12-9-2008

L1 recombination-associated deletions generate human genomic variation

Kyudong Han
Louisiana State University

Jungnam Lee
Louisiana State University

Thomas J. Meyer
Louisiana State University

Paul Remedios
Louisiana State University

Lindsey Goodwin
Louisiana State University

See next page for additional authors

Recommended Citation

Han, K., Lee, J., Meyer, T., Remedios, P., Goodwin, L., & Batzer, M. (2008). L1 recombination-associated deletions generate human genomic variation. *Proceedings of the National Academy of Sciences of the United States of America, 105* (49), 19366-19371. https://doi.org/10.1073/pnas.0807866105

Authors
Kyudong Han, Jungnam Lee, Thomas J. Meyer, Paul Remedios, Lindsey Goodwin, and Mark A. Batzer

L1 recombination-associated deletions generate human genomic variation

Kyudong Han, Jungnam Lee, Thomas J. Meyer, Paul Remedios, Lindsey Goodwin, and Mark A. Batzer[1]

Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, Baton Rouge, LA 70803

Edited by John C. Avise, University of California, Irvine, CA, and approved October 3, 2008 (received for review August 11, 2008)

Mobile elements have created structural variation in the human genome through their *de novo* insertions and post-insertional genomic rearrangements. L1 elements are a type of long interspersed element (LINE) that is dispersed at high copy numbers within most mammalian genomes. To determine the magnitude of L1 recombination-associated deletions (L1RADs), we computationally extracted L1RAD candidates by comparing the human and chimpanzee genomes and verified each of the L1RAD events by using wet-bench analyses. Through these analyses, we identified 73 human-specific L1RAD events that occurred subsequent to the divergence of the human and chimpanzee lineages. Despite their low frequency, the L1RAD events deleted ≈450 kb of the human genome. One L1RAD event generated a large deletion of ≈64 kb. Multiple alignments of prerecombination and postrecombination L1 elements suggested that two different deletion mechanisms generated the L1RADs: nonallelic homologous recombination (55 events) and nonhomologous end joining between two L1s (18 events). In addition, the position of L1RADs throughout the genome does not correlate with local chromosomal recombination rates. This process may be implicated in the partial regulation of L1 copy numbers by the finding that ≈60% of the DNA sequences deleted by the L1RADs consist of L1 sequences that were either directly involved in the recombination events or located in the intervening sequence between recombining L1s. Overall, there is increasing evidence that L1RADs have played an important role in creating structural variation.

LINE-1 | nonallelic homologous | nonhomologous end joining | retrotransposon

Long interspersed elements (LINE-1s or L1s) are universal constituents of mammalian genomes and account for ≈17% of the human genome (1). They have expanded to ≈ 520,000 copies over the last 150 million years (1, 2). Full-length L1s are ≈6 kb long, and encode two ORFs (ORF1 and ORF2), which code for a 40-kDa RNA-binding protein with nucleic acid chaperone activity (3) and a 150-kDa protein with both endonuclease (EN) and reverse transcriptase (RT) activities (4–6). L1s mobilize via an RNA intermediate to integrate themselves into genomic DNA at the target site. However, ≈99.8% of L1s in the human genome are unable to retrotranspose (7), either because of point mutations or structural deficiencies (e.g., 5′ truncations, 5′ inversions, or other internal rearrangements) (8–10). Consequently, only 80–100 retrotransposition-competent L1s capable of autonomous retrotransposition are located in the human genome (7, 11).

Homologous recombination between closely related DNA fragments occurs in all living organisms (12, 13). A recent study of human genomic deletions caused by unequal homologous recombination between two *Alu* elements showed that 492 human-specific deletion events resulted in a total of ≈400 kb DNA being lost since the divergence of the human and chimpanzee lineages (14). Similar to the *Alu* elements, L1s may have been a source of recombination-associated genomic deletion throughout human evolution because of their high copy numbers and relatively long stretches of sequence identity. Surprisingly, only three L1 recombination-associated deletion (L1RAD) events causing human diseases (i.e., glycogen storage disease, Alport Syndrome-Diffuse Leiomyomatosis, and Ellis–van Creveld syndrome) have been reported (15–17). However, there have been no previous systematic studies of the genome-wide impact of this process in the human lineage. Here, we report the identification and characterization of 73 human lineage-specific L1RAD events that have occurred since divergence of the human and chimpanzee lineages (≈6 million years ago) (18, 19).

