Volume 11 Number 21 1983 Nucleic Acids Research

A second major class of Alu family repeated DNA sequences in a primate genome

Gary R.Daniels and Prescott L.Deininger

Department of Biochemistry, Louisiana State University Medical Center, 1901 Perdido Street, New Orleans, LA 70112, USA

Received 3 August 1983; Revised and Accepted 5 October 1983

ABSTRACT

We have analyzed repetitive DNA sequences in a prosimian, *Galago crassicaudatus*, and found that there are two distinct, highly repetitive families of sequences related to the human Alu family. The Type I family is closely analogous to the human Alu Family. The Type II family of repeats, which appears to be present in higher copy number, has a right half that is almost identical to the Type I family. However, the left half of the Type II sequence shows only limited homology to the galago Type I or human Alu families. A comparison of homologous sequences in the left half indicated that they are centered in regions of the Alu family which function as RNA polymerase III promoters. We have also observed at least one example of a Type II left half that was integrated into the genome independent of the Alu family right half sequence. The Type II family appears to be of much more recent evolutionary origin than the Type I and may have arisen by the independent integration of a RNA polymerase III promoter adjacent to the right half of a Type I Alu family sequence.

INTRODUCTION

In the human genome there are approximately 500,000 members of a single repetitive DNA family, the Alu family (1,2,3). These sequences are about 300 nucleotides long and are interspersed throughout the genome (2,4). Structurally, the human Alu family members actually represent a head-to-tail dimer of two approximately 130 base pair monomers. The two halves of the dimer contain about 70% homology to each other with the right half also containing an internal region of 31 base pairs which is not part of the left half (3,7). It is not clear whether these 31 base pairs represent an insertion or a deletion event relative to the evolutionary prototype sequence. The individual members of the human Alu family do not contain exactly the same sequence. Instead they are divergent from a canonical consensus sequence by about 13% (3) which makes them a family of similar, but not identical sequences.

There is strong evidence that the Alu family sequences have been interspersed throughout the genome by means of RNA intermediates (8,9). The Alu family members contain a RNA polymerase III promoter (4) within the left half of the dimer (10) which would allow the formation of the RNA intermediate. The details of the mechanism for Alu family movement are as yet unclear, but upon insertion into a new site in the genome, short direct repeats are formed in the genomic sequence flanking the Alu family member as occurs with most other transposing elements (5,6).

The dimer structure of the Alu family appears to be conserved throughout all lines of primate evolution (11). However, rodents have equivalent sequences to the Alu family which are monomers rather than dimers (5,6,12,13, 13). In the hamster genome there are actually two subfamilies of Alu family equivalent sequences, neither of which involve dimer structure (13,14). These hamster Type I and Type II Alu family equivalent sequences are approximately 140 and 96 bases long, respectively. They share about 50 residues at the 5' end of the Type II repeat with 88% homology and a lesser homology at the 3' end. The Type II Alu-equivalent family makes a RNA polymerase III product, in vitro, whereas the Type I does not. In this sense the hamster Type II and Type I repeats are analogous to the left and right halves of the human dimer, respectively. It is not clear whether these multiple types of Alu family related sequences are common to all lower mammals. The rat genome has a closely analogous Type II family (15). The mouse genome has been found to contain a sequence comparable to the Type I, the B1 repeat, and also to contain another class of repeat of similar copy number and arrangement to the Type II (16). This second mouse repeat, the B2 repeat, shows no major homology with Alu family sequences, however. A repetitive family with homology to the Alu family which is common to all mammals and some non-mammalian species (17,18) is the 7S gene. This gene is thought to be present in a few functional copies with about 500 to 1000 pseudogenes. Its structure is essentially that of an Alu family monomer with an approximately 140 base insert. It is not clear whether this is truly an insert or whether the Alu family sequences arose from a deletion of a 7S progenitor (17).

In all of the studies to date there have been no reports of a monomer family, like the rodent families, in the human genome or the genome of any other primate. Neither does the human dimer type of Alu family member appear to be present in rodents. We have shown that even in the prosimian, Galago crassicaudatus, there is a very human-like dimer Alu family. However, in this report we describe a second type of Alu family within the galago genome which is clearly distinct from the human type dimer organization and present in high copy number.

METHODS

Cloning galago repetitive DNA sequences. Genomic *Galago crassicaudatus* DNA was prepared from liver as previously described (11). The DNA was cleaved with the restriction enzyme *Rsa* I and DNA fragments from 300 to 500 nucleotides in length were isolated from a 1.5% agarose gel by a trough elution procedure (19). This DNA was ethanol precipitated, resuspended, and cloned directly into M13mp8 (20) by blunt-end ligation (21). The M13mp8 had previously been cleaved with *Sma* I and treated with alkaline phosphatase to stop religation. Recombinant phage were plated (22) and then screened by hybridization (23) overnight at 65°C in 5 x SSC (SSC = 0.15 M NaCl 0.015 M citric acid) with nick-translated galago DNA. With genomic DNA as a probe, only highly repetitive cloned sequences were detected.

DNA sequence analysis. Recombinant clones were picked and phage DNA prepared from one ml cultures (24). DNA sequence analysis was carried out by the dideoxy termination method (25) using standard procedures (22,24). DNA sequences were analyzed by computer using the programs of Staden (26) as modified by K. Isono for use on the DEC 10. DNA sequences were compared to each other and the human Alu family consensus sequence (3) to detect related families of sequences.

