doi:10.1016/j.jmb.2005.02.043 J. Mol. Biol. (2005) 348, 791–800

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Alu Retrotransposition-mediated Deletion

Pauline A. Callinan[1]†, Jianxin Wang[2]†, Scott W. Herke[1]† Randall K. Garber[1], Ping Liang[2] and Mark A. Batzer[1]*

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

[2]*Department of Cancer Genetics Roswell Park Cancer Institute Elm and Carlton Streets, Buffalo NY 14263, USA*

**Corresponding author*

Alu repeats contribute to genomic instability in primates *via* insertional and recombinational mutagenesis. Here, we report an analysis of *Alu* element-induced genomic instability through a novel mechanism termed retrotransposition-mediated deletion, and assess its impact on the integrity of primate genomes. For human and chimpanzee genomes, we find evidence of 33 retrotransposition-mediated deletion events that have eliminated approximately 9000 nucleotides of genomic DNA. Our data suggest that, during the course of primate evolution, *Alu* retrotransposition may have contributed to over 3000 deletion events, eliminating approximately 900 kb of DNA in the process. Potential mechanisms for the creation of *Alu* retrotransposition-mediated deletions include L1 endonuclease-dependent retrotransposition, L1 endonuclease-independent retrotransposition, internal priming on DNA breaks, and promiscuous target primed reverse transcription. A comprehensive analysis of the collateral effects by *Alu* mobilization on all primate genomes will require sequenced genomes from representatives of the entire order.

Keywords: short interspersed elements; target primed reverse transcription

Introduction

Alu repeats are the most prolific short interspersed elements (SINEs) in primate genomes, accumulating approximately 1.2 million members over the last 65 million years of evolution.[1] True to their moniker as "genomic parasites", *Alu* elements rely on the cellular machinery of other mobile elements, such as long interspersed elements (LINEs), for their successful transmission through the germline.[2–4] Not all *Alu* elements are capable of using the borrowed commodities, however. Some hypotheses suggest that only a few *Alu* source genes are retrotranspositionally competent.[5,6] Over time, the source *Alu* elements accumulate sequence mutations, and this has resulted in an array of *Alu* subfamilies distinguished by diagnostic mutations.[5,7] Although the peak of *Alu* amplification occurred some 40–60 million years ago, lineage-specific, population-specific and individual-specific insertion events in modern primate genomes are indicated in recent studies.[8–13]

Alu elements are a unique source of genomic instability among primates. As a direct result of their abundance and sequence identity, they promote genetic recombination events that are responsible for large-scale deletions, duplications and translocations.[14–18] Some *Alu*-mediated recombination events that have occurred within and nearby coding regions are instigators of disease. Currently, *Alu*–*Alu* recombination events have been linked to approximately 50 human diseases, including hypercholesterolemia, α-thalassemia and BRCA1-related breast cancer.[6] The disruptive consequences of newly integrated *Alu* insertions within genic regions of the human genome have been documented in several studies. *Alu* elements may disrupt splicing by integrating within introns, alter patterns of gene expression by inserting within promoter regions or regions upstream of genes, or even silence gene function by inserting within the gene itself.[6] Mutagenesis *via* *Alu* insertion accounts for approximately 0.1% of all human diseases and is responsible for cases of familial cancer, metabolic disease and blood disorders.[6]

Recently, novel consequences of *Alu*-induced genomic instability have come to light. An example described by Hayakawa *et al*. documents the deletion of an exon caused by gene conversion of

† P.A.C., J.W. & S.W.H. contributed equally to this work.

Abbreviations used: SINE, short interspersed element; LINE, long interspersed element; ARD, *Alu* retrotransposition-mediated deletion; HuARD, human-specific ARD; TSD, target site duplication; pTPRT, promiscuous target primed reverse transcription.

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

an older *Alu*Sx element to a younger *Alu*Y element, specifically within the human lineage.[19] The consequential loss of the CMP-*N*-acetylneuraminic acid hydroxylase gene produces a biochemical difference between humans and non-human primates. Although other gene conversion-associated deletions are documented in the literature,[13,20] this mechanism has yet to be explored on a large scale.

Alu retrotransposition-mediated deletion (ARD), the focus of this work, is another novel type of genetic instability mediated by *Alu* elements. The initial evidence for this mechanism was derived from studies by Gilbert *et al*.[21] and Symer *et al*.,[22] who independently determined that 10% of L1 integrations within cultured human cells resulted in target site deletions spanning from 1 bp to 70,000 bp. L1 insertions associated with the deletion of target DNA had characteristics not typical of usual L1 integrants. In addition to the lack of target site duplications (TSDs), deletion-inducing L1 elements integrated at non-canonical L1 EN (endonuclease) nick sites and sometimes lacked poly(A) tails.[21–23]

Because *Alu* repeats and LINEs share the mobilization machinery needed to retrotranspose,[2–4] it was presumed that *Alu* elements possessed the same ability to introduce genomic instability through ARD.[21,22] Even though ARD has not been investigated *in vitro*, some examples from natural genomes are present in the current literature.[13,20] In the first case, documented by Salem *et al*.,[20] the insertion of an *Alu*Yg6 into human chromosome 3 was accompanied by a deletion of approximately 300 bp of DNA. The second event involved the insertion of a young Yb7 subfamily member, again associated with a deletion of 300 nt.[13] Given that *Alu* elements have reached copy numbers in excess of one million per haploid genome, it is likely that significant genomic alteration resulting from ARD will be found within the primate order. Despite the intriguing preliminary evidence for this unusual mechanism of genomic instability, no comprehensive study has attempted to quantify the rate of ARD within primate genomes.

In this study, we employ a sensitive computational screening approach to compare the draft genomic sequences of man (*Homo sapiens*) and the common chimpanzee (*Pan troglodytes*) in order to assess the occurrence of deletions associated with *Alu* retrotransposition. Our findings, further supported by wet bench verification methods, indicate that *Alu* retrotransposition may have generated over 3000 deletion events during the course of primate evolution, removing nearly a megabase of DNA in the process.

