Using Alu Elements as Polyadenylation Sites: A Case of Retroposon Exaptation

Chongjian Chen,*† Takeshi Ara,‡ and Daniel Gautheret*

*Institut de Génétique et Microbiologie, Université Paris-Sud 11, CNRS UMR 8621, Orsay, France; †Key Laboratory of Gene Engineering of the Ministry of Education, State Key Laboratory for Biocontrol, Zhongshan University, Guangzhou, People’s republic of China; and ‡Kazusa DNA Research Institute, Kisarazu, Chiba, Japan

Of the 1.1 million Alu retroposons in the human genome, about 10,000 are inserted in the 3’ untranslated regions (UTR) of protein-coding genes and 1% of these (107 events) are active as polyadenylation sites (PASs). Strikingly, although Alu’s in 3’ UTR are indifferently inserted in the forward or reverse direction, 99% of polyadenylation-active Alu sequences are forward oriented. Consensus Alu+ sequences contain sites that can give rise to polyadenylation signals and enhancers through a few point mutations. We found that the strand bias of polyadenylation-active Alu’s reflects a radical difference in the fitness of sense and antisense Alu’s toward cleavage/polyadenylation activity. In contrast to previous beliefs, Alu inserts do not necessarily represent weak or cryptic PASs; instead, they often constitute the major or the unique PAS in a gene, adding to the growing list of Alu exaptations. Finally, some Alu-borne PASs are intronic and produce truncated transcripts that may impact gene function and/or contribute to gene remodeling.

Introduction

With over 1.1 million copies, the 300-nt Alu sequence is the most abundant mobile element in the human genome (Lander et al. 2001). It is a member of the SINE family of retroposons, composed of two similar monomers separated by an A-rich linker. The expansion of Alu elements occurred through successive waves of retrotranspositions that started at least 65 Ma in the early primate lineage and produced distinguishable subfamilies (Kapitonov and Jurka 1996). Once classic examples of “junk DNA,” Alu elements recently gained a more prominent status as several studies established their involvement in cell functions and gene remodeling. Alu elements harbor AU-rich motifs acting upon messenger RNA (mRNA) stability (An et al. 2004), they contain transcription factor–binding sites that regulate gene expression (Polak and Domany 2006), and a significant fraction (5%) of Alu elements form alternative exons that impact the expression of a large number of human genes (Sorek et al. 2002; Lev-Maor et al. 2003; Kreahling and Graveley 2004). Such a phenomenon in which an existing genome sequence or any other trait is co-opted to carry out a novel function has been termed exaptation by Gould and Vrba (1982). It is believed that many more mobile elements in the human genome have been exapted (Bejerano et al. 2006; Lowe et al. 2007), which suggests that mobile elements are an important source of functional diversity.

A possible exaptation of Alu element that has not been explored is a polyadenylation site (PAS). Eukaryotic transcripts are cleaved and polyadenylated at specific locations in the 3’ untranslated region (UTR). Mammalian PASs contain an AATAAA or single base variant signal located about 17 nt upstream of the cleavage site (Graber et al. 1999; Beaudong et al. 2000), flanked by GT-rich or T-rich upstream or, more frequently, downstream enhancer elements (Proudfoot 1991; Colgan and Manley 1997). Several thousand Alu sequences are inserted in the 3’ UTRs of human genes. Each Alu sequence contains three regions that either contain a known polyadenylation signal or can become one through a single base mutation. This raises the question of the effect of Alu insertion on polyadenylation. Could Alu sequences have contributed to PAS destruction/creation during their expansion in primate genomes and how did this affect 3’ UTR and gene structures? A previous analysis of PASs in Alu elements carried out on a small sample of 34 human genes (Roy-Engel et al. 2005) found that most Alu elements within genes were inserted in the reverse orientation. As Alu contains poly(A) signals on its forward strand and not on its reverse strand, this suggested a negative selection against the insertion of possibly deleterious PASs. However, in several in vivo assays, these authors could not obtain efficient polyadenylation at potential poly(A) site locations in the Alu sequence, even after canonical AATAAA signals were inserted to replace the variant signals. This led to the conclusion that Alu sequences lacked the proper environment (such as downstream GT-rich elements) to enable efficient polyadenylation and would most likely constitute alternative, cryptic PASs.

