Cell, Vol. 110, 327–338, August 9, 2002, Copyright ©2002 by Cell Press

Human L1 Retrotransposition Is Associated with Genetic Instability In Vivo

David E. Symer,[1,2,4,8] Carla Connelly,[1] Suzanne T. Szak,[6,9] Emerita M. Caputo,[1] Gregory J. Cost,[1,10] Giovanni Parmigiani,[3,5] and Jef D. Boeke[1,3,7]

[1]Department of Molecular Biology and Genetics
[2]Department of Medicine
[3]Department of Oncology
Johns Hopkins University School of Medicine
725 N. Wolfe Street
Baltimore, Maryland 21205
[4]Department of Medical Oncology
Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins
1650 Orleans Street
Baltimore, Maryland 21231
[5]Department of Biostatistics
Bloomberg School of Public Health
Johns Hopkins University
615 N. Wolfe Street
Baltimore, Maryland 21205
[6]National Center for Biotechnology Information (NCBI)
National Library of Medicine
National Institutes of Health
Bethesda, Maryland 20894

Summary

Retrotransposons have shaped eukaryotic genomes for millions of years. To analyze the consequences of human L1 retrotransposition, we developed a genetic system to recover many new L1 insertions in somatic cells. Forty-two de novo integrants were recovered that faithfully mimic many aspects of L1s that accumulated since the primate radiation. Their structures experimentally demonstrate an association between L1 retrotransposition and various forms of genetic instability. Numerous L1 element inversions, extra nucleotide insertions, exon deletions, a chromosomal inversion, and flanking sequence comobilization (called 5′ transduction) were identified. In a striking number of integrants, short identical sequences were shared between the donor and the target site's 3′ end, suggesting a mechanistic model that helps explain the structure of L1 insertions.

Introduction

LINE retrotransposons (L1s) have successfully populated and modified eukaryotic genomes for hundreds of millions of years (Smit et al., 1995). These mobile elements encode factors needed for autonomous movement via an RNA intermediate. An endonuclease (EN) cleaves target DNA (Feng et al., 1996; Mathias et al., 1991; Moran et al., 1996), whereupon the L1 transcript is reverse transcribed back into cDNA. Both the reverse transcriptase (RT) and EN are encoded by the open reading frame-2 (ORF2) protein. In the human genome, L1s have accumulated over time to hundreds of thousands of copies of various ages and structures, and they now comprise ~17% of the draft sequence (Lander et al., 2001). Another ~15% is made up of *Alu* elements and processed pseudogenes, thought to be mobilized by L1 retrotransposition proteins in *trans* (Esnault et al., 2000; Wei et al., 2001).

Direct insertional mutagenesis by L1, a major form of genetic instability attributed to this retrotransposon, has resulted in diseases including muscular dystrophy, hemophilia, and breast cancer (Ostertag and Kazazian, 2001a). Other endogenous retroelements have been implicated in both homology-dependent and homology-independent genome rearrangements, which could play important roles in the evolution of new species as well as genetic diseases (Hughes and Coffin, 2001). The extent to which genomic instabilities of various types are correlated with active retrotransposition versus post-transposition events involving L1 elements is unknown.

Studies of human L1 biology are severely constrained by the vast number of related, repetitive elements riddling the genome. This situation is confounded by the "sloppy" structure of most L1 copies; many of which belong to subfamilies of various ages, are 5′ truncated and/or inverted, and possess variable length target site duplications (TSDs) and poly(A) tails.

The identification of disease-causing, active L1 copies led to the development of a powerful tissue culture system facilitating study of the retrotransposition mechanism (Dombroski et al., 1991; Moran et al., 1996; Sassaman et al., 1997). Active full-length L1s were marked with engineered reporter genes such as *mneoI* that are specifically expressed only after retrotransposition. This tissue culture system has highlighted a requirement for both ORF1 and ORF2 in retrotransposition (Feng et al., 1996; Moran et al., 1996). Moreover, these proteins appear to act preferentially in *cis*, by direct interaction with their precursor transcript (Wei et al., 2001), in mediating L1 mobilization.

Refined definition of L1 subfamilies (Smit et al., 1995) has allowed development of a PCR-based method to find element copies that are dimorphic in humans, based on specific L1 3′ sequence differences. This method, called L1 display, has allowed analysis of L1 target specificity by comparison of genomic sites before and after L1 integration (Ovchinnikov et al., 2001).

Bioinformatic analyses of the draft human genome sequence have added quantitative, correlative information to these other studies (Boissinot et al., 2000, 2001; Lander et al., 2001; Szak et al., 2002). On a genome-wide level, L1s dating back to the primate radiation have similar sloppy structures without a clear preference for particular genomic compartments, except that GC-rich

[7]Correspondence: email: jboeke@jhmi.edu
[8]Present address: Laboratory of Immunobiology, National Cancer Institute, Frederick, Maryland 21702.
[9]Present address: Biogen, Inc., 14 Cambridge Center, Cambridge, Massachusetts 02142.
[10]Present address: Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, California 94702.

Alu elements might be preferentially avoided (Szak et al., 2002).

Despite these advances, to date relatively few de novo L1 insertions have been structurally analyzed. This is an important deficiency, since studies on other transposons have shown integral involvement in various forms of genetic instability in their host organisms (Moore and Haber, 1996; Nevers and Saedler, 1977; Teng et al., 1996). The link between transposable elements and instability could occur either as part of transposition per se or after the fact. Ongoing determination of a near-finalized human genome sequence enables comprehensive analysis of target sites for new L1 integrants, e.g., relative to genes and other chromosomal features, both before and after the insertion event.

Here, we describe a system that facilitates recovery of progeny L1-*neo* integrants in human tissue culture cells. Most de novo insertions have primary target sequence specificity, and overall lengths and genomic contexts, consistent with primate-specific L1s (Lander et al., 2001; Szak et al., 2002). These findings validate the tissue culture system. However, we also observed a high degree of genomic instability associated with the new insertions: the structures of L1s and their genetic neighborhoods reflect the sloppy nature of retrotransposition itself and not simply posttransposition modifications alone.

Results

A System for Recovering L1-*mneo* Insertions

A tissue culture system has been developed that allows for the study of L1 retrotransposition (Moran et al., 1996). Marked L1 elements are launched from an episomal plasmid; resulting new integrants are marked with a selectable reporter gene. In initial experiments, cells were transfected with pJM101/L1.3, encoding both its own selectable plasmid marker, *hygro* (for stable transfection), and full-length L1.3 marked by *neo*. The latter reporter was modified so that it would be expressed only after retrotransposition: *neo* and its promoter and polyadenylation sequences are oriented antiparallel to L1 (i.e. *neo* is expressed from the minus strand; hence, *mneo*), and this cassette in turn was disrupted by an intron (I) in the sense orientation (designated *mneoI*). For a marked integrant to confer G418 resistance on a host cell (making it G418^R), the entire L1-*mneoI* construct must be transcribed to allow intron removal by RNA splicing. After reverse transcription of the spliced RNA and integration of cDNA into a chromosomal target site, the *mneo* can be expressed. To date, only a small number of de novo integrants has been characterized using this system (Moran et al., 1996).

