-deficient CHO cells (Morrish et al., 2002). Introducing the S228P change into L1.3 (Sassaman et al., 1997) also allowed efficient EN-independent retrotransposition, indicating that this mutation is largely responsible for the inactivity of #3-24 in HeLa cells (Figure S1).

Analysis of genomic sequences flanking the 68 L1Hs elements revealed a number of interesting findings. The poly(A) tails of 25 L1s were interrupted or contained 3′ transductions (Goodier et al., 2000; Holmes et al., 1994; Moran et al., 1999; Pickeral et al., 2000), 17 of which clustered into ''subfamilies'' of L1Hs elements. In one case, we identified an L1 (#2-1) as the likely source element for one of these subfamilies. For #1-3, 3-31, and 1-5, these transductions/interrupted poly(A) tails were identical to those in L1Hs elements that have caused disease-producing mutations (e.g., $L1_{RP}$, LRE3) (Brouha et al., 2002; Kimberland et al., 1999). In other cases, the transductions denote examples of recently amplified subfamilies (Goodier et al., 2000; Lander et al., 2001; Pickeral et al., 2000).

Examining the 5′ genomic flanks showed that the retrotransposition of a full-length L1 from the ABC9 genomic library (#2-24) that integrated on chromosome 10 was accompanied by ~250 bp of an Alu element that maps to chromosome 16. The Alu sequence is in the opposite transcriptional orientation to the L1, 13 bp of unmapped sequence separates the elements, and the whole insertion is flanked by target-site duplications (TSDs) (Figure S2). Thus, though most of the full-length L1Hs elements identified here have been amplified by canonical retrotransposition, recombination- and/or replication-mediated repair processes may facilitate the integration of some elements (Gilbert et al., 2002, 2005; Symer et al., 2002). Additionally, our screen allowed us to resolve possible sequence anomalies in the HGR. For example, one fosmid that lacks a dimorphic L1Hs element (#6-105) actually contains two L1s (a PA2 and pre-Ta element) that likely were collapsed into a harlequin element during the HGR assembly (Figure S2).

Finally, the data also enabled us to examine allelic heterogeneity associated with L1Hs elements. For example, one L1 (#5-70) was present in the HGR but contained a stop codon in ORF2 and was not previously tested for activity (Brouha et al., 2003). Interestingly, #5-70 retrotransposed at ~8% of the level of L1.3, further illustrating how allelic heterogeneity can impact retrotransposon activity (Lutz et al., 2003; Seleme et al., 2006).

Allele Frequencies of Genotyped Elements

The 68 L1Hs elements identified here are dimorphic with respect to presence; thus, we tested if a subset of these L1s represented population-restricted or potentially private alleles. To address this question, we first compiled existing genotyping data (Badge et al., 2003; Myers et al., 2002; Xing et al., 2009). Additional genotyping then was conducted on a subset of the L1s discovered here (26 in total; see Extended Experimental Procedures for selection criteria). The 26 L1s first were genotyped in a CEPH panel of 129 unrelated individuals. Nine L1s absent from the CEPH panel then were genotyped in a Zimbabwean panel of 72 unrelated individuals. Finally, if the element was absent from both panels, it was genotyped on the H952 subset of the HGDP consisting of ~1050 individuals from ~51 worldwide populations (Figure 3A and Table S4) (Cann et al., 2002; Rosenberg, 2006).

Two elements (#3-5 and 3-31) genotyped on the HGDP exist at very low allele frequencies and were only found in Africans. Two other L1Hs elements (#1-5 and 3-24) were absent from the HGDP (Table S4). Element #3-24 (the S228P mutant described above) was found in the ABC10 Yoruban library. Further genotyping revealed that the L1Hs element containing the mutation was present in her mother (but not her father), excluding a de novo origin (Figure 3B). The other putatively ''private'' L1Hs element was from G248 (#1-5), so we could not examine its segregation in a trio. Interestingly, this hot L1 insertion occurred into an intron of the *ABCA1* gene (Figure 3C); mutations in *ABCA1* have been associated with Tangier disease and low serum HDL levels (Frikke-Schmidt, 2010).

The Total Number of Active L1Hs Elements Present in ABC13

To estimate the total number of active L1s in one individual, we carried out in silico genotyping of the 68 L1Hs elements in