Results and Discussion

Identification of L1RAD Events in the Human Genome. To investigate the genome-wide impact of L1RADs on the human genome, we computationally compared the position of L1s in the human genome (hg18) to orthologous positions in the chimpanzee genome (panTro2). After various computational filtrations, a total of 4,786 putative L1RAD candidate loci were retrieved for further examination (see *Materials and Methods* for details). We analyzed and discarded 546 of the 4,786 loci as false positives because of (*i*) insertions of an L1 or other type of repetitive element at the orthologous chimpanzee locus (181 events), (*ii*) computational errors in the alignment of the human and chimpanzee genomes (99 events), and (*iii*) other genomic rearrangements (e.g., translocation, gene conversion, and retrotransposition-mediated deletion) (266 events) (Supporting Information (SI) Fig. S1). Of the remaining 4,240 loci, we found 98 L1RAD candidate loci that did not contain poly (N) stretches (i.e., partially unsequenced regions) in the orthologous chimpanzee locus. The remaining 4,142 loci were ambiguous because of the inclusion of poly (N) stretches in the chimpanzee sequence. So, we investigated these loci based either on target site duplication (TSD) structure or by using rhesus macaque and orangutan reference sequences (rheMac2 and ponAbe2, respectively), that encompass the unsequenced chimpanzee genomic region. For TSD structure analyses, the chimeric L1 created by an L1RAD event in the human was expected to lack matching TSDs whereas the orthologous chimpanzee L1s retain the normal, matching TSD structure as described in the study of *Alu* recombination-mediated deletion (ARMD) (14). By applying the criteria mentioned above, we collected 117 more L1RAD candidates from the 4,142 loci that included partially unsequenced regions of the chimpanzee genome. The 215 putative L1RAD candidates were then examined by using locus-specific PCR to confirm their

Author contributions: K.H. and M.A.B. designed research; K.H., J.L., P.R., and L.G. performed research; M.A.B. contributed new reagents/analytic tools; K.H., J.L., T.J.M., and M.A.B. analyzed data; and K.H. and M.A.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. EU916983–EU917024).

See Commentary on page 19033.

[1] To whom correspondence should be addressed at: Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803. E-mail: mbatzer@lsu.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0807866105/DCSupplemental.

Table 1. Summary of human-specific L1RAD events

Classification	Number of loci
Putative L1RAD events	215
False events	142
Sequence disagreement caused by the chimpanzee genome	98
Sequence disagreement caused by the human genome	2
L1 insertion in the chimpanzee genome	40
Others (e.g. lineage sorting and L1IMD)	2
L1 recombination-associated deletions	73
Nonallelic homologous recombination	55
Nonhomologous end joining	18

status as authentic L1RAD events (Table 1). Six of these loci could not be amplified via PCR because of the presence of other repeat elements in the flanking sequence. These six were examined by either the comparison of the chimeric and prerecombination L1s and/or triple alignment of multiple species (14, 20). The analysis resulted in the recovery of 73 events that were classified as authentic human-specific L1RAD events (Fig. S2 and Table S1).

Impact of Genomic Deletions Associated with Recombination Between Two L1s. Subsequent to the divergence between the human and chimpanzee lineages, the total amount of human DNA deleted by the 73 identified L1RAD events, that have occurred throughout the genome aside from chromosomes 16 and 21 (Fig. S2), is estimated to be 447,567 bp. The size of human-specific L1RADs ranges from 56 to 64,113 bp, with an average length of 6,132 bp and a median length of 3,239 bp (Fig. 1). We investigated the 146 L1s that were involved in the 73 L1RAD events. As expected, most of the L1s (85%) involved in these events are truncated, with only 22 elements that were full-length L1s (>6 kb). Interestingly, 13 elements were shown to be inverted/truncated L1s that were generated by twin priming (9), four of which involved a chimeric L1 that was both 5′ and 3′ truncated. The size distribution of human-specific L1RADs indicates that these events are skewed toward smaller deletion sizes (<4 kb). However, this skew is not as pronounced as the one reported (<0.5 kb) for human-specific L1 insertion-mediated deletions (L1IMDs) (21). In addition, the L1RAD process has deleted 25 times as much human genomic sequence as the L1IMD process. Surprisingly, the largest deleted sequence was ≈64 kb in length, within which only the *LOC469769* pseudogene and two intergenic regions are found in the chimpanzee ortholog. This deletion is fixed in 80 human individuals (see *SI Text*) and is the largest mobile element recombination-associated deletion reported to date. Overall, the size of L1RADs is positively correlated with the size of the longer L1 insert of the two L1s involved in each L1RAD (r = 0.258, P = 0.0275). One explanation of this finding is that, when we analyzed the correlation between the sizes of the two L1s involved in each L1RAD, we found the sizes of the two L1s to be positively correlated (r = 0.431, P = 0.0001) with one another. This implies that longer L1s have a higher probability of possessing more regions of homology with other long L1s than with shorter L1s. This observation, combined with the expectation that larger L1s will be less densely distributed in the genome than smaller L1s, suggests that longer L1s participate in larger deletions. Therefore, we conclude that larger L1s contribute more to overall genomic instability in the human genome than do shorter L1 elements.