In this article, we examine the effect of Alu inserts on polyadenylation at the full genome scale. We first observe based on cdNA/expressed sequence tag (EST) evidence that Alu sequences can be efficiently cleaved and polyadenylated at several locations, almost exclusively on the forward strand. Efficient polyadenylation appears to entail point mutations in the polyadenylation-like signals and the acquisition of an enhancer (GT-rich) downstream region, either through a few mutations or through Alu insertion upstream of a GT-rich region. Using human–mouse gene comparison, we then identified predominant mechanisms for Alu-mediated gain and loss of PASs.

Methods

Identification of TE-PASs

We first identified putative poly(A) sites based on EST/cDNA mapping to genome sequences as described in a previous article (Lopez et al. 2006). In brief, we required that poly(A) sites were supported by at least two ESTs/cDNAs located within 30 nt downstream of a potential signal. To account for internally primed ESTs/cDNAs, we disregarded any cleavage site flanked by an A-rich region (at least 9 As out of 10 nt) in the 50-nt downstream genomic sequence. Note that, although predicted PASs in Alu elements correspond to A-rich regions, these PASs...
cannot be attributed to internal priming because, in each case, polyadenylation signals lie within the A-rich region and EST/cDNA 3’ ends extend beyond this region.

EST sequences were obtained from dbEST v. 05/25/2006, and full-length cDNA sequences were obtained from H-InvDB 3.0 (Imanishi et al. 2004) and FANTOM 3.0 (Carninci et al. 2003). We identified 42,501 PASs associated with 18,005 Ensemble genes (release 38.36) in human (NCBIt36.apr) and 37,892 PASs associated with 18,182 Ensemble genes (release 39.36) in mouse (NCBIt36.jun).

For the present study, we extracted the ±300 nt region around each PAS and used the RepeatMasker program (Smit et al. 1996–2008) to detect repeats in this region. To avoid ambiguous matches of EST/cDNA to human repeats, additional criterions were adopted: We required at least one full-length cDNA supporting each site, and we discarded EST/cDNA producing alignments shorter than 100 nt and with less than 99% identity. For Alu-PASs, manual inspection was further performed to filter out PASs only matched by ESTs/cDNAs from the same clone or ESTs/cDNAs that were totally embedded within Alu elements.

Localisation of Active Polyadenylation Signals in the Alu Consensus

Consensus sequences of Alu subfamilies were downloaded from RepBase for RepeatMasker (http://www.girinst.org). Each PAS-containing genomic Alu was aligned to the AluJo consensus using the Muscle program (Edgar 2004), and the active polyadenylation signal was mapped on the consensus based on its predicted position in the genomic Alu.

Measure of Polyadenylation Efficiency and Signal Strength of Alu-PASs

We defined a major PAS as a PAS with a number of supporting ESTs/cDNAs larger than or equal to the number of supporting ESTs/cDNAs at other PASs in the same gene. All other PASs were considered as minor. We assigned strengths to various polyadenylation signals based on their frequencies in the human genome (Beaudoing et al. 2000), that is, AAUAAA > AUUAAA > 12 other variant signals. In subsequent analyses, each PAS was assigned the strongest signal when several known hexamers were present.

G + T Enrichment of Downstream Enhancer Regions

First, we divided PAS-containing Alu’s into two major groups: mid-PAS-Alu’s and tail-PAS-Alu’s according to the location of their PAS in the middle or tail region. We obtained a human 3’ UTR data set from UTRdb (Mignone et al. 2005) and scanned 3’ UTRs for Alu elements (UTR-Alu’s). We searched UTR-Alu’s for putative polyadenylation signals (with no evidence for PASs) and built two control cohorts containing signals in either the middle or tail position (mid-UTR-Alu’s: 3,264 sequences; tail-UTR-Alu’s: 1,255 sequences). We used these control data sets to assess the G + T enrichment in the 100-nt downstream regions of Alu-PAS. First, we compared the downstream regions of mid-PAS-Alu’s with controls mid-UTR-Alu’s; second, we compared the downstream regions of tail-PAS-Alu’s with controls tail-UTR-Alu’s. G + T contents were calculated using a 20-nt window sliding by 1-nt steps. Observed and expected G + T profiles (fig. 4, top panels) are the average G + T frequencies obtained in the true PAS (PAS-Alu) and in the control PAS (UTR-Alu). The difference between observed and expected G + T frequencies was computed using a $\chi^2$ test (R program) and reported as a P value (fig. 4, bottom panels).

