Active Alu Element "A-Tails": Size Does Matter

Astrid M. Roy-Engel,[1] Abdel-Halim Salem,[2,3] Oluwatosin O. Oyeniran,[1]
Lisa Deininger,[1] Dale J. Hedges,[2] Gail E. Kilroy,[2] Mark A. Batzer,[2]
and Prescott L. Deininger[1,4,5]

[1]*Tulane Cancer Center, SL-66, Department of Environmental Health Sciences, Tulane University–Health Sciences Center, New Orleans, Louisiana 70112, USA;* [2]*Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, Baton Rouge, Louisiana 70803, USA;* [3]*Department of Anatomy, Faculty of Medicine, Suez Canal University, Ismailia, Egypt;* [4]*Laboratory of Molecular Genetics, Alton Ochsner Medical Foundation, New Orleans, Louisiana 70121, USA.*

Long and short interspersed elements (LINEs and SINEs) are retroelements that make up almost half of the human genome. L1 and Alu represent the most prolific human LINE and SINE families, respectively. Only a few Alu elements are able to retropose, and the factors determining their retroposition capacity are poorly understood. The data presented in this paper indicate that the length of Alu "A-tails" is one of the principal factors in determining the retropositional capability of an Alu element. The A stretches of the Alu subfamilies analyzed, both old (Alu S and J) and young (Ya5), had a Poisson distribution of A-tail lengths with a mean size of 21 and 26, respectively. In contrast, the A-tails of very recent Alu insertions (disease causing) were all between 40 and 97 bp in length. The L1 elements analyzed displayed a similar tendency, in which the "disease"-associated elements have much longer A-tails (mean of 77) than do the elements even from the young Ta subfamily (mean of 41). Analysis of the draft sequence of the human genome showed that only about 1000 of the over one million Alu elements have tails of 40 or more adenosine residues in length. The presence of these long A stretches shows a strong bias toward the actively amplifying subfamilies, consistent with their playing a major role in the amplification process. Evaluation of the 19 Alu elements retrieved from the draft sequence of the human genome that are identical to the Alu Ya5a2 insert in the *NF1* gene showed that only five have tails with 40 or more adenosine residues. Sequence analysis of the loci with the Alu elements containing the longest A-tails (7 of the 19) from the genomes of the *NF1* patient and the father revealed that there are at least two loci with A-tails long enough to serve as source elements within our model. Analysis of the A-tail lengths of 12 Ya5a2 elements in diverse human population groups showed substantial variability in both the Alu A-tail length and sequence homogeneity. On the basis of these observations, a model is presented for the role of A-tail length in determining which Alu elements are active.

[The sequence data from this study have been submitted to GenBank under accession nos. AF504933–AF505511.]

The genomes of all higher eukaryotes are littered with copies of mobile elements, including DNA transposons, short and long interspersed elements (SINEs and LINEs), and processed pseudogenes. The human SINE, Alu, is one of the most successful mobile elements, having generated over one million copies in the human genome (Lander et al. 2001). SINEs amplify by using an RNA polymerase III-derived transcript as template, in a process termed retroposition (Rogers and Willison 1983; Weiner et al. 1986). The RNA is reverse transcribed into a DNA molecule that integrates into a new site in the genome, probably using a nicked site in the genomic DNA as a primer (Boeke 1997). Previous studies have indicated that the ORF2 product of L1 elements supplies the endonuclease and reverse transcriptase activities for this integration process (Mathias et al. 1991; Feng et al. 1996). There is a variable-length stretch of homopolymeric adenosine residues or at least an A-rich region at the 3′ end of genomic SINE elements (Deininger et al. 1981). In Alu transcripts, this A stretch is within the transcript because the terminator for the transcription is found in the unique sequences downstream of the Alu element (Shaikh et al. 1997). Thus, each Alu transcript will be very similar in the Alu-containing 5′ end, have a similar 3′ oligo(dA)-rich tail that is encoded by the source gene and then having a unique segment at their 3′ end (Batzer et al. 1990; Shaikh et al. 1997). Thus, although this A-rich region is not a poly(A) tail in the sense used for polyadenylated mRNAs, the term is used in this paper for simplicity. The hypothetical common role of the A-rich regions of LINEs, SINEs and processed pseudogenes in the priming of reverse transcription was one of the first features predicted in these elements (Jagadeeswaran et al. 1981; Weiner et al. 1986). More recently, it has been suggested that the sequence specificity of the L1 endonuclease could nick the genomic integration site to provide a primer that could prime on these A-rich regions for all of these elements (Boeke 1997).

Most Alu amplification occurred >35 million years ago, with the current amplification rate almost 100-fold lower than at the peak of amplification (Shen et al. 1991). This current amplification rate is contributing significantly to human disease (Deininger and Batzer 1999), but must have been

[5]**Corresponding author.**
E-MAIL pdeinin@tulane.edu; FAX (504) 588-5516.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.384802. Article published online before print in August 2002.

more deleterious 40 million years ago (for review, see Batzer and Deininger 2002). Alu elements amplifying at different stages of primate evolution have key diagnostic sequence differences that allow them to be classified into subfamilies (Shen et al. 1991; Deininger et al. 1992). These subfamilies show a generally sequential amplification, with the youngest subfamilies amplifying at a much lower rate. The older subfamilies, which constitute about 85% of the Alu copy number, appear to be incapable of retroposition. One of the most likely explanations for the formation of subfamilies and the changes in amplification rate is that very few Alu elements are capable of amplification at any given time (Deininger et al. 1992). Any element that has made at least one copy is considered a 'source' element. The term 'master' element is used to refer to elements that may have amplified very efficiently over a relatively long period of time, contributing significantly to the pattern of evolution of the elements (Deininger et al. 1992; Deininger and Batzer 1995; Roy-Engel et al. 2001). Elements capable of very high levels of activity were probably rare during evolution.

The presence of a homopolymeric adenosine run is likely a common structure for priming the SINE and LINE retroposition events (Boeke 1997), but it does not explain why so few Alu elements are capable of amplification. The vast majority of the million plus Alu elements present in the human genome have poly(A) stretches, but only a very few are able to retropose (Deininger and Daniels 1986; Batzer et al. 1990; Deininger et al. 1992). Therefore, the mere presence of an A stretch is not sufficient to confer on an Alu element the ability to retropose efficiently. Multiple factors, such as transcriptional capability and specific SINE interactions (possibly through other RNA binding proteins) with the LINE retrotransposition proteins, have been suggested to affect the amplification process (Schmid and Maraia 1992; Batzer and Deininger 2002). Ordinarily, L1 element proteins show a strong *cis* preference for the RNA that encodes them (Wei et al. 2001). Thus, Alu elements probably have a specific mechanism to compete effectively with this *cis* preference. In addition, the presence of an A stretch by itself does not explain why Alu elements amplify so effectively, whereas other polyadenylated RNAs in the cell only form retropseudogenes at much lower rates (Goncalves et al. 2000; Harrison et al. 2002). In this paper, a model is presented in which the length of the Alu A-tail, possibly coupled to interacting with poly(A) binding protein (Sondekoppa Muddashetty et al. 2002; West et al. 2002), may be a critical element that allows Alu elements to compete effectively for the L1 ORF2 product, resulting in their very high amplification rate.

RESULTS

Length Variation of the Alu and L1 "A-Tails"

Some of the most recent Alu and L1 inserts have particularly long A-tails and an evaluation was performed to quantitate this observation for both of these families of mobile elements. The length of the A stretch was operationally defined as the number of bases between the last nucleotide of the Alu consensus sequence and the first non-A base in the 3′ flanking direct repeat (for details, see Fig. 1). Determination of the L1 A-tail length was more empirical and less precise because the direct repeats were not defined for most of the elements. In this case, the A-tail length was determined by counting the number of bases between the last nucleotide of the L1 consensus until reaching two consecutive non-A residues, unless another major poly(A) stretch followed. Under these conditions, any errors would more likely involve a few overestimations in the size length. This study was also limited to only the younger Ta subfamily of L1 elements (Boissinot et al. 2000; Sheen et al. 2000; Ovchinnikov et al. 2001) to minimize difficulties identifying the ends of the A-tail.

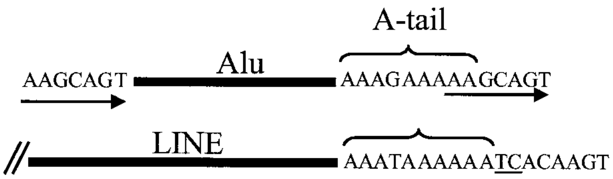

Figure 1 Parameters used to determine the length of the A-tail of an Alu element. Alu elements (solid bar) and their A-tails are flanked by direct repeats (arrows) that are created during their insertion into the genome. For this study, the length of the A-tail was considered as the number of bases between the last nucleotide of the Alu consensus sequence and the 3′ flanking direct repeat. If the direct repeat sequence contained adenosine residues in its 5′ end, they were included in the count. The A-tail length for long interspersed elements (LINES) was determined by counting the number of bases between the last nucleotide of the L1 consensus until reaching two consecutive non-adenosine bases, unless another poly(A) stretch followed. In the example, the A-tail length would be considered to be 9 for the Alu and 10 for the LINE. Note that the A-tails may contain other nucleotide residues.