To study the consequences of L1 retrotransposition further, we modified pJM101/L1.3 so that many individual marked L1 integrants together with flanking human genomic DNA could be cloned in bacteria for sequence analysis. The structure of one such modified donor plasmid, pDES89 (Figure 1A), differs from previously described L1-*mneoI* plasmids in that *ori*, the plasmid origin of replication, and *dhfr*, a gene conferring trimethoprim resistance (Tmp^R) in *E. coli*, have been added. We call these added sequences the bacterial selection cassette

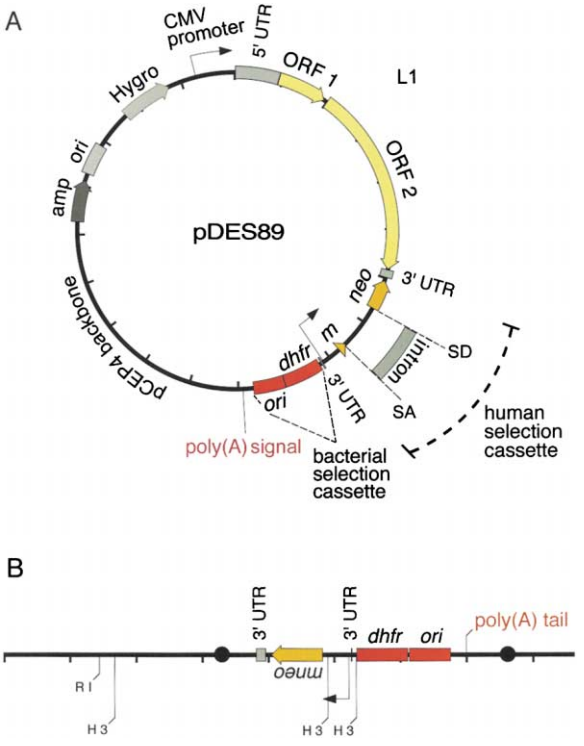

Figure 1. Recovery of New L1 Integrants

(A) pDES89 encodes L1.3 with both open reading frames (ORF1 and ORF2, in yellow arrows). L1 is marked by *mneoI*, minus strand *neo* (orange arrows) interrupted by an artificial intron (*I*), and a bacterial selection cassette including *ori* and the Tmp-selectable gene, *dhfr* (red rectangles). This L1 launch plasmid also includes a strong upstream CMV promoter and a selectable *hygro* marker for stable transfection. Tick marks represent 1 kb DNA intervals.

(B) A recoverable genomic L1 integrant, flanked by target site duplications (TSDs, black circles). After retrotransposition, with splicing of the intron (from SD, splice donor, to SA, splice acceptor) and cDNA integration, the antisense *mneo* reporter can be expressed, rendering cells G418^R. Possible RE sites that would release 5′ and 3′ junctions with genomic DNA (horizontal black line) are shown for EcoRI, R I; and HindIII, H 3.

(Figure 1A). Another L1 donor plasmid used in our study, pGC109, differs from pDES89 in the orientation of this *ori* and other sequences.

New L1-*mneo* integrants could be recovered using restriction endonucleases (REs) chosen to release linearized fragments containing the integrant together with one or both genomic flanks. We used EcoRI, HindIII, and XmaI in separate experiments. Resulting fragments were then ligated under dilute conditions to form intramolecular circles and transformed into *E. coli*. Resulting Tmp^R plasmids were examined by a combination of RE digests to identify unique isolates. Multiple isolates of the same insertion were often recovered from the same pool, and the same insertions were often recovered using different REs, suggesting that the pools were of relatively low complexity.

A screen for retrotransposed integrants used DNA sequencing to identify those isolates with poly(A) tails directed by the strong SV40 polyadenylation signal at the 3′ end of the donor construct. A typical new integrant is shown in Figure 1B and includes a characteristic po-

ly(A) tail and target site duplications (TSDs). Note that this system differs slightly from that of Gilbert et al. (2002 [this issue of *Cell*]), in that our system does not require correct splicing, integration, and expression of an intact *mneo* for recovery. Indeed, a large number of integrants lack full-length *neo*, probably because multiple L1 integrants can occur frequently in individual cells, and only one correctly expressed *neo* is needed per $G418^R$ host cell (Wei et al., 2000).

In separate experiments, both HeLa and HCT116 cell pools were transfected with pDES89, pGC109, or empty vector controls. L1 was allowed to retrotranspose under transient or stable transfection conditions (Moran et al., 1996; Wei et al., 2000), with selection on Hygromycin (for stable maintenance of the episomal backbone) and/or G418 (for spliced *mneoI* expression).

To compare retrotransposition frequencies, we transfected cells with equimolar concentrations of pDES89, pJM101/L1.3, and empty plasmids. Despite comparable transfection efficiencies as demonstrated by similar numbers of Hygromycin-resistant cells, the rate of retrotransposition (assayed by the number of $G418^R$ colonies) was ~90% lower for pDES89 than pJM101/L1.3 (Moran et al., 1996). Since most extant L1 integrants are 5′ truncated, we attribute this decrease in $G418^R$ colony numbers to a similar truncation process operating on the longer donor element used in our recovery system; the added bacterial selection cassette increases the length from 5′ to 3′ ends of L1 by ~3.5 kb.

We initially studied insertions in HeLa cells because virtually all prior L1 tissue culture experiments have been performed in that system (Moran et al., 1996). We recovered a high fraction of integrants showing evidence for various forms of genetic instability associated with new L1 integrants, including aberrant splicing of *mneoI*, addition of extra 5′ terminal nucleotides, and inverted internal sequences. Therefore, we studied HCT116 cells as well, because the latter has a relatively stable near-diploid karyotype (Lengauer et al., 1997) and thus in principle should more closely resemble normal human cells. Nevertheless, HCT116 cells are colon cancer cells with microsatellite instability. A total of 42 L1-*mneo* insertions were recovered, 11 from HeLa and 31 from HCT116 cells. We recovered both 5′ and 3′ junctions for 38 of these, only 3′ junctions are available for the remainder.

De Novo Integrants Faithfully Model Preexisting L1s

In a parallel study using the computer programs RepeatMasker (Smit and Green, 2001) and TSDFinder, Szak et al. (2002) identified over 72,000 L1s that have intact 3′ ends and accumulated almost exclusively since the time of the primate radiation. On average, these elements share ~76% ± 9% sequence identity with L1.3; of these, over 16,000 (~23%) have evaluable TSDs. This latter subset has 88% ± 7% nucleotide identity with L1.3 and includes 845 L1Hs-Ta elements specific to humans. Both the 3′ intact L1s and the L1Hs-Ta elements form a basis for comparison with our de novo integrants.

To analyze L1 target site specificity, we aligned primary genomic sequences of unoccupied de novo target sites (plus strand) so that 10 nucleotides both to the left and right of the putative L1 EN minus strand nick site are presented. Nucleotide frequencies at some of these 20 aligned positions are strongly nonrandom. There is a strong preference in the plus strand for two to three Ts immediately to the left of, and for five As immediately to the right of, the corresponding target nick site on the minus strand (Figure 2A, top). These biases recede to background nucleotide frequencies at the left and right edges of the histogram, where A+T content is ~60%. Strikingly similar biases in target site nucleotide frequencies were observed in endogenous 3′ intact L1s in the draft sequence (Figure 2A, bottom). To quantify possible differences between the data sets, nucleotide frequencies were compared using a set of chi square tests for each position. The patterns are statistically indistinguishable, based on a sensitive test for a difference between them that used the maximum of these 20 chi squares ($p = 0.99$).

The consensus sequence for the first-strand nicking site is therefore presented for the plus strand as 5′-TTAAAA-3′, both in the human genome and for de novo integrants (Figure 2A), confirming extensive previous studies on the target specificity of L1 EN (Cost and Boeke, 1998; Feng et al., 1996; Jurka, 1997; Szak et al., 2002). The actual physical substrate for the first nicking reaction by EN is thought to be the antiparallel (minus) strand, i.e., 3′-AA ^ TTTT-5′.

