Integration site preferences of the Alu family and similar repetitive DNA sequences

Gary R.Daniels and Prescott L.Deininger

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

Received 26 August 1985; Accepted 13 November 1985

ABSTRACT

Numerous flanking nucleotide sequences from two primate interspersed repetitive DNA families have been aligned to determine the integration site preferences of each repetitive family. This analysis indicates that both the human Alu and galago Monomer families were preferentially inserted into short d(A+T)-rich regions. Moreover, both primate repeat families demonstrated an orientation specific integration with respect to dA-rich sequences within the flanking direct repeats. These observations suggest that a common mechanism exists for the insertion of many repetitive DNA families into new genomic sites. A modified mechanism for site-specific integration of primate repetitive DNA sequences is provided which requires insertion into dA-rich sequences in the genome. This model is consistent with the observed relationship between galago Type II subfamilies suggesting that they have arisen not by mere mutation but by independent integration events.

INTRODUCTION

Interspersed repetitive DNA sequences have been discovered in the genomes of all vertebrate species studied to date (1-4). Many of these repetitive DNA families are present in extremely high copy number. In the case of the human Alu family, the 500,000 copies are dispersed so that, there is one every 5 kilobases on the average (5). Similarities in the structural features of a number of the short, interspersed repeated DNA sequences (SINEs, 3) indicate that they are dispersed throughout the genome by a common mechanism (6). The majority of these SINEs have a precisely defined 5' terminus, a somewhat variable oligo dA-rich 3' terminus, and are flanked by terminal direct repeats of variable lengths (Figure 1). Moreover, most of these SINE families contain an internal RNA polymerase III promoter which directs the initiation of transcription to coincide exactly with the 5' end of the repetitive DNA sequence (7-9).

Several laboratories have proposed a model using the above structural features to account for the high copy numbers and dispersal of these repetitive elements throughout the genome (10,11). The model is summarized in Figure 1 with an additional step included to reflect the analysis presented in

this paper. The first step of the proposed model for transposition of repetitive elements such as the human Alu family requires that RNA polymerase III transcription initiates precisely at the 5' end of the sequence, proceeds through an internal oligo dA-rich region, and terminates in a flanking region which codes for four or more uridine (U) residues (7). In the second step, the U-rich sequence at the 3' end of the transcript anneals to the oligo A-rich region to prime complementary DNA (cDNA) synthesis by some form of reverse transcription. Alternatively, this step may occur as a bimolecular reaction in which the Alu transcript is primed by a second RNA molecule (not shown in Figure 1). The cDNA copy of the repetitive DNA sequence is then inserted between a pair of staggered nicks at the new genomic site. In the last step of the integration process, DNA repair generates the short, direct repeats which flank each end of the repetitive sequence at its new location. Analysis of Alu family integration sites demonstrates that the flanking direct repeats have formed without a deletion of sequences that were present at the site before the integration (reviewed in 6).

The steps involved in cDNA integration into new genomic sites are not clearly defined. Recently, Van Arsdell and Weiner have proposed two possible mechanisms for the integration of U2 pseudogenes into the human genome (12). One model requires a displacement reaction between the 3' end of the cDNA and an activating group at the 5' end of the staggered nick to initially join the DNA molecules in a blunt fashion. The alternate model involves limited base pairing between the 3' end of the cDNA and a 3' overhanging region of the staggered chromosomal break. We have analyzed integration data on several primate families of SINES, including the human Alu family, and find that our data is consistent with their first model. However, there are some differences in the transposition of the SINES relative to the U2 pseudogenes with respect to the integration of the cDNA. In addition, there are some sequence preferences for the genomic integration site which help to explain some of the complexities of repeated DNA formation.

METHODS

Monomer family members were isolated from genomic DNA of Galago crassicaudatus (GAL 32,38,39) and Galago senegalensis (GSE 9,18,20, 32,36,40, 41,43,55) and cloned into M13mp8 as previously described (13). Recombinant phage were plated (14) and then screened by hybridization (15) overnight at 65°C in 5XSSC (SSC = 0.15 M NaCl and 0.015 M citric acid) with nick-translated Galago crassicaudatus DNA (GAL clones) or with nick-translated GAL 39 DNA (GSE

clones). Recombinant clones were picked and phage DNA prepared from one ml cultures (16). DNA sequence analysis was accomplished by the dideoxy termination method (17) using the [³⁵S]deoxyadenosine 5'-(α -thio)triphosphate labeling method (18). A consensus sequence for the galago Monomer family was developed by alignment of individual sequences (19). Approximately 65 bases flanking each end of individual Monomer family members including the direct repeats were compared in the present analysis.