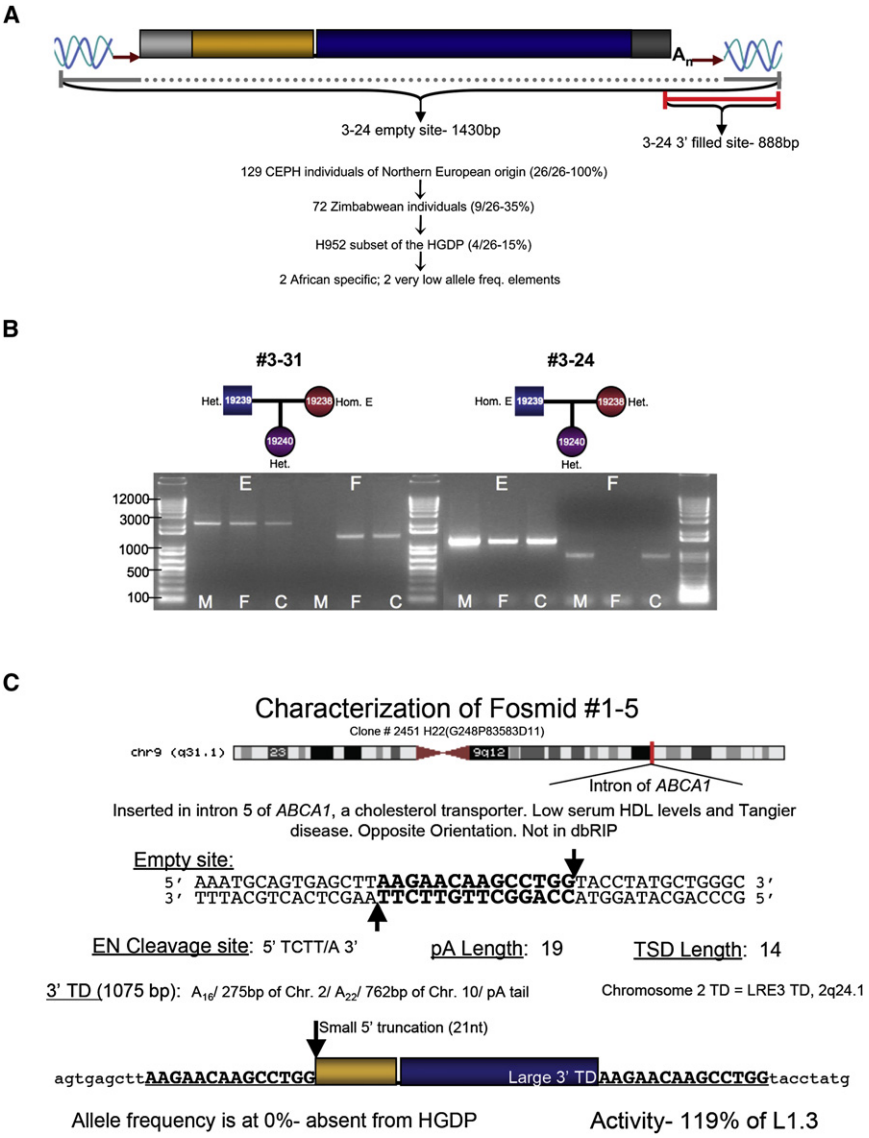

Figure 3. Allele Frequencies of L1Hs Alleles in the Population

(A) Genotyping assays: L1s were queried in panels of individuals for their absence (solid gray lines) or presence (red line). Genotyping of 26 elements in the three panels allowed the discovery of population restricted or potentially "private" L1Hs elements. The expected amplicon sizes are diagrammed for element #3-24.

(B) Pedigrees showing the inheritance of two elements typed in the ABC10 trio: Genotyping gels show the heritability of #3-31 (African specific) and #3-24 (absent from the HGDP). E and F at the top of the gel image indicate PCR results for empty and filled sites. M, F, and C at the bottom of the image indicate lanes for the mother, father, and child of the trio.

(C) Example datasheet for the G248 element #1-5: Empty site: insertion site in the HGR. EN cleavage site: the endonucleolytic cleavage site used by L1 EN to initiate retrotransposition. pA length: the approximate L1 poly(A) tail length; 3′ transductions and interrupted poly(A) tails also are annotated. TSD length: the length of the target site duplication flanking the L1Hs element (underlined lettering). Table S2 contains datasheets for each L1 in this study. Table S3 contains L1Hs insertion locations with respect to genes. Figure S2 displays a noncanonical L1Hs insertion and documents a possible sequence anomaly in the HGR.

ABC13, the last library examined in our subtractive scheme. We identified 20 regions containing distinct L1 insertions identified in the first five individuals that corresponded to insertion fosmids in the ABC13 HGSV track (http://hgsv.washington.edu/) of the UCSC genome browser (Figure 4A; Table S4) (Kent et al., 2002; Kidd et al., 2008). PCR genotyping confirmed that ABC13 contained 18 of these 20 elements (Figure 4B) and was homozygous with respect to presence for 3 of the elements. This result suggests that in silico genotyping could be used as a screening tool to identify L1Hs elements present at low allele frequencies in the population (Table S4).

Adding the 18 L1Hs elements identified by in silico genotyping to the 7 novel L1Hs elements identified in the ABC13 genome through our fosmid screens revealed that this individual contains 25/68 L1Hs elements identified in this study. Additional genotyping revealed that this individual contains 2 of the hot L1s characterized in a previous study (Table 1) (Brouha et al., 2003). Combining these numbers with our retrotransposition data indicates that the ABC13 genome contains 14 potentially hot L1Hs elements, and that at least 3 of these elements are present in a homozygous state.

Estimates of L1 Age

Our data suggest that, on average, the 68 L1Hs elements identified here are present at lower allele frequencies, are more active, and may be evolutionarily younger than those in previous studies (Brouha et al., 2003). To test this hypothesis, we derived maximum likelihood estimates for the ages of Ta-1 L1Hs elements in our dataset and that of Brouha et al. (Brouha et al., 2003; Marchani et al., 2009). This analysis revealed that the Ta-1 L1Hs elements identified here are significantly younger (1.0 million years [MY] 95% confidence interval [C.I.] 0.98–1.01 MY) than those reported previously (2.01 MY 95% C.I. 2.00–2.02 MY) (Marchani et al., 2009) (1.73 MY 95% C.I. 1.69–1.77 MY) (Brouha et al., 2003).

