The Methylation-Driven Logic of Alu Lineage

Abstract

Here we decipher the logic underlying a key functional motif that defines the landscape of Alu repetitive elements. Alu positions 7-10 lie within the A Box, a site necessary for its transcription by Pol III and subsequent retrotransposition. We show that this site originated as a mother motif (CGCG) that gave rise to four daughter and ultimately four granddaughter motifs. The sequences of the progeny are all dictated by methylation-driven cytosine deamination of the mother. These finding provide the mechanistic basis for diversification and expansion of Alu, a major event in primate evolution.

Summary

Alu repeats are short interspersed elements [1-4] that are of great interest because they arose along with primates ~65 million years ago, because they are restricted to the primate lineage, and because they are the most abundant mobile elements in primates [5-8]. Consequently, Alu is a potential “reason” that primates differ from other species. Alu continues to expand throughout the human genome, even today as an Alu is inserted at a new position in the human genome every twenty or so births. The 1.1 million Alus in the genome have an immense influence on human biology [5]. Insertion of Alu elements cause numerous human genetic diseases ranging from cancer to familial hypercholesterolemia [5]. Over the last decade it has also become clear that Alus can contain tens of thousands of functional transcription factor binding sites [9]. For this reason, Alus are believed to have played a role in establishing larger gene regulatory networks [9-11]. The role of Alu also extends to the RNA world since bioactive Alu elements are transcribed to free monomeric Alu-RNAs (small cytoplasmic: scAlu) [5] and because dozens of microRNAs are transcribed through Alu-dependent RNA polymerase III (Pol III) transcription [12].

The mechanistic logic underlying the expansion of the lineage of Alu remains unclear. It was originally hypothesized that Alu elements arose from a single master gene [13], but this notion has evolved to the current belief that a small subset of Alus remain competent for retrotransposition, and therefore serve as source genes for further propagation [5]. Thus, Alu elements that were at one time, or still are, active can be considered master or “source” genes [5,14]. It is not possible to define source Alu with the well-known JSY hierarchical classification scheme. This classification system is based on 20-30 diagnostic positions that undergo neutral mutation [15-17], which allows for stratification of Alus into families of different age, the J (oldest), S (intermediate) and Y (youngest), and makes it possible to understand the molecular archeology of Alu. Almost by definition, however, the JSY classification scheme is independent of the changes to functional sites brought about by evolutionary pressure. Hypermutable CG doublets are a profound example of this because they are excluded from the sequence analyses used for JSY classification. The C of ‘CG’ is a preferred site of methylation, an epigenetic process linked to chromosome remodeling and gene silencing.

Indeed, Alus contain nearly one-third of the genome’s CpG sites and they play a major role in epigenetic silencing [18-22].