To determine the possible effects of the elimination of ancestral genomic sequences during the 73 human-specific L1RAD events, we compared the prerecombination sequences (i.e., orthologous chimpanzee sequences) with the human genome. This analysis showed that ≈27% of the L1RAD events were located within predicted or known RefSeq genes. When compared with the ARMD events, the density of L1RAD events within genic regions was relatively low (Table 2). This result is not unexpected because 66% of *Alu* elements are located in intronic regions whereas only 58% of L1s are located in intronic regions (22). In other words, the universal distribution of L1s is biased toward gene-poor regions relative to their *Alu* counterparts. Nevertheless, one L1RAD event generated exonic deletions in two genes annotated as putatively functional in the chimpanzee genome. One of the two genes, *LOC745816*, encodes a hypothetical protein. The other, *LOC457712*, is a model chimpanzee gene similar to a sorting nexin (*SNX*) *25* gene. In the human lineage, SNX 25 is one of the cellular tracking proteins (23) and has been predicted to encode phox homology (PX), PX associated, and a regulator of G protein domains (24). However, the role of SNX 25 is currently unclear.

L1RADs Created by Two Different Mechanisms. To analyze the recombination junction, sequence alignment between prerecombination and postrecombination L1s involved in the 73 L1RAD events was performed by using BioEdit (25). We found that the L1RAD events were generated by two different mechanisms. Among them, 55 L1RADs were generated by nonallelic homologous recombination (NAHR). In this mechanism, the two prerecombination L1s still present in the chimpanzee genome have recombined into a single chimeric L1 in the human genome. This recombination occurs at a point within the identical sequence shared by the two L1s that averages 40 bp in length. In the human genome, the resulting chimeric L1 is recognized as a single element by RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker), and only careful analysis of L1 alignments and TSDs demonstrate its chimeric nature. The other

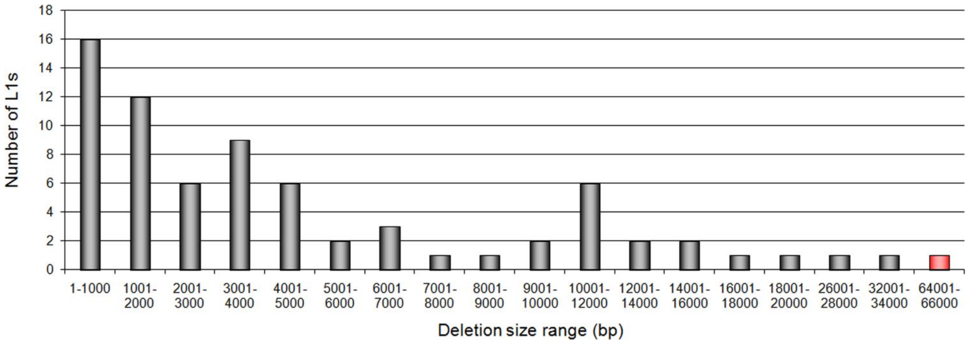

Fig. 1. Size distribution of the L1RADs. The size distribution of DNA sequences deleted by human-specific L1RAD events is displayed. The largest deleted sequence is 64,113 bp, represented by a red bar.

Table 2. Comparison of the L1RADs and ARMDs

Feature	L1RAD	ARMD*
Total events	73	492
Total deletion size, kb	≈450	≈400
Maximum deletion size, kb	64	7.3
Mean of deletion size, bp	6132	806
Median of deletion size, bp	3239	486
GC content (neighboring 20 kb)	38%	45%
Gene density (neighboring 4 Mb)	1 gene/105 kb	1 gene/66 kb
Located within genes	27%	60%
Exonic deletions (loci)	2	3

*The features of ARMD events are derived from Sen *et al.* (14).

mechanism resulting in L1RAD events is nonhomologous end joining (NHEJ) between two L1s, a process shown to be responsible for the other 18 L1RADs. It is known that NHEJ typically involves microhomology between sequences (12, 26, 27). After the alignment of two prerecombination L1s and their chimeric L1 recombination product, we identified the microhomologies between the two prerecombination L1s. These microhomologies were found to range from 1 to 6 bp. Unlike the chimeric L1s generated by the NAHR-L1RAD events, the NHEJ-L1RAD process produced two different contiguous L1s rather than a single chimeric L1.