A similar comparison was made for L1 length distributions. A histogram showing lengths of preexisting L1s with intact 3′ ends and TSDs, and of human-specific L1Hs-Ta family members, is shown in Figure 2B. These bimodal distributions of lengths are consistent with but extend previous analyses. A large majority of L1 elements is 5′ truncated, with increasing numbers at shorter lengths (Figure 2B). This part of the length distribution has been attributed to poor processivity of L1 RT. The fraction of elements not truncated, i.e., full-length or near full-length, comprises those elements that could remain active due to retention of the 5′ untranslated promoter. However, most of these full-length genomic L1s likely contain point mutations, rendering them inactive. Human-specific Ta elements are more likely to be nearly full-length, perhaps because they have not yet been subjected to "purifying selection" (Boissinot et al., 2001).

Again, a remarkably similar pattern of lengths was observed in our recovery system (Figure 2B). Chi square analysis demonstrated that the de novo versus 3′ intact length distributions are statistically indistinguishable (chi square = 14.72 with 12 degrees of freedom, $p = 0.74$). However, significantly more Ta elements are near full-length (Figure 2B and Table 1; Szak et al., 2002). This disparity between frequencies of full-length de novo and Ta L1s could reflect a technical limitation of our recovery system: long DNA fragments are substantially more difficult to recover, due to biases imposed by intramolecular ligation (Revie et al., 1988) and bacterial transformation (Hanahan, 1983), both of which favor shorter recovered molecules.

Comparison of other features of de novo versus extant L1s shows significant differences. First, as shown in Figure 2C and Table 1, the poly(A) tails of our de novo integrants are significantly longer and more frequent than those found flanking 3′ intact or L1Hs-Ta elements in the genome (mean length 88 ± 27 A's versus mean

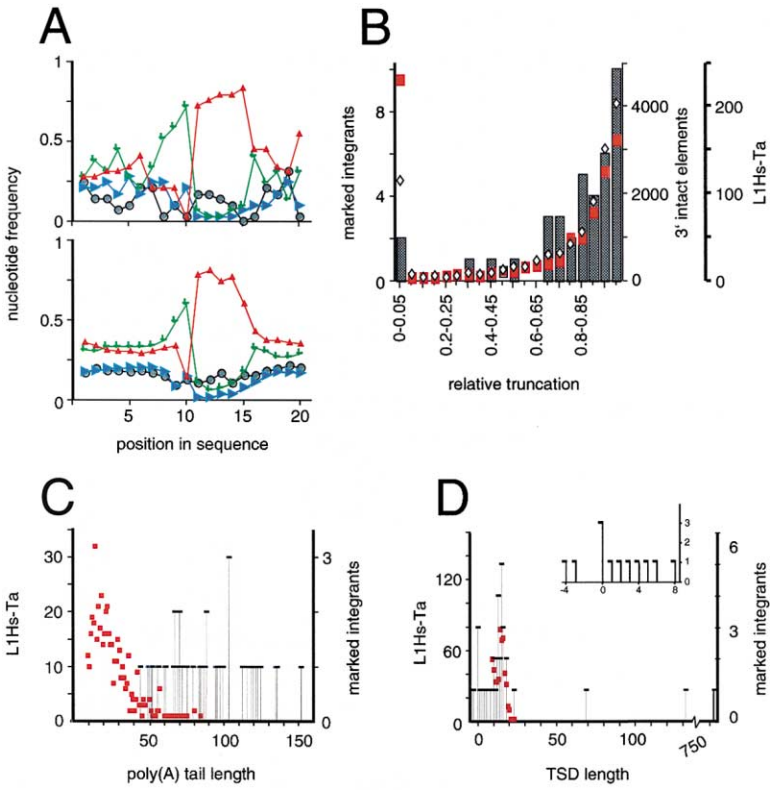

Figure 2. Comparison of De Novo Versus Extant L1 Integrants

(A) Frequency distribution of nucleotides at the target site. Plus strand sequences were aligned about the first nick site at the TSD's left junction, between positions 10 and 11. The average nucleotide frequency at each alignment position (10 bp left and 10 bp right of the first nick site) is presented: A, red upward triangles; T, green pluses; C, blue sideways triangles; G, gray circles. Frequencies for (upper image) de novo integrants (n = 29) and (lower image) preexisting 3′ intact L1s (n = 16,266) with evaluable TSDs are presented.

(B) 5′ truncations of L1s. The length distributions of (left axis, shaded rectangles) de novo, marked L1s (n = 38); (right axis, open diamonds) extant, 3′ intact L1s (n = 16266); and (right axis, red squares) L1Hs-Ta elements (n = 845) in the human genome are shown. Full-length integrants are counted as having 0–0.05 truncation. De novo integrant lengths were normalized to account for the added length of the human selection cassette, after correcting for the bacterial selection cassette's length (required for recovery). The lengths of integrants with 5′ inversions were counted as the sums of the inverted and direct segments.

(C) Poly(A) tail length distributions. The number of integrants with various poly(A) tail lengths is presented for preexisting L1Hs-Ta (left axis, red squares, n = 845) and for de novo (vertical marks, n = 42) L1s in the human genome.

(D) TSD length distributions. A histogram showing the number of TSDs of various lengths is presented for (left axis, red squares) *L1Hs*-Ta (n = 845) and (right axis, vertical marks) de novo L1s in the human genome. For L1Hs-Ta, TSDs < 9 nt are not reported based on our scoring algorithm (Szak et al., 2002). This algorithm would also not detect the 758 nt direct repeats in the hybrid element. (Inset) Expanded *x*-axis showing new integrant target site deletions and short TSDs.

18 ± 10 [3′ intact] versus mean 27 ± 13 [L1Hs-Ta]). Moreover, de novo poly(A) tails are exclusively homopolymeric runs of As. We identified no cases of "patterned" tails such as arrays of the tetranucleotide TAAA that commonly occur in the genome (Szak et al., 2002), despite the fact that our screen for tails by DNA sequencing was not predicated upon any particular poly(A) pattern. This result could be attributed to the strong

Table 1. Structural Summary of L1s

	Element type		
		Genomic	
Feature	L1-*neo* insertions, %	3′ intact[1], %	L1Hs-Ta, %
Full-length	5.3	5.1	28
5′ truncated	95	95	72
5′ inverted	0	0.07	0.12
5′ inverted and 5′ truncated	16	12	12
With TSDs of any length	84	N.A.	N.A.
>8 nt	68	23	62
≤8 nt	16	N.A.	N.A.
Without TSDs	16	77	38
Unknown bases at 5′ end	11	N.A.	N.A.
"Pure" poly(A) tails >10 nt	98	20	57
Patterned tails	0	13	6.6
Target site deletion	16	N.A.	N.A.
<10 nt	7.9		
≥10 nt	7.9		
Target site inversion	2.6	N.A.	N.A.
5′ transduction	5.3	~0	~0
3′ transduction	N.A.	8.6	15

N.A., data not available by this determination.
[1] data for comparison (Szak et al., 2002).

Table 2. Target Site Summary for New L1-*mneo* Insertions

Target type	Element type: De novo L1-*mneo* insertions, %	Element type: Genomic 3′ intact[1], %
Insertions in: L1 elements	26	13
Alu elements	7.1	3
α-satellites	4.8	0.5
LTR and DNA elements	4.8	7.2
predicted genes	50	17
Same orientation as gene	38	38
Opposite orientation	62	62
exons	0	N.A.
between genes	50	83
Percent GC content[2]	40.6 ± 5.4	35

N.A., data not available by this determination.
[1] Data for comparison (Szak et al., 2002).
[2] Mean ± standard deviation determined from 20 kb intervals in human genome draft assembly.