RESULTS AND DISCUSSION

We decided to compare the sequences flanking members of several repetitive DNA families to determine whether there are specific sequence requirements or preferences for the integration of a new family member. We initially studied the direct repeats themselves because several groups have observed a strong bias toward dA residues in the direct repeats (4,6,13,20). Those results led to the analysis of flanking regions spanning 50 bases on each side of the direct repeats. Our initial sequence comparisons were focused on the human Alu family because more data is available for this family than for any other. We also analyzed our collection of sequences for an independent prosimian family, the galago Monomer (19), confirming many of the observations on the human Alu family.

The direct repeats are dA-rich at their 5'end.

We have chosen the repetitive DNA sequences and defined the direct repeat regions by several criteria. The sequences themselves were either collected by computer analysis of GENBANK or from sequencing studies being carried out in our laboratory. A small number of repetitive DNA sequences without identifiable direct repeats (6 human Alu and 4 galago Monomer family members) has been eliminated from our analysis since their integration sites cannot be precisely defined. Since both of the primate families considered in this study have a precisely defined 5' terminus, we have assumed that the 3' end of the direct repeat should be positioned next to the first base of each repetitive sequence. The 5' end of the direct repeat was more difficult to define, especially when it contained a dA-rich sequence. The problem in this case was to determine whether the dA-rich sequence should be included as part of the 5' end of the direct repeat or part of the variable oligo dA-rich region located at the 3' end of the Alu family member (see Figure 1). For consistency in this analysis, we have considered all such dA-rich stretches to be part of the direct repeat, and for reasons discussed later, we do not think it has a large effect on the conclusions.

Figure 1. Mechanism for Alu family integration into a new genomic site (10,11). A transcript of an Alu family sequence is generated by RNA polymerase III initiating transcription at the 5' terminus and terminating in a uridine-rich sequence. Reverse transcription of the self-primed cRNA produces a cDNA with a T-rich 5' terminus. The cDNA is subsequently annealed and ligated to a staggered DNA nick containing a dA-rich sequence at one of its 5' ends. DNA repair processes fill in missing nucleotides to form a newly integrated Alu family sequence flanked by tandem direct repeats (depicted as short arrows).

To determine the base composition of the direct repeats, we have aligned 32 direct repeats flanking 36 human Alu family members as shown in Figure 2A. We have positioned the 5' direct repeats (left DRs) so that their 3' ends are aligned and the 3' direct repeats (right DRs) with their 5' ends aligned. This allows us to look for features which are specific for the 3' and 5' ends, respectively, in sequences which are heterogeneous in length. The results are

striking, with a very high abundance of dA residues at the 5' end of the direct repeats. An example of the asymmetrical distribution of the base composition is demonstrated by the right DRs which are 64% dA-rich in the first 5 positions at the 5' end. This abundance of dA gradually decreases until all four bases are present in approximately equal amounts at the 3' end. In addition, all but 8 of the direct repeats actually begin with a dA residue and 18 out of the 32 have two or more dA residues at the immediate 5' end. In Figure 2B, the nucleotide composition at each position in the left and right direct repeats is plotted to show the distribution across the aligned sequences. The points for dA composition were joined to show the decreasing gradient of dA richness which occurs from the 5' to 3' end of the direct repeats. As mentioned earlier, we cannot be absolutely sure that all of these 5' dA-rich stretches should actually be included in the direct repeats. However, if they are not, then this striking abundance of dA residues must be found immediately adjacent to the direct repeats in the 5' flanking region (see discussion below).

To confirm our results on the human Alu family direct repeats, we carried out a similar analysis on the direct repeats flanking the galago Monomer family (Figure 3A and 3B). The Monomer family of repetitive sequences, which are less than half of the length of the human Alu family, have all of the structural features characteristic of SINEs and are present in high copy number (19). Analysis of twelve Monomer family members having identifiable direct repeats also demonstrates the striking dA-richness found predominantly at the 5' end of the direct repeats (Figure 3A and 3B). The only significant difference in the direct repeats flanking Alu and Monomer family members was that the latter family showed a higher degree of dA richness (55% dA for the galago Monomer compared to 44% for the human Alu family).