We determined the length of the A-tail from different subfamilies of Alu and L1 elements and compared them to the most recent Alu insertion events. Those elements that resulted in a human disease were considered as the very recent inserts. Several of these are probably de novo events, and most of them should have occurred within a relatively small number of generations because there is probably some selective pressure against the disease-causing alleles that would lead to their rapid loss from the population fairly quickly (Prak and Kazazian 2000). L1 elements that were generated de novo in tissue culture using the L1 retrotransposition assay were also included (Moran et al. 1996). For simplicity, all of these most recent elements will be referred to as the "disease" elements. The Alu elements analyzed were divided into the following subfamilies: Alu J and S were pooled to represent the 'old' Alu elements, whereas Alu Yb8, Alu Ya5, and Alu Ya5a2 were studied separately as members of young subfamilies (Roy et al. 2000; Carroll et al. 2001; Roy-Engel et al. 2001). Details of the sources of the Alu and L1 elements analyzed are indicated in Methods.

Some of the results of our analysis of the A-lengths of the different Alu and L1 elements are shown in Table 1. The A-tail was significantly longer in the disease inserts, with a difference of almost twofold in the mean length relative to any of the Alu subfamilies. Although, L1 elements have longer A stretches than do Alu elements in general, the same trend in A-length difference was observed. The longer L1 A-tails are consistent with their being generated from a polyadenylated mRNA rather than from an internal A-region as in the SINEs. In addition, the frequency distribution of the A-tail length in the different subfamilies was determined (Fig. 2). Again, there is a striking difference between the disease elements and the other older groups of Alus. The lengths of the A stretches in the disease Alu elements are all over 40 bp, and in the L1 elements only one is <40 bp with the rest being over 53 bp in length. Some of the A-lengths in this group reach as high as 97 bases for Alu and almost 180 bases for L1. For the groups

Table 1. Mean Length of the 3′ "A-Tail" of Different Alu and L1 Elements

Repeated element group	Total analyzed	Length A-tail[a]
Alu J and S	100	21 ± 8
Alu Yb8	79	28 ± 9
Alu Ya5	235	26 ± 9
Alu Ya5a2	23	30 ± 11
Alu "disease"[b]	16	58 ± 19
L1 Ta	50	41 ± 20
L1 "disease"[b,c]	14	77 ± 33

[a] Mean ± standard deviation.
[b] Elements found to cause disease by insertion.
[c] Includes some elements that were generated as new inserts from the L1 tissue culture assay and two L1 elements causing disease in mice.

with a large *n* value (Ya5 and J/S), the data roughly fit a Poisson distribution with a slight skew to the longer A-stretch lengths. This indicates that, once the elements are within the genome, various forces act on the A-tails until they shrink to an equilibrium value. Because the data represent Alu elements that have inserted at a variety of time points, the skewed distribution may indicate that some elements have not yet reached equilibrium. Overall, older Alu elements appear to have shorter A-tails than do their younger counterparts. This correlation was previously observed in LINEs when comparing the A-tail length of a group of older L1-GAG elements to the younger L1H-Ta group (Ovchinnikov et al. 2001).

Alu Elements With Long A-Tails Are Uncommon in the Human Genome

To evaluate the abundance and the age of the Alu elements with long A stretches in the human genome, we performed a search of the human genome draft sequence. Only the Alu elements with 40 or more adenosine residues, with no other intervening bases, were retrieved. Fewer than 1000 Alus were identified from the draft sequence (Table 2). The 1000 Alus with long A-tails represents approximately 0.1% ($1000/1.1\times10^6$) of the total number of Alu elements present in the human genome. The use of a more stringent parameter of 50 or more adenosine residues to perform the search reduces the total number by about 10-fold (Table 2).

The distribution of these long A-tailed Alus roughly divides the elements to about half that clearly belong to the currently active Alu subfamilies Y, Ya5, or Yb8 and the other half that belong to the older, inactive Alu subfamilies J, Sx, Sg, or Sc. Evaluation of the proportion of long A-tailed Alu members within each subfamily (using the previously estimated total copy number for each subfamily [Roy-Engel et al. 2001]) showed that the youngest Alu subfamilies have the highest percentage of long A-tail members. The group of most active Alu subfamilies, Ya5/Ya5a2/Ya8, contained the most long A-tail elements with about 4.1% (113/2750) of their members in this category, followed by Alu Yb8/9 with about 3.0% (38/1900), Alu Y with only 0.17% (313/~190,000), and Alu S and J with 0.06% ($539/9\times 10^5$). There is an inverse correlation between the age of the Alu subfamily and the proportion of the members with long A-tails in the genome, indicating that loss of A stretches may be a primary, but not the only, inactivating feature in the older subfamilies.

Analysis of the Candidates for the Potential Source Gene of the Alu Ya5a2 in *NF1* Patient Loci

If the A-tail length is a critical parameter for Alu retroposition activity, we would predict that the source Alu element responsible for the de novo insertion of an Alu element inactivating the *NF1* gene in a neurofibromatosis patient (Wallace et al. 1991) should have an identical sequence to the inserted Alu, as well as possessing a long A-tail. Because the NF1 insertion was in a paternal chromosome, the long A-stretch source Alu would have to be found either in the father's DNA or in the DNA of the patient. A query of the draft of the human genome database in search of the Alu elements with 100% sequence identity to the *NF1* Alu retrieved a total of 19 elements. We had previously identified 13 matches to the NF1 insertion event, which is a perfect match to the Ya5a2 consensus, using a less complete version of the human sequence available at the time (Roy et al. 2000).

Of the 19 Ya5a2 elements, only 5 had A-tails of 40 or more bases in the GenBank database. These five and the two with the next longest A-tails were selected to evaluate those specific loci in the genomes of the NF1 patient and the father. The loci were analyzed using polymerase chain reaction (PCR)-generated clones to determine insertion presence or absence of the element and the length of the A stretch. More than one sample from each PCR was sequenced to increase the probability of detecting differences (if any) between the two alleles of each subject or potential size variations generated by the cloning procedure. Only two loci (Table 3) contained Alu elements with the long A-tail in either the patient (35 bases) or the father (53 bases). The 35-base allele may represent the other paternal allele, a maternal allele, or, less likely, the shortening of the allele from one generation to another. In the other case, the locus contained two alleles with very different A-tail lengths in the patient. One of the A-tails was in the 24–26 bp range (the heterogeneity may result from either mosaicism within the patient or, more likely, minor variation created in the PCR and cloning). This allele was similar to the paternal allele(s). However, the patient also had an allele with an A stretch of 45 bases (comparable to the one observed in the draft human genome), which may be an allele inherited from the mother. Because exhaustive measures were not taken to definitively analyze all alleles in these samples, it is impossible to say that there were no other long A-tail alleles. These data show the presence of at least two alleles present in the father or the patient with an A-tail of sufficient length to have served as the source element for the disease-causing insertion. In addition, the A-tail length of a specific Alu element can differ between individuals, indicating a high variability in the population. Evaluation of the degree of insertion polymorphism associated with these elements showed that the three fixed loci with long A-tails had at least one clone comparable to the A-tail length observed in the draft human genomic sequence (Table 3).

Population Diversity of A-Tail Length and Composition

The polymorphism in A-tail lengths from 12 Ya5a2 containing loci was evaluated by sequence analysis from at least 25 and up to 72 PCR-generated clones from multiple individuals from different populations, with the exception of one locus (Ya5a2AD9) in which there are data from only five individuals. These data show extensive population diversity in A-tail length (Fig. 3A). A number of the loci show a similar length