Comparative Analysis of Alu-PASs

We obtained the pairwise alignment of human and mouse genomes from the UCSC Genome Bioinformatics site (Karolchik et al. 2003). We mapped the ±100-nt region around each PAS-containing Alu in human to the alignment and retrieved the corresponding mouse genome positions. When the Alu-PAS region was not aligned to a mouse sequence in the UCSC database, we retrieved the flanking exons on both sides of the human PAS and manually sought the corresponding region in mouse based on the presence of syntenic flanking exons. After identifying the orthologous mouse region, we retrieved predicted mouse PAS in this region from the AltPAS database (Lopez et al. 2006). Data in table 4 were derived by comparing the number of PASs in the orthologous human and mouse regions. When the human PAS was in an intron and the orthologous region did not contain any PAS in mouse, we considered the Alu-PAS as a case of premature termination.

Results

Polyadenylation-Active Alu Sequences in 3’ UTRs

A total of ~4.3 million transposable elements (TEs) from four major classes (SINEs, LINEs, LTR elements, and DNA transposons) cover about 43% of the human genome (Li et al. 2001). More than 25,000 TEs are located in the 3’ UTR of genes. The “UTR-TE” column in figure 1 presents the proportions of TEs from eight major families inserted in 3’ UTR regions, in either the sense or antisense orientation relative to the gene. In order to assess active PASs in TEs, we used the AltPAS database that compiles EST/UTR-based PAS predictions in the complete human genome (Lopez et al. 2006). To keep false positives to a minimum, we only retained poly(A) sites supported by at least two cDNAs or sequence tags including one full-length cDNA. The “PAS-TE” column in figure 1 represents polyadenylation-active TEs. Overall, only 6% of TEs (~1,500 TEs) gave rise to active PASs. Among PAS-TEs, we observed strong biases in the orientation of TEs from the L1, L2, LTR, Alu, and MIR families, reflecting the different polyadenylation potentials of each transposon’s forward and reverse strands.

Transposons of the Alu family are distinctive on two counts. First, they seem to be the least apt to form active PASs, based on the ratio of active sites versus total TEs. Polyadenylation-active Alu’s only represent 1% of 3’ UTR-Alu’s, whereas active TEs from other families represent 4–14% of 3’ UTR-TEs. Therefore, Alu sequences
appear not to provide a favorable environment for polyadenylation compared with other TEs. The most favorable TE is the L1 sequence (14% of 3’ UTR L1s are polyadenylation active), especially on its negative strand, which is known to contain an efficient PAS (Han et al. 2004). Another specificity of Alu sequences is the intensity of the orientation bias ($P = 3.9 \times 10^{-21}$), which is the highest overall. This suggests that even though Alu’s are generally not prone to form PASs, there is a considerable difference between Alu strands with respect to polyadenylation abilities.

We refined our list of predicted polyadenylation-active Alu’s by visual inspection of each cDNA/genome alignment. We discarded predictions that rested on short cDNA/EST sequences in order to eliminate any potential independent Alu transcript. Also, some of the predicted sites were not in the expected orientation relative to the flanking gene. Overall, 24 predicted sites were discarded (18% false positives), which suggests a generally low false-positive rate for the other PAS-TEs in figure 1. We finally retained 107 bona fide Alu-borne PASs, including 106 on the forward strand and one on the reverse strand. We will thereafter refer to these PASs as “Alu-PASs.”

Polyadenylation-Active Alu Sequences Are Cleaved at Two Specific Locations

Known mammalian polyadenylation signals comprise the major hexamer AATAAA plus at least 12 single-nucleotide variants (Lopez et al. 2006). Alu sequences contain three sites that potentially harbor such hexamers, either directly or through a single base mutation (fig. 2). The first site is located around position 96 and may contain an

---

**Fig. 1.**—Total TEs in human 3’ UTRs (UTR-TE) and polyadenylation-active TEs (PAS-TE). Pie charts represent the fraction of sense and antisense TEs, relative to host gene orientation. $P$ values indicate the significance of the orientation bias in polyadenylation-active TEs.