Here we consider the possibility that methylation-driven cytosine deamination is a guiding event in the propagation and functional diversification of Alu. Specifically we focus on positions 7-10 of Alu (labeled according to homology with the sequence of the 7SL gene). These positions lie within the Alu A box, which functions along with a downstream B box to create the canonical binding site for Pol III [23]. In the human 7SL gene, a presumed ancestor of Alu [6,24], positions 7-10 are CGCG, a motif that can be viewed as two adjacent and hypermutable CpG sites. By the process of cytosine methylation and subsequent deamination, these sites could be converted to CpG (or CpA on the antisense strand) [18,22]. Some of the first evidence for this possibility came from a study by Zemojtel et al. [9], who showed that this particular CGCG in Alu is deaminated to create a vast number of functional binding sites for p53. Thus, full development of the methylation-driven deamination landscape of a “mother” CGCG sequence, shows that it would be converted to a select set of only four daughter motifs (one deamination) and four additional granddaughter motifs (two deaminations). One the other hand, random mutation of a four nucleotide stretch would theoretically yield a normal distribution of 256 possible sequences. To distinguish between these two extremes, we tallied the motifs at position 7-10 in a set of 8,422 representative dimeric human Alu sequences extracted from the human genome (Supplemental Figure 1). While 1,844 Alus contained motifs that are outside of the deamination landscape, the vast majority of the Alus (6,538) contained either the mother CGCG, or one of the eight daughter or granddaughter motifs that arise by deamination (Figure 1a and 1b). Essentially same distribution was observed in the entire human genome when all 1,091,321 Alus were analyzed. This result leads to the inescapable conclusion that CpG deamination has been the major determinant in the sequence at Alu positions 7-10. It is almost inconceivable that such a constrained distribution could have arisen in any other manner.
If methylation-driven deamination shaped the sequence at positions 7-10 in the Alu population, one would also anticipate a time-dependent conversion of the mother into the daughter, and ultimately the granddaughter motifs. To look for such a trend, we examined the distribution of motifs in the J (oldest), S (intermediate) and Y (youngest) families of Alu [5,17]. The 6,538 test Alus were segregated into J, S and Y families with the online CENSOR tool, and the fraction of mother, daughter and granddaughter motifs in each Alu family was calculated. In the Alu J family, the mother CGCG motif represents only a small fraction of the elements (6%). A substantial number of the J Alus have been converted to daughter motifs (29%), and an even higher fraction (65%) into granddaughter motifs. In the Alu S family, which originated somewhere between 65 and 35 million years ago, 15% of the 7-10 motifs are mother, 43% daughter and 42% granddaughter motifs. In the Y family, which arose only 2 million years ago [5,16], the mother CGCG motif is 38%, the daughter motifs represent 42%, but granddaughter motifs are only 20% (Figure 1c). Thus, we observe an age-dependent increase in mother elements, and a concomitant decrease in granddaughter motifs. This trend can certainly be taken as strong evidence that the mother CGCG has a higher rate of retrotransposition than the progeny. If either of the progeny were more active, one would expect to observe that the frequency of progeny would be higher than that of the mother in younger Alus as well as their overall sequence conservation (e.g. fossil-Alu-monomers with CGCG-motif should be mostly preserved in primates; Figure S3). While more complicated scenarios could conceivably explain the observed distributions, the simplest interpretation of the data is that methylation-driven deamination of positions 7-10 shaped the Alu landscape during primate evolution.

There is a strong line of reasoning to suggest that the mother CGCG motif is predominant in source Alus. Since Pol III transcription of Alu is necessary for retrotransposition, any active Alu must support interaction with Pol III. A key site for this interaction is the A box, which encompasses positions 7-10 in Alu [23]. Interestingly, work in the early 1990s showed the
granddaughter motifs CATG and TGCA are unable to support Pol III transcription [19], so these progeny are unlikely to be active. In our search of the literature, we found only one daughter motif (CGTG) reported to support Pol III transcription [25]. Along similar lines, the 40 human Alu elements with demonstrated retrotransposition activity in cells, all contain the mother CGCG motif [26]. While no comprehensive study has been conducted on the efficiency by which all daughter and granddaughter motifs can support Pol III transcription, and/or retrotransposition, the available biochemical data are entirely consistent with the idea that Alus with the mother CGCG are most active, and that the progeny, particular granddaughters, are less active. The fact that most Alus of the Y family, which are presumed to be the most active, contain the CGCG motif is also consistent with the idea that Alus with the mother motif are optimal "source" Alus.

Based on the observations put forth here, we propose a new three-dimensional projection of the phylogeny of Alus that takes into account both the methylation-driven logic described here, and the canonical JSY classification scheme (Figure 2). This projection essentially illustrates how extant Alus containing mother, daughter and granddaughter motifs have arisen, and how they project onto a flat landscape representing the JSY classifications. This cone-like projection begins to map the positional coordinates for each UTR. A database containing this information was compiled from UTRs (710,598: 65.11%) contained the CGCG motif or one of the eight potential methylation driven derivatives of this motif. The frequency of CGCG, which we call the mother (M), was found in less than 10% of all Alus, whereas the majority contained a motif derived from deamination at one site, which we call daughters (D), or at both sites, which we call granddaughters (G). Furthermore, of the 35% of Alus that lack one of the eight noted motifs, nearly half of these (45%) contain a motif that can be derived by a single point mutation in one of the G motifs. Consequently, the majority of the sequence landscape at 7-10 in Alus has been shaped by methylation driven cytosine deamination (manuscript in preparation).