Results

Alu retrotransposition-mediated deletions

To detect lineage-specific ARD events, data from the National Center for Biotechnology (NCBI) draft sequence of the human genome were compared to the draft genomic sequence of *P. troglodytes* (for program details, see Materials and Methods). The program was designed to detect lineage-specific *Alu* elements in one genome that are associated with extra (non-homologous) genomic sequences in the other (see alignment Figure 1(a)). To eliminate the presence of *Alu* gene conversion-mediated deletions in our dataset, manual verification of the sequence was performed (see Materials and Methods). The remaining putative ARD events were verified as authentic deletions rather than independent insertions through polymerase chain reaction (PCR) amplification of the locus in outgroup taxa (gorilla, orangutan and African green monkey; see Materials and Methods) (Figure 1(b)).

In total, 19 young *Alu* insertion events specific to the human lineage were associated with deleted target site DNA; in the chimpanzee genome, 14 such events were recovered (Table 1). Among the human data, we recovered the two ARD events detected in prior studies,[13,20] thereby validating our computational methods. One of the human-specific ARD events, HuARD9, could not be verified experimentally due to a lack of unique flanking sequence, but it was included in the total *Alu* insertion number due to its structural authenticity. Our data indicate that humans possess 1.36 times more detectable ARD events than do chimpanzees. Adjusting this number to account for polymorphisms missed by sampling a single sequenced genome as described,[10] we conclude that ARD levels in the human genome are approximately 1.1 times greater than in the chimpanzee (Table 1).

Levels of *Alu* retrotransposition-mediated deletion polymorphism

To assess the level of polymorphism in *H. sapiens* for ARD events, we used PCR to amplify loci from 20 unrelated individuals from each of four geographically diverse populations (80 total individuals). In all, 11% (2/18) (one locus, HuARD9, could not be amplified) of the tested loci were polymorphic; this value translates to a polymorphism rate of 19% after an adjustment for single genome sampling (Table 1). The polymorphism level obtained appears to be lower than that typical for recently integrated *Alu* elements. Of the 18 events, 14 were insertions of elements from either the *Alu*Yb or the Ya5 lineages, which have insertion polymorphism rates of 20–25% across diverse human populations.[8–10,12,24] This is a conservative estimate of polymorphism for these subfamilies, considering the figures are unadjusted for single genome sampling. We believe that the reduced polymorphism in our dataset is a result of the relatively small sample sizes as compared to the previous analyses of thousands of young *Alu* insertions.

Using a DNA panel of 12 unrelated chimpanzee individuals, all 14 chimpanzee loci were amplified

(a)

```
Human  GGTAAGGCGT GACGACAAGT TTTACAAGTT TTTAGTGAGA GTGCAATGGA AATAACAAAT CAGGctgagg  70
Chimp  GGTAAGGCGT GACGACAAGT TTTACAAGTT TTTAGTGAGA GTGCAATGGA AATAACAAAT CAGGgactga

Human  caggagaatg gcgtgaaccc gggaagcgga gcttgcagtg agccgagatt gcgccactgc agtccgcagt 140
Chimp  tctcacagta ataaatgact ggctggaaga ctctagtgat ggaatctttt tgcagggcca actaatgaat

Human  ccggcctggg cgacagagcg agactccgtc tcaaaaaaaa aaaaaaaaaa acaaaaaaaa aa~~~~~~~~ 210
Chimp  gcaggacttt gggatattta tgtgaagatg agacaggctg aaggtatgaa ccttaataca aggaaaaata

Human  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ 280
Chimp  aggaacaaga gaagcaatat tcagagctat caatgagagg gagatattgg agaatagtta agagaactag

Human  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ 350
Chimp  ctttacatgt tttgtggaag gaatgaatta ggaaaacatc agactcaact agtgtgtgtg caaattgata

Human  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ 420
Chimp  ggcattctca tgccatgcct ggtagaaatg gaaaatggta caaaccttct ggtaaacaat tttgtgatac

Human  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ 490
Chimp  attatcaaga acttcaaaat gttaaaatct tttgacataa taactctact tagaaaattt gtacaaagaa

Human  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ 560
Chimp  ttacacatag aaatgcttgt cccattttct ctcatcaaaa ttattcacaa tagtgaaaat tttgaagtat

Human  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ 630
Chimp  tgcccttagg attttgcagc taggtaaatg ataacataca attatccaaa gtaattttaa ctgagaataa

Human  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ 700
Chimp  taatattcag tgagaaagca gaggtgtgtg tgtattatat aattatgtac tgtattttgt gacataactg

Human  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ 770
Chimp  aatgtctaat aatattcttg ctgtatataa agacaggtct taagttttat ggaagtttct tggaaatttt

Human  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ 840
Chimp  ccctatggta tcatgaaatt ccctggatat ccattatata atatttcctg ttcatgaaat tagaacatta

Human  ~~~~~~~~AAC ACTGATCCTA GAAGAGTATG TCAATGGTCA ACTATGCCT 890
Chimp  ctgctatAAC ACTGATCCTA GAAGAGTATG TCAATGGTCA ACTATGCCT
```

(b)