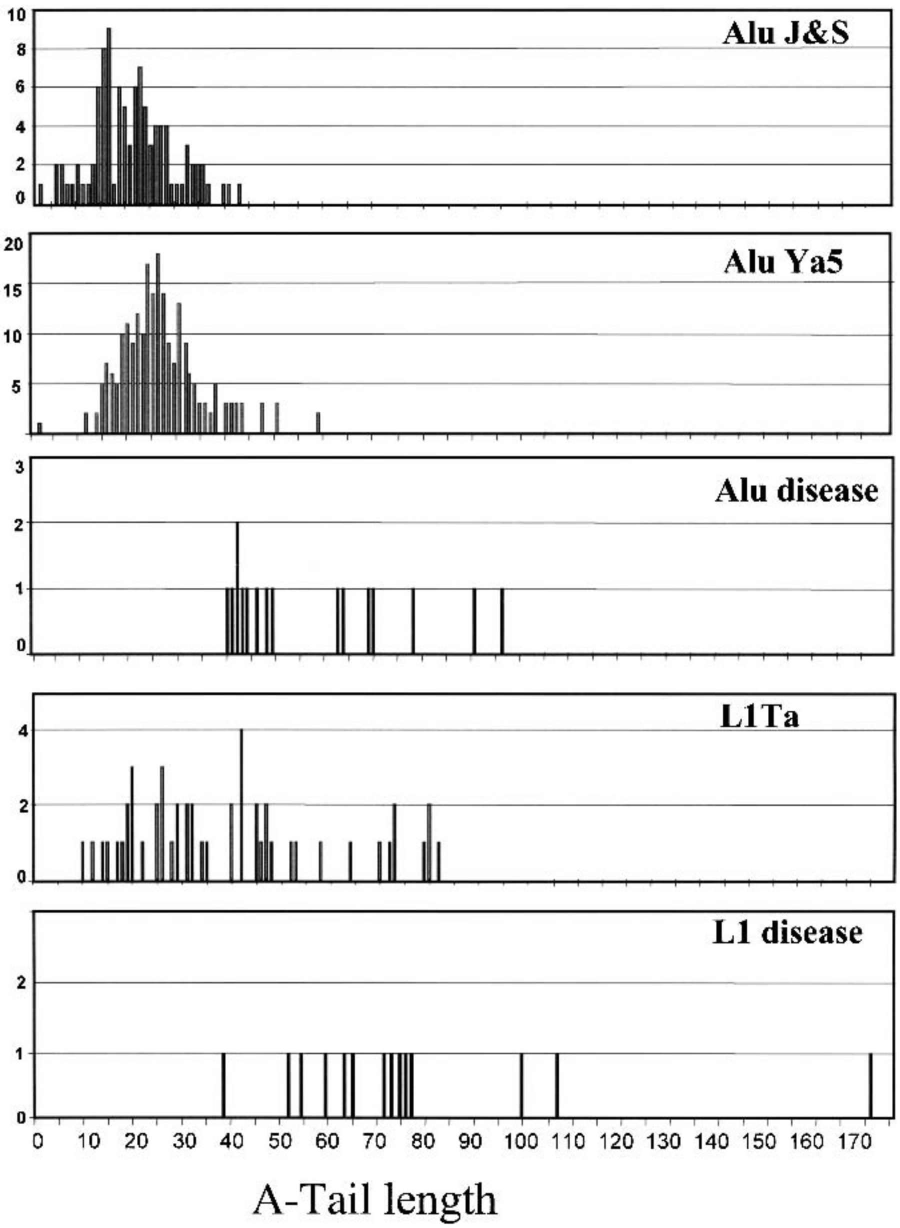

Figure 2 Histogram of the distribution of the length of A-tail in different subgroups of Alu and L1 elements. Distribution of the A-tail length of Alu elements belonging to old subfamilies (Alu J and S), the young subfamily Alu Ya5, and those reported to cause disease (see text for definition), plus the L1 elements from the young Ta (Ta0 and Ta1) and those linked to disease. Note: The scale for the *Y*-axis (number of elements) is different for each group.

distribution to those seen between different elements shown in Figure 2. However, a few of the loci showed a significant number of alleles with A stretches longer than 40. Several appeared to be most consistent with A-tail length distributions centered around two major allele lengths (i.e., Ya5a2AD5, AD11, and AD12). Comparison of the distribution of the Alu Ya5 elements and the combined data of all of the analyzed Ya5a2 loci shows the same distribution, with a more pronounced skew to longer A-tails for the younger Alu Ya5a2 elements (Fig. 3B).

In addition to the changes in length, there are also other mutations appearing in these sequences, which break up the perfect nature of the A stretch. The A-tails for almost half of the Ya5a2 elements analyzed have almost no sequence heterogeneity, with a few rare Alus that have one non-A base. In contrast, the other loci contain Alu elements in which the majority, if not all, contained at least one non-A base in the A stretch. An example of the A-tails of the elements present in one of these loci is shown in Figure 4. Some of the changes appear consistently in individuals from the same population group, indicating a possible founder effect (see Fig. 4). It seems likely that this sequence heterogeneity may also influence the length stability of the elements in which disruption of the A stretch with other bases may prevent length reduction. In some cases, these base changes may serve as nuclei for subsequent microsatellite expansion (Arcot et al. 1995). The A-tail variability may indicate a more dynamic Alu amplification than previously thought, in which different individuals may have different active elements at a given time.

DISCUSSION

The Length of the Alu A-Tail and Retroposition Capability

Only a few Alu elements are currently capable of serving as source elements (Batzer and Deininger 1991; Batzer et al. 1992; Liu and Schmid 1993). There are a number of potential features that may limit the retroposition activity of Alu elements and subfamilies. Among these factors are any sequence differences within the Alu element, either of a subfamily or random nature (Liu and Schmid 1993; Chu et al. 1995; Shaikh et al. 1997). The subfamily changes may influence the interaction of the Alu elements with the retroposition machinery, resulting in a selection against the older subfamilies (Matera et al. 1990; Englander et al. 1993; Kim et al. 1994; Labuda and Zietkiewicz 1994; Shaikh and Deininger 1996; Shaikh et al. 1997). The random mutations may result in loss of expression of the older elements, either through promoter mutations or through changes in RNA stability (Alemán et al. 2000; Batzer and Deininger 2002). However, transcription itself only partially explains more than a small part of the limitation in source elements. The majority of Alu RNA transcripts are still transcribed from subfamilies that appear to be incapable of amplification (Shaikh et al. 1997). The 3′ end of the Alu tran-

Table 2. Subfamily Distribution of Alu Elements With Long "A-Tails" Retrieved From the Human Genome Draft

	A-tail with ≥40 A		A-tail with ≥50 A	
Alu subfamily	Total members	Percent (%)	Total members	Percent (%)
J	74	7.2	7	5.6
S	465	45.5	51	40.5
Y/Yc1	313	30.6	34	27.0
Ya5/Ya5a2	113	11.1	26	20.6
Yb8/Yb9	57	5.6	8	6.3
Total	1022	100	126	100

scripts is the most variable, both in terms of length and homogeneity of the A stretch, and as the unique sequences between the A stretch and the RNA polymerase III terminator. Thus, interactions occurring with this 3′ end would be likely determinants causing selection of a few active master elements.

Both the length and the homogeneity of the A stretch appear to be important for efficient retroposition. Sequence comparison of the BC1 locus, the known master gene for the rodent ID family of SINEs (Kim et al. 1994) from different rodent species relative to the accumulated copy number (Kass et al. 1996) strongly supports this observation (Fig 5). Although the sequence of the BC1 gene and flanking regions in general are well conserved (Deininger et al. 1996), as are their relative expression levels (Sapienza and St Jacques 1986), the 3′ A stretch varies greatly between the different species of rodents. There is a close correlation between A-richness of these BC1 loci and the copy numbers of ID family members in these species (Fig. 5). Of particular note are the presence of only 200 ID copies in guinea pig and the very short A stretches in the BC1 gene, compared with the long A stretches in both *Peromyscus* species and their high copy numbers. Rat has a copy number closer to 125,000, but previous studies showed that the BC1 RNA gene generates only about 12,000 of the copies (Kim et al. 1994). This apparent requirement for long A-tails in addition to a minimum amount of pure adenosine stretches for efficient retroposition may be determined by the interaction with poly(A) binding protein, as discussed in the following paragraph.

Previous reports hypothesize that poly(A) polymerase may occasionally act on RNA polymerase III transcripts and create the A stretches on SINEs (Chen et al. 1998; Borodulina and Kramerov 2001). However, it seems unlikely that the 3′ end of an Alu transcript would first degrade back to the A stretch followed by a polyadenylation reaction before retroposition. Studies of Alu RNAs have shown no evidence for such polyadenylation being a significant factor (Shaikh et al. 1997), indicating that the Alu A stretches are coded in the genomic elements (Batzer et al. 1990). For all of the new disease-causing insertions to have A stretches of 40 bases or more, their source elements must have had longer A stretches. Therefore, the 1000 Alus with A-tails longer than 40 bp in the initial human genome sequence are possible candidates to serve as source elements. The 126 with 50 A-tail lengths are even better candidates, because this is probably the length for the most efficient, cooperative binding of two molecules of poly(A) binding protein (Smith et al. 1997). However, a long A stretch alone is not sufficient to allow Alu amplification. Elements with mutations mitigating transcription capacity or the interaction with L1 supplied factors, for instance, would not be capable of retroposition even with a long A-tail.

This latter observation would help to explain the apparent amplification inactivity of the old Alu subfamilies despite the presence of some elements with long A-tails. Several of the

Table 3. "A-Tail" Length of Several Alu Elements With 100% Identity to the Ya5a2 Alu Element Insert in the NF1 Gene

		A-tail length			
Locus	Accession number	Draft[a]	Patient	Parent[b]	Allele frequency[c]
Ya5a2AD11	AC090071	53	35	51	Fixed present
Ya5a2AD12	AC004057	46	26–28/45	24–31	Fixed present
Ya5NBC208	AL132992	37	38	34	Fixed present
Ya5a2AD7	AL138681	43	27–28	26–29	Intermediate
Ya5a2AD8	AL162713	53	23	31	Intermediate
Ya5NBC220c	AC007611	39	24	24–25	Intermediate
Ya5NBC239	AL133284	41	Empty[d]	Empty[d]	Low

[a] Size of A-tail of the Alu element retrieved from the human genome draft sequence.
[b] Only DNA from the father of the patient was available for polymerase chain reaction (PCR) analysis.
[c] Allele frequency was classified as fixed present, fixed absent, low, intermediate, or high frequency insertion polymorphism. Fixed present: Every individual tested had the Alu element in both chromosomes. Low frequency insertion polymorphism: The absence of the element from all individuals tested, except for one or two homozygous or heterozygous individuals. Intermediate frequency insertion polymorphism: The Alu element is variable as to its presence or absence in at least one population. High frequency insertion polymorphism: The element is present in all individuals in the populations tested, except for one or two heterozygous or absent individuals.
[d] No Alu insert present in the locus; sequence confirmed the PCR fragment to be the preinsertion site.