The maximum likelihood estimated age (Marchani et al., 2009) (1.0 MY) of the L1s reported here differs significantly from that calculated using the ad hoc method, which uses sequence divergence within subfamilies of elements to determine age (Carroll et al., 2001) (1.18 MY old). These two methods are known to be respectively robust (the maximum likelihood method) and sensitive (the ad hoc method) to the presence of multiple active lineages in the dataset (i.e., departures from the master gene

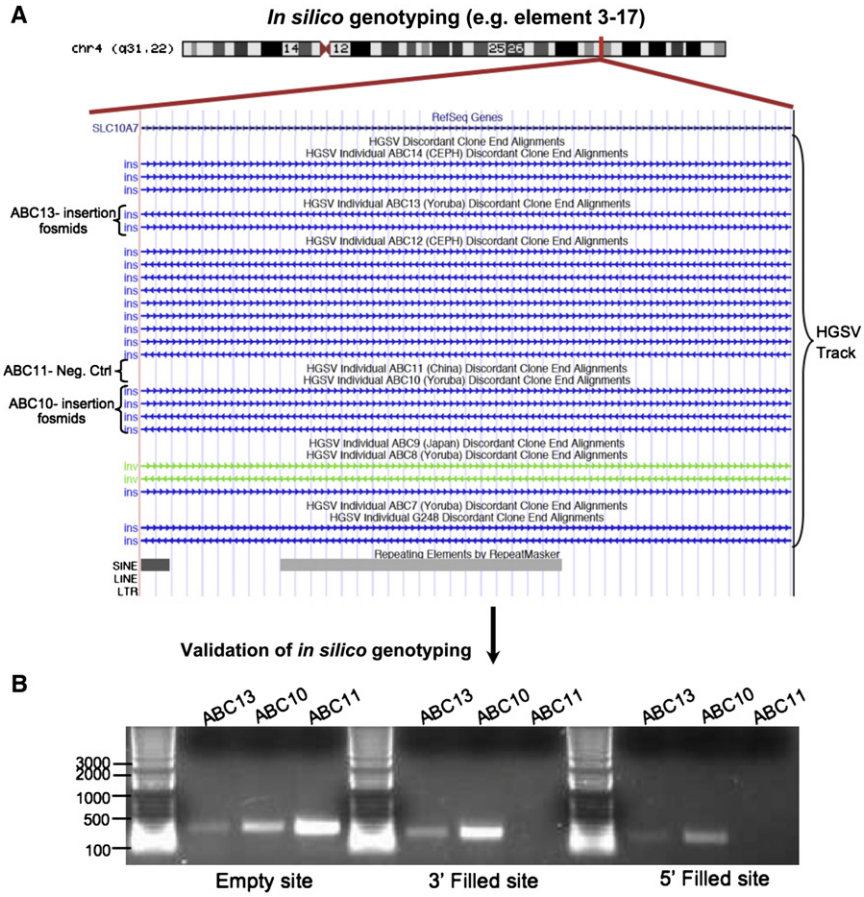

Figure 4. An Estimate of the Number of Active L1Hs Elements in an Individual (ABC13) Genome

(A) In silico genotyping: The last library in our study, ABC13, was examined in silico (see text) for the presence of insertion fosmids mapping to the location of L1Hs elements found in other individuals. Element 3-17 is used as an example. All blue lines represent insertion fosmids in the genomes of the eight individuals on the HGSV track (http://hgsv.washington.edu/) of the UCSC genome browser (http://genome.ucsc.edu) (Kent et al., 2002). The ABC7, 8, and 14 libraries were not investigated in this study.

(B) PCR validation: The elements identified in silico were genotyped using similar schemes to that shown in Figure 3A to validate the predictions from the HGSV track of the UCSC browser. Element 3-17 is used to illustrate the genotyping. ABC10 and ABC13 are heterozygous with respect to the L1Hs insertion. ABC11 lacks the L1Hs insertion. Table S4 displays genotyping results for all elements in this study.

model of L1 evolution) (Cordaux et al., 2004). The difference in these two estimates may indicate that members of multiple active L1Hs subfamilies are present in our dataset and suggests that the true age of the L1s may be younger than either calculation suggests. Indeed, the above data are consistent with the hypothesis that the HGR is strongly biased in favor of older, fixed L1Hs elements.

We next used a neighbor-joining approach, rooted with an intact chimpanzee L1 element, to generate a phylogenetic tree of the 68 full-length L1Hs elements (Figure 5, see Experimental Procedures). As predicted, pre-Ta elements were located near the root of the tree. Interestingly, two known ($L1_{RP}$ and LRE3) and five other currently amplifying subfamilies clustered together on the tree (Figure 5; see groups of colored elements), even though the interrupted poly(A) tail/transduction sequences themselves were excluded from the sequence alignments.

DISCUSSION

We have developed a systematic process to identify novel, dimorphic, active L1Hs elements in genomes of individuals from diverse geographic populations. Many of the newly identified L1Hs elements exist at low allele frequencies in the population, and four L1Hs elements represent "rare" alleles, three of which appear to be restricted to Africans. Sequence-based age estimates further reveal that these L1Hs elements appear to be, on average, evolutionarily younger than those identified in previous studies (Brouha et al., 2003; Marchani et al., 2009). These data are consistent with the notion that full-length active L1s are systematically underrepresented in available genome reference sequences (Badge et al., 2003; Boissinot et al., 2004; Brouha et al., 2003; Sassaman et al., 1997; Sheen et al., 2000; Xing et al., 2009).

Our study has underscored the effectiveness of fosmid paired-end libraries in the discovery of novel, active L1Hs elements. Though a number of technologies have been developed to identify polymorphic L1s (Badge et al., 2003; Bennett et al., 2004; Boissinot et al., 2004; Brouha et al., 2003; Moran et al., 1996; Myers et al., 2002; Sheen et al., 2000; Xing et al., 2009), the approach described here is not reliant upon PCR fidelity, readily allows the identification of active L1Hs elements, and makes sequencing of genomic flanking sequences, poly(A) tails, and L1-mediated transductions relatively straightforward. Thus, we predict that the fosmid-based approach likely will be superior to second-generation, low-coverage genome sequencing methodologies (e.g., many individual genomes characterized in the 1000 genomes project; http://www.1000genomes.org/page.php) for comprehensively identifying and characterizing rare L1 alleles in individual genomes. Indeed, recently published genome sequences highlight the difficulties in detecting and unambiguously mapping highly repetitive insertions (relative to a reference genome), including L1Hs elements (Bentley et al., 2008; McKernan et al., 2009; Wang et al., 2008; Wheeler et al., 2008).