RESULTS

Our experimental approach was to randomly create a large number of Alu family containing clones in a form which facilitates DNA sequence analysis. To do this we chose to cleave genomic galago DNA with the restriction enzyme *Rsa* I. This enzyme did not cut the human Alu family sequences and our preliminary data indicated that the same might be true for galago. The cleaved DNA was then size fractionated to yield fragments of 300 to 500 nucleotides in length. The fractionated DNA fragments could contain a 300 nucleotide-long Alu family member with only a minimum of flanking sequences, thus facilitating the sequence analysis of the clones. These fragments were blunt-end cloned directly into M13mp8 and screened by hybridization to nick-translated genomic galago

```
                             10         20         30         40         50
                                  T                G
CONS:     .......... TGGCTTGGCG CC GTAGCAC AGTGGTTA G GCGCCAGCCA CATACACCTA
                                  C                T

GAL 12    AACCAAAGGG .........A ..C...A... .........CA .....G.... ...G....A.
                                          A
GAL 21    AACCCATTTT .....C.... ..T..GC.T. .....C..A. .......... ..........
                                          ^
GAL 25    AAATGCAATC A........A T.C....... .........CA .......... ...G.T.AG.
                                          A
GAL 39    AGTAAATGAT G..XX...T. ..T..X..T. .AG.AG..G. .T.......C ....TG..AG
                                          ^
GAL 34    AGTTAATGCA .....CA... G.CA...... ........T. .T........ ........A.
GAL 20    GAGTAGCAGA G..XX...T. ..T..C.GT. .....A..G. ..A.TG...X ....T...G.
GAL  5    CAAAACTACT .....CA... .TC....... ........T. ..A....... ........A.
GAL 40            AC ...XX..... ..TA.GA.T. ....AG..G. ..A.TG...X X...T...A.
GAL  1                           .C... ..X.....T. TT..TG.... ........C.
GAL 27                                    C.GC.XA .T........ .........G
GAL 35                                              A......C T...T.TTC.
GAL  7                                                          ...T...
GAL 16                                                          .......
GAL  6                                                          ....G..
GAL 33                                                          .......
GAL 30                                                          .......
GAL 26                                                            ...CT

                  60         70         80         90        100        110
CONS:     GGGTGGTGGG TTCGAACCCA GCCGGGCCCA GCCAAACAAC AATGACAACT GCAACCAAAA

                                                                       CAACA
GAL 12    ..C..T.... ..T.....TX ...T...... ..A......X XX........ .....A.C..
                                                                        ^
                                                                       AAAAA
GAL 21    A......... ......T... ...T.....T ........T. ..A..TGG.. ..........
                                                                        ^
GAL 25    .A.....A.. .......... ....A....T ..T.....G. .......... ......XX..
GAL 39    A......... ...A...A.. ....C.A... AXX....TG. ..A..XXXXX XXXXXXXX..
GAL 34    .......... ..T....TTX ...A...... ..T....... .......... C..G.XXXG.
GAL 20    ......A... ...A...T.G A...XXX... .T.....TG. .GCAXXXXXX XXXXXXXX..
                                              A
GAL  5    .TC....... ..T......X ...T...... ..T....... .......... .T...XXX..
                                              ^
GAL 40    T.......T. .......... ....X....X XXX....TG. ....XXXXXX XXXXXXXX..
                                                                       CAACAA
GAL  1    ..C...A.A. ..T....... ...TX..... ..T.....G. .GC.....X. ....TA.C..
                                                                        ^
GAL 27    A.C....... ......T... ...T.....T .......... .....TGG.. AT.....X..
GAL 35    .......... ...A....TG ....C....X XXX....TG. ..CA.T...X XXXXXXXX..
                             CC
GAL  7    A.....C... ......T... ......A..C .......... .C....GG.A ......XX..
                              ^
                             CC
GAL 16    A......... ......T... ....A.A..C A......... ......GX.. ......XX..
                              ^
GAL  6    A.....C... ..T..CT... ...T....TC A......... .......G.. ......XX..
GAL 33    ......C... ..T...T... .........T .......... .....TGG.. ......XX..