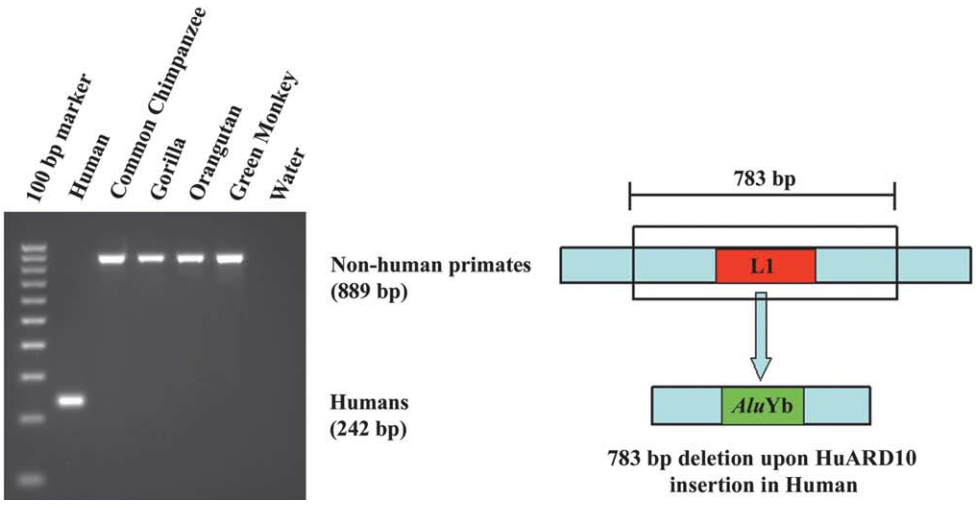

Figure 1. Example of an *Alu* retrotransposition-mediated genomic deletion. (a) Sequence alignment of HuARD10 (a 5′-truncated *Alu* element) in human and chimpanzee. Black upper case letters indicate shared flanking unique sequence. The human-specific *Alu* insertion is featured in red; the extra portion in chimpanzee (representing that sequence deleted in human) is shown in blue. (b) Agarose gel chromatograph of a phylogenetic PCR analysis with an adjacent diagram depicting the insertion of the HuARD10 element and the deletion of 783 bases of DNA including a LINE element.

Table 1. Retrotransposition-mediated deletion frequency and polymorphism levels within the human and chimpanzee lineages

	Human	Chimp	Human to chimp ratio
Observed deletion events total	19	14	1.36
PCR tested	18	14	–
Fixed present	16	9	–
Polymorphic loci	2	5	–
Polymorphic fraction	0.11	0.36	0.31
Adjusted polymorphic loci	4	10	
Adjusted polymorphic fraction	0.19	0.53	0.36
Adjusted deletion events total	21	19	1.11

successfully by PCR. We determined the *Alu* insertion polymorphism to be 36% (five polymorphic loci; Table 1), similar to the polymorphism level of 37% reported recently by Hedges *et al.*, who used the same DNA panel.[10] After adjusting the value for sampling from a single sequenced genome, our chimpanzee diversity rose to 53%, again similar to the adjusted 59% polymorphism level reported.[10]

However, we found that two highly variable chimpanzee DNA donors accounted for four of the five polymorphic loci represented in the dataset. Another study from our laboratory has found these two chimpanzee genomes to be highly polymorphic.[25] Although information on sub-species membership for these chimpanzees is unavailable, recent nucleotide diversity data suggest that central African chimpanzees possess between 1.5 and 2.5 times more variability than do other chimpanzee subspecies.[26,27] Without these two individuals, our chimpanzee insertion polymorphism levels would have appeared considerably lower. Therefore, care should be taken when assessing polymorphism using small datasets and DNA of unknown subspecies membership. Further research to identify the four putative subspecies of chimpanzee through genetic testing will improve primate genomic diversity sampling strategies.

From our PCR screening of 160 human chromosomes (80 human individuals) and 24 chimpanzee chromosomes (12 chimpanzee individuals), we did not detect evidence of individual variation in the presence/absence of extra sequence alongside the newly inserted *Alu* elements.

Nucleotides lost through *Alu* retrotransposition-mediated deletion

The number of nucleotides deleted per retrotransposition event varied considerably within and between species. The number of nucleotides eliminated from the human genome totaled 8550 bp, with a range of 1546 bases between the largest and the smallest deletion (Table 2). Deletions associated with *Alu* retrotransposition occurring in chimpanzee totaled 466 bp (range 204 bp), considerably fewer bases than in human, even considering the smaller quantity of chimpanzee-specific insertion events.

Table 2. Genomic alteration through *Alu* retrotransposition-mediated deletion

	Human	Chimpanzee
Base-pairs deleted	8550	466
Mean	450	33
Range	1546	204

Target site duplications

Target site duplications were absent from the ARD loci detected in human and chimpanzee genomes, consistent with previous examples of L1 retrotransposition-mediated genomic deletions. Potential TSDs were present in only one ARD event, HuARD15. However, the sequences were not a perfect match. Given that HuARD15 is a young *Alu* element (0.6% diverged from consensus), there has been insufficient time for originally perfect TSDs to mutate to the current sequences, suggesting that this element did not possess TSDs from the integration process. Therefore, we conclude that hallmarks identified from retrotransposition-mediated deletion events using a cell culture system to study L1 retrotransposition[21] closely mirror the characteristics of ARD events *in vivo*.

Cleavage site preferences

In our data set, only eight out of the 33 ARD events (HuARD7, HuARD15, HuARD19, ChARD3, ChARD6, ChARD7, ChARD9 and ChARD12) possessed an integration site sequence similar to that preferred by L1 endonuclease, the endonuclease purportedly used by *Alu* elements during mobilization (Table 3).[3,4] The remaining 25 events exhibited non-canonical integration sites that may indicate L1 EN-independent integration, as postulated in previous studies (Table 3).[21–23] However, these non-canonical integration sites may be characteristic of L1 EN-dependent nicking, followed by promiscuous target primed reverse transcription (pTPRT; see below).