SV40 polyadenylation signal present in our L1 donor construct, in contrast to the weak native L1 polyadenylation signal (Moran et al., 1999). It also corroborates previous findings that the length of poly(A) tails is inversely correlated with the elements' age (Ovchinnikov et al., 2001). It is possible that poly(A) tail length decreases with time in evolution due to "slippage" during replication. The establishment of inherently unstable, very long homopolymeric tracts of As may lead to posttranspositional genetic instability.

A second discrepancy involves the range of TSD lengths (Figure 2D). We observed several de novo integrants with extremely short TSDs of one to two nucleotides (as well as short deletions; see below). Such TSDs have not been quantified in preexisting L1s, due to statistical uncertainties about the occurrence of short duplications. We also identified integrants with TSDs as long as 69, 132, and possibly 758 nucleotides, respectively. By contrast, the longest TSD identified in the 3′ intact L1s collected genome-wide was only 60 nucleotides (nt) (Szak et al., 2002). The establishment of long direct repeats in relatively close proximity could lead to posttranspositional genetic instability by recombination with loss of intervening sequences, resulting in gradual progressive shortening of observed mean TSD lengths.

To extend this comparison further, we analyzed other structural aspects and the genomic context of the integrants. As shown in Table 1, for many parameters the tissue culture recovery system generated integrants with structures very comparable to preexisting L1s. However, the de novo integrants include short target site deletions (Figure 2D, inset and Table 1), a result that cannot be reliably compared to extant L1s, due to statistically insignificant sequence fluctuations confounding analysis of TSDs shorter than 9 nt (including target site deletions; Szak et al., 2002).

Looking more broadly at genomic targets for L1, we observed a modest preference for L1-*mneo*s integrating into preexisting repetitive elements (Table 2), including L1s and SINE elements (*Alu*). Of the 3′ intact L1s with TSDs, only ~3% are directly flanked by *Alu*s (Szak et al., 2002). By contrast, 7% of our new integrants fell within preexisting *Alu*s (Table 2); however, this difference is not statistically significant. A similar tendency was observed in targeting L1 elements. Notably, the new insertions hit a very wide age-range of SINEs and LINEs, suggesting no specific preference. De novo L1 targeting frequencies into different repetitive families approximate the composition of the human genome (Lander et al., 2001).

Two new integrants hit α-satellite DNA (Table 2), one unambiguously in the centromere of chromosome 3. Although L1s and other repetitive elements have been described in constitutive heterochromatin (Santos et al., 2000), there is no good current estimate for the frequency of preexisting L1s (Table 2), due to difficulties in assembling reliable sequences in this compartment (Lander et al., 2001).

We did not observe any hotspot for de novo insertions. This negative result might be based simply on the number of integrants recovered. Interestingly, in a recent study, a new hotspot for HIV integration has been found in a 2 kb region on chromosome 11q13, involving ~1% of all HIV insertions in this small target (F. Bushman, personal communication). Many more L1-*mneo* integrants are needed to achieve similar statistical power. Additionally, we found ~50% of new L1 insertions hit predicted genes (annotated by at least two independent algorithms as per http://genome.cse.ucsc.edu/). A majority is oriented opposite to the genes, and all are in noncoding sequences (Table 2). Relatively fewer existing L1s are found within predicted genes, although increased sensitivity for predicted genes afforded by our manual analysis of de novo integrant neighborhoods might explain this difference.

Retrotransposition Is Associated with Various Forms of Genetic Instability

Overall, the results (Figure 2, Tables 1 and 2) strongly suggest that recovered L1-*mneo* integrants faithfully reflect the native mechanisms of retrotransposition operating since the primate radiation. However, the new integrants appear to have significantly longer and more frequent poly(A) tails, a wider range of TSD lengths including occasional very long direct repeats, and a modest propensity for predicted genes and repeats including centromeric, α-satellite DNA. These comprise potential secondary sources of genetic instability, i.e., after trans-

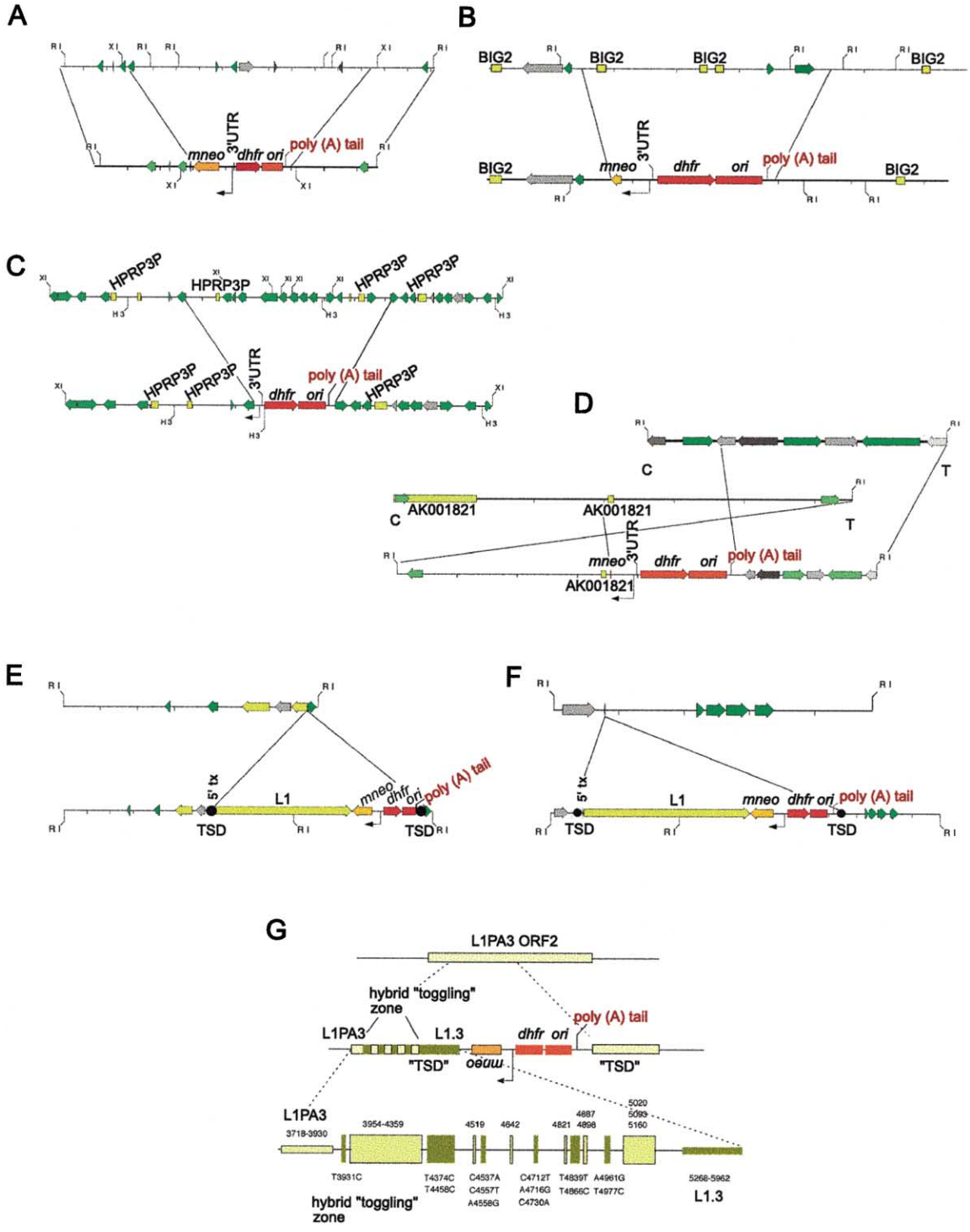

Figure 3. Genetic Instability in the Context of L1 Retrotransposition

Pre- and postinsertional loci (top) and (bottom) respectively are presented for individual recovered genomic integrants. Exons (yellow squares), DNA repetitive elements and LTR elements (gray squares), SINEs (green arrows), L1s (yellow arrows), the human selection cassette *mneo*(orange arrow), bacterial selection cassette (red), TSDs (black circles), and RE sites for EcoRI (R I), HindIII (H 3), and XmaI (X I) are annotated. Tick marks represent 1 kb intervals.