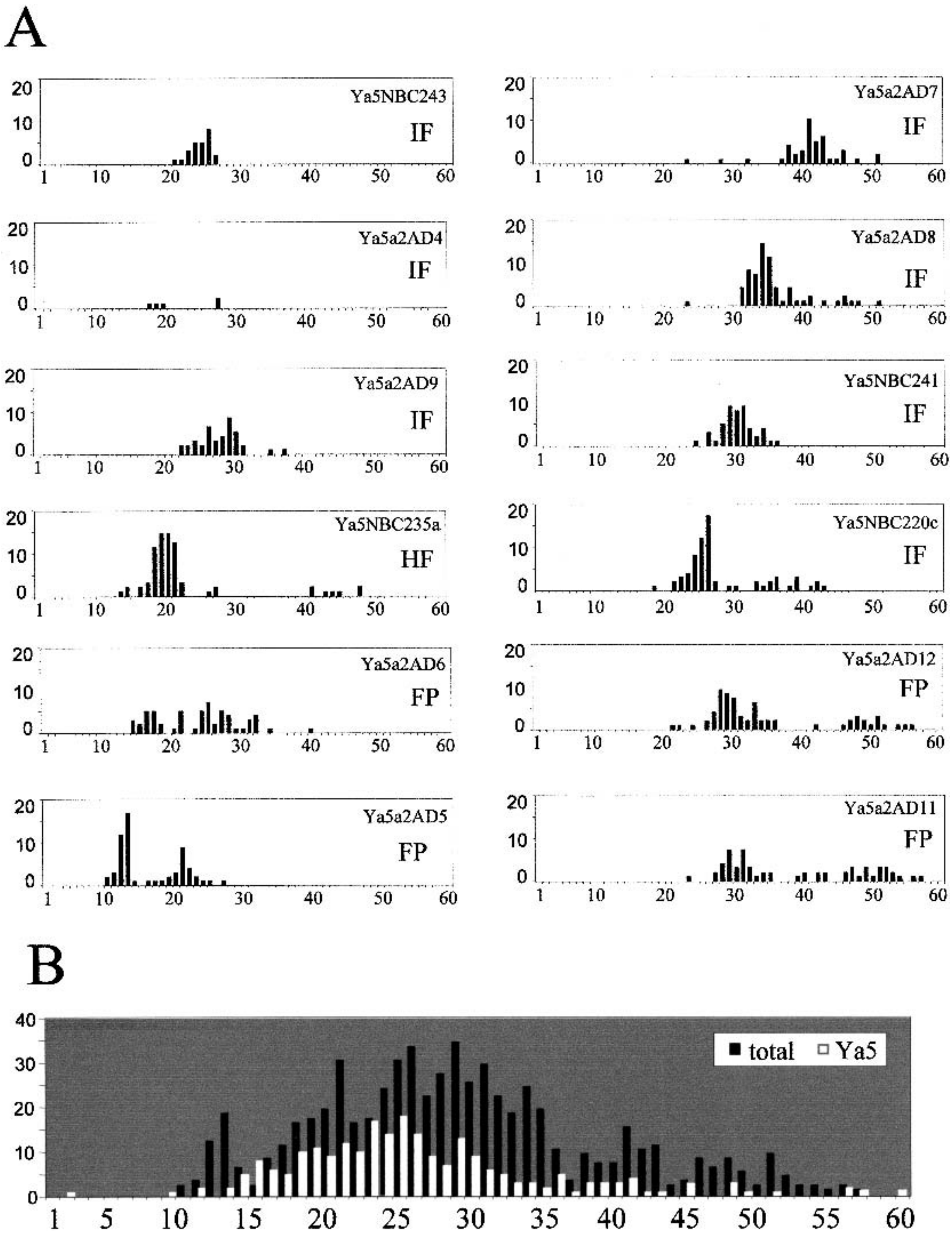

Figure 3 A-tail length variation within the population. (*A*) Histograms of the A-tail length distribution of the selected Ya5a2 element from the same locus of different individuals. The histogram reflects the A-tail length (*X*-axis) and the amount of elements for that length (*Y*-axis). The allele frequency for each locus is indicated as fixed present (FP), intermediate (IF), or high frequency (HF); see definitions in Table 3. (*B*) Histogram of the distribution of the Ya5 elements retrieved form the genome draft sequence (white bars) and the combined data of all the individual Ya5a2 elements (black bars) from different populations analyzed, as listed earlier.

characterized older Alu elements are highly unlikely to generate pol III transcripts because of 5′ truncation and multiple mutations within their internal A and B boxes. In addition, these older elements have accumulated multiple mutations (both CpG and non-CpG), which may reduce or abolish retropositional capability. To calculate the expected influence of random mutations decreasing the relative retroposition efficiency of the older subfamilies, we used the proportion of Alu

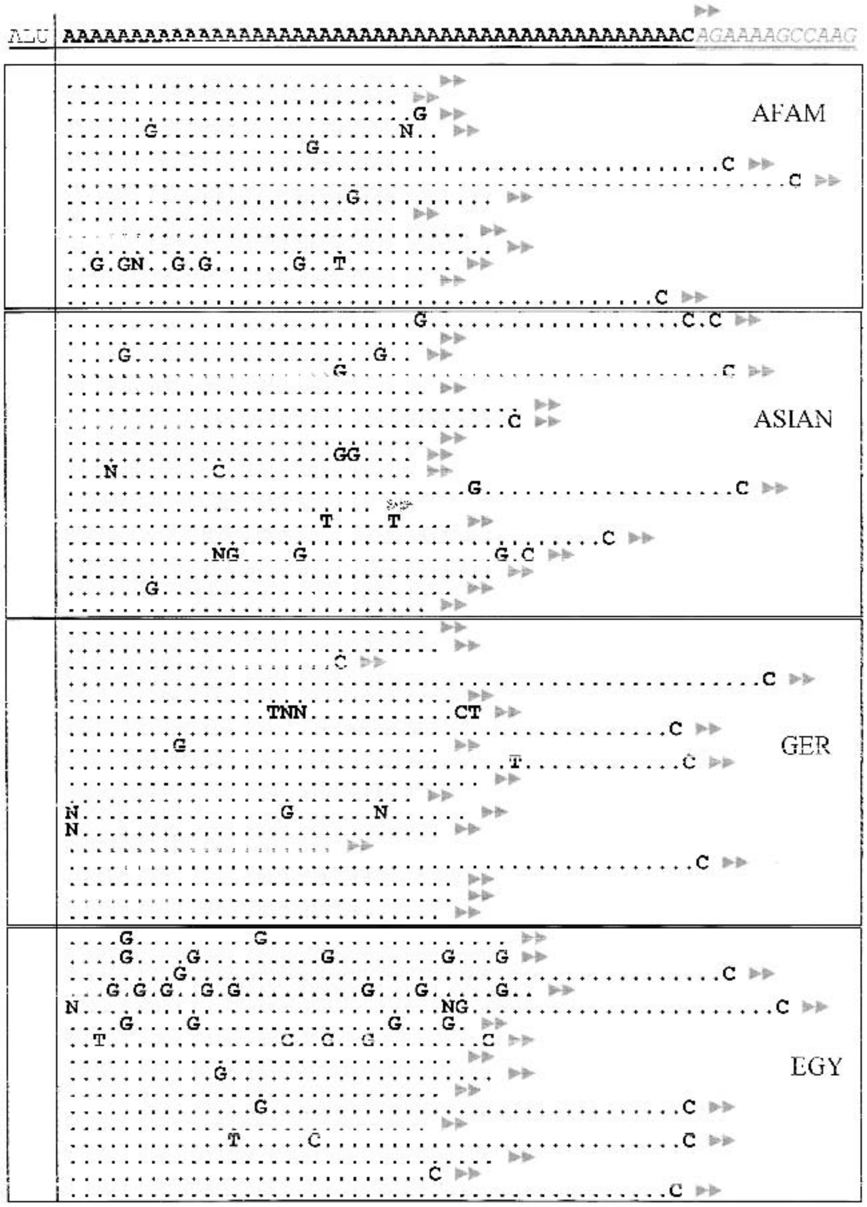

Figure 4 Sequence alignment of the A-tails of the Alu Ya5a2AD12 from different individuals. The sequence of the A-tail from the Alu element obtained from the human genome draft is shown on the top line. The sequence of the Alu body is not shown but is represented by the word ALU. The 3′ direct repeat is shown in gray italics and is represented by double arrowheads. Nucleotide substitutions at each position are indicated with the appropriate nucleotide or with an N when the nucleotide could not be clearly identified. The four populations analyzed are indicated inside each boxed group: African American (AFAM), Asian (ASIAN), German (GER), and Egyptian (EGY).