Genomic location

Alu insertions associated with genomic deletion localized to 12 of the 24 human chromosomes, and to 11 of the 25 chromosomes in chimpanzee (Figures 2 and 3, respectively). Deletions within gene-rich (typically GC-rich) regions would most likely be detrimental to the survival of an organism. Therefore, we would expect ARDs to be located in more AT-rich regions of the genome. To investigate this hypothesis, 10,000 nucleotides directly surrounding each element were analyzed for GC content using sequence analysis software (DNAstar

Table 3. *Alu* element integration sites

Human retrotransposition-mediated deletion		Chimpanzee retrotransposition-mediated deletion	
Locus name	Target integration site[a]	Locus name	Target integration site[a]
1	5′-aaat/a	1	5′-aagt/a
2	5′-gaat/a	2	5′-aacc/a
3	5′-tttt/t	3	5′-tttt/a[b]
4	5′-tttc/t	4	5′-acac/c
5	5′-ttga/t	5	5′-ttat/t
6	5′-ttct/g	6	5′-ttct/a[b]
7	5′-tttc/a[b]	7	5′- tctt/a[b]
8	5′-gccc/t	8	5′-tttt/g
9	5′-gtct/t	9	5′-ttct/a[b]
10	5′- atgc/t	10	5′-gttt/g
11	5′-ttgt/t	11	5′-ttcc/a
12	5′-tgta/t	12	5′-ttct/a[b]
13	5′-aaat/t	13	5′-gaat/a
14	5′- ttca/t	14	5′-tact/a
15	5′-tctt/a[b]		
16	5′-tttt/t		
17	5′-cttc/t		
18	5′-tatc/t		
19	5′-tttc/a[b]		

[a] Target integration sites are presented on the anti-sense strand in the 5′ to 3′ direction.
[b] Typical L1 EN nick sites.

v.5). The young deletion-associated *Alu* inserts in the human genome were more common in regions with lower GC content (~38% GC; genome-wide average 42% GC), similar to chimpanzee-specific *Alu* element insertions (36.4% GC; genome-wide average 40% GC). Thus, our dataset indicates that deletions in the human and chimpanzee genomes are tolerated better in regions with higher AT content, rather than in regions of high GC content. Approximately 75% of the genomic deletions detected in our study occurred within the introns of genes, rather than between genes. In one instance, a 1002 bp deletion at the HuARD6 locus induced the functional loss of a retroviral transforming gene, *c-rel*, within the human lineage. Research indicates that *c-rel* may have important roles in regulating cell proliferation and differentiation.[28]

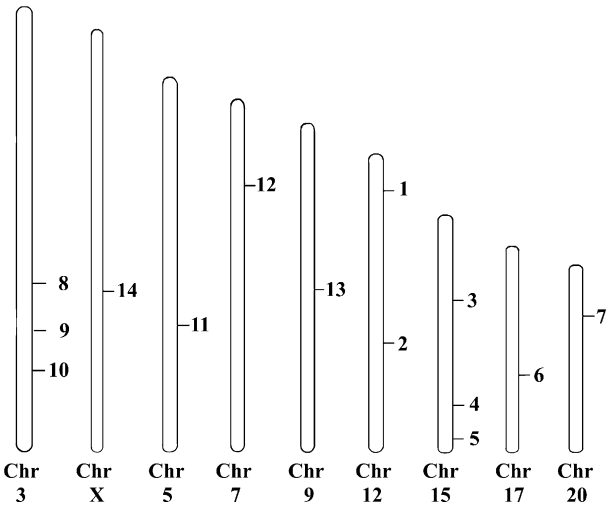

Figure 2. *Alu* retrotransposition-mediated deletions (ARD) within the chimpanzee genome. A partial schematic of the chimpanzee genome including those chromosomes occupied by ARDs. The labels indicate the name.

Unusual loci: internal priming

Within our human dataset, we found an example of a tail-less *Alu* repeat element. A member of the *Alu*Ya5 subfamily, the element (HuARD8) lacked approximately 20 bp of its 3′-end as well as the characteristic oligo(dA)-rich tail. This *Alu* element inserted at a non-canonical integration site and induced a small target site deletion of 21 bp. Plausible explanations for these unusual structural characteristics include internal priming and, alternatively, deletion of the tail *via* unequal recombination subsequent to the element's insertion. Internal priming appears more plausible than does A-tail recombination, given that the lineage-specific element has resided in the human genome only briefly. This hypothesis is supported by evidence that shows tail-less *Alu* sequences in only four elements (0.1%, one Yb8 and three Ya5) out of over 4000 lineage-specific *Alu* elements that have been analyzed in the human genome.[12,13,24,29] Therefore, to determine if internal priming could account for the tail-less nature of HuARD8, we used the 3′-end of the Ya5 consensus sequence to simulate the missing portion of the *Alu* RNA transcript. Using this approach, we found 11 bases at the 3′-end of the reconstructed HuARD8 RNA transcript to be complementary to the putative primer-binding site located within the first 25 bases downstream of the nick site (Figure 4). These data suggest that internal priming occurred during this particular *Alu* integration/deletion event.

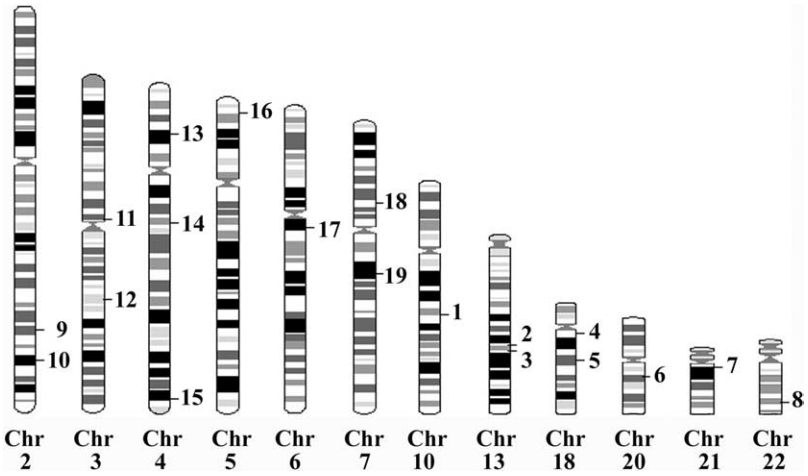

Figure 3. *Alu* retrotransposition-mediated deletions (ARDs) within the human genome. A partial schematic of the human genome including the chromosomes occupied by ARDs. The labels indicate the locus name.