Our analysis revealed that many active L1s cluster in small subfamilies. In the strictest sense, these data argue against a master gene model (Deininger et al., 1992) and instead support

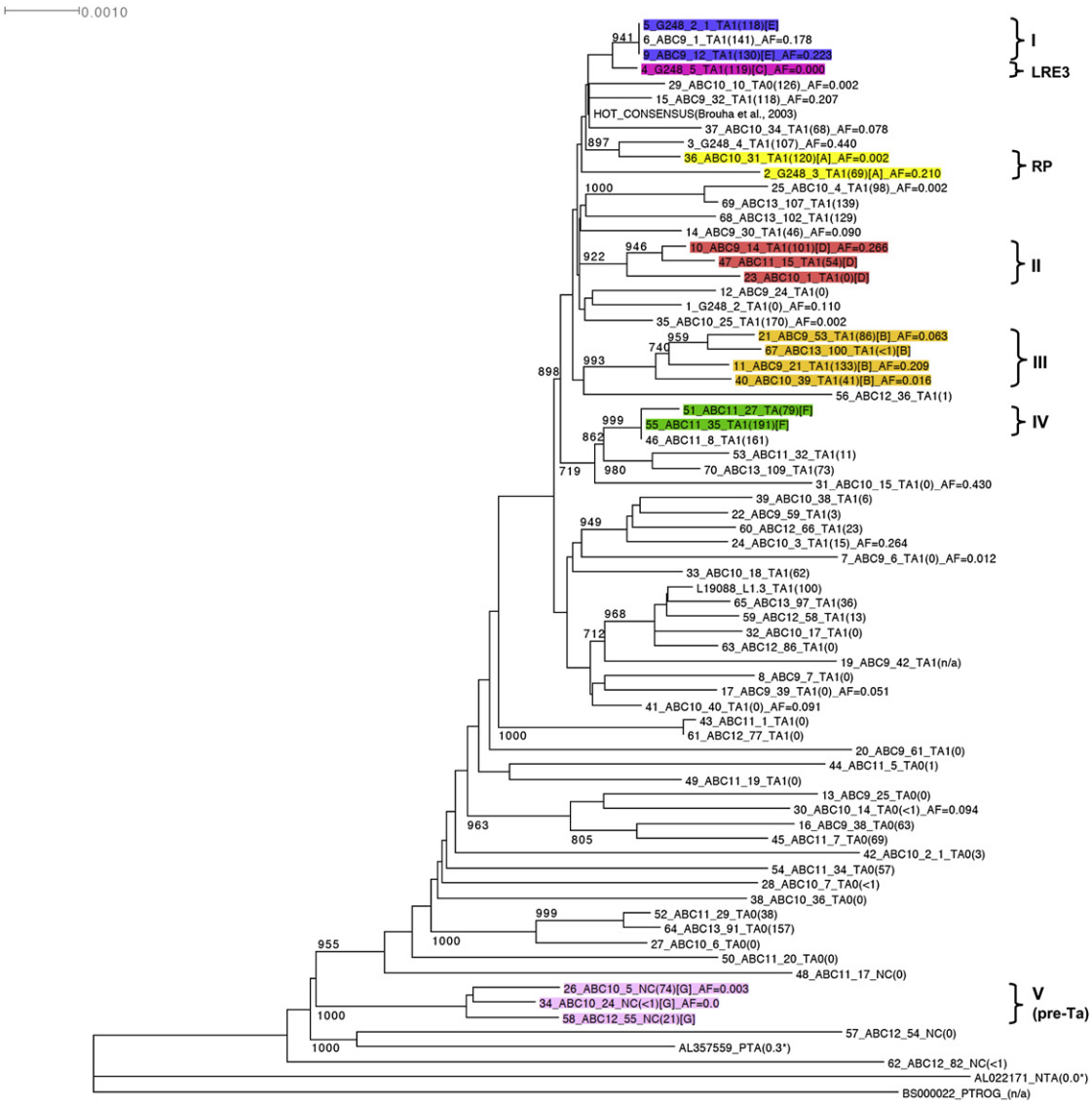

Figure 5. Phylogenetic Tree of the L1Hs Elements Identified in This Study

The tree is a single neighbor-joining tree (with branch lengths corrected using the Kimura 2 parameter model of nucleotide substitution) with 68 full-length elements from our study. The numbers at particular nodes indicate the number of times that node was observed in 1000 bootstrap replicates of the dataset. Only bootstrap values exceeding 70% are shown. The brackets at the right side indicate previously described "transduction subfamilies" ($L1_{RP}$ [labeled RP in the Figure] and LRE3) and distinct L1Hs subfamilies currently capable of amplifying in human genomes (I–V) (Goodier et al., 2000; Pickeral et al., 2000). Those subfamilies are highlighted in the same color to show their clustering on the tree. Retrotransposition activity (% relative to L1.3) as well as allele frequency (e.g., AF = 0.012), if determined, are appended to the sequence identifiers. Element #4-17 contains ACG characters in its 3′ UTR, which are diagnostic for pre-Ta L1s; however, the element clusters with the Ta0 subfamily. Activities for elements AL357559 and AL022171 were previously determined (Brouha et al., 2003). n/a = an L1 element not assayed for retrotransposition. The tree and age estimates use sequences indicated in Table S5.