transcripts originating from elements of the older Alu S and J subfamilies and the younger Alu Y subfamily (Shaikh et al. 1997). The most actively amplifying subfamily, Alu Ya5, contributed <1% of the transcripts. A "transcript enrichment" factor was calculated by using the percent of each Alu subfamily and the percent of actual transcripts they contribute (Table 4). For example, the Alu Ya5 contributes ~ 0.3% of the total number of Alus in the genome and contributes 0.8% of the Alu transcripts; thus the enrichment factor would be 2.64 (0.8/0.3), indicating the detection of 2.6-fold more transcripts than that expected just by copy number. We must caution that this transcript enrichment was for expression in cultured cells and not germ cells, and therefore is only an approximation. Multiplication of elements in each subfamily with long A-tails by their relative subfamily transcription rates generates a first correction to the expected amplification capability of these long A-tailed Alus (Table 4). When comparing the actual numbers (observed) of Alu elements in each subfamily from our disease group with our estimates (expected), based on the product of the number of long A-stretch Alus and their relative transcription rate, a reasonable correlation is obtained, with the exception of the expected amplification of some of the older, inactive Alu subfamilies. Thus, the length of the A stretch and transcription rate may be the primary factors, but are not the only factors required for efficient retroposition. Other factors (such as loss of secondary structure, interaction with proteins, etc.) that might negatively impact the amplification capabilities of the older subfamilies in addition to this simple model need to be considered in this model. For example, the ability to interact with proteins, such as the L1 factors, are potential explanations for the discrepancy. In this case, variations digressing away from the consensus sequence more commonly found in older Alu elements may have an important impact in the RNA–protein interaction.

Alu A-Tail Length Instability

The long A stretches of the newly inserted elements are not evolutionary stable. A clear example of this instability can be observed with the Alu that inserted in the *eya1* locus, where the A-tail went from A_{97} to A_{31} in one generation (Abdelhak et al. 1997). Our own data show that even the Ya5a2 subfamily, which has an estimated average age of 0.62 million years (0.28–1.08 with 95% confidence) (Roy et al. 2000), is quickly approaching an A-tail distribution similar to that of the older subfamilies. In these latter studies, some cases presented two size distributions of alleles. The longer alleles might be consistent with the original length of the A stretch following insertion. Thus, it is likely that even a period of about one million years has not completely eliminated the initial long A stretches in some loci. Alternatively, the A-tail length may be more dynamic than this simple model of long A stretches shrinking to a smaller size distribution. Other events may allow the A-stretches to amplify to longer lengths (Wang et al. 2001). The

```
                                        3'end                                                                        species copy#
ctgg-AAAAAAAAAAAAAAAAAAAAAAAAAgAA--------gAcAAAAcAAcAAAAAAAgAccAAAAAAAAAcAAggtAActgg-cAcAcA--cAAccttt   C   37000
ctgttAAAAAAAAAAgAccAAAAAAAAAAgAAAAAAAAAAAgAcAAAAtAAcAAAAAA-gAccAAAAAAAAAcAAggtAActgg-cAcAcA--cAAccttt   D   25000
ccg--AAAAAAAAAAAAAAAAAAAAAA--------------gAcAAAAtAAcAAAAA--gAccAAAAAAAAAcAAggtAActgg-cAcAcA--cAAccttt   R  *12000
ctgg-AAAAAAAAAAAAAAAAAAAAA---------------gAcAAAAtAAcAAAAA--gAccAAAAAAAAAcAAggtAActgg-cAcAcA--cAAccttt   M   12000
ctgg-AAAAAAAAAAAAAAAAAAAAA---------------gAcAAAAtAAcAAAAA--gAccAAAAAAAAAcAAggtAActAg-cAcAcA--cAAccttt   G    9000
ctgg-AAAAAAAAAAA-------------------------gAcAAAAtAAcAAAAA--gAccAAAAAAAAAcAAggtAActAg-cAcAcA--cAAccttt   H    2500
ctcc-AAAcctgAAAAAcAAAAAAtccctAtAAAAAAA---cAcAAAAtAtAAAAAA--gAccAAAAAcAAAcAAggtAActgcgcAcAcAAAcAAccttt   GP    200
```

Figure 5 Sequence alignment of the A-rich regions of the BC1 RNA gene of different rodent species and the copy number. The sequences of the 3′ ends of the BC1 RNA genes, containing the A-rich stretches, are shown with dashes used to improve the alignments. The rodent sequences are as follows: C, *Peromyscus californicus* (Accession #U33852); D, deer mouse, *Peromyscus maniculatus* (#U33851); R, rat, *Rattus norvegicus* (#M16113); M, mouse, *Mus musculus* (# U01310); G, gerbil, *Meriones unguiculatus* (#U33852); H, hamster *Mesocricetus auratus* (#U01309); and GP, guinea pig, *Cavia porcellus* (#U01304). Variations of the adenosine residues between the species are indicated in bold. Rectangles enclose the longer homopolymeric A stretches within the variant region. BC1 copy number for each species (Kass et al. 1996) is indicated at the right. The asterisk represents the copy number only for the rat BC1 type 1 (Kim et al 1994).

general equilibrium may favor the shorter alleles, but occasional growth of A stretches could result in activation of a previously inactive allele, affecting some of the current views on Alu amplification dynamics. These changes in A-tail length could be caused by either strand slippage during replication, unequal recombination, or even possibly gene conversion from an Alu with a long A stretch. Most data on simple-sequence repeats would be consistent with strand slippage being the primary mode of change (Calabrese et al. 2001).

Several potential factors may lead to the overall shortening of A stretches. First, these sequences tend to replicate poorly and therefore either result in selection for shorter alleles or somehow favor shortening by strand slippage. Alternatively, poly(A) stretches have previously been reported to form unusual DNA structures (Cubero et al. 2001; Hizver et al. 2001; McConnell and Beveridge 2001), and this can inhibit progression of the DNA polymerase. Also, Alus in transcribed regions may result in a selection bias if they influenced transcription or transcript stability in some way. One such example was observed in the chloroplasts of *Chlamydomonas* (Lisitsky et al. 2001). However, when using a reporter construct, the incorporation of Alus with short (20–30 bp) versus long (40–58 bp) A-tails in the transcript showed no significant changes in expression of that reporter gene (data not shown). Whatever the mechanistic cause, however, there is no question that A stretches are quite unstable in evolutionary terms and tend to approach the smaller size distribution relatively rapidly. As these elements get older, the A stretches accumulate mutations, resulting in higher complexity within these sequences. These changes may stabilize the sequence as it reduces the simple sequence nature of the region. Alternatively, microsatellites may form in these regions (Economou et al. 1990; Arcot et al. 1995; Jurka and Pethiyagoda 1995), perhaps resulting in different amplification dynamics. One extreme example is the formation of a triplet repeat in the middle A-rich region of an Alu in the frataxin gene, which can lead to massive, disease-causing amplification of the triplet repeat (Montermini et al. 1997).

Mechanistic Model for the Role of A-Tail Length in Alu Retroposition

Our data support the critical importance of the length of the Alu A stretch for retroposition capability. There is essentially no evidence of activity from the massive number of the older Alu elements where the vast majority contain modest-length A stretches. This indicates that there is a critical threshold length of A stretch for retroposition efficiency, but does not provide evidence for whether that retroposition capability continues to increase rapidly with increasing length beyond that threshold. None of the disease inserts shown in Figure 2 had an A-tail shorter than about 40 bp. If the priming for reverse transcription were to occur randomly at different positions on the A stretch in the RNA, we would expect some of the priming events would create shorter A stretches. The com-

Table 4. "A-Tail" Length and Transcriptional Activity of the Different Alu Subfamilies

Alu subfamily	% Alu transcripts[a]	% Total Alus	Transcript enrichment[b]	Long A-Tail Alus (%)[c]	Transcription/ A-tail Factor	Expected (%)[d]	Observed (%)[e]
S + J	66	82	0.80	58 (46)	36.9	4 (23)	0 (0)
Y	33	17	1.96	34 (27)	53.0	5 (33)	5 (31)
Ya5	0.8	0.3	2.64	26 (21)	54.4	5 (34)	6 (38)
Yb8	0.5[f]	0.2	2.50[f]	8 (6)	15.9	2 (10)[f]	5 (31)
Total	100	100	—	126 (100)	160.2	16 (100)	16 (100)

[a] Determined using previous data obtained from the isolation and sequencing of cDNAs derived from primary Alu transcripts (Shaikh et al. 1997).

[b] Transcript enrichment is the increase in transcript proportion relative to copy number, also referred to in the Results section as transcription rate.

[c] Data from Table 2 using the numbers of Alu elements retrieved from the human draft genome sequence with A-tail with ≥50 A.

[d] Expected is obtained using the percentage of the transcription/A-tail factor (the product of the transcript enrichment and percentage of long A-tail members) to estimate the number of Alu elements from each subfamily when there are a total of 16.

[e] Subfamily distribution of the Alu elements observed in 16 disease-causing insertions.