The 5' genomic flanking regions are d(A+T)-rich sequences.

Because the direct repeat data suggested some degree of sequence preference for integration, we extended our analysis to include flanking regions outside of the direct repeats. Figure 4 shows the nucleotide composition and alignment of 50 bases flanking each side of the human Alu family direct repeats shown in Figure 2A. This analysis of the base composition demonstrates some preference for the integration site. There is a strong bias for d(A+T)-rich sequences in the 5' flanking region at positions -1 to -10 directly adjacent to the 5' direct repeat (Figure 4). The composition of d(A+T) residues in this 10 base region averages 72% and is 5% greater than that found in the combined direct repeats. Unlike the direct

A

ALU FAMILY DIRECT REPEATS

5' direct repeat 3' direct repeat

ACTH A	AATCACTGGGTA		AATCACTGGGCA
ACTH B	AAGAAACTG		AACAAACAG
ACTH C	AAAAGACCGGGCT		AAAAGACCGGGTT
ACTH D	CACACTGA		CATACTGT
ACTH F	AATGAATATTGAG		AAAGAATATTGAA
ALPHA 2 GLOBIN	AATAAACTAAAATC		AATAAACTAAAATC
ALPHA 2 DIMER	AAAGTGATGGTTT		AAAGTCATGATTT
PSEUDO ALPHA	AAACAGTTGCGGG		AAACACTTGGGGG
3'ALPHA1 GLOBIN	TAAAAATT		TAAAAATT
3'BETA GLOBIN	AAGATCTACTCTC		AAAATCTACTCTC
DELTA GLOBIN	AAGACCTTATCCT		AAGACCTTATCT
DELTA A GAMMA	TTCTT		TTCTT
EPSILON GLOBIN A	ATCTGCAATG		ATCTGCAATA
EPSILON GLOBIN B	AAATGGATGGAGAC		AAATGGATGTAGAA
G-GAMMA GLOBIN	AAGATTCACCTGTTTAGA		AAGATTCACCTGTTTAGG
INSULIN	AAAACAAGCAGGAGA		AAAACAAGCAGGAGG
PJP53	AAAAATGTTTAGATAAG		AAAAAAGTTTAGATAAA
PROTHROMBIN A&B	AAGACCAATACCAGG		AAGACCAACCCAGG
PROTHROMBIN C	AGAAGTAAAGGACCCATG		AGAAGTAAACAACCCATC
PROTHROMBIN D	AAAAAGCAGGAGGCA		AAAAAGTGTGAGGCA
T KINASE A	AAAAGAATGTTGG		AAAAGAATGGTGC
T KINASE B	AGAAAAGCTGTA		AGAAAAACTCTG
T KINASE D	ATAAATAACTGGTTTTC		ATAAATAACTGGTTTTC
T KINASE E	AAAAAGGTACAAT		AAAAAAGTAAAAAC
T KINASE G&H	ACAAGATCAA		ACAAGATCAA
T KINASE I&J	CTTTTTTCT		CTTTTTGTGT
B TUBULIN A	GAGGA		GAGGA
B TUBULIN B	TAAAACAGGGATT		TAAAACAGGGATT
B TUBULIN C	CTCA		CTCA
B TUBULIN D	GTGGCCTCA		GTGGACGTA
B TUBULIN E&F	AGAATATT		AGAATATT
B TUBULIN J	AAATAATAACAAAT		AAATAATAACAATT

B

repeats, this d(A+T) richness is not biased strongly towards dA residues. Moreover, the relative d(A+T) richness of the 5' flanking region gradually decreases with distance from the 5' direct repeat until all four bases are present in approximately equal proportions (Figure 4). For example, positions -45 to -50 in the 5' flanking region are 43% d(A+T) while positions -1 to -5 are 72% d(A+T). Since the human genome is approximately 60% d(A+T) (21), it is possible that the first 20 to 30 bases shown in the 5' flanking region are slightly d(G+C) rich in their composition. Surprisingly, we did not observe any significant base preferences exhibited in the Alu family 3' flanking region. Apparently all bases are represented equally in this region.