GAL 30    ......CAXX XXX...A.TG ...T.A...T A.T..G.... .....T..G. .A....XX..
                             CC
GAL 26    AA....C..A ......T... ......T..C .......... ......GG.. ..........
                              ^
GAL  4                                                 .X... .....XXX..

                 120        130        140        150        160        170
CONS:     AAAATAGCCG GGCGTTGTGG CGGGCGCCTG TAGTCCCAGC TACTTGGGAG GCTGAGGCAA

GAL 12    .T........ ..TA...... ....T..... .......... .......... ..........
GAL 21    .........A ...A...... T......... .G........ .......... ..A......G
GAL 25    .......... ..T...A... .A........ ..A......A .........A .T...A...G
GAL 39    ....
GAL 34    .......... .......... .A..T..... .......... ...C...... ..........
GAL 20    ........T. ........A. TA..TT...A .......... .......... ..........
GAL  5    .........T ...A...... .A..T..... .......... .......... .......A..
GAL 40    ...G...... .AT....... T...A..... .......... ....C....A ..........
                                                        GGG
GAL  1    ........T. ...A...... .A...A...C .......... ...AA..... ..........
                                                         ^
GAL 27    ........T. ...A....A. .A...A...A ..A....... .........A ..G......G
GAL 35    ...G...T.. ..T....... T...T....A .......... ....CA.A.. ..........
GAL  7    ....C..... .......... .......X.. .G........ .G........ ..A......G
GAL 16    .......... .........X .......... .G........ .......... ..G......G
GAL  6    ..T....... A..A...... T...T....A .G........ ....C..... .........G
GAL 33    .....G.... .........X ..A....... .G........ .......... ..A..A...G
GAL 30    ....C...A. ..T..CA... T...T..... ......T... .....A.... ..........
GAL 26    .......... .......... .......... .G........ .......... ..G.......
                                                     GGGGGTGGGG G
GAL  4    ...GC..T.A .AT....... .A...A.... .......... .......... ..........
                                                        ^         ^
GAL  3                                        A.T... .......... .........G
```

```
                180        190        200        210        220        230
CONS:   GAGAATCGCT TAAGCCCAGG AGTTGGAGGT TGCTGTGAGC TGTGATGCCA CAGCACTCTA

GAL 12  ......T... .G......A. .......... ....A..... .......... .......G..
GAL 21  ....C.
GAL 25  .......... ........A. ....T..... .......... ...A.C.T
GAL 34  ......T... ........A. .......... .........T .......... AG.T
GAL 20  .........C ........A. ..C....... .......... .......... .T........
GAL  5  ......T... .....T..A. CA..T..... .........T ....G..... ..........
GAL 40  ..X....A.C .......... ..C...C... .......... ........A. T...G..T..
GAL  1  ......X.T. TTGC...A
                     ^
GAL 27  ......T... .G...TT... .......... .......... .........T .....T....
GAL 35  ..A...T..C .......... ..C....... .......... .......... TG........
GAL  7  ....C..... .G........ .......... .......... ...A...... .G........
GAL 16  ....C..... .G........ .......... .......... ...A...... .G.T
GAL  6  ....C..... .G........ .......... .......... .....A.... ...T
GAL 33  ....T.T... .G..T..... ....A..... .A........ .......... ..........
GAL 30  .......A.. ....T...A. ....T...A. .......... .A........ .X........
GAL 26  ....C..... .G........ .......... .......... ...A...... .G........
GAL  4  .......A.C ........A. .......... .......... .A........ T...XXXX..
GAL  3  A...C..... .G........ .A........ .......... .......T.. T.........

                240        250        260        270        280
                      A
CONS:   CCCAGGGGGA CA CTTGAGA CTCTGTCTCT AAAAAAAAAA AAAAAA.... ..........
                      G

GAL 20  ..A....T.. ..AAG..... ....A..... .......... ..G.GTAGGA GAGTTTAAAA
GAL  5  ..A....CA. ..TAG..... XX.......X .......... TC....TTAC AATGAGGT
GAL 40  .TG....CA. ..AAC..... ....X..... ....T...T. ..T...TAAA A
GAL 27  .......CA. ..G......G .........X .......... .G..CCTGCC AGCCTTGGTG
GAL 35  .TG.A..C.. ..AAGCA... ...C...... .......... ..GG.TGAAG ACTTTACATT
GAL  7  ........X. ..G......G .........C .......... ......AAAA AGAAACCTTG
GAL 33  ..T....... ..G......G .......A.X .......... .....GAAGT
GAL 30  .......T.. ..A....... A........X .......... ......AGAT TCTCTAGCTT
GAL 26  .......... A.G......G ....A....X .......... ......GAAA ATTATTGCAG
                                     T
GAL  4  .TG.A..... ..AAG..... .........^ .......... ..G...TTTT CAACTTTACA
                                C
GAL  3  ......A... T.G.C....G ..^......X .......... G....GAAAC ACCAGTAACA
```