**Fig. 2.**—Consensus sequences of major Alu subfamilies with potential poly(A) signal regions boxed. Potential signals have at most 1-nt difference from a known poly(A) signal hexamer. The 3’ poly(A) tails were shortened for clarity.
AACATA sequence (AluJ subfamily) that can give rise to known signals AACAAA or AATATA; the second is located in the central A-rich region, which contains two known signals (ACTAAA and AATACA) and AAAAAA hexamers that can mutate to AATAAA; and the third one is located in the tail region, which contains AAAAAA hexamers. We will refer to these sites as Loc96-PAS, mid-PAS, and tail-PAS, respectively. In figure 3, we show the distribution of all active PASs along the consensus Alu sequence. As expected, all active sites (with one exception) occur at signal-containing locations. The locations most often used for polyadenylation are the tail region (63 events), followed by the middle region (42 events). We did not observe any confirmed PAS at Loc96, but a few isolated ESTs/cDNAs indicate that this position could form a cryptic site. Finally, one PAS is present at position 199 on the forward strand and one is on the reverse strand in the middle A-rich region (hence T rich on the complementary strand).

Alu-borne PASs are not cryptic sites. Table 1 presents the number of minor, major, or unique PASs observed at the main Alu locations. Of the Alu-PASs, 56% are alternative, minor sites. This does not differ from the overall ratio of minor PASs in the human genome: 57% of PASs in the Alt-PAS database (Lopez et al. 2006) are minor sites.

Emergence of Polyadenylation Activity through Mutations in Signal and Enhancer Region

Most polyadenylation-active Alu sequences use an AATAAA signal. Upstream of mid-PASs, predominant signals are AATAAA in AluJ and ACTAAAAATACA in AluS (supplementary fig. S1A, Supplementary Material online). The former signal likely results from a single-nucleotide mutation, whereas the latter is the direct AluS consensus that contains the two known signals ACTAAA and AATACA. Upstream of tail-PAS, predominant signals are AATAAA in all Alu families (supplementary fig. S1B, Supplementary Material online), implying an A → T mutation in the poly(A) sequence. The nature of the signal appears to be determinant in the selection of the polyadenylation-active site. Table 2 shows that the active site is most often the one with the strongest polyadenylation hexamer (i.e., canonical AATTTAAA vs. variant signal).

Efficient polyadenylation is also favored by mutations toward higher G + T contents in the downstream regions or by Alu insertion in G + T-rich environments. Downstream GT-rich or T-rich regions are known to act as polyadenylation enhancers (Proudfoot 1991; Colgan and Manley 1997). We analyzed the downstream regions of Alu-PAS for such sequence biases. G + T frequencies were consistently higher at Alu-PAS than at non-PAS control Alu sequences (fig. 4). This enrichment was mild in absolute terms (less than 5%) but was particularly significant around 40 nt past the polyadenylation signal, at both mid-PAS and tail-PAS (fig. 4). After inspection of aligned Alu elements, we could not identify specific mutations responsible for the G + T enrichment. However, the +40 region corresponds precisely to the average position of downstream enhancers.

Table 1

<table>
<thead>
<tr>
<th>Number of Major and Minor Polyadenylation Sites Observed at Two Alu Locations</th>
<th>Mid-PAS</th>
<th>Tail-PAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major or unique</td>
<td>19</td>
<td>27</td>
</tr>
<tr>
<td>Minor</td>
<td>23</td>
<td>36</td>
</tr>
</tbody>
</table>

Table 2

| Number of Alu-PASs as a Function of the Relative Strength of Polyadenylation Signals within the Same Alu Sequence (strength: AATTTAAA > ATTAAA > other variant signals) |
|---|---|---|
| Actual PAS Location | Mid Stronger Occurrence (%) | Equal Strength Occurrence (%) | Tail Stronger Occurrence (%) |
| Mid-PAS | 26 (61.9%) | 12 (28.6%) | 4 (9.5%) |
| Tail-PAS | — | 6 (12.0%) | 44 (88.0%) |
Dynamics of PAS Formation through \textit{Alu} Insertion