Discussion

Our study offers the first genome-wide attempt to quantify the contribution of genomic ARD to the instability of the primate genome. Using computational comparisons supported by wet-bench methodologies, we provide evidence from the genomes of human and chimpanzee for 33 independent ARD events that have deleted approximately 9000 bases of DNA during the last five million years. These deletions may have been created independently of the *Alu* insertions or as a direct result of the insertion process. However, as we found no non-deleted allele across 80 human genomes and 12 chimpanzee genomes, we believe that it is highly unlikely that the deletions were created independently of the insertion of the mobile elements. Therefore, we conclude that the deletion of adjacent genomic sequences occurred prior to or, more likely, tightly associated with the insertion of the *Alu* elements. Further, our study indicates that *Alu* elements are able to use non-typical insertion sites in order to proliferate.

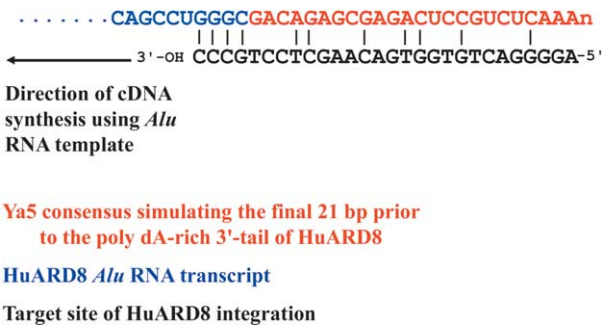

Figure 4. Internal priming of HuARD8. To determine if internal priming could account for the tail-less nature of HuARD8, we used the 3′-end of the Ya5 consensus sequence to simulate the missing portion of the *Alu* RNA transcript. This diagram indicates that 11 bases at the 3′-end of the reconstructed HuARD8 RNA transcript are complementary within the first 25 bases downstream of the nick site.

Insertion frequency and polymorphism of *Alu* retrotransposition-mediated deletion events *in vivo*

We determined that the human genome has suffered approximately 1.1 times more ARD events than has the chimpanzee. The direction of this adjusted *Alu* insertion ratio agrees with other comparisons of human and chimpanzee sequence data,[10,30] although it is somewhat lower than the insertion ratios of 1.8–2.0 detected in those studies. However, our program specifically searched for rare ARD events, so we would not necessarily expect to fully replicate results gathered from larger datasets. It is likely then, that our small data set in human did not fully capture the true level of polymorphism associated with these *Alu* element insertions, which would lead to a lower adjusted *Alu* insertion ratio. This bias would occur because sampling a single genome misses ~50% of the polymorphic insertion events that are present in the species as a whole.[10]

Although our chimpanzee sample size was smaller than that for human, we still obtained chimpanzee *Alu* insertion polymorphism levels consistent with other published studies.[10] By comparing the chimpanzee polymorphism rate to that of human, we determine chimpanzees to be three times more diverse than humans, in terms of ARD events. However, this comparison of polymorphism is skewed upwards by the low level of human *Alu* insertion polymorphism captured in our data.

The rate of retrotransposition-mediated deletions in primate genomes

We estimate that 0.28% (14 ARD events/5000 total chimpanzee-specific *Alu* insertion events; 0.38%, if adjusted for single genome sampling) of all *Alu* insertions in chimpanzee are non-typical and involve deletions of genomic material during retrotransposition. The ARD rate in humans is about 0.21% (19 ARD events/9000 total

human-specific *Alu* insertion events; 0.23%, if adjusted for single genome sampling), similar to the rate within the chimpanzee genome. For each species, the total number of lineage-specific *Alu* elements is based on the results of a previous study.[10]

The estimated frequencies of ARD in our data are lower than previously published reports of between 0.8% and 8%.[20–22] However, those studies generated biased estimates of retrotransposition-mediated deletion frequency in native genomes by using retrotransposition assays in cell culture from L1 element integrations,[20–22,31] or by studying exclusively one or two small *Alu* subfamilies.[20] These biases are outlined as follows. First, cell-culture assays do not assess the viability of cells suffering the effect of large deletions. Second, the effect of natural selection on the afflicted genome is essentially ignored under experimental conditions, thereby skewing estimates of deletion event frequency in naturally occurring genomes. Third, cells grown in culture may suffer from genomic repair insufficiencies that provide many more opportunities for mobile element integration and genomic deletion. Finally, deletion events drawn from small subfamilies of *Alu* elements rather than from the entire *Alu* family of elements might provide non-representative frequency estimations. The genome-wide search in this study provides a relatively unbiased estimate of tolerable ARD in primate genomes.

The size of deleted sequence *in vivo*

It is intriguing that human deletions are, on average, approximately 400 bp larger per deletion event than those found in chimpanzee. However, there is no mechanism known to account for this consistent disparity. In any event, the largest deletions retrieved from the genome sequence comparison accounted for 1556 (human) and 210 (chimpanzee) nucleotides. These deletions are small in comparison to those detected by L1 retrotransposition assays in human cells in prior studies,[21,22] which found deletions of up to 11,000 bp (and even 70,000 bp, empirically unconfirmed) that were presumably generated upon genomic integration of LINE cDNA transcripts. Whether such massive deletions are tolerable at the organismal level can be determined only by examining existing genomes, and our data suggest that they are not. Further studies to investigate whether human-specific L1 retrotransposition-mediated deletion events *in vivo* are smaller than those found *in vitro* will be informative.

Different mechanisms of *Alu* retrotransposition-mediated deletion

L1 EN-dependent retrotransposition-mediated deletion

The *Alu* insertions recovered during our study possess features uncommon to typical *Alu* elements, including the absence of surrounding TSD sequences and unusual target site preference. Experimental retrotransposition assays have documented similar characteristics within deletion-producing L1 element integrations.[21–23] From these *in vitro* studies, two putative mechanisms were put forward to explain the unique hallmarks of retrotransposition-mediated deletion. The first mechanism presumes that slight variations in L1 EN nicking can account for the absence of TSDs in addition to the insertion site deletions.[21,23] The authors proposed that L1 EN sometimes nicks the second strand a few base-pairs to the left of its initial nick site on the bottom strand, creating a substrate for exonuclease 5′–3′ digestion at the target site. L1 EN-dependent nicking is evident through L1 integration site preferences for sequences such as 3′- A/TTTT.[21,22] Our data suggest that L1 EN-dependent retrotransposition-mediated deletion, as determined through analysis of integration site preference, may account for 25% of the combined ARD events in native human and chimpanzee genomes.