Figure 1. A consensus sequence for the galago Type II Alu family. The sequence of nineteen cloned galago Type II repetitive sequences are aligned and a consensus sequence (CONS) is created. In the few positions where one base was not the most prevalent, both bases were included. Approximately ten bases of flanking sequence are included and the numbering begins at the left end of the repeated DNA sequence. Deletions in a clone sequence are marked with an X and insertions are marked above the rest of the sequence for that clone.

DNA. Under these conditions, clones containing repeats present in high copy number in the galago genome would be detected. Although this approach could easily select against the presence of larger repetitive DNA sequences, it should yield a fairly random population of interspersed repeated sequences of 300 nucleotides in length and less.

Using the procedure described above, we detected 40 clones which contained highly repetitive elements out of approximately 500 genomic clones. Sequence analysis of these forty clones showed that six of the clones represented repetitive sequences analogous to the human Alu family. In galago we will refer to this as the Type I Alu family and they are described in detail in the accompanying paper (27). Another 18

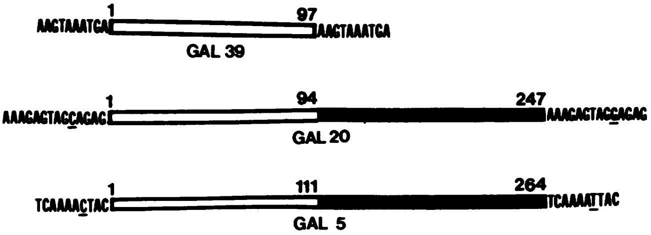

Figure 2. Direct repeats flanking Type II family members. The direct repeats that flank three galago Type II Alu family members are presented. Underlined bases are mismatches within the direct repeats. The overall structures of the repeats are represented schematically. The solid heavy line represents the right half sequences and the open lines represent the left halves. The numbers correspond to the exact lengths of these regions for each specific clone.

clones, referred to as the Type II Alu family, showed excellent homology to the right half of the Alu family, but contained a different left half of about 100 bases in length (Figure 1). One other clone, GAL 39 (Figure 1), contained only the left half of a Type II sequence with no attached Alu family homologous sequence. This DNA sequence was flanked by direct repeats (Figure 2) suggesting that it was inserted into the genome independently and did not result from the insertion of a Type II Alu family member with the subsequent deletion of the right half. Figure 2 also shows some of the direct repeats seen flanking the typical Type II Alu family sequences. Analysis of the sequences within these direct repeats showed 42.3% A and 23.8% T residues suggesting the possibility for a preference for A+T rich integration sites with a bias towards A on one strand. This is consistent with the sequences of similar direct repeats (5,6).

The Alu family clones of both types represented approximately 5% of the clones screened. Since we did not sequence the entire insert of all 40 clones, we may not have detected all Alu family containing clones so that this number may be an underestimate. Also, any highly divergent family members may not have been detected by our screening procedure. This estimate is in close agreement with the proportion of the human genome present as Alu family (3-7%, 1,2,3). However, in the case of galago, only one-fourth (6/25) of these clones represented a human-like dimer Alu family (Type I). The other three-fourths represent members of the Type II family. The presence of occasional Rsa I cleavage sites in the two families may have affected their relative proportions somewhat in our analysis. However, there was no major difference between the two families in this respect as seen by sequence analysis, so that it appears that the Type II family is actually represented in several-fold excess over the Type I family in the galago genome.

We can learn a great deal about the heterogeneity of the galago Type II Alu family from Figure 1. First of all, the overall heterogeneity is about 14% in each clone relative to the consensus sequence. With only one exception, the clones are diverged from the consensus within the range of 10.5% to 19.5%. The specific values are 18.5%,13%, 13.5%,14%,14%,11%,12%,10.5%,17.5%,13%,14%,9.5%,16.5%,15.5%, 11.5%,11%, 19.5%,28% and 18% for the clones in numerical order. Clone GAL 39 with a divergence of 28% represents a repeated DNA sequence with only the left half of the Type II sequence. If we partition the sequences at position 114 to separate left and right halves we see that the left half has almost 19% divergence while the right half has only 8%. This is similar to what was observed for the galago Type I family (27) with 20% and 13% divergence for the left and right halves, respectively.

There are at least two subfamilies of sequences which stand out from the remainder of the clones. We initially identified these subfamilies because of identical insertions or deletions in three separate clones. One example is demonstrated by clones GAL 20,39, and 40 which all share a two base deletion at position 4 and almost identical deletions at position 96. In addition, between positions 1 and 110, there are five point mutations which are common to all three clones and numerous mutations common to two of the clones at a time. One unusual feature of this subfamily is that clone GAL 39 represents only left half, monomeric sequences. A subfamily which is even more striking involves clones GAL 7,16 and 26. These clones all share an unusual two base insert at position 70. About half of the remaining mutations relative to the consensus are common to all three of these clones and many more occur in a paired manner. This subfamily deviates from the consensus by an average of 11.5%, but the individual clones differ from each other by less than 3.5%. In the region from 171 to 213 these clones do not differ from each other at a single base. There may be other subfamilies in addition to these with less specific variances or which are less abundant so that they are not as obvious in this study.

We look at the distribution of the heterogeneity of primate Alu

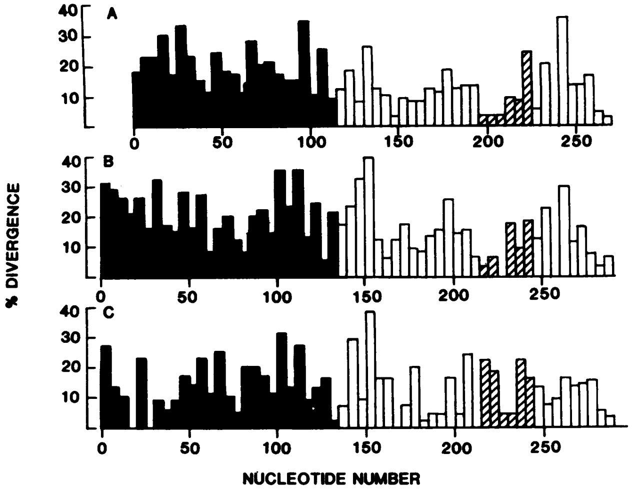

Figure 3. Divergence as a function of position in primate Alu families. The percentage divergence of individual clones from their respective consensus sequence is plotted as a function of position: (A) the galago Type II family, (B) the galago Type I family (27) and (C) the human Alu family (3). The percentage divergence is calculated for each five base unit of the consensus sequence. Insertions or deletions are each calculated as a single mutational event. The blackened area corresponds to left half sequences and the cross-hatched area corresponds to the region in the human Alu family that is not present in the left half of the dimer.