[f] Because of the lack of transcript detection, an estimation was made on the basis of the AluYa5 subfamily copy number.

plete absence of short A-tails in the inserts implies that not only do the source RNA(s) have a long A stretch, but also that priming occurs either at, or very near, the 3′ end of the A stretch. They could be preferentially adjacent to the downstream unique sequences, or at least constrained from priming in the first 40 bases of the A stretch in the RNA. Recently, the cytoplasmic poly(A) binding protein (PABP) has been identified as part of the rodent SINE (BC1) ribonucleoprotein (RNP) complex (Sondekoppa Muddashetty et. al 2002; West et. al 2002) and is also complexed with other SINE RNAs. PABP binds poly(A) stretches in a cooperative manner, in which one molecule of PABP binds ~25 adenosine residues (Smith et al. 1997). Therefore, the longer the A-tail, the more PABP molecules bind, increasing the stability of the protein–RNA interaction. Our A-tail data would be consistent with a requirement for two PABP molecules binding and protecting the first 40–50 bases from the priming event. It is therefore tempting to also consider whether PABP may be an important factor in the efficiency of the SINE retroposition process.

In addition to PABP, SINE RNAs also bind other proteins that bind to the SINE portion of the RNA. Among these are the SRP9/14 proteins that bind the 5′ end of the Alu RNA (Hsu et al. 1995; Chang et al. 1996). SRP9/14 is an important component of the signal recognition particle (SRP), an RNP complex that interacts with the ribosome and targets proteins to the rough endoplasmic reticulum (for reviews, see Walter and Johnson 1994; Bovia and Strub 1996). For Alu elements, these proteins in conjunction with PABP may help target the Alu RNA to the ribosome in close proximity to the L1 RNA, where it may effectively compete for the L1 translation products necessary for the retroposition process. Other SINEs are unlikely to bind SRP9/14 because they are derived from tRNA genes rather than 7SL RNA. However, the proteins that bind to them may serve a similar function.

We propose a model in which the PABP influences SINE retroposition (Fig 6). Many other RNAs in the cell have long A-tails and bind PABP. However, they are almost all mRNA molecules that are capped. The cap may inhibit their retroposition by two possible mechanisms. In one case, the cap targets the mRNA for translation by their own ribosomes, isolating them from the ribosomes that are translating the L1 RNA. The second influence may relate to the other proteins in the cap-binding complex, such as eIF4G and eIF4E. PABP interacts directly with eIF4G, which effectively circularizes the RNA (for reviews, see Gingras et al. 1999). However, the polymerase III-transcribed RNAs are not capped, leaving the A stretch/PABP end free to interact with other molecules, possibly including the cap complex of the L1 RNA (Fig. 6). The proximity of the Alu RNA to the L1 RNA and translation machinery may allow it to efficiently compete for the necessary components required for its retroposition, as proposed previously (Boeke 1997). This might allow the Alu elements to use the *cis* preference of the L1 translation products (Wei et al. 2001) for their own benefit. In this scenario, almost any RNA polymerase III-transcribed RNA could retropose if it acquired an A stretch long enough to bind PABP effectively. In addition, perhaps pol II-transcribed genes with normal polyadenylation that "escape" the cap complex control could also retropose, for example, 5′ truncated retropseudogenes or endogenous viruses. Although this model is highly speculative, it now seems critical to incorporate A-stretch length into considerations of SINE formation and activity.

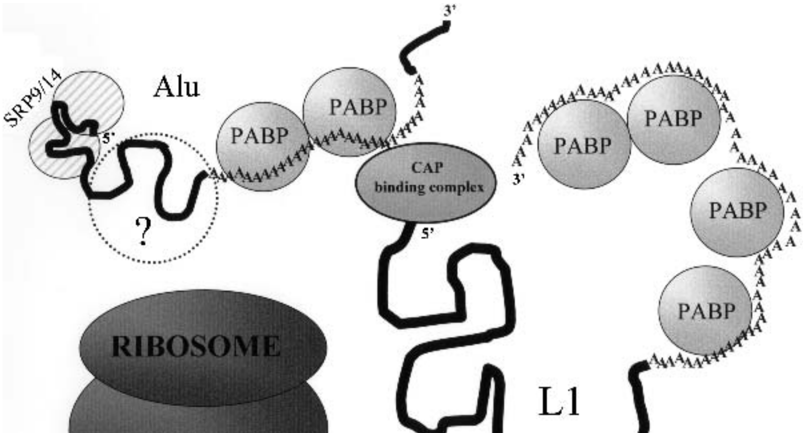

Figure 6 Alu retroposition model. Poly(A) binding protein (PABP) interacts with the A-tails of the Alu and L1 RNA. The SRP9/14 proteins bind the 5′ end of the Alu RNA, which may help target the RNA to the ribosome and translational machinery. The cap complex of L1 RNA interacts with the PABP in the A-tail, circularizing the RNA. In contrast, Alu RNA, being a pol III, has no cap and thus is not circularized, allowing the A-tail/PABP end free to interact with the L1 cap complex. Because the Alu RNA is in close proximity to the L1 RNA when translation of the retroposition factors occurs, it allows the Alu to compete for the required components for its own retroposition. Details of the model are presented in the Discussion.

METHODS

Data Mining and Analysis of the A-Tail Length

The source for members of the Alu Ya5, Ya5a2, and Yb8 subfamilies came from previously generated databases obtained from BLAST searches of the human genome (Roy et al. 2000; Carroll et al. 2001; Roy-Engel et al. 2001). To generate a database of Alu members belonging to older subfamilies, we performed analysis of the elements from three genes, FAA (AC005360), NCF1 (AF184614), and RRM1 (AF107045). They were retrieved from the NCBI GenBank database and analyzed by RepeatMasker2 from the University of Washington Genome Center server (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker). The Alu elements from the Sc, Sx, Sg, Sq, Sp, Jo, and Jb subfamilies present in these genes were randomly selected for the analysis of the A-tail length. The list containing the majority of the disease-causing Alu inserts has been previously published (Deininger and Batzer 1999). The Alu elements found in MLVI-2 (Economou-Pachnis and Tsichlis 1985), PROGINS (Rowe et al. 1995), and ACE (Cambien et al. 1992) were excluded, because they are either not clearly causative of disease or were found by linkage analysis rather than by causing a disease. In addition, the full sequence for the Alu elements found in Btk (Lester et al. 1997) or the Factor IX (Wulff et al. 2000) was not available. An additional three examples of disease-causing Alu inserts were found in the following genes: one each in factor IX (Li et al. 2001), factor VIII (Sukarova et al. 2001), and PBGD (Mustajoki et al. 1999). A total of 16 disease Alu elements were analyzed.

A database of 50 Ta L1 elements was collected from the human genome using BLAST on the nr database, selecting for perfect matches to the 26 bases of the Ta consensus sequence (5′-GATGACACATTAGTGGGTGCAGCGCA-3′) with the expected value of $-e = 1.0e-40$. For recent LINE inserts, the following L1 elements were used: JH-25 (Woods-Samuels et al. 1989), APC (Miki et al. 1992), $L1_{\beta\text{-thal}}$ (Divoky et al. 1996), $L1_{XLCDM}$ (Yoshida et al. 1998), $L1_{RP}$ (Schwahn et al. 1998), two causing disease in mouse $L1_{ORI}$ and $L1_{SPA}$ (Naas et al. 1998), and seven L1 inserts generated from the L1 tissue culture assay (Moran et al. 1996; Wei et al. 2001). All the databases are available on our Web site (http://129.81.225.52).

To search for putative source genes with 100% identity to the Ya5a2 Alu element that inserted into the *NF1* gene (Wallace et al. 1991), we searched the GenBank human genome draft database with the BLAST program, as previously described (Roy et al. 2000). To isolate Alus with long A-tails, we performed BLAST searches on the GenBank draft human database with low complexity filter disabled and the maximum number of results returned (-v -b flags) set at 5000 to retrieve more hits and an expected value of $e-18$; otherwise standard settings were used. The oligos used were TCTC(A × 40 or A × 50). Elements containing an unbroken stretch of 40 or 50 A's were extracted from GenBank accessions and analyzed by RepeatMasker.

Details of the parameters used to determine the length of the A-tails from the Alu and LINE elements are shown in Figure 1.