(A–C) Three instances of chromosomal deletions on chromosomes 10q25, 20q13, and 1q32, respectively. Note that in cases (B) and (C) several exons of known genes are eliminated.

(D) A chromosomal inversion on chromosome 12q23. C, centromere; T, telomere as per draft human genome sequence.

(E–F) Examples of 5′ transduction (*5′ tx*), each involving mobilization of ~50 nt upstream of the bona fide L1 start site. (Sequence data are provided in the Supplemental Data, Table S1 available at www.cell.com/cgi/content/full/110/3/327/DC1 and at www.bs.jhmi.edu/mbg/boekelab/

position. As shown in Figure 3, we found numerous additional examples of various forms of genetic instability in our collection of de novo integrants.

Chromosomal Deletions and an Inversion

A relatively frequent category of instability involves the formation of target site deletions ranging from 1 bp (Figure 2D) to >11 kb (Figures 3A–3C) long. Each is spanned by a continuous, albeit 5′ truncated, portion of a new L1-*mneo*, complete with poly(A) tail and where possible, a spliced *mneo* reporter. (In Figures 3B and 3C, we could not assess *mneoI* splicing due to 5′ truncation.) This strongly suggests that each arose directly during retrotransposition and not after a secondary recombination event. Large deletions were also found in another set of L1 insertions identified in the companion paper (Gilbert et al., 2002 [this issue of *Cell*]). The presence of the larger chromosomal deletions (Figures 3A–3C) has been confirmed in 3 of 3 cases by the recovery of identical structures using independent REs and/or confirmatory PCR and DNA sequencing using cell pool DNA as template. Two of these resulted in the loss of coding exons (Figures 3B and 3C).

Large sequence losses associated with L1 retrotransposition are surprising, because genome-wide analyses have not documented them in significant numbers. The difficulty in determining how many native L1s are associated with small or large deletions probably stems from the fact that genomic preintegration sequences are typically lacking (although dimorphisms in human populations or genomic duplicons could provide such information). Additionally, large deletions could be lost by purifying selection. By contrast, the tissue culture system gives us the unique opportunity to analyze target sequences both before and after the integration event. Many of the 3′ intact L1s analyzed by Szak et al. (2002) lack a TSD longer than 8 nt (Table 1), raising a question about whether some of them could comprise a significant class of L1 integrants that have target site deletions.

Another previously undocumented form of instability identified is a chromosomal inversion (of a segment on chromosome 12), associated with a new L1-*mneo* insertion (Figure 3D). This inversion is similar to the insertion-associated deletions in that L1-*mneo* is situated precisely at the inversion breakpoint, indicating that its insertion was associated with the inversion. The two chromosomal sequences spanned by the L1 insertion directly at the inversion breakpoint are ~120 kb apart both in the draft human sequence and in assembled BAC clones. We were concerned, given the well-documented chromosomal instability of HeLa cells that this inversion might have arisen independently of L1 retrotransposition. Although the likelihood of L1 targeting the unique position at the breakpoint is vanishingly small, we screened for this inversion junction lacking the integrated L1 by PCR. As expected, it could not be detected. Moreover, unoccupied genomic sequences extending past both integrant junctions, as predicted by the draft sequence, were confirmed by PCR and sequencing. Unfortunately, despite extensive PCR assays designed to detect a clean reciprocal inversion, the second inversion breakpoint could not be identified (data not shown).

5′ Transduction

We observed two cases of full-length L1-*mneo* insertions. Each extends an additional ~50 nucleotides upstream of the 5′ end of the donor L1 (Figures 3E and 3F), and has a characteristic TSD, poly(A) tail, and no detected errors in *mneoI* splicing or reverse transcription of the full-length, ~9 kb transcript. The additional 5′ sequence begins at the transcription start site of the strong upstream CMV promoter in pDES89 (Figure 1A). The two sequences differ in length by a single 5′ nucleotide, G. Their 5′ endpoints lie ~20 bp downstream of the CMV promoter TATAA box. These integrants provide experimental evidence for "5′ transduction," in which flanking 5′ sequences become part of the transposon transcript and hence are mobilized by an adjacent non-L1 promoter. This could lead to acquisition of a new promoter, but only if it consists of downstream elements. Genomic L1s have probably acquired new 5′ ends by a similar mechanism (Lander et al., 2001). This mode of mobilizing genomic sequences is the converse of 3′ transduction, in which readthrough L1 transcripts proceed past the elements' own weak polyadenylation signals up to stronger downstream signals, thereby mobilizing 3′ flanking sequences (Moran et al., 1999).

Formation of a Hybrid Element

A hybrid L1 element was formed during another L1-*mneo* retrotransposition event (Figure 3G). In this integrant, we infer that transposition into position 5270 of a preexisting L1PA3 element on chromosome 8q24 was initiated normally, because it contains a long de novo poly(A) tail and spliced *neo* gene. However, the left flank of this insertion is a hybrid, alternating between the L1PA3 and newly integrated L1.3 sequences. The origins of each DNA segment within the hybrid zone can be determined unambiguously by frequent single nucleotide polymorphisms (SNPs); de novo L1.3 sequences are intermingled with homologous L1PA3 sequences over several kilobases (Figure 3G). Exact transition points (or sites of "toggling") within the hybrid molecule cannot be assigned precisely, because the two intermingled elements are largely conserved, differing only at SNPs. There are 13 unambiguous transitions back and forth between the donor and target elements within this toggling zone. This integrant therefore contains ~758 bp direct repeats (which differ only at the SNPs; Figure 3G). Given uncertainties about this integrant's left boundary, these direct repeats could also be considered a peculiar type of inexact "TSD".

Extra Nucleotides

Several examples of so-called "untemplated bases" at the left junction between L1-*mneo* and genomic DNA were identified. We prefer use of the term "extra nucleo-

boeke_lab_homepage) In both cases, robust TSDs (black circles) were identified.
(G) A hybrid element with 13 toggling events switching between target L1PA3 and homologous integrant L1.3. (Sequence data are provided in the Supplemental Data, Table S1 available at above website.)

Table 3. L1 Integrants with 5′ Inversions

Name	Length of inverted segment	Length of direct segment	5′ nt position in PDES89	Number of contiguous nt copied in inversion	Number of contiguous nt overlapping at junction	Twin Priming nt match	Twin Priming 3′ overlap
7A1	1040	1803	6137	0	2	3/5	3
7C2	1069	2368	5545	2	2	2/5	1
7G1	1424	1842	5719	4	1	2/5	1
7E1	90	1716	7178	4	2	5/5	5
8E3	46	1550	7387	3	2	4/5	4
17F8	825	2288	5872	6	3	4/5	2

tides," as in some cases these may have been templated by another sequence. These insertions range in length from a single nucleotide to >100 bp long. In one intriguing integrant, three extra nucleotides flank the 5′ junction between the new L1 and the genomic target. Upon close inspection, these same three untemplated bases are adjacent to six contiguous bases, all shared between the target sequence and ~100 bp upstream in the donor L1 sequence. This raises the possibility that the RNA template may be scanned for regions of microhomology with the target site as the template for RT. In this case, template switching or skipping by RT (e.g., from L1 RNA to flanking DNA) could result in templating of the putatively untemplated three nucleotides.