The analysis of L1 subfamilies involved in the NAHR-L1RAD events shows that the number of elements from each L1 subfamily is proportional to their genome-wide copy number (Fig. 2*A*). This is an expected result as NAHR events occur via the mispairing of two closely related L1 sequences that share a relatively long stretch of sequence identity. Additional evidence supporting this observation comes from the fact that 83% of NAHR-L1RAD events resulted from elements in the L1PA2 to L1PA7 subfamilies (Fig. 2*A*). These subfamilies are relatively young and exist in high copy number in the human genome (28, 29). By contrast, NHEJ-L1RAD events showed no relationship between the number of elements in each subfamily involved in the events and their genome-wide copy number for each L1 subfamily (Fig. 2*B*). NHEJ has been shown to be one of the repair mechanisms for double-strand breaks (DSBs). Thus, it may be hypothesized that the two L1s involved in a NHEJ-L1RAD event were present in the flanking region of DSB(s), and that a pair of short complementary L1 sequences (i.e., microhomology between two L1s) is associated with end-binding to bridge the DNA lesion.

Genomic Environment and Distribution of L1RADs. L1s tend to be found in regions of low GC content relative to the ≈41% average of the human reference genome (1). Consistent with this observation, recent L1 insertions also show a preference for AT-rich DNA (30) because of either the L1 EN cleavage site or the greater selective pressure operating in GC-rich regions. To characterize the genomic environment of human-specific L1RAD events, we estimated the neighboring GC content, gene density, and local chromosomal recombination rate of L1RAD loci. The GC content of neighboring L1RAD loci was determined by extracting 20 kb of flanking sequences (±10 kb in either direction) for each L1RAD locus from the human genome. The GC content of this sequence, excluding the chimeric L1 (i.e., the postrecombination L1) itself, was then calculated by using in-house Perl scripts. The resulting GC content for the flanking regions of the human-specific L1RADs averaged ≈38%. Our results show that the L1RAD loci seem to be located in AT-rich areas of the human genome which is congruent with findings that most L1s exist in GC-poor regions (36–38%) of the human reference genome (1). To measure gene density in the neighborhood of human-specific L1RAD events, we retrieved 4 Mb of sequence flanking L1RAD events (±2 Mb in either direction) and determined the number of known and predicted human RefSeq genes there. The gene density of these loci was estimated to be, on average, one gene per 105 kb and their distribution is skewed toward low gene density (median is one gene per 121 kb) (Fig. S3). When compared with the average gene density for the entire human genome (one gene per 94 kb), this finding indicates that human-specific L1RAD events tend to be found in regions of low gene density. This trend is correlated with the location of L1s, which predominate in gene-poor heterochromatin (31), but those observations likely reflect either the L1 insertion preference (30) or selective pressures against deleterious L1s (32) during genome evolution. Next, we investigated whether L1RADs were correlated with the local chromosomal recombination rate. We analyzed the recombination rate, as calculated by UCSC's BLAST-like alignment tool (BLAT) browser in the human genome, for each chimeric L1 element, but could not find any correlation between L1RADs and the local chromosomal recombination rates. The correlation among four parameters (GC content, gene density, deletion size, and recombination rate) reported above can be found in Table