families in more detail in Figure 3. By plotting heterogeneity as a function of position we see the overall high level of divergence within the left half. There is also appreciable heterogeneity in the right half, with two major regions that seem to be more highly conserved. The first of these regions is found between positions 145 and 175 (Figure 3A). This is the region which has been shown to be highly conserved in not only the human (3) and galago Type I (27) Alu families, but also in rodent Alu families (4). In addition, because of the large amount of data here, we are confident that the region from positions 196 to 220 shows even less heterogeneity (Figures 1 and 3A). This region shows only 5% divergence in the Type II Alu family and it can also be seen that the same region in the human and galago Type I Alu families is highly conserved (Figure 3B,C). This region is found entirely within the 28 base region in galago and 31 base region in human which is not found within the left half of these Alu families (3).

The similarities and differences in the two types of galago Alu

```
                            10         20         30         40         50         60
Galago Type I       TGCCTGGGCA TGGTGGCTCX ACTCATGTAA TCCTAGCACT CTGGGXAGGC CAAGGCAGGT
                                       ** * * *   ** ***      * *   **   *  *********
Galago Type II                    TGGCTTG GCGCCTGTXX XXXXAGCACA GTGGTTAGGG CCCCAGCCAC
                                       ** * *     ** ****     * *   **   *   ********
Human               TGCTGGGXCG TGGTGGCTCX ACACCTGTAA TCCCAGCACT TTGGGXAGGC CGAGGTGGGT

                            70         80         90        100        110        120
Galago Type I       GGATTGCTTG AGCTCAGGAG TTTGAGACCA GCCTGAGXCA AGAXXGCGAG ACCCCAXTCT
                    ** *** * * * * *** *   *  **       * * *   ***  *** *   **  * **
Galago Type II      ATAXCACCTA GGGTGGTGGG TTCGAACCCA GCCCGGGCCA GCCXXXXXAA ACAACAATGA
                    ** *     * *   *** *     * **       *  *    * ******     ** ** **
Human               GGATCACCTG AGGTCAGGAG TTCAAGACCA GCCTGGXCCA ACATGGTGAA ACCCCGXTCT

                           130        140        150        160        170        180
Galago Type I       CTACTAAAAA TAGAAAAATT AGCTGGGCAT GGTGGCAGGT GCCTGTAGTC CCAGCTACTT
                     *   **  * * *     *     *    *  *     *  *
Galago Type II      CAACTGCAAC CAAAAAAAAT AGCCGGGCGT TGTGGCGGGC GCCTGTAGTC CCAGCTACTT
                     *   **  * * *     *             *      *          *           *
Human               CTACTAAAAA TACAAAAATT AGCCGGGCGT GGTGGCGCGC GCCTGTAATC CCAGCTACTC

                           190        200        210        220        230        240
Galago Type I       GGGAGGCTGA GGCAAGAGGA TCGCTTGAGC CCAAGAGTTT GAGGTTGCTG TGAGCTGTGA
                                       *           *       *     *
Galago Type II      GGGAGGCTGA GGCAAGAGAA TCGCTTAAGC CCAGGAGTTG GAGGTTGCTG TGAGCTGTGA
                                   *           * *         *          *       * *
Human               GGGAGGCTGA GGCAGGAGAA TCGCTTGAAC CCAGGAGGTG GAGGTTGCAG TGAGCCGAGA

                           250        260        270        280        290        300
Galago Type I       TXXXGCCACA GCACTCTAGC CAGGGTGACA GAGTGAGACT CTGTCTCXAA AAAAAAAAAA
                                       *          *     **                *
Galago Type II      TXXXGCCACA GCACTCTACC CAGGGGGACA GCTTGAGACT CTGTCTCTAA AAAAAAAAAA
                     ***     *       * *   *   **     ***         **    *
Human               TCGCGCCACT GCACTCCAGC CTGGGCAACA GAGCGAGACT CCATCTCXAA AAAAA
```

Figure 4. A comparison of human and galago Alu families. The consensus sequences for galago Type I (27), Galago Type II (Figure 1) and the human Alu family (3) are aligned. Both the human and Type 1 sequences are compared to the Type II consensus, with stars indicating mismatches. Deletions in one consensus versus another are marked with an X.

family and the human Alu family consensus sequences are pointed out in Figure 4. Position 138 marks what we consider as the boundary between the left and right halves of the different Alu families. The right halves are very homogeneous, with the Type II consensus sequence showing about 10% divergence from the Type I and about 16% divergence from the human Alu family. The Type I consensus sequence also shows about 16% divergence from the human sequence. Not surprisingly the two galago Alu family members are more closely related to each other than to human.