5′ Inversions

Another category of genetic instability observed is the well-known inversion of 5′ segments of L1 elements (Table 3). Previous estimates for existing L1s suggested that 8 to 12% include this kind of endogenous rearrangement (Ostertag and Kazazian, 2001a; Szak et al., 2002). We found ~16% of our de novo integrants contain 5′ inversions. As is the case for almost all inverted genomic L1s, these new inverted integrants are all 5′ truncated and have shorter inverted 5′ segments relative to the noninverted 3′ segments (Table 3; Szak et al., 2002).

Microhomology between L1 RNA and the 3′ End of the TSD

In analyzing junctions between genomic TSDs and L1 sequences, we noted numerous cases of microhomology that make precise assignment of the 5′ boundary of the L1 integrant ambiguous. We calculated whether these observed microhomologies were significantly different from what is expected by chance, using methods outlined in an evaluation of viral/host junction sequences (Roth et al., 1985). To simplify analysis, we initially considered only cases with TSDs; no events with target site deletions, inversions, or toggling were included. Additionally, we did not count past any gaps or mismatches, despite several compelling cases in which the microhomology could be extended substantially further 5′ and/or 3′ past them. Proceeding from the unambiguous 5′ end of each new integrant, we counted the identical nucleotides shared between the TSD and the donor L1 RNA template. The increased A+T content of target sequences where L1 elements tend to insert (Figure 2A) was considered in calculating the expected number of identical nucleotides overlapping by chance (Roth et al., 1985). Thus, we used the average %GC content of 20 kb windows flanking de novo integrants, and in a second calculation, the nucleotide frequencies centered on each ambiguous boundary (20 nt window; data not shown). The observed number of matching, overlapping nt is shown in Figure 4A; the distribution of microhomology lengths is significantly skewed to values longer than those expected by chance (chi square test, $p < 0.001$). These results indicate that there is a significantly higher degree of complementarity between the right boundary of TSDs and the L1 cDNA template than expected by chance.

We also counted the number of overlapping nucleotides when these strict rules were liberalized, i.e., including L1 integrants with target site deletions and inversions (Figure 4A, "observed max") and considering sequences adjacent to these junctions (as TSDs in such cases are undefined). These results provide more evidence for frequent (albeit not ubiquitous) microhomology between the target site and the L1 template (Figure 4A).

Discussion

The opportunity to analyze several dozen de novo L1 element insertions in the context of a near-complete draft human genome sequence confirms that our L1 recovery system faithfully mimics L1 elements already present. However, it has also yielded new insights into the consequences of retrotransposition, both to the L1 element and to the target chromosome. These consequences are not merely posttranspositional; they also appear to include a wide range of "collateral damage" directly related to retrotransposition itself.

This analysis also provides an opportunity to examine preferred sites of L1 insertion; these reflect the composition of the human genome remarkably well (Table 2). It is estimated that >90% of the genome is present in the August, 2001 draft assembly, and indeed we could assign 93% of the new insertions to unique genomic locations. The targets represented by this collection of insertions are a microcosm of the human genome overall (Table 2).

Significant (albeit not well quantified) regions of the genome are packaged in heterochromatin (Lander et al., 2001). Our recovery system yielded limited evidence for insertions into this compartment; we recovered two insertions into α-satellite DNA. An important caveat is that at least one L1 element must express the *mneo* reporter per surviving cell (Wei et al., 2000), and *neo* expression might not occur in constitutive heterochromatin. On the other hand, ~74% of the recovered insertions lack an intact *mneo* gene, highlighting a key advantage of this

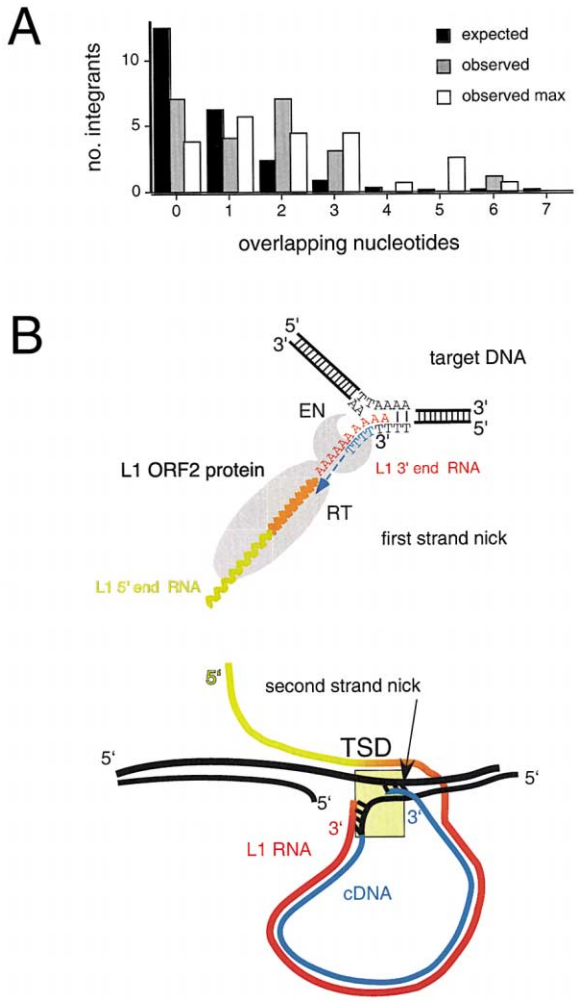

Figure 4. Evidence and a Model for Microhomologous Annealing at the Target Site

(A) A histogram comparing the number of expected versus observed L1 integrants with variable overlapping nucleotide identities shared between the 3′ end of the TSD and the 5′ truncated flank of L1. Expected numbers (black rectangles) were calculated as described (Roth et al., 1985) and recalculated based on average nucleotide frequencies flanking the target (see text). The observed number (shaded) of integrants with various numbers of overlapping nt is significantly biased toward longer overlaps ($n = 22$; $p < 0.001$ by chi square test). With a liberalized definition of overlaps allowed ("observed max"; open rectangles), we observed still more bias toward long overlaps (data for $n = 35$ normalized by 22/35).

(B) Microhomology between the L1 template and the TSD. (Top): A representation of TPRT. L1 EN (gray symbols) nicks at a consensus target site, allowing a putative RNA:DNA hybrid to form there (red and black, respectively). The 3′OH of the nicked substrate forms a primer that is extended using the RNA template (red at 3′ end), forming first-strand cDNA (blue). (Bottom): We show a possible interaction between a 3′ segment of first-strand cDNA (blue) and the complementary strand at the 3′ end of the TSD. Alternatively, a corresponding 5′ segment of the RNA template might interact with the TSD, forming an R loop.

recovery system: the human reporter cassette need not be active for recovery (Figures 1 and 2). Thus, our results suggest that the large blocks of constitutive heterochromatin in the human genome, while not highly preferred, nevertheless can be targets for new integration.