PCR Amplification

PCR amplification of the different loci containing the candidate Alu elements was performed in 20-µl reactions using an MJ Research PTC 200 thermal cycler with the following conditions: 1X Promega buffer, 1.5 mM $MgCl_2$, 200 µM deoxyribonucleoside triphosphates, 0.25 µM primers, and 1.5 U *Taq* polymerase (Promega) at 94°C for 2 min; 94°C for 20 sec, X°C for 30 sec (X is the annealing temperature indicated as follows for each primer pair), and 72°C for 1 min, for 30 cycles; 72°C for 20 min. The final PCR products were cloned into pGEMTeasy Vector System I (Promega). The following primers were used to amplify the *NF1*-selected loci [accession number]: Ya5a2AD7 [AL138681]: 5′-TGGACACGCATGAAAGAAACCCTACC-3′ and 5′-TGGATTAATTCATTCAACTTCACTACAA-3′at 58°C; Ya5a2AD8 [AL162713]: 5′-TTAGCTGCCATCAACCACCCTTATCA-3′ and 5′-CATGTTTGCTGTCTCCTACTGTCTTCTGC-3′ at 53°C; Ya5AD11 [AC090071]: 5′-TTCTGCTCCAAATAATAACTACCT-3′ and 5′-AGAATTTAAGACCCTATGGCACCTC-3′ at 50°C; Ya5AD12 [AC004057]: 5′-CAGGCCAGGAGTTTTGAAATTTTATC-3′ and 5′-CATGCCTCTTCTCACCTGTTTCAACCA-3′ at 58°C; Ya5a2AD4 [AC023932]: 5′-TGACGGGAGAAGTTAACAGT-3′ and 5′-TGCACCTGACAGAAAATCT-3′ at 55°C; Ya5a2AD5 [AC026839]: 5′-GCATTAAGAATGTGGACCAT-3′ and 5′- TGTAGTTGGAAGCCCTTAAT-3′ at 60°C; Ya5a2AD6 [AL355580]: 5′-ATTATTCCTAGTGAGGGGAATGAA-3′ and 5′-ACTTCCTATGATTCCATCTCCAA-3′ at 60°C; and Ya5a2AD9 [AC079456]: 5′-TTTATTGCCAGAAGCTTTCGT-3′ and 5′-AAGCGCTTCCTTCATAAATCA-3′ at 60°C. For Ya5NBC243, Ya5NBC235a, Ya5NBC239, Ya5NBC241, Ya5NBC220c, and Ya5NBC208, primers and conditions have been previously published (Roy et al. 2000). DNA sequences (~600 individual sequences) from the Alu elements listed earlier were determined in a diverse panel of human genomes. These DNA sequences were derived from Alu-containing PCR products cloned using TOPO TA cloning kits (Invitrogen) and sequenced using an ABI 3100 automated DNA sequencer. These sequences were assigned GenBank accession numbers AF504933–AF505511.

ACKNOWLEDGMENTS

A.M.R. was supported by a Brown Foundation fellowship from the Tulane Cancer Center. Parts of these data were used as the foundation for the research project presented by O.O.O. in the International Science Fair in Kentucky, May 2002. This research was supported by the Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000–05)-05 (M.A.B. and P.L.D.), (2000–05)-01 (M.A.B. and P.L.D.), and (2001–06)-02 (M.A.B.), National Institutes of Health RO1 GM45668 (P.L.D.) and GM59290 (M.A.B.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Abdelhak, S., Kalatzis, V., Heilig, R., Compain, S., Samson, D., Vincent, C., Levi-Acobas, F., Cruaud, C., Le Merrer, M., Mathieu, M., et al. 1997. Clustering of mutations responsible for branchio-oto-renal (BOR) syndrome in the eyes absent homologous region (eyaHR) of EYA1. *Hum. Mol. Genet.* **6:** 2247–2255.

Alemán, C., Roy-Engel, A.M., Shaikh, T.H., and Deininger, P.L. 2000. Cis-acting influences on Alu RNA levels. *Nucleic Acids Res.* **28:** 4755–4761.

Arcot, S.S., Wang, Z., Weber, J.L., Deininger, P.L., and Batzer, M.A. 1995. Alu repeats: A source for the genesis of primate microsatellites. *Genomics* **29:** 136–144.

Batzer, M.A. and Deininger, P.L. 1991. A human-specific subfamily of Alu sequences. *Genomics* **9:** 481–487.

———. 2002. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3:** 370–379.

Batzer, M.A., Kilroy, G.E., Richard, P.E., Shaikh, T.H., Desselle, T.D., Hoppens, C.L., and Deininger, P.L. 1990. Structure and variability of recently inserted Alu family members. *Nucleic Acids Res.* **18:** 6793–6798.

Batzer, M.A., Bazan, H.A., Kim, J., Morrow, S.L., Shaikh, T.H., Arcot, S.S., and Deininger, P.L. 1992. Large-scale subcloning of bacteriophage λ ZAP clones. *Biotechniques* **12:** 370–371.

Boeke, J.D. 1997. LINEs and Alus—The polyA connection. *Nat. Genet.* **16:** 6–7.

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.

Borodulina, O.R. and Kramerov, D.A. 2001. Short interspersed elements (SINEs) from insectivores. Two classes of mammalian SINEs distinguished by A-rich tail structure. *Mamm. Genome* **12:** 779–786.

Bovia, F. and Strub, K. 1996. The signal recognition particle and related small cytoplasmic ribonucleoprotein particles. *J. Cell Sci.* (Pt 11) **109:** 2601–2608.

Calabrese, P.P., Durrett, R.T., and Aquadro, C.F. 2001. Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. *Genetics* **159:** 839–852.

Cambien, F., Poirier, O., Lecerf, L., Evans, A., Cambou, J.P., Arveiler, D., Luc, G., Bard, J.M., Bara, L., Ricard, S., et al. 1992. Deletion polymorphism in the gene for angiotensin-converting enzyme is a potent risk factor for myocardial infarction. *Nature* **359:** 641–644.

Carroll, M.L., Roy-Engel, A.M., Nguyen, S.V., Salem, A.H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M., et al. 2001. Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J. Mol. Biol.* **311:** 17–40.

Chang, D.Y., Hsu, K., and Maraia, R.J. 1996. Monomeric scAlu and nascent dimeric Alu RNAs induced by adenovirus are assembled into SRP9/14-containing RNPs in HeLa cells. *Nucleic Acids Res.* **24:** 4165–4170.

Chen, Y., Sinha, K., Perumal, K., Gu, J., and Reddy, R. 1998. Accurate 3′ end processing and adenylation of human signal recognition particle RNA and alu RNA in vitro. *J. Biol. Chem.* **273:** 35023–35031.

Chu, W.M., Liu, W.M., and Schmid, C.W. 1995. RNA polymerase III promoter and terminator elements affect Alu RNA expression. *Nucleic Acids Res.* **23:** 1750–1757.

Cubero, E., Luque, F.J., and Orozco, M. 2001. Theoretical studies of d(A:T)-based parallel-stranded DNA duplexes. *J. Am. Chem. Soc.* **123:** 12018–12025.

Deininger, P.L. and Batzer, M.A. 1995. SINE master genes and population biology. In *The impact of short, interspersed elements (SINEs) on the host genome* (ed. R. Maraia), pp. 43–60. R.G. Landes, Georgetown, TX.

———. 1999. Alu repeats and human disease. *Mol. Genet. Metab.* **67:** 183–193.

Deininger, P.L. and Daniels, G. 1986. The recent evolution of mammalian repetitive DNA elements. *Trends Genet.* **2:** 76–80.

Deininger, P.L., Jolly, D., Rubin, C., Friedmann, T., and Schmid, C.W. 1981. Base sequence studies of 300 nucleotide renatured repeated human DNA clones. *J. Mol. Biol.* **151:** 17–33.

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

Deininger, P.L., Tiedge, H., Kim, J., and Brosius, J. 1996. Evolution, expression, and possible function of a master gene for amplification of an interspersed repeated DNA family in rodents. In *Progress in nucleic acid research and molecular biology* (eds. W.E. Cohn and K. Moldave), pp. 67–88. Academic Press, San Diego.

Divoky, V., Indrak, K., Mrug, M., Brabec, V., Huisman, T.H.J., and Prchal, J.T. 1996. A novel mechanism of β-thalassemia. The insertion of L1 retrotransposable element in β globin IVSII. *Blood* **88:** 148a.

Economou, E.P., Bergen, A.W., Warren, A.C., and Antonarakis, S.E. 1990. The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. *Proc. Natl. Acad. Sci.* **87:** 2951–2954.

Economou-Pachnis, A. and Tsichlis, P.N. 1985. Insertion of an Alu SINE in the human homologue of the Mlvi-2 locus. *Nucleic Acids Res.* **13:** 8379–8387.

Englander, E.W., Wolffe, A.P., and Howard, B.H. 1993. Nucleosome interactions with a human Alu element. Transcriptional repression and effects of template methylation. *J. Biol. Chem.* **268:** 19565–19573.

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

Gingras, A.C., Raught, B., and Sonenberg, N. 1999. eIF4 initiation factors: Effectors of mRNA recruitment to ribosomes and regulators of translation. *Annu. Rev. Biochem.* **68:** 913–963.

Goncalves, I., Duret, L., and Mouchiroud, D. 2000. Nature and structure of human genes that generate retropseudogenes. *Genome Res.* **10:** 672–678.

Harrison, P.M., Hegyi, H., Balasubramanian, S., Luscombe, N.M., Bertone, P., Echols, N., Johnson, T., and Gerstein, M. 2002. Molecular fossils in the human genome: Identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* **12:** 272–280.

Hizver, J., Rozenberg, H., Frolow, F., Rabinovich, D., and Shakked, Z. 2001. DNA bending by an adenine–thymine tract and its role in gene regulation. *Proc. Natl. Acad. Sci.* **98:** 8490–8495.

Hsu, K., Chang, D.Y., and Maraia, R.J. (1995). Human signal recognition particle (SRP) Alu-associated protein also binds Alu interspersed repeat sequence RNAs. Characterization of human SRP9. *J. Biol. Chem.* **270:** 10179–10186.