Supplemental Figure 1. Computational Extraction of Alu’s

To obtain a representative set of Alus we focused on the regions upstream of human genes. We created a database containing the untranslated regions (UTRs) extending from the translation start site (ATG) to -5,000 bp in 17,532 human genes. These sequences were extracted using the genome browser at UCSC (http://genome.ucsc.edu/cgi-bin/hgGateway, 2006 assembly - hg18, [S1]) and were then deposited in a Structured Query Language (SQL)-database. Each UTR in the database was matched to all coding sequences from the RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq). This allowed us to define first ATG-codon of each coding sequence and then to create positional coordinates for each UTR. A database containing this information was compiled and it contains ~92 megabases (~3% of the human non-coding genome).

Alu sequences were computationally extracted from UTRs in this database using "collocation" searches like those used in statistical language processing to classify segments of text based on the presence of two or more separate words [S2]. By way of analogy then, we sought to identify "Alu paragraphs" by searching for "words" common to all Alus. Three sequence motifs were used as the search words; the sequence that encodes the SRP [S3,S4] binding site in Alu RNA, and the motifs encoding the A- and B-boxes [S5,S6]. The A-box is positioned seven nucleotides upstream of the SRP binding motif and the B-box is positioned 46 nucleotides downstream yielding an Alu signature of "[A-box]−[SRP]−[B-box]" (Supplemental Figure 1, top panel).

All sequences corresponding to the SRP binding site were extracted from the database of UTRs. Then, within this set, we identified all sequences that contained an A- and B-box at a fixed distance from one another (as indicated in the top panel). This procedure extracted 12,440 Alus, which included both dimeric (~300 nucleotides) and monomeric Alu sequences (~150 nucleotides) (Supplemental Figure 1, Venn diagram). Since Alu monomers might include aberrant or truncated sequences, we elected to exclude them from the analysis. This was accomplished by a second collocation search for the poly-A tail that exists between the monomers of full-length Alus. As a result of this filtering, we obtained dimeric 8,422 Alu sequences that were used for this study (Supplemental Figure 1, Venn diagram, blue). Alu-[7-10]-tailing was also performed on all human Alus (hg19: 1,091,321). It showed essentially the same distribution. The majority of the Alus (710,598: 65.11%) contained the CGCG motif or one of the eight potential methylation driven derivatives of this motif. The frequency of CGCG, which we call the mother (M), was found in less than 10% of all Alus, whereas the majority contained a motif derived from deamination at one site, which we call daughters (D), or at both sites, which we call granddaughters (G). Furthermore, of the 35% of Alus that lack one of the eight noted motifs, nearly half of these (45%) contain a motif that can be derived by a single point mutation in one of the G motifs. Consequently, the majority of the sequence landscape at 7-10 in Alus has been shaped by methylation driven cytosine deamination (manuscript in preparation).

Supplemental Figure 2. Using online CENSOR for traditional (JSY)-age Alu's annotation