a model in which multiple active source L1Hs elements (including members of both the pre-Ta and Ta subfamilies) are currently retrotransposing in modern human genomes (Cordaux et al., 2004). We cannot formally exclude a "stealth" model, where L1s in unfavorable expression contexts sometimes give rise to new retrotransposition-competent source elements that can be expressed from a more favorable genomic context (Han et al., 2005). However, the most parsimonious explanation of our data is that multiple source L1Hs elements and subfamilies with limited "life spans" exist in the genome. We posit that hot L1Hs elements must give rise to new, active progeny at a faster rate than they are inactivated by cellular mutational processes (see Figure 6 for model); this can lead to a scenario where small numbers of currently active L1Hs lineages may outcompete older L1s for limiting reagents, such as host factors (Boissinot and Furano, 2001). This competition scenario both supports and extends current lineage succession models and could potentially explain the monophyletic history of L1s and the

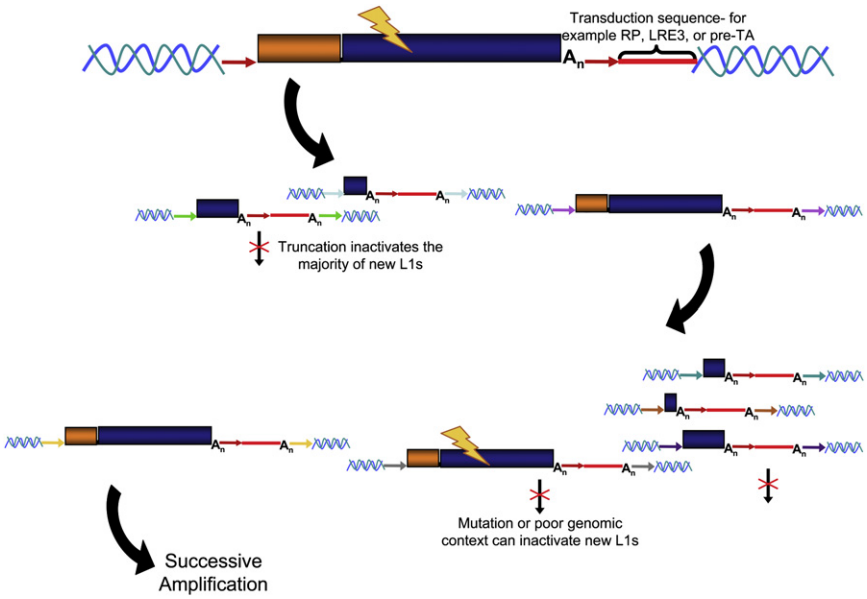

Figure 6. Multiple Source Loci Model for Continued L1Hs Activity

An element (source locus) that is both active and in a conducive genomic environment can retrotranspose. Shown here is an example of a progenitor element that can be associated with subsequent members of a "family" through the use of interrupted poly(A) tails and/or 3′ transduced sequence (3′ red arrow and line). Distinct elements are marked by distinguishing TSDs specific for their new integration site (different colored horizontal arrows). There are many of these families active in human genomes, such as $L1_{RP}$, LRE3, and the five families noted in Figure 5. Although host processes (lightning bolt) may inactivate some older elements, some of their descendents may retain the ability to retrotranspose and could harbor the 3′ transduction/interrupted poly(A) tail.

appearance of a replication-dominant L1Hs subfamily (Boissinot et al., 2000; Cordaux et al., 2004; Seleme et al., 2006).

Our dataset is still relatively small, and it remains difficult to estimate the actual number of hot L1s in the extant population. However, our ability to readily identify rare hot L1s in the genomes of geographically diverse individuals strongly suggests that these highly active L1Hs elements are more abundant in the population than previously appreciated. Indeed these results are in general agreement with the studies published by Iskow et al. (2010 [this issue of *Cell*]) and Huang et al. (2010 [this issue of *Cell*]).

The active L1Hs elements identified here also have the potential to impact modern human genomes by retrotransposing flanking genomic sequences to new chromosomal locations and by serving as substrates for nonallelic homologous recombination (reviewed in Cordaux and Batzer, 2009; Moran et al., 1999). The proteins encoded by these L1s also may promote the retrotransposition of Alu elements and noncoding RNAs (Bennett et al., 2008; Dewannieux et al., 2003; Garcia-Perez et al., 2007). Indeed, our data support the hypothesis that hot L1s are actively retrotransposing in modern-day human genomes and suggest that some of the L1 alleles identified here could serve as source elements for disease-producing L1 insertions.

EXPERIMENTAL PROCEDURES

Creation of Fosmid Libraries and Identification of Insertion-Containing Fosmids

Genomic DNA from the six individuals was obtained from transformed lymphoblastoid cell lines (available from the Coriell Cell Repository). The DNA was hydrodynamically sheared, end-repaired, size selected for 40 kb fragments by pulsed field gel electrophoresis, and ligated into fosmid vectors (Donahue and Ebling, 2007). Agencourt Biosciences Corporation constructed all libraries, with the exception of the G248 library, which was constructed as part of the human genome project finishing effort. From each library, approximately 1 million individual cloned fragments were arrayed into 384-well plates. End-sequence pairs were obtained from both ends of each DNA fragment using standard capillary sequencing and were mapped back to the HGR. Insertion-containing fosmids were identified as the subset of fosmids containing an apparent insert that was ~3 standard deviations smaller than the library mean (Kidd et al., 2008; Tuzun et al., 2005).

Screening of Fosmid Clones for LINE-1 Insertions

Insertion-containing fosmids identified in silico were screened for L1Hs elements in the following manner. First, all insertion fosmids were subjected to allele-specific oligonucleotide hybridization to identify characters in the 5′ UTRs of newer L1 subfamilies (Badge et al., 2003; Boissinot et al., 2000). This protocol was adapted from "hybridization of bacterial DNA on filters" (Sambrook et al., 1989). Fosmid DNAs were prepared according to the Very Low-Copy Plasmid/Cosmid Purification protocol for the QIAGEN-tip 100 Midi prep kit (QIAGEN). Those DNAs were subjected to Southern blotting followed by ATLAS (Badge et al., 2003) and/or direct sequencing to identify L1Hs elements that were absent from the HGR. Sequences flanking the L1Hs elements then were used as probes in BLAT searches at the UCSC genome browser (http://genome.ucsc.edu/) to determine the insertion site in the HGR (Kent, 2002; Kent et al., 2002). Detailed protocols for each step of the screening process, as well as the number of fosmids positive at each stage of the analysis, can be found in the Extended Experimental Procedures.