Ancient \textit{Alu} elements are more frequently utilized as PASs than more recent elements. \textit{Alu} expansion occurred through successive waves of retrotransposition that started from the \textit{Alu} monomers/7SL in early mammals and later gave rise to the various primate \textit{Alu} families \textit{AluJ}, \textit{AluS}, and \textit{AluY} (Batzer and Deininger 2002). Polyadenylation-active \textit{Alu}’s are found across all families (table 3); however, the most ancient \textit{Alu}’s seem to have been more successful at forming active PASs. \textit{Alu}-PASs are relatively more frequent in \textit{AluJ} than in \textit{AluS} or \textit{AluY} (table 3, \(P = 3.2 \times 10^{-5}\)). This may reflect a higher propensity of \textit{AluJ} to form active PASs or, more likely, that older \textit{Alu} families have had more time to evolve the features of active PASs.

To explore the mechanisms of \textit{Alu}-PAS fixation, we compared polyadenylation signals in human and mouse orthologs. As \textit{Alu} sequences arose after the human/mouse divergence, we expected this comparison to reveal how \textit{Alu} elements may have replaced or supplemented preexisting PASs in the common ancestor. We identified 82 pairs of human–mouse orthologs where we could compare PAS occurrences. Of 69 \textit{Alu}-PASs located in 3′ UTRs, 38 (55%) caused additional PASs in human, whereas 31 (45%) did not change the number of sites in the \textit{Alu} insertion region (table 4). The majority of “additional” PASs (87%) are minor sites, consistent with \textit{Alu} insertion providing an alternative, nonessential PAS. Conversely, \textit{Alu}-PASs that do not change the number of sites are mostly major sites (71%), suggesting that many \textit{Alu} sequences have substituted major PASs in the ancestral genes. Interestingly, 17 \textit{Alu} sequence are inserted precisely at the position of a mouse PAS. Figure 5A presents a likely mechanism for such insertion events, which we named “PAS breaks,” in which an \textit{Alu} sequence literally breaks into an existing PAS and functionally replaces it. In support of this mechanism, \textit{Alu} elements tend to integrate at 5′-TTAAAA-3′/5′-TTTTAA-3′ or other...
AT-rich sites (Jurka 1997). Obviously, major polyadenylation signals ATTAAA and AATAAA may easily form such integration hot spots. Alu integration involves the duplication of the AT-rich region as shown in figure 5A, eventually causing duplication of the preexisting polyadenylation signal. Ensuing sequence mutations can cause loss of one or both PAS or create new PAS within the Alu sequence at the middle or tail position. A visual inspection of the 17 cases of PAS breaks revealed that the different outcomes actually occur, with the Alu insertion resulting in a variable number of PASs in humans. Supplementary table S1 (Supplementary Material online) provides the gene names and coordinates of the corresponding sites in human and mouse. Note that this relatively high incidence of Alu integration within preexisting PASs also explains why tail-PAS in Alu sequences are followed by strong GT-rich enhancers (fig. 4B), as these enhancers were probably present as part of the existing PAS before Alu insertion. Alu integration may also occur in intronic sequences and cause premature transcript termination. Thirteen Alu-PASs are located in intronic sequences that contain no PASs in the mouse ortholog (table 4). The consequence of intronic Alu-PAS integration is the synthesis of a shortened transcript, as shown in figure 5B. Details on intronic Alu-PASs are provided in supplementary table S2 (Supplementary Material online). Interestingly, although Alu-PASs in 3′ UTRs are either major or minor sites, intronic Alu-PASs correspond predominantly to minor isoforms (12 of 13 cases). Therefore, in most cases, the major transcript retains the same exon structure as in the mouse ortholog. The only exception is Ensembl gene ENSG00000187860, a gene of unknown product, in which the truncated transcript is the major isoform. These intronic Alu-PASs are reminiscent of the Line-1 (L1) retroposon insertions causing “gene breaks” in several human genes (Wheelan et al. 2005). As L1 contains both a PAS and a promoter, L1 insertion has led in some cases to the breaking of a gene into two smaller genes. We did not observe such events associated with intronic Alu’s, but the fact that a truncated transcript can become the major isoform may eventually lead to exon loss in the host gene.