The left half of the Type II sequence does not show this close relationship with either the Type I or human Alu families (Figure 4). Attempts to optimize the homolgy by making insertions at six different positions yielded only a 50% homology between consensus sequences. This small amount of homology is largely centered in three locations. One is the A-rich region from 125 to 138 which seems to be a standard feature of interspersed repeated DNA sequences (5,6) and the other two are around 35 to 50 and 80 to 99. The latter of these areas spans the region thought to be important for RNA polymerase III promoters (10) and for reasons discussed later, we feel that this is also true of the region from 35 to 50. In spite of these similarities, the overall structure of the left halves of the Type I and II Alu families is quite different. Most striking is that when the sequences are lined up for maximum homology (Figure 4), the Type II sequence begins 14 bases within the Type I sequence. In addition there are several other regions with multiple base deletions in the Type II consensus versus the Type I consensus (positions 29 and 104 in Figure 4). These are somewhat less striking than the truncated left end because some of the individual Type I members also show variable deletions in these general regions (27). The region from position 52 to about 79 shows essentially no significant homologies between the Type II and Type I Alu family sequences. There is also a sequence near the right end of the Type II left half which includes a distinctive region with a variable number of CAA units (Figure 1, positions 83-92). Although the significance of this sequence is unknown the sequence AAACNNCAA is highly conserved and diagnostic of the galago Type II Alu family.

In the accompanying paper we showed that the galago Type I sequences contain species-specific differences relative to the human Alu family sequences. The positions discussed in that paper as being galago Type I specific are also found in the galago Type II family, at least in the right halves where a comparison can be made. However, there are several positions where Alu family type-specific differences occur in galago. Most of the differences in the consensus sequences of Figure 4 result from minor variations in one base or another and are likely not to be significant. However at position 199 (Figure 4), the Type II has an A in 18 out of 18 clones while at the equivalent position the Type I has 6 out of 7 G's with only one A. Another case is at position 144 where the Type II Alu family has 13 C's, 3T's and one A, while the Type I has 5T's, only one C and one A. There are also several less convincing positions where the consensus sequences lean towards different bases. It is clear that the two galago types are very similar in their right halves, but seem to have several distinct differences. This suggests either a relatively recent formation of the Type II Alu family relative to galago-human divergence, or a sequence correction mechanism whereby the two types of galago Alu family maintain a similar, but not identical spectrum of sequences.

DISCUSSION

Although multiple types of Alu family repeats have been found in the hamster genome (13,14) and are quite possibly common to the lower mammals, finding a new Alu family in a primate genome is unprecedented. This is particularly true since the human Alu family has been extensively studied by DNA sequence analysis without finding any alternate forms of this family of sequences.

For reasons discussed below, we believe that this Type II Alu family has arisen quite recently on an evolutionary time scale. We also believe that it has spread itself by means of a RNA intermediate which is produced using a RNA polymerase III promoter within its new sequence component. This new family not only seems to be much more active at amplifying and transposing itself than the Type I family, but also may be replacing that family within the galago genome.

The apparent lack of the galago Type II Alu family in the human genome suggests that the first Type II member was created after the divergence of the human and galago lines. Preliminary Southern blots indicate that this family is not present in the genomes of several monkeys tested, nor in another prosimian, the lemur. In contrast, the Type I sequences were clearly present before this split (27). Consistent with this observation is the relative divergence of these two families from their consensus sequence which is 14% and 17% for the Type II and Type I families, respectively. This also indicates that the Type II repeat is a more recently evolved Alu family. If we combine its more recent origin with its higher copy number in the galago cells, the Type II Alu family also appears to amplify and spread itself more efficiently than the Type I.

It appears likely that the human Alu family has amplified itself by means of RNA intermediates and that the RNA polymerase III promoter in its left half is responsible for the transcription of this intermediate (8,9). It is interesting, then, that it is the left half sequence which is different in the galago Type II family. Preliminary evidence indicates that at least three of the Type II clones, GAL 20,25 and 34, also contain active in vitro RNA polymerase III promoters (data not shown). We therefore propose that the Type II family represents the fusion of a new RNA polymerase III promoter (within the left half sequence) with the right half of a Type I Alu family sequence. The right half Type I sequence would then be transcribed in conjunction with the left half Type II sequence and allow for the rapid spread of the Type II sequence. If this model is true, then the fusion of these two sequences caused the Type II Alu family to amplify more efficiently than either of its parent sequences.

The galago Type II Alu family differs from the Type I in that the left half is appreciably shorter as well as that many portions of the left half sequences are consistently non-homologous. A large portion of the limited homology that exist between the Type I and II left halves appears to be associated with a RNA polymerase III promoter. The most essential portions of the human Alu family promoter were mapped within the region from -106 to -7 (10) which corresponds to positions 64 to 91 in Figure 4, a region which is homologous between the Type I and II Alu families. The Alu family RNA polymerase III promoter mentioned above contains a region highly homologous to the sequence G_ATCRANNC which is the consensus for the "B" box of the tRNA promoter (28-30). The tRNA promoter has been found to contain a second region, the "A" box, near the 5' end of all tRNA genes (30). The "A" box consensus sequence RRYNNARYGG (28-30) begins about 15 bases downstream from the 5'end of the transcript. In the human Alu family there is an "A" box homology 6 bases from the 5'end of its transcript (position 6 in Figure 4). This sequence is almost completely absent at the equivalent position in the galago Type II sequence. However, the Type II sequence AGCACAGTGG, starting at position 35 precisely matches the "A" box canonical sequence. Moreover, the position of this sequence corresponds to the first region of major homology between the different Alu family types (Figure 4). This gives the galago Type II Alu family a classic RNA polymerase III promoter (28-30) with an "A" box fifteen bases from the end of the repeat which we assume will also correspond to the 5'end of transcription, and a 34 base spacer to the "B" box. This suggests that all of the observed homology in the left half of the Type II Alu family can be explained as essential features of a RNA polymerase III promoter and not necessarily as any direct evolutionary relationship between Alu family members.