L1 EN-independent retrotransposition-mediated deletion

In contrast to the studies by Gilbert *et al.*[21] and Symer *et al.*,[22] 75% of the ARD events in our data did not occur at typical AT-rich L1 EN cleavage sites. This result provides an argument for the existence of an L1 EN-independent integration mechanism for *Alu* elements, similar to that previously suggested for L1.[23] In this second model of retrotransposition-mediated deletion, it is likely that reverse transcriptase exploits existing breaks in the genome for TPRT initiation, not

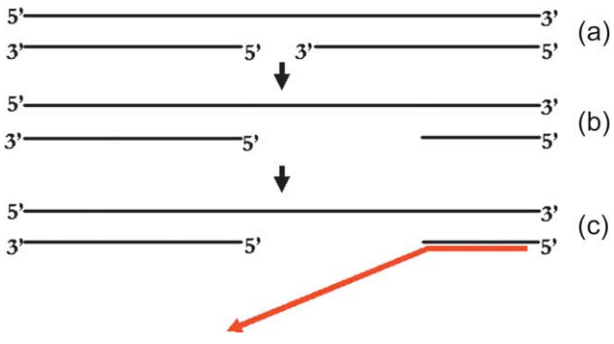

Figure 5. Model of genomic deletion mediated by promiscuous TPRT. (a) In this model, genomic breaks lead to the unwinding of the double DNA strand. (b) Removal of the unwound DNA strand may be resolved by mechanical force, or through enzymatic degradation. (c) Following this, TPRT is initiated from binding sites downstream of the initial break. This particular mechanism can account for the integration of elements at non-canonical sites without TSDs, in addition to the generation of target site deletions. However, the exact means by which the second strand breaks and the lesions are resolved are unknown factors in this model.

depending on L1 EN for the initial nick. The 3′ overhangs are presumably created prior to host repair of the lesion, generating the characteristic target site deletion. Thus, it appears that, similar to L1 elements, *Alu* repeats may be able to facilitate the patching of lesions in the genome. Whether EN-free insertion indicates a true function of retroelements or just a fortuitous portal into the genome is unknown. Regardless, confirmation of the L1 EN-independent integration of *Alu* elements requires further investigation using cell culture-based *Alu* retrotransposition assays within DNA repair-deficient cells.[4,23]

Promiscuous TPRT: a new model for *Alu* retrotransposition-mediated deletion

Here, we introduce a new mechanism to explain the unique characteristics associated with ARD events (Figure 5). The alternative priming system, promiscuous target primed reverse transcription (pTPRT), is named after the promiscuous initiation of reverse transcription from sites downstream of genomic breaks. In this model, genomic breaks lead to the unwinding of the double DNA strand, binding of *Alu* RNA transcript at a downstream homologous region and initiation of reverse transcription. Removal of the unwound DNA strand may be resolved by mechanical force or through enzymatic degradation. This particular mechanism can account for the integration of elements at non-canonical sites without TSDs, in addition to the generation of target site deletions. However, the exact means by which the second strand breaks and the lesions are resolved are unknown factors in this model.

Internal priming of *Alu* elements

We recovered one example of a 3′-truncated *Alu* repeat element (HuARD8) in the human dataset. Similar sequence hallmarks have been attributed to the mechanism of internal priming and were documented previously within L1 element *in vitro* assays and in the human genome sequence.[23,32] We determined that internal priming is consistent with the sequence hallmarks of HuARD8; further, regions of homology existed between the site of integration and the 3′-end of the simulated *Alu* Ya5 transcript, making internal priming possible. Although the primer binding site was not 100% complementary to the RNA transcript, empirical evidence suggests that initiation of cDNA synthesis does occur, if less efficiently, with RNA transcripts having a low level of homology to the site of integration.[33,34] Hence, we believe this study provides the first published analysis of internal priming in the reverse transcription of an *Alu* repeat element.

The mechanism of internal priming is a potential alternative to the L1 EN-independent integration mechanism presented earlier. *Alu* and L1 elements do not require L1 EN to nick at AT-rich sites because the RNA transcript can bind internally at the site of genomic breaks, even without 100% homology. Although this mechanism is exploited rarely (less than 0.1% of events), it represents an effective way for *Alu* elements to enter the genome by DNA breaks.

Contribution of *Alu* retrotransposition-mediated deletion to primate genomic instability

We have provided the first genome-wide study to quantify the contribution of ARD to the instability of the human and chimpanzee genomes, with an estimate of approximately 0.21–0.28% of all *Alu* element integrations over the last five million years being responsible for target site genomic deletions. If we assume the occurrence of ARD has been constant throughout the evolution of all primate orders, approximately 2520–3360 of all *Alu* insertion events (1.2 million) have eliminated around 687,960–917,280 bp (753,480–1,244,880, if adjusted for single genome sampling) of DNA from primate genomes (based on a human-chimpanzee average of 273 bp per deletion event). Even conservative amplification rates of one *Alu* insertion every 250 births[35] suggest that ARD could induce significant future changes to the overall architecture of primate genomes.

Although only one ARD event (*c-rel*) appears to have caused a coding difference between humans and chimpanzees over the last five million years of evolution, the potential contribution of ARD to primate genomic instability as a whole is undeniable. The true extent of collateral effects caused by *Alu* mobilization will require sequencing the genomes of representative members throughout the primate order.