**Discussion**

Of the ∼10,000 Alu elements located in the 3′ UTRs of human genes, only a small subset is active as PASs (Alu-PASs). Almost all Alu-PASs are oriented in the same direction as their host genes, consistent with the presence of motifs resembling polyadenylation signals on the forward Alu strand and their absence on the reverse strand. In their most common form, Alu sequences are probably inapt for polyadenylation on either strand, as only a tiny fraction (1%) of the Alu sequences in 3′ UTRs are polyadenylation active. However, the acquisition of just a few point mutations in hexamers resembling poly(A) signals and/or in flanking GT-rich regions can produce efficient signal + enhancer combinations in forward strand Alu’s. Reverse strand Alu’s, on the other hand, lack all the features required for polyadenylation and are essentially unfit for this task. With one exception, none of the thousands of reverse Alu’s in 3′ UTRs could turn into a PAS in the 65-My period since the onset of their propagation.

An alternative explanation for the predominance of plus strand Alu-PAS may lie in the PAS break mechanism shown in figure 5A. Jurka’s model for transposon insertion at an AATAAA site (21) involves the formation of a staggered nick exposing a poly(T) on the reverse DNA strand. This poly(T) could serve as a primer for Alu insertion by binding the poly(A) tail of an Alu transcript, thus forcing insertion in the positive orientation. This alternative explanation, which is not exclusive of the “fitness” explanation above, would only concern the <20% fraction of Alu-PASs that appear to be inserted within a previously existing PAS.

Earlier studies have suggested that Alu-borne PASs would form alternative, minor sites or would be under negative selection (Roy-Engel et al. 2005). We found that the fraction of major or unique sites is the same in Alu-PAS as in the general PAS population. Although some Alu elements may have cryptic PAS activity, others are most likely under selection to act as major PASs. Good examples of such selections are PAS break events where Alu insertions obliterated and replaced existing PASs (fig. 5A). This
argues for a genuine functionalization of certain Alu elements as PASs, which clearly fits the definition of exaptation (Gould and Vrba 1982). Alu-PASs forming major/unique poly(A) sites can thus be considered as exapted. While this article was being submitted for publication, further support for Alu functionalization as PASs came from a global study of transposon-based PASs by Bin Tian’s group (Lee et al. 2008). The authors established using comparative genomics that TEs in general could be gradually fixed and optimized for polyadenylation. They confirmed the preferred usage of mid-PAS and tail-PAS in Alu and also identified additional sites upstream of Alu sequences that use an enhancer-like element in the Alu 5’ region.

The formation of PASs by Alu sequences is reminiscent of Alu exonization, the conversion of intronic Alu elements into functional exons (Lev-Maoz et al. 2003; Kreahling and Graveley 2004), Alu exons are mostly alternative, minor forms, but single-nucleotide mutations can convert alternative or silent Alu elements into constitutive exons, and this can lead to the development of human disease or produce proteins with new functionalities. Likewise, Alu-PASs may occur only as cryptic sites that do not confer any selective advantage, but they could turn deleterious if they were to become major sites. For instance, most intronic Alu-PASs are minor sites, but their conversion into major sites through random mutations would produce a truncated transcript as the major gene product. Some of the cryptic Alu-PASs are therefore a latent source of disease. Considering that some Alu sequences are still active in the human genome (Mills et al. 2007), this possibility of deleterious functionalization should not be overlooked. As with Alu exons, however, the emergence of alternative PAS can also be a source of functional diversity. Alternative PASs lead to the expression of transcripts with different UTRs in different tissues (Beaudoin et al. 2000; Zhang et al. 2005; Sandberg et al. 2008). Alu insertion in 3’ UTR may thus contribute to the emergence of differential expression events.

Supplementary Material
Supplementary figures S1 and tables S1 and S2 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
This work was supported in part by the European Commission FP6 Program (contract number LSHG-CT-2003-503329), C.C. was supported by a studentship from the China Scholarship Council.

Literature Cited
An HJ, Lee D, Lee KH, Bhak J. 2004. The association of Alu repeats with the generation of potential AU-rich elements (ARE) at 3’ untranslated regions. BMC Genomics. 5:97.
Polak P, Domany E. 2006. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics. 7:133.

Dan Graur, Associate Editor
Accepted October 27, 2008