Jagadeeswaran, P., Forget, B.G., and Weissman, S.M. 1981. Short, interspersed repetitive DNA elements in eukaryotes: Transposable DNA elements generated by reverse transcription of RNA pol III transcripts? *Cell* **26:** 141–142.

Jurka, J. and Pethiyagoda, C. 1995. Simple repetitive DNA sequences from primates: Compilation and analysis. *J. Mol. Evol.* **40:** 120–126.

Kass, D.H., Kim, J., and Deininger, P.L. 1996. Sporadic amplification of ID elements in rodents. *J. Mol. Evol.* **42:** 7–14.

Kim, J., Martignetti, J.A., Shen, M.R., Brosius, J., and Deininger, P.L. 1994. The rodent BC1 RNA gene is a master gene for ID element amplification. *Proc. Natl. Acad. Sci.* **91:** 3607–3611.

Labuda, D. and Zietkiewicz, E. 1994. Evolution of secondary structure in the family of 7SL-like RNAs. *J. Mol. Evol.* **39:** 506–518.

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.

Lester, T., McMahon, C., VanRegemorter, N., Jones, A., and Genet, S. 1997. X-linked immunodeficiency caused by insertion of Alu repeat sequences. *J. Med. Gen. Suppl.* **34:** S81.

Li, X., Scaringe, W.A., Hill, K.A., Roberts, S., Mengos, A., Careri, D., Pinto, M.T., Kasper, C.K., and Sommer, S.S. 2001. Frequency of recent retrotransposition events in the human factor IX gene. *Hum. Mutat.* **17:** 511–519.

Lisitsky, I., Rott, R., and Schuster, G. 2001. Insertion of polydeoxyadenosine-rich sequences into an intergenic region increases transcription in *Chlamydomonas reinhardtii* chloroplasts. *Planta* **212:** 851–857.

Liu, W.M. and Schmid, C.W. 1993. Proposed roles for DNA methylation in Alu transcriptional repression and mutational inactivation. *Nucleic Acids Res.* **21:** 1351–1359.

Matera, A.G., Hellmann, U., and Schmid, C.W. 1990. A transpositionally and transcriptionally competent Alu subfamily. *Mol. Cell. Biol.* **10:** 5424–5432.

Mathias, S.L., Scott, A.F., Kazazian Jr., H.H., Boeke, J.D., and Gabriel, A. 1991. Reverse transcriptase encoded by a human transposable element. *Science* **254:** 1808–1810.

McConnell, K.J. and Beveridge, D.L. 2001. Molecular dynamics simulations of B ′-DNA: Sequence effects on A-tract-induced bending and flexibility. *J. Mol. Biol.* **314:** 23–40.

Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K.W., Vogelstein, B., and Nakamura, Y. 1992. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res.* **52:** 643–645.

Montermini, L., Andermann, E., Labuda, M., Richter, A., Pandolfo, M., Cavalcanti, F., Pianese, L., Iodice, L, Farina, G., Monticelli, A., et al. 1997. The Friedreich ataxia GAA triplet repeat: Premutation and normal alleles. *Hum. Mol. Genet.* **6:** 1261–1266.

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

Muddashetty, R.S., Khanam, T., Kondrashov, A., Bundman, M., Iacoangeli, A., Kremershothen, J., Cuning, K., Barnekow, A., Huttenhoffer, A., Tiedge, H., et al. 2002. Poly(A) binding protein is associated with neuronal BC1 and BC200 ribonucleoprotein particles. *J. Mol. Biol.* (in press).

Mustajoki, S., Ahola, H., Mustajoki, P., and Kauppinen, R. 1999. Insertion of Alu element responsible for acute intermittent porphyria. *Hum. Mutat.* **13:** 431–438.

Naas, T.P., DeBerardinis, R.J., Moran, J.V., Ostertag, E.M., Kingsmore, S.F., Seldin, M.F., Hayashizaki, Y., Martin, S.L., and Kazazian, H.H. 1998. An actively retrotransposing, novel subfamily of mouse L1 elements. *EMBO J.* **17:** 590–597.

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.

Prak, E.T. and Kazazian Jr., H.H. 2000. Mobile elements and the human genome. *Nat. Rev. Genet.* **1:** 134–144.

Rogers, J.H. and Willison, K.R. 1983. A major rearrangement in the H-2 complex of mouse t haplotypes. *Nature* **304:** 549–552.

Rowe, S.M., Coughlan, S.J., McKenna, N.J., Garrett, E., Kieback, D.G., Carney, D.N., and Headon, D.R. 1995. Ovarian carcinoma-associated TaqI restriction fragment length polymorphism in intron G of the progesterone receptor gene is due to an Alu sequence insertion. *Cancer Res.* **55:** 2743–2745.

Roy, A.M., Carroll, M.L., Nguyen, S.V., Salem, A.H., Oldridge, M., Wilkie, A.O., Batzer, M.A., and Deininger, P.L. 2000. Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res.* **10:** 1485–1495.

Roy-Engel, A.M., Carroll, M.L., Vogel, E., Garber, R.K., Nguyen, S.V., Salem, A.H., Batzer, M.A., and Deininger, P.L. 2001. Alu insertion polymorphisms for the study of human genomic diversity. *Genetics* **159:** 279–290.

Sapienza, C. and St Jacques, B. 1986. ‘Brain-specific’ transcription and evolution of the identifier sequence. *Nature* **319:** 418–420.

Schmid, C.W. and Maraia, R. 1992. Transcriptional regulation and transpositional selection of active SINE sequences. *Curr. Opin. Genet. Dev.* **2:** 874–882.

Schwahn, U., Lenzner, S., Dong, J., Feil, S., Hinzmann, B., van Duijnhoven, G., Kirschner, R., Hemberger, M., Bergen, A.A., Rosenberg, T., et al. 1998. Positional cloning of the gene for X-linked retinitis pigmentosa 2. *Nat. Genet.* **19:** 327–332.

Shaikh, T.H. and Deininger, P.L. 1996. The role and amplification of the HS Alu subfamily founder gene. *J. Mol. Evol.* **42:** 15–21.

Shaikh, T.H., Roy, A.M., Kim, J., Batzer, M.A., and Deininger, P.L. 1997. cDNAs derived from primary and small cytoplasmic Alu (scAlu) transcripts. *J. Mol. Biol.* **271:** 222–234.

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

Shen, M.R., Batzer, M.A., and Deininger, P.L. 1991. Evolution of the master Alu gene(s). *J. Mol. Evol.* **33:** 311–320.

Smith, B.L., Gallie, D.R., Le, H., and Hansma, P.K. 1997.

Visualization of poly(A)-binding protein complex formation with poly(A) RNA using atomic force microscopy. *J. Struct. Biol.* **119:** 109–117.

Sukarova, E., Dimovski, A.J., Tchacarova, P., Petkov, G.H., and Efremov, G.D. 2001. An Alu insert as the cause of a severe form of hemophilia A. *Acta Haematol.* **106:** 126–129.

Wallace, M.R., Andersen, L.B., Saulino, A.M., Gregory, P.E., Glover, T.W., and Collins, F.S. 1991. A de novo Alu insertion results in neurofibromatosis type 1. *Nature* **353:** 864–866.

Walter, P. and Johnson, A.E. 1994. Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. *Annu. Rev. Cell Biol.* **10:** 87–119.

Wang, T., Lerer, I., Gueta, Z., Sagi, M., Kadouri, L., Peretz, T., and Abeliovich, D. 2001. A deletion/insertion mutation in the *BRCA2* gene in a breast cancer family: A possible role of the Alu-polyA tail in the evolution of the deletion. *Genes Chromosomes Cancer* **31:** 91–95.

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 versus trans complementation. *Mol. Cell Biol.* **21:** 1429–1439.

Weiner, A., Deininger, P., and Efstradiatis, A. 1986. The reverse flow of genetic information: Pseudogenes and transposable elements derived from nonviral cellular RNA. *Annu. Rev. Biochem.* **55:** 631–661.

West, N.C., Roy-Engel, A.M., Imataka, H., Sonenberg, N., and Deininger, P.L. 2002. Shared components of SINE RNPs. *J. Mol. Biol.* (in press).

Woods-Samuels, P., Wong, C., Mathias, S.L., Scott, A.F., Kazazian Jr., H.H., and Antonarakis, S.E. 1989. Characterization of a nondeleterious L1 insertion in an intron of the human factor VIII gene and further evidence of open reading frames in functional L1 elements. *Genomics* **4:** 290–296.

Wulff, K., Gazda, H., Schroder, W., Robicka-Milewska, R., and Herrmann, F.H. 2000. Identification of a novel large F9 gene mutation—An insertion of an Alu repeated DNA element in exon e of the factor 9 gene. *Hum. Mutat.* **15:** 299.

Yoshida, K., Nakamura, A., Yazaki, M., Ikeda, S., and Takeda, S. 1998. Insertional mutation by transposable element, L1, in the DMD gene results in X-linked dilated cardiomyopathy. *Hum. Mol. Genet.* **7:** 1129–1132.

WEB SITE REFERENCES

http://ftp.genome.washington.edu/cgi-bin/RepeatMasker; RepeatMasker.

http://129.81.225.52; Deininger Lab Web page.

Received June 20, 2002; accepted in revised form July 3, 2002.