An analogous alignment of the galago Monomer flanking sequences demonstrated some differences, but reinforced many of the observations made for the human Alu family flanking regions (Figure 5). In both the galago and human 5' flanking regions, DNA sequences immediately adjacent to the 5' direct repeat, had a higher d(A+T) composition than any other region excluding the direct repeats. The total d(A+T) composition in this region was 70% for galago Monomer sequences compared to 72% for human Alu family members. For both Alu and Monomer repetitive DNA sequences, we also observed a gradient of increasing dA richness in the 5' flanking region beginning 20 bases upstream from the 5' direct repeat and continuing into it. These results suggest that the preferred integration sites for human and galago repetitive sequences are extremely d(A+T) rich and extend over a 20 base region upstream from the insertion site.

Figure 2. Sequence comparison and base composition of tandem direct repeats flanking human Alu family DNA. A) The nucleotide sequences from 32 direct repeats flanking 36 human Alu family members were aligned in positive orientation to compare the nucleotide distribution at each position in the 5' and 3' direct repeats. Dotted lines between the direct repeats represent the positions of Alu family members. Four sets of two tandemly repeated Alu family sequences, Prothrombin A&B, T Kinase G&H, T Kinase I&J, and B Tubulin E&F were considered as a single sequence since each set contained only one tandem direct repeat. The nomenclature for each Alu family member refers to a human gene in close proximity to the Alu repetitive sequence. In most cases the Alu family sequence was located in the 5' and 3' flanking region or intervening sequences of the human gene. Clustered Alu family members were listed in alphabetical order according to their occurrence along the gene. Direct repeats flanking human Alu family members from the following genes were used in this analysis: ACTH (30), Alpha 2 and Alpha 2 Dimer Globin (22), Pseudo Alpha Globin (31), 3' Alpha 1 Globin (32), 3' Beta Globin (33), Delta and Delta A Gamma Globin (33,34), Epsilon Globin (35,36), G-Gamma Globin (37), Insulin (38), pJP53 (39), Prothrombin (23), Thymidine Kinase (40), and Beta Tubulin (41). B) The relative abundance of each type of nucleotide is plotted as a function of its position within either the 5' or 3' direct repeat. Symbols for nucleotides are dA(●), dG(▲), dC(□), and T(○).

Figure 3. Comparison of sequence and nucleotide composition of the galago Monomer family direct repeats. A) The nucleotide sequences of direct repeats flanking 12 Monomer family members isolated from the genomes of *Galago crassicaudatus* (GAL 32,38,39) and *Galago senegalensis* (GSE 9,18,20,32,36,40,41,43,55) were aligned in positive orientation to compare the nucleotide distribution at each position in the 5' and 3' direct repeats. Dotted lines represent the position for each Monomer family sequence. Dashes represent unknown bases in the 3' direct repeat of GSE 55 which were lost in a *Rsa* I restriction site subcloning of this sequence. B) The plot represents the relative abundance of each nucleotide as a function of its position within either the 5' and 3' direct repeat. Symbols for nucleotides are dA(●), dG(▲), dC(□), and T(○).

In contrast to the human data, the galago Monomer family 3' flanking sequences are slightly d(A+T) rich (Figure 5). The galago 3' flanking region averaged 62% d(A+T) compared to 53% d(A+T) for the human Alu 3' flanking sequences. However, we do not know whether this difference reflects a difference in the base composition of the galago genome, or some additional sequence preference of the Monomer family that is not shared with the human Alu family. We do not as yet understand the local preference for d(A+T)-rich sequence near the integration site. It is possible that such sites are more prone to the formation of staggered nicks or that local denaturation and/or breathing of DNA strands is helpful to some aspect of the integration process (Figure 1).

Both the dA-rich direct repeats and the d(A+T)-rich 5' flanking regions seem to be preferred for integration, but are definitely not required. At least for the Alu family (Figure 2A), there are several direct repeats without dA residues and some flanking sequences that are actually d(G+C) rich.

Adjacent integrations of repetitive sequences.

One conclusion that might be made from this analysis is that regions of the genome that are generally rich in d(A+T) residues would be more likely to have preferred integration sites. In addition, since many SINEs contain an oligo dA-rich 3' end, one might expect this oligo-dA region to serve as a preferential integration site for a second repetitive DNA sequence of the same or different type. This appears to be substantiated by the relatively high abundance of human Alu family members which have integrated in a tandem fashion (22,23). In our analysis of 36 human Alu family members, four had integrated into the dA-rich region of another, forming a tandem Alu family dimer. Moreover, there are numerous examples of Alu family members and other SINEs integrating adjacent to each other in a similar manner (reviewed in 4). Many of these adjacent repetitive sequences share a set of direct repeats which flank the pair suggesting that they transposed as a single unit into the new site. Thus, it is possible that the process of preferential integration into dA-rich sequences could lead to the fusion of repeats into larger repetitive units. This is almost certainly the origin of the dimeric structure found for the human Alu family (24).