We had previously noted that some regions of the human Alu family sequence showed less divergence than others (3). This presumeably reflected evolutionary constraint on those regions. One of the most notable of those regions was also conserved as far back as the rodent Alu families and had homology with the papovavirus origins of DNA replication (4). We have plotted the relative divergence of the Type II Alu family as a function of position in Figure 3. We have also included the same analysis for the galago Type I (27) and the human Alu families (3). The divergence profiles are remarkably alike. First of all, the right halves of the Alu families are appreciably more homogeneous than the left halves (19% versus 8% in galago Type II, 20% versus 13% in galago Type I and 14% versus 12% in human for left and right halves of the Alu families, respectively). This suggests more stringent functional constraints maintaining the right half sequences. Secondly, there are specific regions which maintain a very high degree of homogeneity, most of which are in the right half as well. As mentioned previously, the region with homology to the papovavirus origins of DNA replication (position 153 to 195 in human Figure 3C) is well conserved. The oligo-A region at the extreme right is also very homogeneous. Perhaps the most striking homology is found in the region from 195 to 220 in the galago Type II sequences (5% mismatch). This corresponds to the sequences present in the right half of the Alu family which are not present in the human or galago Type I left half. Thus, in all of the Alu families in Figure 3, the region which we had previously described as an "insert" (3,7) shows the most remarkable sequence conservation. Since the left half of the human Alu family seems to be all that is necessary for the RNA polymerase III promoter (10), this data suggests that the right half sequences share a functionally better conserved role than the left half containing the promoter function. At present we have no firm data on what the function of these sequences may be.

One point that also suggests a correction mechanism acting between these sequences is that at some positions where the Type II sequence is very homogeneously different from the human Alu family, the Type I will be heterogeneous and have a gradient of sequences ranging from the human type of Alu family to the galago Type II. The best example of this is shown in Figure 5, which depicts the region at the junction of the left and right Alu family halves. The consensus for the galago Type II sequence (Figure 1) and the human Alu family (3) are presented along with the data on the Type I family (27). Some of the Type I clones agree well with the human sequence and others with the Type II suggesting

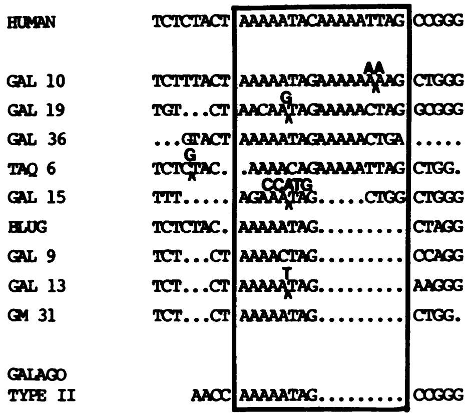
HUMAN TCTCTACT AAAAATACAAAAATTAG CCGGG
AA
GAL 10 TCTTTACT AAAAATAGAAAAAAAAG CTGGG
G
GAL 19 TGT...CT AACAATAGAAAAACTAG GCGGG
GAL 36 ...GTACT AAAAATAGAAAAACTGA .....
G
TAQ 6 TCTCTAC. .AAAACAGAAAAATTAG CTGG.
CCATG
GAL 15 TTT..... AGAAATAG.....CTGG CTGGG
BLUG TCTCTAC. AAAAATAG......... CTAGG
GAL 9 TCT...CT AAAACTAG......... CCAGG
T
GAL 13 TCT...CT AAAAATAG......... AAGGG
GM 31 TCT...CT AAAAATAG......... CTGG.
GALAGO
TYPE II AACC AAAAATAG......... CCGGG

Figure 5. A comparison of a region of the galago Type I repeats with galago Type II and human. The consensus sequence at the junction of the left and right halves of the human (top) and galago Type II (bottom) Alu families are presented. The sequence of this region from individual galago Type 1 clones is presented in the boxed region showing the region where some of the clones resemble the human Alu family and some resemble the galago Type II sequence.

that, if the human consensus represents something closer to the prototype Alu family, the Type II family is gradually converting the Type I family. The reverse does not seem to be true but, since there appears to be three times as many Type II sequences as Type I, it is not surprising that they would dominate. If there is a conversion process it is clearly very slow and inefficient as evidenced by the heterogeneity within the families and the fact that there are specific differences existing between the families. Another explanation for the data in Figure 5 might be that the change from AAAAATAGAAAAATTAG to AAAAATAG could be the result of a homologous recombination event between the two closely related halves of that sequence.