L1 Is an Engine for Genetic Change

Along with small target site deletions, we observed two types of gross chromosomal rearrangements at L1 target sites, i.e., large deletions and a chromosomal inversion. Small deletions are likely due to a subtle variation in the usual mechanism of target nicking and cDNA integration, called target-primed reverse transcription, TPRT (Luan et al., 1993; Ostertag and Kazazian, 2001a). We speculate that the second-strand nick at the L1 target site is made a few bp to the left of the initial nick rather than a few bp to the right as is typically believed to be the case. Notably, a few small target site deletions have been observed with de novo insertions of the I factor, a LINE-like element from *Drosophila* (Jensen et al., 1995).

We offer two alternative explanations for the gross chromosomal rearrangements. First, DNA damage and nicking at two associated target sites may occur through some L1 EN-independent mechanism. In this case, the new L1 integrant might span this damage by cooperating with some underlying host repair process. Although the usual retrotransposition mechanism, TPRT, is thought to depend on L1 EN cleavage of target DNA, L1 EN-independent DNA nicks can also be used by L1 RT as substrates in target extension reactions (G.J.C. and J.D.B., unpublished data) and in tissue culture cells (Morrish et al., 2002). In yeast, both endogenous Ty1 retrotransposons and L1 RTs help repair double-strand chromosomal breaks by inserting retrotransposon DNA (Moore and Haber, 1996; Teng et al., 1996). An L1 EN-independent mechanism for large deletions has been suggested by Morrish et al. (2002), who observed their formation by an EN mutant element. However, those integrants are structurally distinct from those described here in that they lacked poly(A) tails, the L1 3′ end, and evidence for initial target site cleavage by L1 EN.

The L1 integrants associated with chromosomal deletions and the inversion have intact poly(A) tails and 3′ ends (Figures 3A–3D), suggesting that the L1 retrotransposition machinery was directly associated with their formation. They may have been formed by a distinct subversion of the canonical EN-dependent TPRT reaction. For the deletions, we propose that the second nick may have occurred many kb to the left of the first nick rather than a small number of bp to its right as in a standard insertion. Similarly, L1 EN itself may have triggered the chromosomal inversion by nicking twice on the same strand, ~120 kb apart, with the resulting integrated L1 structure resolved as a chromosomal inversion.

TPRT: Variations on a Theme

Recently a new mechanism has been proposed for 5′ L1 inversions (Tables 1 and 3), called "twin priming" (Ostertag and Kazazian, 2001b). According to this model, both nicked target DNA strands interact with different parts of the same template RNA, thereby serving as antiparallel substrates in two TPRT reactions. The two antiparallel cDNAs are then joined, resulting in a 5′ inversion. This differs from the usual situation, without L1 inversion, where second-strand priming is expected to use the first cDNA strand as template. An explicit prediction of the twin-priming model is that the 3′ end of

the TSD (upper strand) anneals to internal homologous sequences in the RNA, directly adjacent to the internal inversion junction. Our set of de novo integrants with 5′ L1 inversions provides an unbiased opportunity to test this model. As shown in Table 3, all of the inversion junctions share at least one nucleotide of overlap with the upper strand of TSD; one has 4 contiguous bases at this flank and one 5 in a row. We also observed overlapping nucleotides at the internal junction of these inversions (Table 3). Thus, these findings provide significant additional support for the twin-priming model.

We observed significant microhomology at the 3′ end of the TSD and the corresponding position (i.e., the segment just beyond the last nucleotides copied as cDNA) in L1s lacking an inversion. Put another way, the regions of micro-complementarity between the primer generated by the second EN cut site and the cDNA strand template are more frequent than expected by chance, even after adjusting for base composition (Figure 4A).

Generalizing from proposed mechanisms of L1 retrotransposition, including TPRT and twin priming, there are multiple possible interpretations of this microhomology (Figure 4A). The TPRT model suggests that the 3′-OH formed by L1 EN at the first nick site (consensus 3′ AA ^ TTTT 5′) is elongated by L1 RT, using the L1 RNA poly(A) tail as template (Figure 4B, top). The RNA probably interacts with the nicked DNA at this site, via its poly(A) tail forming a short RNA:DNA hybrid (however, there is no direct evidence for this). We propose that the second nick (which has far less specificity than the first, Figure 2A) can similarly find a region of complementarity on the cDNA at which plus strand synthesis can be primed (Figure 4B, bottom). Clearly this is not obligatory, as such complementarity is not always observed (Figure 4A). However, when such complementarities occur, they may stabilize the primer-template complex and facilitate plus strand priming.

Alternatively, these microhomologies may represent regions near the DNA target site that form an R loop with the RNA template. Such an R loop (in which invading RNA displaces a DNA strand at the target site) could impede the progress of the RT, leading to premature termination of polymerization and formation of a truncated element. Alternatively, such an R loop could help specify the second site of cleavage on the displaced single-stranded DNA by L1 EN to form the plus strand primer. The latter seems unlikely, as L1 EN is unable to cleave single-stranded DNA in vitro (Cost and Boeke, 1998). On the other hand, this last model is attractive because we have obtained individual integrants with imperfect but extended complementarity between the L1 RNA at its 5′ terminus and the target site, both up- and downstream of the TSD (Figure 4A). Such complementary RNA could potentially bridge the nicked DNA target site, until new L1 cDNA could be incorporated in its repair.

Some additional support for the latter model comes from the hybrid element (Figure 3G), wherein the sequence toggles back and forth 13 times between L1.3 and target L1PA3 sequences. We speculate that this hybrid zone may have arisen subsequent to the formation of a long donor/target heteroduplex, which could be either RNA/DNA (i.e., R-loop anchoring) or cDNA/DNA. In any case, the heteroduplex appears to have undergone mismatch repair in discrete patches differing only in which strand was used as the template. Another possible explanation, that this unusual integrant arose from multiple homologous recombination events within this region, independent of integration per se, seems extremely unlikely, due to the precision of integration (there are only 2 nucleotides out of ~1300 in this toggling zone not accounted for by L1PA3 or L1.3). Interestingly, an active hybrid mouse L1 has recently been described, although its etiology is unknown (Saxton and Martin, 1998), and hybrid human L1 formation has also been described by Gilbert et al. (2002 [this issue of *Cell*]).

In summary, the recovery system has provided many new L1 integrants, which match preexisting L1s in the human genome in several important respects (Figure 2, Table 1). Unique to the de novo integrants, however, is the opportunity to analyze target sites before and after integration, and to analyze integrants' structures themselves after only a few weeks propagation in tissue culture, rather than millions of years of human evolution. These findings show that L1 can contribute directly to genetic instability in human cells through previously undescribed mechanisms. We are obliged to describe the link between L1 retrotransposition and the various observed forms of instability as an "association" rather than propose causality, since other precedent causes of genetic instability (e.g., DNA damage, nicking) could lead to increased substrates for L1 retrotransposons in a DNA repair capacity. Such processes could lead to increased recovery of genomic rearrangements, with L1 literally right in the middle, but with L1 retrotransposition not causing such rearrangements per se.

Experimental Procedures

L1 Donor Plasmids

The episomal plasmid that encodes full-length L1.3, pJM101/L1.3 (Moran et al., 1996), was modified to allow for recovery of human L1 integrants by selection first in human tissue culture cells and then for genomic fragments containing the marked, retrotransposed L1 cDNA in bacteria. To minimize changes to pJM101/L1.3, whose ability to mediate L1 retrotransposition has been well characterized (Moran et al., 1999, 1996; Wei et al., 2000), we inserted the bacterial selectable reporter for Tmp resistance, *dhfr*, and bacterial origin of replication *ori*, downstream of the antisense *mneoI* cassette and upstream of the SV40 strong polyadenylation signal, forming pDES89. Additional details are provided in the Supplemental Data, available at www.cell.com/cgi/content/full/110/3/327/DC1 and at www.bs.jhmi.edu/mbg/boekelab/boeke_lab_homepage.