Cloning of L1s

In general, L1Hs elements were cloned directly from insertion-containing fosmids by digestion with AccI (Sassaman et al., 1997). The restricted DNA was separated on a 0.8% agarose gel, and the ~6 kb L1-containing restriction fragment was cloned into an L1 expression vector. This method captures the vast majority of the L1Hs sequence, leaving only the first ~35 bp and last ~50 bp of the original L1 5′ and 3′ UTRs present in the cloning vector, respectively. One element, #2-42, was refractory to this cloning procedure, as it contains a polymorphism near the 3′ end of ORF2 that creates an additional AccI site. The PDH L1.3 mutant was generated by site-directed mutagenesis. Each L1Hs element was sequenced in its entirety. Detailed protocols for the creation of each construct are included in the Extended Experimental Procedures.

L1 Retrotransposition Assays

We used a modification of a transient transfection protocol to conduct retrotransposition assays in HeLa and CHO cells (Moran et al., 1996; Morrish et al., 2002; Wei et al., 2000). Briefly, cells in 6-well dishes were transfected using the Fugene 6 agent (Roche) with 1 μg of plasmid (containing the indicator cassette) per each well. Cells were fed with media ~24 hr post-plating and daily from 72 hr or 5 days with media containing either 400 μg/ml G418 or

10 μg/ml blasticidin, respectively. Fourteen days post-transfection, cells were fixed and stained with 0.1% crystal violet. Colonies were counted in the appropriate wells, and these counts were normalized to green fluorescent protein (GFP) transfection efficiency. Detailed protocols for culture and assay conditions are found in the Extended Experimental Procedures.

Genotyping and Panels

The genomic locations of L1Hs insertions were compared to a database of human retrotransposon insertion polymorphisms (dbRIP; http://dbrip.brocku.ca/) (Wang et al., 2006). PCR genotyping assays were designed for a subset of L1Hs elements that were not completely annotated in dbRIP. Genotyping initially was conducted on a CEPH panel of 129 unrelated individuals of Northern European ancestry. If a L1Hs element was absent from the CEPH panel, it was genotyped on a panel containing genomic DNAs from 72 unrelated Zimbabwean individuals. Finally, if an L1Hs element was absent from both genotyping panels, it was genotyped on the H952 subset (Rosenberg, 2006) of the HGDP (Cann et al., 2002) (see Figure 3A). In silico genotyping was conducted using the HGSV track of the UCSC genome browser (Kent et al., 2002; Kidd et al., 2008). Details about these analyses are in the Extended Experimental Procedures.

Estimation of L1 Element Age

Sequences of the 69 full-length L1 elements were classified into subfamilies using the L1Xplorer analysis website (Penzkofer et al., 2005). Ta-1, Ta-0, and Non-Canonical (NC) (Brouha et al., 2003) elements were separately aligned using Muscle 3.52 (Edgar, 2004) on the Phylomen web server (http://phylemon.bioinfo.cipf.es/cgi-bin/home.cgi) (Tarraga et al., 2007). Raw alignments were manually refined using Jalview to remove all indels, all variable CpG sites, and the L1 polypurine tract (Waterhouse et al., 2009). Maximum likelihood estimates of the age (T) of each group, the sampling variance of T, and its 95% C.I. were calculated using the mleT script (Marchani et al., 2009) running under Matlab 7.2−2007a (The Mathworks Inc., Natick, MA, USA). The subroutine CountMutations (Marchani et al., 2009) was also utilized to calculate the number of substitutions in the datasets to enable the "ad hoc" subfamily age estimation method (Marchani et al., 2009).

Phylogenetic Tree

The sequences of the 69 elements were aligned as described above. An intact chimpanzee element (BS000022_PTROG) was used to root the tree. The alignment also includes an intact Ta-1 L1 (L19088_L1.3), a non-Ta L1 (AL022171_NTA), a pre-Ta L1 (AL357559), and the "Hot Consensus" L1 element from Brouha et al. (2003). Raw alignments were manually refined using Jalview (Waterhouse et al., 2009) to remove large indels and truncated elements; this led to the exclusion of #6-113 due to a large 5′ UTR deletion.

A single neighbor-joining tree of the 68 remaining full-length elements was constructed using the PHYLIP package (Felsenstein, 1989). Branch lengths were corrected using the Kimura 2 parameter model (Kimura, 1980). To assess the reliability of the phylogeny, 1000 bootstrapped resamples of the multiple alignment were made using the seqboot program of the PHYLIP package (Felsenstein, 1989). The neighbor-joining tree derived from the full dataset was manually annotated with bootstrap values using Dendroscope (Huson et al., 2007) (Figure 5). Only bifurcations that occurred in more than 70% of bootstrap resamples are labeled.

ACCESSION NUMBERS

Accession numbers for all elements are tabulated in Table S5. Two L1Hs elements (accession numbers (#1-5) GU477636 and (#6-102) GU477637) were recently posted in GenBank.

SUPPLEMENTAL INFORMATION

Supplemental Information contains Extended Experimental Procedures, two figures, and five tables and can be found with this article online at doi:10.1016/j.cell.2010.05.021.