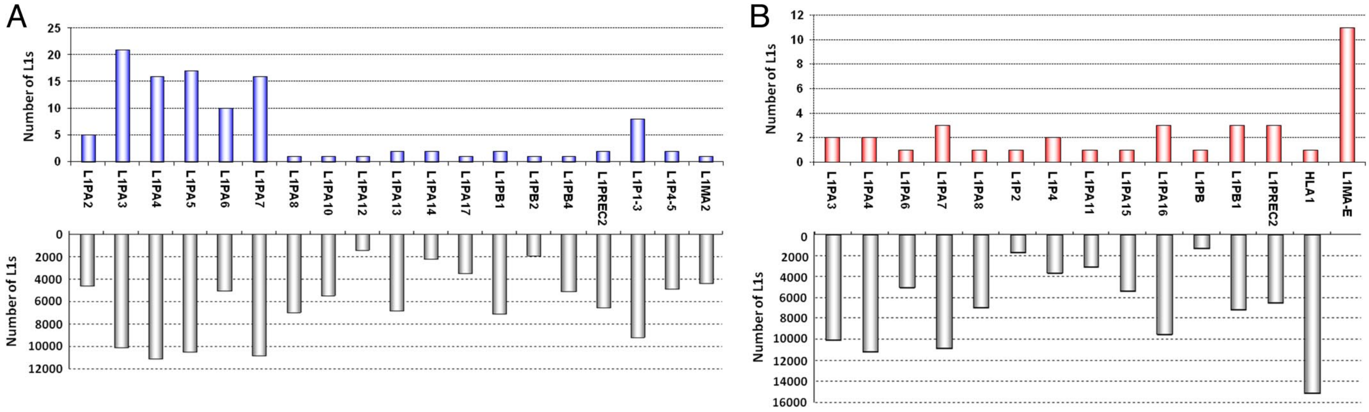

Fig. 2. L1 subfamily composition involved in L1RAD events. (*A*) Human-specific NAHR-L1RAD events. Shown are the number of L1 elements involved in 55 human-specific NAHR-L1RAD events (blue bars) versus total number of L1 elements in each subfamily in the human genome (gray bars). (*B*) Human-specific NHEJ-L1RAD events. Shown is the number of L1 elements involved in 18 human-specific NHEJ-L1RAD events (red bars) versus total number of L1 elements in each subfamily in the human genome (gray bars). The category labeled L1MA-E is comprised of subfamilies L1MA1, L1M2a, L1MA8, L1MB3, L1MC2, L1MD2, L1ME1, L1ME3B, and L1ME4a.

S2. Furthermore, to investigate whether the L1 density around each L1RAD region can affect the local chromosomal recombination rate, we performed a correlation test for relationship between the recombination rate in the 100-kb windows of L1RAD events (±50 kb in either direction) and the total number of L1s in the windows. Whereas the *r*-value was negative, the P value represented no significant correlation between the two factors ($r = -0.029$; $P = 0.8089$). However, because L1RADs are rare events, it may be difficult to find robust correlation between these variables, even if it does exist.

Human-Specific L1RAD Polymorphism. To estimate the polymorphism rates of L1RAD in humans, we analyzed 35 human-specific NAHR-L1RAD loci by using a panel of genomic DNA, from 80 human individuals (see *SI Text* and Table S3). Our results show that the polymorphism level of human-specific L1RAD is 20% (Table S4), which is similar to the polymorphism rate of human-specific ARMD events (15%) (14). These observations suggest that genomic deletions associated with recombination between retrotransposons have generated structural variation between humans.

Comparison of Human-Specific L1RADs with ARMDs. Despite the remarkable copy number of L1s in the human genomes, the frequency of human-specific L1RADs is not as high as that of human-specific ARMDs (73 vs. 492 events, Table 2). The observed difference between these two similar processes is caused by several reasons. First, L1s are monomer sequences whereas *Alu* elements are dimeric consisting of left and right monomers. Although each monomer evolved from 7SL-RNA independently (33), their 5′ ends are fairly homologous. The particular dimeric structure of each *Alu* element, a free left *Alu* monomer (FLAM) and a free right *Alu* monomer (FRAM), could contribute to an increase in opportunities for recombination between two *Alu* elements. FLAMs and FRAMs exist in 54,965 and 21,730 copies, respectively, in the human genome, as estimated by UCSC table browser (http://genome.brc.mcw.edu/cgi-bin/hgTables?command = start). Furthermore, it was observed that ≈25% of all human ARMD events are caused by the dimeric structure of *Alu* elements (14). The second reason accounting for the observed difference between ARMD and L1RAD frequencies involves the average distance between L1s (one insertion every 6 kb), which is twice that observed between *Alu* elements (one insertion every 3 kb). This implies that L1RAD events could be more deleterious as compared with ARMDs because of their potential to cause relatively large deletions in the host genome. This may result in an increase in the selective pressure against L1RAD events. The third factor impacting the relative frequencies of ARMDs and L1RADs is the observation that L1s tend to be located in less recombinogenic areas of the genome. A study of comparative recombination rates has shown that the recombination rate is highly correlated with CpG fraction, GC content, and polypurine/polypyrimidine tract fraction in 5-Mb, nonoverlapping windows of the human genome, but is negatively correlated with the density of LINEs in the human, mouse, and rat genomes (34). These observations imply that the high density of LINEs could counteract or decrease chromosomal recombination. Finally, the presence or absence of a recombination hotspot between L1s might contribute to the observed differences in the levels of recombination-associated deletions between *Alu* elements and L1s. To investigate the existence of possible recombination hotspots involved in L1RADs, we aligned prerecombination L1s from the chimpanzee genome with the chimeric L1s recovered from the human genome. The alignment windows contained identical regions (5 to 366 bp in length) between the two L1s. However, no recombination hotspot on L1 sequences emerged (Fig. S4). By contrast, the studies of human- and chimpanzee-specific ARMD processes revealed a recombination hotspot consisting of 22–24 bp on *Alu* sequences (14, 20) that could account for some of the difference in frequency of ARMDs compared with L1RADs. Given these four reasons, it is not surprising that the number of L1RAD events is lower than that of ARMD events. This may also explain why only a few examples of genetic disorders caused by L1RAD events have been reported.