One last point worth considering, is the presence of a monomer unit in the Type II left half sequences (GAL 39, Figure 2). To our knowledge this has not been observed in primates for any other Alu family. The relatively high divergence of GAL 39 relative to the consensus does suggest that the left half sequence may exist as a family of sequences that may be evolving separately from the related Type II Alu family sequences. This clone may even represent a member of an independent repetitive DNA family, one member of which fused with a Type I Alu family member to form the original Type II family. Alternatively, monomer left half sequences may only be the result of a mistake in the transposition of the Type II member. Truncations of sequences which are amplified using a RNA intermediate are not unusual (31). The higher divergence could then be the result of a lessening of functional constraints on the Alu family member once it had lost the right half sequence. The elucidation of the relationship between monomer and dimer Alu families in the primate genome will require additional information.

ACKNOWLEDGEMENTS

We wish to acknowledge that the galago tissue was provided by the California Primate Research Center. We would like to thank Ms. A. Jennifer Diniak for her technical assistance. This work was supported by a grant from the National Institutes of Health, GM 29848.

Abbreviations:

N	any deoxynucleotide
R	purine
Y	pyrimidine

REFERENCES

1. Rinehart, F.P., Ritch, T.G., Deininger, P.L and Schmid, C.W. (1981) Biochemistry 20, 3003-3010.
2. Houck, C.M., Rinehart, F.P., and Schmid, C.W. (1979) J. Mol. Biol. 132, 289-306.
3. Deininger, P.L. Jolly, D.J., Rubin, C.M., Friedmann, T. and Schmid, C.W. (1981) J. Mol. Biol. 151, 17-33.
4. Jelinek, W., Toomey, T., Leinwand, L., Duncan, C.H., Biro, P.A., Choudary, P.L., Weissman, S., Rubin, C., Houck, C., Deininger, P.L. and Schmid, C.W. (1980) Proc. Nat. Acad. Sci., U.S.A. 77, 1398-1402.
5. Schmid, C.W. and Jelinek, W.R. (1982) Science 216, 1065-1070.
6. Jelinek, W.R. and Schmid, C.W. (1982) Ann. Rev. Biochem. 51, 813-844.
7. Deininger, P., Jolly, D., Friedmann, T., Rubin, C., Houck, C. and Schmid, C. (1980) in Mechanistic Studies of DNA Replication and Recombination (B. Alberts and C. Fred Fox, eds.) Academic Press, Inc., New York. pp. 369-378.
8. Van Aarsdell, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Maser, T. and Gestland, R.F., (1981) Cell 26, 11-17.
9. Jagadeeswaran, P., Forget, B.G., Weissman, S.M. (1981) Cell 28, 141-142.
10. Fuhrman, S.A., Deininger, P.L., LaPorte, P., Friedmann, T. and Geiduschek, E.P. (1981) Nucl. Acids Res. 9, 6439-6456.
11. Deininger, P.L. and Schmid, C.W. (1979) J. Mol. Biol. 127, 437-460.
12. Krayev, A.S., Kramerov, D.A., Skryabin, K.G., Ryskov, A.P., Bayev, A.A. and Georgiev, G.P. (1980) Nucl. Acids Res. 8, 1201-1215.

13. Haynes, S.R., Toomey, T.P., Leinwand, L. and Jelinek, W.R. (1981) Mol. Cell. Biol. 1, 573-584.
14. Haynes, S.R. and Jelinek, W.R. (1981) Proc. Natl. Acad. Sci. U.S.A. 78, 6130-6134.
15. Leinwand, L., Wydro, R. and Nadel-Ginard, B. (1982) Mol. Cell. Biol. 2, 1320-1330.
16. Krayev, A.S., Markusheva, T.V., Kramerov, D.A., Ryskov, A.P., Skryabin, K.G., Bayev, A.A. and Georgiev, G.P. (1982) Nucl. Acids Res. 10, 7461-7475.
17. Ullu, E., Murphy, S. and Melli, M. (1982) Cell 29, 195-202.
18. Ullu, E., Esposito, V. and Melli, M. (1982) J. Mol. Biol. 161, 195-201.
19. Kang, B.R., Lis, J. and Wu, R. (1979) in Methods in Enzymology (Wu, R., ed.) Vol. 68, pp. 176-183 Academic Press, New York.
20. Messing, J. and Vieira, J. (1982) Gene 19, 268-275.
21. Winter, G., Fields, S., Gait, M. and Brownlee, G. (1981) Nucl. Acids Res. 9, 237-245.
22. Messing, J., Crea, R. and Seeburg, P.H. (1981) Nucl. Acids Res. 9, 309-321.
23. Benton, W.D. and Davis, R.W. (1977) Science 196, 180-182.
24. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A. and Roe, B. (1980) J. Mol. Biol. 143, 161-178.
25. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci., U.S.A., 75, 5463-5467.
26. Staden, R. (1980) Nucl. Acids Res. 8, 3673-3694.
27. Daniels, G., Fox, M. Loewensteiner, D., Schmid, C. and Deininger, P. (1983) accompanying article.
28. Traboni, C., Ciliberto, G. and Cortese, R. (1982) EMBO J. 1, 415-420.
29. Gauss, D.H. and Sprinal, M. (1982) Nucl. Acids Res. 10, rl-r23.
30. Ciliberto, G., Raugei, G., Castanzo, R.F., Dente, L. and Cortese, R. (1983) Cell 32, 725-733.
31. Bernstein, L.B., Mount, S.M. and Weiner, A.M. (1983) Cell 32, 461-468.