Materials and Methods

DNA samples

DNA cell lines used in this study were obtained from the following sources: DNA samples from the African-American, European, and Asian populations were isolated as described.[8–10,12,24,36] DNA for the South American population group (HD 17 and HD 18) and for a lowland gorilla (*Gorilla gorilla gorilla*, AG05253A) was purchased from the Coriell Institute for Medical Research. African green monkey (*Chlorocebus aethiops* ATCC CCL70) and orangutan (*Pongo pygmaeus* ATCC CR6301) DNA was obtained from the American Type Culture Collection. A chimpanzee panel comprising 12 unrelated chimpanzees of unknown subspecies membership was obtained from the Southwest Foundation for Biomedical Research.

Computational analysis

The human July 2003 freeze and the *Pan troglodytes* November 2003 freeze from the University of California Santa Cruz were analyzed in this study†. To identify ARD events, 100 bases of 5′ and 3′ *Alu* flanking sequence in human were extracted and joined together into 200 bp

† http://genome.ucsc.edu

fragments. These 200 bp fragments were used as queries against the common chimpanzee genomic sequence using the Parcel BlastMachine at the Genome Core Facility at Columbia University. Due to random sequence match at the ends, we often see that the matches for the 5′ flanking region extend past the first 100 bp and the matches for the 3′ flanking region start before the 101 bp position. Therefore, the end-point for the 5′ flanking sequence and the start-point for the 3′ flanking sequence have to be re-adjusted in order to obtain the correct start-points and end-points in the target sequence. Following this, the sequences in the target chimpanzee genome between the 5′ and 3′ flanking sequences were extracted and used to compare with the corresponding human *Alu* sequences using the bl2seq program. To identify an ARD event, the corresponding criteria were met: (1) bl2seq did not produce a match between the query and the target sequence; or (2) bl2seq produced one or several hits (from deleted, unrelated *Alu* fragments) but the aligned region(s) were at least 5 bases away from at least one end of the target sequence. Then the computational comparison was reversed, comparing the chimpanzee genome against the human target sequence. Manual verification was performed using the Blast Like Alignment Search Tool (BLAT) and Basic Local Alignment Search Tool (BLAST) software.[37,38] This eliminated instances of deletion due to *Alu* gene conversion deletion, which appear as a replacement of an *Alu* in one lineage over another, accompanied by deleted sequence in the derived state. All of the manually verified ARD candidates were subjected to experimental verification using the PCR analyses of the loci.

To determine whether deleted sequences in the human or chimpanzee genome contained coding or regulatory regions, the experimentally verified deleted sequence data retrieved from the computational comparison above were queried against BLAT[37] and TRANSFAC software†.

Polymerase chain reaction analysis

To authenticate the ARD events, oligonucleotide primers were designed within the 400–1000 nucleotide long flanks surrounding each locus of interest using the primer design software Primer3 (Whitehead Institute for Biomedical Research, Cambridge, MA, USA‡). Primer sequences, annealing temperatures, PCR product sizes and chromosomal locations are located in the publications section of our website§. Each locus was amplified from the genomes of 80 humans (20 from each of four geographically diverse populations), 12 chimpanzees, one western lowland gorilla, one orangutan and one green monkey.

PCR analysis was performed in 25 μl reactions using 10–30 ng of DNA, 200 nM each oligonucleotide primer, 200 μM dNTPs in 50 mM KCl, 1.5 mM $MgCl_2$, 10 mM Tris–HCl (pH 8.4) and 2.5 units of Taq DNA polymerase. Each sample reaction was subjected to an initial denaturation step of 94 °C for 120 seconds, followed by 32 amplification cycles of 30 seconds at 94 °C, 30 seconds at the specific annealing temperature and 60 seconds at 72 °C, followed by one round of extension at 72 °C for five minutes. The PCR products were separated on a 2% (w/v) agarose gel and stained with ethidium bromide. Following separation, DNA fragments were visualized with UV fluorescence to assess the status of each locus.

DNA sequence analysis

To verify the existence of the ARDs, individual PCR products were either sequenced using chain termination sequencing methodology[39] with ABI Big Dye v.3.1 (ABI Biosystems) after gel extraction and cloning with the TOPO-TA cloning vector (Invitrogen), or directly from PCR products purified by the Wizard gel and PCR clean-up system as directed by the manufacturer (Promega). All sequenced PCR products were analyzed on an Applied Biosystems 3100 automated DNA sequencer. DNA sequence data were analyzed using the Seqman program in the DNAstar suite and aligned with BioEdit.

Data Bank accession numbers

The sequences of the orthologous non-human primates loci analyzed in this study have been assigned GenBank accession numbers AY881293–AY881325, and AY900585–AY900619.

† www.gene-regulation.com

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

§ http://batzerlab.lsu.edu

Acknowledgements

This research was supported by the Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000–05)-01 (to M.A.B.), National Science Foundation BCS-0218338 (to M.A.B.) and EPS-0346411 (to M.A.B.) and the State of Louisiana Board of Regents Support Fund (to M.A.B.) National Institutes of Health RO1 GM 59290 and a development fund from Roswell Park Cancer Institute (to P.L.).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2005.02.043

References

1. International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Sinnett, D., Richer, C., Deragon, J. M. & Labuda, D. (1992). Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences. *J. Mol. Biol.* **226**, 689–706.
3. Boeke, J. D. (1997). LINEs and Alus—the polyA connection. *Nature Genet.* **16**, 6–7.
4. Dewannieux, M., Esnault, C. & Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nature Genet.* **35**, 41–48.
5. Deininger, P. L., Batzer, M. A., Hutchison, C. A., III & Edgell, M. H. (1992). Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8**, 307–311.
6. Deininger, P. L. & Batzer, M. A. (1999). Alu repeats and human disease. *Mol. Genet. Metab.* **67**, 183–193.