The online classification of Alu elements was performed by a
program available on a public server ("CENSOR", www.girinst.org/censor) dedicated to analysis of repetitive elements as deposited in RepBase [S7] and originally classified by RepeatMasker (another well-known tool for library-based repeat identification by A. F. Smit, R. Hubley and P. Green; http://www.repeatmasker.org). RepBase Update, the most comprehensive database of repetitive element consensus sequences, is also available for use in the CENSOR or BLASTN search step. RepBase Update is compiled and maintained by the Genetic Information Research Institute [S8,S9]. Both programs (RepeatMasker and CENSOR) perform similarity searches based on local alignment using precompiled libraries of consensus or representative sequences of repeat families in RepBase. CENSOR, like RepeatMasker, is designed to locate and mask regions in genomic sequences that correspond to known repetitive elements [S10]. CENSOR uses the fast and sensitive similarity search program WU-BLAST (W. Gish; http://blast.wustl.edu). Optionally, the BLASTN or BLASTX programs of the WU-BLAST package can be used directly instead of CENSOR. The authors are grateful to Jerzy Jurka and the GRI for providing the RepBase Update database and the CENSOR program. The CENSOR program is being developed and maintained by Oleksiy Kohany at GRI. The authors also thank the TIGR Plant Repeat Database Team for making their databases freely available. The CENSOR server permits screening repeats in DNA sequences from all eukaryotic species represented in the database by comparing them to the most recent version of RepBase Update and returning output back to the user via the browser (if used in a web/remote host-mode, see a snapshot of the web-CENSOR-input interface) or after download of CENSOR program on a local PC/CPU (access of the compressed *tar.gz file requires user registration). We submitted as input the sets of Alu-sequences as grouped in the M/D/G 'families' all compiled in FASTA-format and uploaded to CENSOR in a batch mode (i.e. uploaded from a local file). Detailed information about CENSOR input/output is presented in "Help/Information" main menu at (www.girinst.org/censor). CENSOR can be run in three different sensitivity modes. The WU-BLAST parameter settings corresponding to these modes are listed in the online Tutorial. Certain parameters (word size, E-value threshold, gap penalties) of the direct WU-BLAST searches can also be adjusted by the user (see the online Tutorial for details). We used the default options of CENSOR. After the web-based CENSOR returned its output for each of the M/D/G Alu-submissions, the tabular annotations were extracted and converted into MS-Excel files, where Alus in each M/D/G group were segregated into J, S and Y families as reported by CENSOR.

Supplemental Figure 2: Using online CENSOR for traditional (JSY)-age Alu’s annotation.

Supplemental Figure 3. Unity of JSY-MDG models with phylogeny of M-FAMs.

Supplemental Figure 3: Unity of JSY-MDG models with phylogeny of M-FAMs.

A simplified branching order of the major Alu subfamilies is shown. Branching points correspond to the average age of individual JSY subfamilies in million of years (Myr), all presumably originating from 7SL RNA [24]. The results of the present study suggest branching of the Alu lineages based on CpG mutations at position 7-10 (Figure 2). These mutations in a Mother (M) Alu give rise to Daughter (D) and Granddaughter (D) Alus. Because deamination at CpG is much faster than non-CpG mutations, an alignment of the two evolutionary processes results in MDG branching for each of the JSY lineage. Daughters that mutate at
The Methylation-Driven Logic of Alu Lineage

References

1. Houck CM, Rinehart FP, Schmid CW (1978) Fractionation of renatured repetitive human DNA according to thermal stability, sequence length, and renaturation rate. Biochim Biophys Acta 518(1): 37-52.

2. Houck CM, Rinehart FP, Schmid CW (1979) A ubiquitous family of repeated DNA sequences in the human genome. J Mol Biol 132(3): 289-306.

3. Britten RJ, Kohne DE (1968) Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science 161(3841): 529-540.

4. Schmid CW, Jelinek WR (1982) The Alu family of dispersed repetitive sequences. Science 216(4550): 1065-1070.

5. Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. Nat Rev Genet 3(5), 370-379.

6. Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J (2007) Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. Trends Genet 23(4), 158-161.

7. Kazaian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303(5664): 1626-1632.

8. Han K, Konkel MK, Xing J, Wang H, Lee J, et al (2007) Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. Science 316(5822): 238-240.

9. Zemotjel T, Kielbasa SM, Arndt PF, Chung HR, Vingron M (2009) Methylation and deamination of CpGs generate p53-binding sites on mammalian DNA. Science 312(5774): 745-747.

10. Tomilin NV (2008) Regulation of mammalian gene expression by retroelements and non-coding tandem repeats. Bioessays 30(4): 338-348.