Consequences of L1RAD in the Human Genome. L1 is an autonomous retrotransposon (35) and ≈1,850 copies of L1 are specific to human genome (30). Nevertheless, only 80–100 L1s are retrotransposition-competent (7, 11). These retrotransposition-competent L1s can cause human diseases by disrupting or altering functional gene expression, an event termed insertional mutagenesis. Not only these active L1s but also other inactive L1s are a source of genomic instability because they could provide sequence identity through which recombination may occur, resulting in deletions and other genomic rearrangements.

Our study represents the first genome-wide analysis of L1RAD events within the human lineage. Despite their low frequency, the L1RAD events removed ≈450 kb of human genomic DNA, an amount that is larger than the combined effects of *Alu* retrotransposition-mediated deletion (≈9 kb deleted) (36), L1IMD (≈18 kb deleted) (21), and ARMD events (≈400 kb deleted) (14) in the last 6 million years. Therefore, it appears that the impact of the L1RAD is much higher than reported retrotransposon-associated deletion mechanisms in contributing to the fluidity of the human genome. However, we could not rule out the role that the different genomic environments in which *Alu* and L1 elements tend to be found as a contributor to the relative frequencies of their associated deletion events. *Alu* elements tend to be found in gene-rich regions whereas L1s are more commonly found in gene-poor regions. We may therefore expect *Alu*-associated deletions to be more often selected against and to also have a smaller total deletion size as compared with L1-associated deletions. If ARMDs cause larger deletions than L1RADs, the evidence of these events has been erased by the strong purifying selection present in gene-rich regions.

To better show the effect of this deletion process in context with L1 copy number increase, we estimated the L1-associated sequence turnover rate in the human genome after the divergence of the human and chimpanzee lineages. Approximately 1.65 Mb (900 bp, average size of L1s; ≈1,850, copy number of human-specific L1s) of human genome sequence was added by the insertion of human-specific L1s whereas ≈450 kb of sequence was deleted by L1RAD events during the same time period. Thus, L1RAD can be said to have counteracted ≈27% of the L1-mediated increase in genome size. However, reciprocal recombination-associated deletions should produce concurrent duplications even though these two recombination products (deletions and duplications) could have different evolutionary fates in the host genome. If we consider L1 recombination-associated duplication as another regulator of L1 copy number increase, the turnover rate will be decreased. Interestingly, the majority of genomic sequences deleted by L1RADs were L1 sequences (≈60%) (Table S5). L1s are often found clustered in the genome, likely because of their insertion mechanism specific target site preference. Thus, in many cases, the deleted regions that existed between the two prerecombination L1s contained sequence from other L1s. In addition, each L1RAD event deleted portions of the two prerecombination L1s. Since the divergence of human and chimpanzee lineages, the L1RAD process has not only played a significant role in counteracting the increase in genome size caused by new L1 insertions, but it may also regulate the overall copy number of L1s in the genome.

L1RADs are also involved in DSB repair, a function that may be important for genomic stability and cell survival in the