ACKNOWLEDGMENTS

We thank Professor Sir Alec Jeffreys FRS for access to CEPH and Zimbabwean DNA samples and Professor Mark Jobling for access to HGDP DNA samples. We thank Dr. Elizabeth Marchani for advice on maximum likelihood age estimates and Dr. José Luis Garcia-Perez for plasmid JJ105/L1.3. We thank Dr. Garcia-Perez and members of the Moran lab for helpful comments. C.R.B. was supported in part by NIH training grants T32GM7544 and T32000040. J.M.K. was supported by a National Science Foundation Graduate Research Fellowship. Work in the laboratory of E.E.E. was supported by grant HG004120. P.C. and C.M. were supported by a Wellcome Trust Project Grant (075163/Z/04/Z) to R.M.B. and Professor Sir Alec Jeffreys FRS. J.V.M. is supported by NIH grants GM066695 and GM060518. The University of Michigan Cancer Center Support Grant (5P30CA46592) helped defray sequencing costs incurred in this study. J.V.M. and E.E.E. are Investigators of the Howard Hughes Medical Institute. E.E.E. is a member of the SAB for Pacific Biosciences.

Received: January 15, 2010
Revised: March 23, 2010
Accepted: May 13, 2010
Published: June 24, 2010

REFERENCES

Athanikar, J.N., Badge, R.M., and Moran, J.V. (2004). A YY1-binding site is required for accurate human LINE-1 transcription initiation. Nucleic Acids Res. *32*, 3846–3855.

Babushok, D.V., and Kazazian, H.H., Jr. (2007). Progress in understanding the biology of the human mutagen LINE-1. Hum. Mutat. *28*, 527–539.

Badge, R.M., Alisch, R.S., and Moran, J.V. (2003). ATLAS: a system to selectively identify human-specific L1 insertions. Am. J. Hum. Genet. *72*, 823–838.

Bennett, E.A., Coleman, L.E., Tsui, C., Pittard, W.S., and Devine, S.E. (2004). Natural genetic variation caused by transposable elements in humans. Genetics *168*, 933–951.

Bennett, E.A., Keller, H., Mills, R.E., Schmidt, S., Moran, J.V., Weichenrieder, O., and Devine, S.E. (2008). Active Alu retrotransposons in the human genome. Genome Res. *18*, 1875–1883.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Boeke, J.D. (2003). The unusual phylogenetic distribution of retrotransposons: a hypothesis. Genome Res. *13*, 1975–1983.

Boissinot, S., and Furano, A.V. (2001). Adaptive evolution in LINE-1 retrotransposons. Mol. Biol. Evol. *18*, 2186–2194.

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

Boissinot, S., Entezam, A., Young, L., Munson, P.J., and Furano, A.V. (2004). The insertional history of an active family of L1 retrotransposons in humans. Genome Res. *14*, 1221–1231.

Brouha, B., Meischl, C., Ostertag, E., de Boer, M., Zhang, Y., Neijens, H., Roos, D., and Kazazian, H.H., Jr. (2002). Evidence consistent with human L1 retrotransposition in maternal meiosis I. Am. J. Hum. Genet. *71*, 327–336.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., and Kazazian, H.H., Jr. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. USA *100*, 5280–5285.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.

Carroll, M.L., Roy-Engel, A.M., Nguyen, S.V., Salem, A.H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M., et al. (2001). Large-scale

analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J. Mol. Biol. *311*, 17–40.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. *10*, 691–703.

Cordaux, R., Hedges, D.J., and Batzer, M.A. (2004). Retrotransposition of Alu elements: how many sources? Trends Genet. *20*, 464–467.

Cost, G.J., and 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.

Deininger, P.L., Batzer, M.A., Hutchison, C.A., 3rd, and Edgell, M.H. (1992). Master genes in mammalian repetitive DNA amplification. Trends Genet. *8*, 307–311.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. *35*, 41–48.

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

Donahue, W.F., and Ebling, H.M. (2007). Fosmid libraries for genomic structural variation detection. Curr. Protoc. Hum. Genet. *Chapter 5*, Unit 5 20.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Felsenstein, J. (1989). PHYLIP- Phylogeny Interference Package (Version 3.2). Cladistics *5*, 164–166.

Feng, Q., Moran, J.V., Kazazian, H.H., Jr., and Boeke, J.D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell *87*, 905–916.

Freeman, J.D., Goodchild, N.L., and Mager, D.L. (1994). A modified indicator gene for selection of retrotransposition events in mammalian cells. Biotechniques *17*, 46.

Frikke-Schmidt, R. (2010). Genetic variation in the ABCA1 gene, HDL cholesterol, and risk of ischemic heart disease in the general population. Atherosclerosis *208*, 305–316, Published online June 11, 2009.

Garcia-Perez, J.L., Doucet, A.J., Bucheton, A., Moran, J.V., and Gilbert, N. (2007). Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. Genome Res. *17*, 602–611.

Gilbert, N., Lutz, S., Morrish, T.A., and Moran, J.V. (2005). Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol. Cell. Biol. *25*, 7780–7795.

Gilbert, N., Lutz-Prigge, S., and Moran, J.V. (2002). Genomic deletions created upon LINE-1 retrotransposition. Cell *110*, 315–325.

Goodier, J.L., Ostertag, E.M., and Kazazian, H.H., Jr. (2000). Transduction of 3′-flanking sequences is common in L1 retrotransposition. Hum. Mol. Genet. *9*, 653–657.

Grimaldi, G., and Singer, M.F. (1983). Members of the KpnI family of long interspersed repeated sequences join and interrupt alpha-satellite in the monkey genome. Nucleic Acids Res. *11*, 321–338.

Han, K., Xing, J., Wang, H., Hedges, D.J., Garber, R.K., Cordaux, R., and Batzer, M.A. (2005). Under the genomic radar: the stealth model of Alu amplification. Genome Res. *15*, 655–664.