11. Moohijizen P, Kulkki JK, Dunn DS, Schibeci D, Barrero R, et al. (2010) The transcript repeat element: the human Alu sequence as a component of gene networks influencing cancer. Funct Integr Genomics.

12. Gu TJ, Yi X, Zhao XW, Zhao Y, Yin JQ (2009) Alu-directed transcriptional regulation of some novel miRNAs. BMC Genomics 10: 563.

13. Shen MR, Batzer MA, Deininger PL (1991) Evolution of the master Alu gene(s). J Mol Evol 33(4): 311-320.

14. Cordaux R, Hedges DJ, Batzer MA (2004) Retrotransposition of Alu elements: how many sources? Trends Genet 20(10): 464-467.

15. Jurka J, Mikolajевич A (1991) Reconstruction and analysis of human Alu genes. J Mol Evol 32(2): 105-121.

16. Kapitonov V, Jurka J (1996) The age of Alu subfamilies. J Mol Evol 42(1): 59-65.

17. Batzer MA, Deininger PL, Hellmann-Blumberg U, Jurka J, Labuda D, et al. (1996) Standardized nomenclature for Alu repeats. J Mol Evol 42(1): 3-6.

18. Ehrlich M, Wang RY (1981) 5-Methylcytosine in eukaryotic DNA. Science 212(4501): 1350-1357.

19. Liu WM, Schmid CW (1993) Proposed roles for DNA methylation in Alu transcriptional repression and mutational inactivation. Nucleic Acids Res 21(6): 1351-1359.

20. Jones PA, Takai D (2001) The role of DNA methylation in mammalian epigenetics. Science 293(5532): 1068-1070.

21. Klose RJ, Bird AP (2006) Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci 31(2): 89-97.

22. Mazin AL (2009) Suicidal function of DNA methylation in age-related genome disintegration. Ageing Res Rev 8(4): 314-327.

23. Paolella G, Lucero MA, Murphy MH, Baralle FE (1983) The Alu family repeat promoter has a tRNA-like bipartite structure. EMBO J 2(5): 691-696.

24. Ullu E, Tschudi C (1984) Alu sequences are processed 7SL RNA genes. Nature 312: 171-172.

25. Murphy MH, Baralle FE (1984) Construction and functional analysis of a series of synthetic RNA polymerase III promoters. J Biol Chem 259(16): 10208-10211.

26. Bennett EA, Keller H, Mills RE, Schmidt S, Moran IV, et al. (2008) Active Alu retrotransposons in the human genome. Genome Res 18(12): 1875-1883.

Supplemental References

S1. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12(16): 996-1006.

S2. Manning CD, Schütze H (1999) Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts, USA.

S3. Strub K, Fornalaz M, Bui N (1999) The Alu domain homolog of the yeast signal recognition particle consists of an Srp14p homodimer and a yeast-specific RNA structure. RNA 5(10): 1333-1347.

S4. Halic M, Gartmann M, Schlenker O, Mielke T, Pool MR, et al. (2006) Signal recognition particle receptor exposes the ribosomal translocon binding site. Science 312(5774): 745-747.

S5. Paolella G, Lucero MA, Murphy MH, Baralle FE (1983) The Alu family repeat promoter has a tRNA-like bipartite structure. EMBO J 2(5): 691-696.

S6. Murphy MH, Baralle FE (1984) Construction and functional analysis of a series of synthetic RNA polymerase III promoters. J Biol Chem 259(16): 10208-10211.

S7. Jurka J, Kapitonov VV, Pavlicek A, Kloukovsk P, Kohna O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110(1-4): 462-467.

S8. Jurka J (1998) Repeats in genomic DNA: mining and meaning. Curr Opin Struct Biol 8(3): 335-337.

S9. Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16(9): 418-420.

S10. Jurka J, Kloukovsk P, Dagman V, Pelton P (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. Comput Genet 20(1): 119-121.