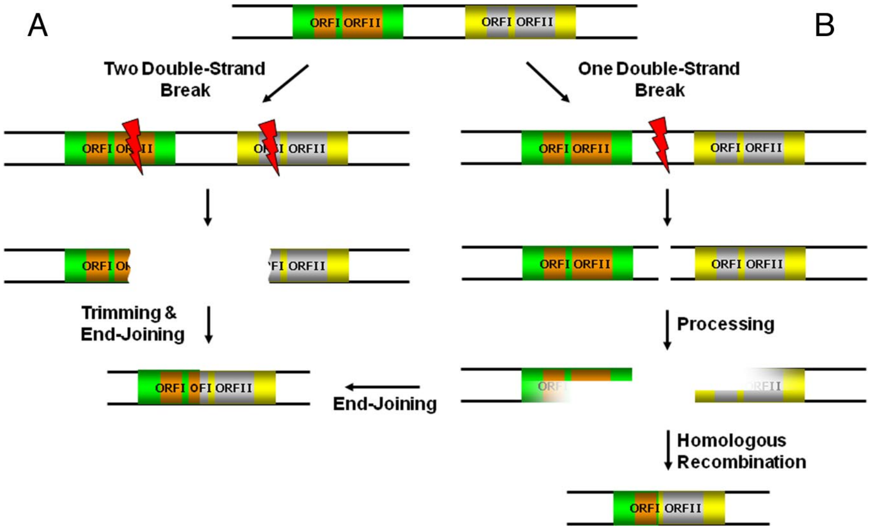

Fig. 3. Models for DSB repair mediated by L1RAD mechanisms. Parallel black lines represent double strand DNA and green and yellow boxes indicate two L1s which are involved in the L1RAD process. The thunderbolts indicate DSB events. (*A*) Two DSB repairs by a NHEJ-L1RAD. Two DSBs occur inside the L1s and the broken ends might be trimmed, which results in the removal of nucleotide sequence at each DSB site. Microhomology between two L1 sequences allows for the repair of the DSBs via NHEJ, resulting in two different contiguous L1s in the post repair genome. (*B*) One DSB repair by either NHEJ- or HR-L1RAD. One DSB occurs between two L1s and the broken ends might be processed by 5′-3′ exonuclease. Next, microhomology or a longer homologous stretch between two L1 sequences allows the DSB to be repaired, forming two different contiguous L1s or a single chimeric L1, respectively.

host genome. When DSBs occur in L1-rich regions, L1s would represent first-class material that could be used to restore the DNA lesion site. Whether one DSB occurs between two L1s or two DSBs occur inside the L1s, the broken ends are resected by 5′-3′ exonuclease. During this process, two DSB ends are bound by short complementary L1 sequences and the gaps are filled by DNA synthesis and ligation (Fig. 3*A*). In our study, we found 18 NHEJ-L1RAD events that could be the result of DSB repair mediated by the mechanism described above. Also, some NAHR-L1RAD loci could be created by homology-mediated DSB repair rather than by unequal cross-over (Fig. 3*B*).

In summary, our study suggests that L1 recombination-mediated genomic deletions are a significant source of human genetic variation, along with the genomic alterations caused by other mobile elements. Additionally, we believe that L1s could be involved in the restoration of DSBs occurring near or within L1 sequences.

Materials and Methods

Computational Search and Manual Inspection for Human-Specific L1RAD Loci. To computationally screen the human genome for potential L1RAD events, we modified the technique described in a previous study of human lineage-specific ARMD events (14). The technique uses the flanking sequences of each human L1 to locate the orthologous locus in the chimpanzee genome. In-house Perl scripts were used to calculate the positions of the 20 kb of flanking sequence, both upstream and downstream, for every L1 locus in the human genome. The orthologous chimpanzee loci of these human L1 flanking regions were then located by using UCSC's liftOver utility (http://genome.brc.mcw.edu/cgi-bin/hgLiftOver). Next, the position of the gap between the chimpanzee flanking regions for each locus was calculated, and the nibFrag utility bundled with the BLAT software package (http://genome.ucsc.edu/cgi-bin/hgBlat) was used to generate sequence for these chimpanzee gap loci, each of which corresponded to the orthologous site known to contain an L1 in the human genome.

The chimpanzee ortholog of an authentic L1RAD locus should be larger than the size of the putative human chimeric L1. Therefore, any chimpanzee gap locus whose size was equal-to-or-greater-than the size of the original human L1 plus 50 bp was considered a candidate locus worthy of further scrutiny, and was subjected to RepeatMasker analysis. We would expect that the pre-deletion locus, as represented by the chimpanzee gap sequences, would have L1 insertions at the beginning and end of the deleted region. Furthermore, to undergo L1RAD, these two L1 insertions would be found in the same orientation. To filter our RepeatMasked candidates based on these criteria, more in-house Perl scripts were used, screening out all loci except those for which the chimpanzee gap locus contained same-orientation L1 insertions as the first and last annotated repeats. In all, this computational filtering process produced a subset of all human L1s (547,171 in the human genome) that was feasible to screen manually (4,786 candidates).

Analysis of Flanking Sequences. For each L1RAD locus, 10 kb of flanking sequence upstream and downstream were collected by using a combination of in-house Perl scripts and the nibFrag utility bundled with the BLAT software package. The GC content of the flanking regions of each L1RAD locus was calculated by using the combined 20 kb of flanking sequence via another in-house Perl script, which excluded N's from the analysis. All scripts used are available from the authors on request. For the gene density analysis, we counted the number of genes by using the National Center for Biotechnology Information Map Viewer utility, run on Build 36.3 of the *Homo sapiens* genome (http://www.ncbi.nlm.nih.gov/mapview/map search.cgi?taxid = 9606). The neighboring 2 Mb of sequence 5′ and 3′ to each chimeric human L1 element was analyzed, and the number of genes found within this combined 4 Mb were noted.