Holmes, S.E., Dombroski, B.A., Krebs, C.M., Boehm, C.D., and Kazazian, H.H., Jr. (1994). A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. Nat. Genet. *7*, 143–148.

Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics *8*, 460.

Huang, C.R.L., Schneider, A.M., Lu, Y., Niranjan, T., Shen, P., Robinson, M.A., Steranka, J.P., Valle, D., Civin, C.I., Wang, T., et al. (2010). Mobile Interspersed repeats are major structural variants in the human genome. Cell *141*, this issue, 1171–1182.

International HapMap Consortium. (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.

Iskow, R.C., McCabe, M.T., Mills, R.E., Torene, S., Pittard, W.S., Neuwald, A.F., Van Meir, E.G., Vertino, P.M., and Devine, S.E. (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. Cell *141*, this issue, 1253–1261.

Kazazian, H.H., Jr., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., and Antonarakis, S.E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature *332*, 164–166.

Kent, W.J. (2002). BLAT–the BLAST-like alignment tool. Genome Res. *12*, 656–664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56–64.

Kimberland, M.L., Divoky, V., Prchal, J., Schwahn, U., Berger, W., and Kazazian, H.H., Jr. (1999). Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. Hum. Mol. Genet. *8*, 1557–1560.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. *16*, 111–120.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science *318*, 420–426.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lavie, L., Maldener, E., Brouha, B., Meese, E.U., and Mayer, J. (2004). The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. Genome Res. *14*, 2253–2260.

Lutz, S.M., Vincent, B.J., Kazazian, H.H., Jr., Batzer, M.A., and Moran, J.V. (2003). Allelic heterogeneity in LINE-1 retrotransposition activity. Am. J. Hum. Genet. *73*, 1431–1437.

Marchani, E.E., Xing, J., Witherspoon, D.J., Jorde, L.B., and Rogers, A.R. (2009). Estimating the age of retrotransposon subfamilies using maximum likelihood. Genomics *94*, 78–82.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., et al. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. *19*, 1527–1541.

Moran, J.V., DeBerardinis, R.J., and Kazazian, H.H., Jr. (1999). Exon shuffling by L1 retrotransposition. Science *283*, 1530–1534.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., and Kazazian, H.H., Jr. (1996). High frequency retrotransposition in cultured mammalian cells. Cell *87*, 917–927.

Morrish, T.A., Garcia-Perez, J.L., Stamato, T.D., Taccioli, G.E., Sekiguchi, J., and Moran, J.V. (2007). Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. Nature *446*, 208–212.

Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A., and Moran, J.V. (2002). DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat. Genet. *31*, 159–165.

Myers, J.S., Vincent, B.J., Udall, H., Watkins, W.S., Morrish, T.A., Kilroy, G.E., Swergold, G.D., Henke, J., Henke, L., Moran, J.V., et al. (2002). A comprehensive analysis of recently integrated human Ta L1 elements. Am. J. Hum. Genet. *71*, 312–326.

Ovchinnikov, I., Troxel, A.B., and Swergold, G.D. (2001). Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res. *11*, 2050–2058.

Penzkofer, T., Dandekar, T., and Zemojtel, T. (2005). L1Base: from functional annotation to prediction of active LINE-1 elements. Nucleic Acids Res. *33*, D498–D500.

Pickeral, O.K., Makalowski, W., Boguski, M.S., and Boeke, J.D. (2000). Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res. *10*, 411–415.

Rosenberg, N.A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann. Hum. Genet. *70*, 841–847.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual, Second Edition (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory press).

Sassaman, D.M., Dombroski, B.A., Moran, J.V., Kimberland, M.L., Naas, T.P., DeBerardinis, R.J., Gabriel, A., Swergold, G.D., and Kazazian, H.H., Jr. (1997). Many human L1 elements are capable of retrotransposition. Nat. Genet. *16*, 37–43.

Seleme, M.C., Vetter, M.R., Cordaux, R., Bastone, L., Batzer, M.A., and Kazazian, H.H., Jr. (2006). Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. Proc. Natl. Acad. Sci. USA *103*, 6611–6616.

Sheen, F.M., Sherry, S.T., Risch, G.M., Robichaux, M., Nasidze, I., Stoneking, M., Batzer, M.A., and Swergold, G.D. (2000). Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res. *10*, 1496–1508.

Skowronski, J., Fanning, T.G., and Singer, M.F. (1988). Unit-length line-1 transcripts in human teratocarcinoma cells. Mol. Cell. Biol. *8*, 1385–1397.

Smit, A.F., Toth, G., Riggs, A.D., and Jurka, J. (1995). Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J. Mol. Biol. *246*, 401–417.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. (2002). Human l1 retrotransposition is associated with genetic instability in vivo. Cell *110*, 327–338.

Tarraga, J., Medina, I., Arbiza, L., Huerta-Cepas, J., Gabaldon, T., Dopazo, J., and Dopazo, H. (2007). Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics. Nucleic Acids Res. *35*, W38–W42.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. (2005). Fine-scale structural variation of the human genome. Nat. Genet. *37*, 727–732.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M.A., and Liang, P. (2006). dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum. Mutat. *27*, 323–329.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y., et al. (2008). The diploid genome sequence of an Asian individual. Nature *456*, 60–65.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191.

Wei, W., Morrish, T.A., Alisch, R.S., and Moran, J.V. (2000). A transient assay reveals that cultured human cells can accommodate multiple LINE-1 retrotransposition events. Anal. Biochem. *284*, 435–438.

Weichenrieder, O., Repanas, K., and Perrakis, A. (2004). Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. Structure *12*, 975–986.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876.

Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, Q., Kirkness, E.F., Levy, S., Batzer, M.A., et al. (2009). Mobile elements create structural variation: analysis of a complete human genome. Genome Res. *19*, 1516–1526.