Recovery and Confirmation of New L1 Integrants

HeLa cervical cancer cells (obtained from Dr. John Moran, University of Michigan) and HCT116 colon cancer cells (from Dr. Christoph Lengauer, Johns Hopkins University) were grown as previously described (Lengauer et al., 1997; Moran et al., 1996). Cells were transfected with purified, supercoiled plasmids using Fugene-6 reagent (Roche) and Opti-Mem II serum free medium (Gibco) according to the manufacturer's instructions, and selected using Hygromycin-B (Calbiochem) and/or Geneticin (Invitrogen).

$G418^R$ cells were grown to confluence and harvested. Genomic DNA was prepared using DNAeasy columns (Qiagen) and restricted to completion using EcoRI, HindIII, or XmaI (New England Biolabs). Fragments were ligated under extremely dilute conditions to favor intramolecular circularization; ~500 ng restricted genomic DNA was incubated with 5 U T4 DNA ligase in a volume of 500 μl ligation buffer at 16° overnight. The ligation mixture with added glycogen (New England Biolabs) was then ethanol precipitated and resuspended in water; this was used to transform electrocompetent

DH10B cells (Electromax, Invitrogen) by electroporation in 1 mm gap cuvettes in a BTX Electro Cell Manipulator 600 (Genetronics, San Diego, CA). Transformed cells were selected on M9 minimal plates supplemented with 0.5 g/100 ml casamino acids, 200 ng/mL folic acid, and 25 μg/mL TMP at 37° for 48 hr. Individual colonies were picked and grown on 2× LB liquid medium or 1× LB plates with 25 μg/mL Tmp prior to DNA isolation.

Tmp^R plasmids were screened for uniqueness using RE digests including the initial cutter (i.e., EcoRI, HindIII, or XmaI), along with HincII and StyI. Unique L1 integrant candidates were screened by sequencing through the SV40 polyadenylation signal at their 3′ end (Figure 1B, see Supplemental Data, available at above URL). Candidates containing either poly(A) tails or flanking genomic DNA identifiable by BLAST searching were analyzed by systematic DNA sequencing using a panel of primers spanning the entire L1-*mneo-dhri-ori* construct (see Supplemental Data, available at above URL). For the six recovered cases of 5′ inversions, antiparallel primers were used to obtain unambiguous inversion junction sequences. Chromosomal locations, neighboring genomic structures, and percent G+C content of an overlapping 20 kb window were determined using the UCSC "Golden Path" web browser and BLAT alignment algorithm at http://genome.cse.ucsc.edu/, with the August 6, 2001 draft assembly.

For comparisons between de novo and extant L1s in the genomic draft sequences, we used RepeatMasker (Smit and Green, 2001) and TSDfinder (Szak et al., 2002) to annotate preexisting L1s. L1Hs-Ta subfamily members were identified from these 3′ intact L1 elements (Szak et al., 2002) using RepeatMasker. All statistical analyses were performed using the statistical software package R. All p-values were computed by Monte Carlo simulations. Chi square analysis was performed as described (Agresti, 1990).

Acknowledgments

We would like to dedicate this study to the memories of Gloria J. Symer (mother of D.E.S.) and Daniel Boeke (father of J.D.B.).

We thank Drs. Tony Furano for helpful discussions, and John Moran and Christoph Lengauer for providing reagents. We also thank Yolanda Eby for expert technical support. We gratefully acknowledge funding in part from the Howard Hughes Medical Institute Physician Postdoctoral Fellowship (D.E.S.) and grant CA16519 from the National Institutes of Health (J.D.B.).

Received: May 3, 2002
Revised: June 17, 2002

References

Agresti, A. (1990). Categorical Data Analysis (New York: Wiley).

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., and Furano, A.V. (2001). Selection against deleterious LINE-1-containing loci in the human lineage. Mol. Biol. Evol. *18*, 926–935.

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.

Dombroski, B.A., Mathias, S.L., Nanthakumar, E., Scott, A.F., and Kazazian, H.H., Jr. (1991). Isolation of an active human transposable element. Science *254*, 1805–1808.

Esnault, C., Maestre, J., and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. *24*, 363–367.

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

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

Hanahan, D. (1983). Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. *166*, 557–580.

Hughes, J.F., and Coffin, J.M. (2001). Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat. Genet. *29*, 487–489.

Jensen, S., Gassama, M.P., and Heidmann, T. (1995). Retrotransposition of the *Drosophila* LINE I element can induce deletion in the target DNA: a simple model also accounting for the variability of the normally observed target site duplications. Biochem. Biophys. Res. Commun. *15*, 111–119.

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. USA *94*, 1872–1877.

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.

Lengauer, C., Kinzler, K.W., and Vogelstein, B. (1997). Genetic instability in colorectal cancers. Nature *386*, 623–627.

Luan, D.D., Korman, M.H., Jakubczak, J.L., and Eickbush, T.H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell *72*, 595–605.

Mathias, S., Scott, A., Kazazian, H., Boeke, J., and Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. Science *254*, 1808–1810.

Moore, J.K., and Haber, J.E. (1996). Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. Nature *383*, 644–646.

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.

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

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.

Nevers, P., and Saedler, H. (1977). Transposable genetic elements as agents of gene instability and chromosomal rearrangements. Nature *268*, 109–115.

Ostertag, E.M., and Kazazian, H.H., Jr. (2001a). Biology of mammalian L1 retrotransposons. Annu. Rev. Genet. *35*, 501–538.

Ostertag, E.M., and Kazazian, H.H., Jr. (2001b). Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res. *11*, 2059–2065.

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.

Revie, D., Smith, D.W., and Yee, T.W. (1988). Kinetic analysis for optimization of DNA ligation reactions. Nucleic Acids Res *16*, 10301–10321.

Roth, D.B., Porter, T.N., and Wilson, J.H. (1985). Mechanisms of nonhomologous recombination in mammalian cells. Mol. Cell. Biol. *5*, 2599–2607.

Santos, F.R., Pandya, A., Kayser, M., Mitchell, R.J., Liu, A., Singh, L., Destro-Bisol, G., Novelletto, A., Qamar, R., Mehdi, S.Q., et al. (2000). A polymorphic L1 retroposon insertion in the centromere of the human Y chromosome. Hum. Mol. Genet. *9*, 421–430.

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.

Saxton, J.A., and Martin, S.L. (1998). Recombination between subtypes creates a mosaic lineage of LINE-1 that is expressed and actively retrotransposing in the mouse genome. J. Mol. Biol. *280*, 611–622.

Smit, A.F., and Green, P. (2001). RepeatMasker. http://repeatmasker.genome.washington.edu/

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

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., and Boeke, J.D. (2002). Molecular archeology of L1 insertions in the human genome. Genome Biol., in press.

Teng, S.-C., Kim, B., and Gabriel, A. (1996). Retrotransposon reverse transcriptase-mediated repair of chromosomal breaks in *Saccharomyces cerevisiae*. Nature *383*, 641–644.

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.

Wei, W., Gilbert, N., Ooi, S.L., Lawler, J.F., Ostertag, E.M., Kazazian, H.H., Boeke, J.D., and Moran, J.V. (2001). Human L1 retrotransposition: *cis*-preference vs. *trans*-complementation. Mol. Cell. Biol. *21*, 1429–1439.