ACKNOWLEDGMENTS. We thank Dr. J. Kim for his useful comments during preparation of the manuscript and J. A. Walker for her help throughout this project. This work was supported by National Science Foundation grant BCS-0218338 (to M.A.B.), and National Institutes of Health Grant RO1 GM59290 (to M.A.B.).

1. Lander ES, *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
2. Smit AF, Toth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol* 246:401–417.
3. Martin SL, Bushman FD (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol Cell Biochem* 21:467–475.
4. Mathias SL, Scott AF, Kazazian HH, Jr, Boeke JD, Gabriel A (1991) Reverse transcriptase encoded by a human transposable element. *Science* 254:1808–1810.
5. Feng Q, Moran JV, Kazazian HH, Jr, Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87:905–916.
6. Moran JV, *et al.* (1996) High frequency retrotransposition in cultured mammalian cells. *Cell* 87:917–927.

7. Sassaman DM, *et al.* (1997) Many human L1 elements are capable of retrotransposition. *Nat Genet* 16:37–43.
8. Kazazian HH, Jr, Moran JV (1998) The impact of L1 retrotransposons on the human genome. *Nat Genet* 19:19–24.
9. Ostertag EM, Kazazian HH, Jr (2001) Twin priming: A proposed mechanism for the creation of inversions in l1 retrotransposition. *Genome Res* 11:2059–2065.
10. Myers JS, *et al.* (2002) A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* 71:312–326.
11. Brouha B, *et al.* (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci USA* 100:5280–5285.
12. Jeggo PA (1998) Identification of genes involved in repair of DNA double-strand breaks in mammalian cells. *Radiat Res* 150:S80–91.
13. Gebow D, Miselis N, Liber HL (2000) Homologous and nonhomologous recombination resulting in deletion: Effects of p53 status, microhomology, and repetitive DNA length and orientation. *Mol Cell Biol* 20:4028–4035.
14. Sen SK, *et al.* (2006) Human genomic deletions mediated by recombination between Alu elements. *Am J Hum Genet* 79:41–53.
15. Burwinkel B, Kilimann MW (1998) Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J Mol Biol* 277:513–517.
16. Segal Y, *et al.* (1999) LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. *Am J Hum Genet* 64:62–69.
17. Temtamy SA, *et al.* (2008) Long interspersed nuclear element-1 (LINE1)-mediated deletion of EVC, EVC2, C4orf6, and STK32B in Ellis-van Creveld syndrome with borderline intelligence. *Hum Mutat* 29:931–938.
18. 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.
19. Goodman M, *et al.* (1998) Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol* 9:585–598.
20. Han K, *et al.* (2007) Alu recombination-mediated structural deletions in the chimpanzee genome. *PLoS Genet* 3:1939–1949.
21. Han K, *et al.* (2005) Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res* 33:4040–4052.
22. Sela N, *et al.* (2007) Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biol* 8:R127.
23. Worby CA, Dixon JE (2002) Sorting out the cellular functions of sorting nexins. *Nat Rev* Mol Cell Biol 3:919–931.
24. Carlton J, Bujny M, Rutherford A, Cullen P (2005) Sorting nexins–unifying trends and new perspectives. *Traffic* 6(2):75–82.
25. Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Series* 41:95–98.
26. Moore JK, Haber JE (1996) Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Mol Cell Biol* 16:2164–2173.
27. Bentley J, Diggle CP, Harnden P, Knowles MA, Kiltie AE (2004) DNA double strand break repair in human bladder cancer is error prone and involves microhomology-associated end-joining. *Nucleic Acids Res* 32:5249–5259.
28. Khan H, Smit A, Boissinot S (2006) Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res* 16:78–87.
29. Giordano J, *et al.* (2007) Evolutionary history of mammalian transposons determined by genome-wide defragmentation. *PLoS Comput Biol* 3:e137.
30. Lee J, *et al.* (2007) Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene* 390:18–27.
31. Korenberg JR, Rykowski MC (1988) Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* 53:391–400.
32. Boissinot S, Entezam A, Furano AV (2001) Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol* 18:926–935.
33. Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J (2007) Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet* 23:158–161.
34. Jensen-Seaman MI, *et al.* (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res* 14:528–538.
35. Skowronski J, Fanning TG, Singer MF (1988) Unit-length line-1 transcripts in human teratocarcinoma cells. *Mol Cell Biol* 8:1385–1397.
36. Callinan PA, *et al.* (2005) Alu Retrotransposition-mediated Deletion. *J Mol Biol* 348:791–800.