7. Batzer, M. A. & Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Rev. Genet.* **3**, 370–379.
8. Callinan, P. A., Hedges, D. J., Salem, A. H., Xing, J., Walker, J. A., Garber, R. K. *et al*. (2003). Comprehensive analysis of Alu-associated diversity on the human sex chromosomes. *Gene*, **317**, 103–110.
9. Carroll, M. L., Roy-Engel, A. M., Nguyen, S. V., Salem, A. H., Vogel, E., Vincent, B. *et al*. (2001). Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J. Mol. Biol.* **311**, 17–40.
10. Hedges, D. J., Callinan, P. A., Cordaux, R., Xing, J., Barnes, E. & Batzer, M. A. (2004). Differential Alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res.* **14**, 1068–1075.
11. Xing, J., Salem, A. H., Hedges, D. J., Kilroy, G. E., Watkins, W. S., Schienman, J. E. *et al*. (2003). Comprehensive analysis of two Alu Yd subfamilies. *J. Mol. Evol.* **57**, S76–S89.
12. Otieno, A. C., Carter, A. B., Hedges, D. J., Walker, J. A., Ray, D. A., Garber, R. K. *et al*. (2004). Analysis of the human Alu Ya-lineage. *J. Mol. Biol.* **342**, 109–118.
13. Carter, A. B., Salem, A.-H., Hedges, D. J., Nguyen Keegan, C., Kimball, B., Walker, J. A. *et al*. (2004). Genome wide analysis of the human Yb lineage. *Hum. Genomics*, **1**, 167–168.
14. Bailey, J. A., Liu, G. & Eichler, E. E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73**, 823–834.
15. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P. *et al*. (2004). Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
16. Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y. *et al*. (2004). Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951.
17. Chen, S. J., Chen, Z., Font, M. P., d'Auriol, L., Larsen, C. J. & Berger, R. (1989). Structural alterations of the BCR and ABL genes in Ph1 positive acute leukemias with rearrangements in the BCR gene first intron: further evidence implicating Alu sequences in the chromosome translocation. *Nucl. Acids Res.* **17**, 7631–7642.
18. McNeil, N. (2004). Alu elements: repetitive DNA as facilitators of chromosomal rearrangement. *J. Assoc. Genet. Technol.* **30**, 41–47.
19. Hayakawa, T., Satta, Y., Gagneux, P., Varki, A. & Takahata, N. (2001). Alu-mediated inactivation of the human CMP- N-acetylneuraminic acid hydroxylase gene. *Proc. Natl Acad. Sci. USA*, **98**, 11399–11404.
20. Salem, A. H., Kilroy, G. E., Watkins, W. S., Jorde, L. B. & Batzer, M. A. (2003). Recently integrated Alu elements and human genomic diversity. *Mol. Biol. Evol.* **20**, 1349–1361.
21. Gilbert, N., Lutz-Prigge, S. & Moran, J. V. (2002). Genomic deletions created upon LINE-1 retrotransposition. *Cell*, **110**, 315–325.
22. Symer, D. E., Connelly, C., Szak, S. T., Caputo, E. M., Cost, G. J., Parmigiani, G. & Boeke, J. D. (2002). Human l1 retrotransposition is associated with genetic instability *in vivo*. *Cell*, **110**, 327–338.
23. Morrish, T. A., Gilbert, N., Myers, J. S., Vincent, B. J., Stamato, T. D., Taccioli, G. E. *et al*. (2002). DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nature Genet.* **31**, 159–165.
24. Roy-Engel, A. M., Carroll, M. L., Vogel, E., Garber, R. K., Nguyen, S. V., Salem, A. H. *et al*. (2001). Alu insertion polymorphisms for the study of human genomic diversity. *Genetics*, **159**, 279–290.
25. Han, K., Xing, J. *et al*. (2005). Extended retrotranspositonal quiescence supports a back seat driver model of Alu evolution. *Genome Res*. In the press.
26. Yu, N., Jensen-Seaman, M. I., Chemnick, L., Kidd, J. R., Deinard, A. S., Ryder, O. *et al*. (2003). Low nucleotide diversity in chimpanzees and bonobos. *Genetics*, **164**, 1511–1518.
27. Fischer, A., Wiebe, V., Paabo, S. & Przeworski, M. (2004). Evidence for a complex demographic history of chimpanzees. *Mol. Biol. Evol.* **21**, 799–808.
28. Bishop, J. M. & Varmus, H. (1982). Functions and origins of retroviral oncogenes. In *Molecular Biology of the Tumor Viruses: RNA Tumor Viruses* (Weiss, R., Teich, N., Varmus, H. & Coffin, J., eds), pp. 999–1108, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
29. Garber, R. K., Hedges, D. J., Herke, S. W., Hazard, N. W. & Batzer, M. A. (2005). The Alu Yc1 subfamily: sorting the wheat from the chaff. *Cytogenet. Genome Res*. In the press.
30. Liu, G., Zhao, S., Bailey, J. A., Sahinalp, S. C., Alkan, C., Tuzun, E., Green, E. D. & Eichler, E. E. (2003). Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13**, 358–368.
31. Kazazian, H. H., Jr & Goodier, J. L. (2002). LINE drive. retrotransposition and genome instability. *Cell*, **110**, 277–280.
32. Ovchinnikov, I., Troxel, A. B. & Swergold, G. D. (2001). Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. *Genome Res.* **11**, 2050–2058.
33. Chambeyron, S., Bucheton, A. & Busseau, I. (2002). Tandem UAA repeats at the 3′-end of the transcript are essential for the precise initiation of reverse transcription of the I factor in *Drosophila melanogaster*. *J. Biol. Chem.* **277**, 17877–17882.
34. Luan, D. D. & Eickbush, T. H. (1995). RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol. Cell. Biol.* **15**, 3882–3891.
35. Deininger, P. L. & Batzer, M. A. (1993). Evolution of retroposons. *Evolut. Biol.* **27**, 157–196.
36. Roy, A. M., Carroll, M. L., Kass, D. H., Nguyen, S. V., Salem, A. H., Batzer, M. A. & Deininger, P. L. (1999). Recently integrated human Alu repeats: finding needles in the haystack. *Genetica*, **107**, 149–161.
37. Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.
38. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
39. Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, **74**, 5463–5467.

Edited by J. Karn

(Received 11 January 2005; received in revised form 17 February 2005; accepted 18 February 2005)