Since dA-rich regions are preferred integration sites, insertion events could also occur in the dA-rich middle region of a human Alu family sequence as well as the dA-rich 3' end. Examples of an internal integration into an Alu family member have been demonstrated for both human (25) and galago (13) SINE families. The galago Type II Alu family is a composite structure

VALU FAMILY 3' FLANKING SEQUENCE

[illegible]

containing a Monomer-like sequence of about 100 bases at its 5' terminus and a typical Alu family (called Type I in galago, 20) right-half sequence of approximately 160 bases at its 3' end (see Figure 6). In a previous report (13), we have speculated that the Type II Alu family may have arisen by the independent integration of a Monomer sequence adjacent to the right half of a Type I Alu family sequence. We had also observed that the Type II Alu family contained distinct subfamilies of sequence which shared common point mutations, insertions, and deletions relative to the Type II consensus sequence (Figure 6).

Analysis of that data in light of the findings in this paper suggests that the subfamilies actually represent independent integrations of a "Monomer-like" repeat into the central dA-rich region of Type I Alu family members forming two very closely related, but independent families of Type II sequence. Figure 6 shows a portion of the sequence data for the two major Type II subfamilies and a schematic to demonstrate the Type II family formation. The most prominent difference between these subfamilies is found around the junction of the left (Monomer-like sequence) and right halves (under-lined sequence homologous to the Type I right half as shown in Figure 6). At position 108 in the Type II consensus sequence both the Type II_A and Type II_B subfamilies have deletions relative to the consensus which extend upstream for 2 and 14 bases, respectively. The precise location of these deletions relative to the dA-rich Type I right-half sequence homology (positions 109-116) is completely consistent with two separate and independent integrations of a Monomer family member into a galago Type I Alu family sequence to form two separate subfamilies, Type II_A and Type II_B (see Figure 6). In the case of the Type II_B subfamily, a shorter dA-rich region has

Figure 4. Nucleotide composition of sequence flanking human Alu family members. The nucleotide sequences of 36 human Alu family members containing tandem direct repeats (see Figure 2) were aligned in positive orientation to determine the composition of their flanking regions. The alignment consists of the first 50 nucleotides flanking the direct repeats on the 5' and 3' sides of each Alu family member. Four sets of two tandemly repeated Alu family sequences were considered as a single flanking sequence since each set contained only one tandem direct repeat (see Figure 2). This analysis did not include adjacent Alu family sequences which were present in some flanker regions. The percent composition for each nucleotide was calculated in blocks of 5 bases each along the sequence and are tabulated below the block in which they were counted. The d(A+T) richness of each block is also shown. The base composition of the 5' and 3' flanking direct repeats which average about 12 bases per Alu family sequence were counted as a single unit to give their overall composition. The nomenclature used to describe each human Alu family sequence was given in Figure 2.

MONOMER FAMILY 5' FLANKING SEQUENCE													MONOMER FAMILY 3' FLANKING SEQUENCE												
	-50	-40	-30	-20	-10	5' dir. rep.	3' dir. rep.	+10	+20	+30	+40	+50													
GAL 32	ACTCTGCAT	CATTTAATT	TTAGTATATC	TGTTCTGAA	GTGACACAA			AAATAAAGA	AAACACAC	AAAAAAGG	T														
GAL 36	CTCCGATAG	ACATTAAGT	AGACAGCAT	CATTCTTAAT	AAAAAGATC			CAAAAAAGA	ATGTCACG	GGT															
GAL 38	ACATCT	CAATCTAGC	AGATATGAA	TATTAATGTC	TTACACAAAT			GTGAAGCCT	ATATTACCTA	ATTGATCTA	AAACATTCGA	ATTCTATATA													
GAL 39	ACATCT	CAATCTAGC	AGATATGAA	TATTAATGTC	TTACACAAAT			GTGAAGCCT	ATATTACCTA	ATTGATCTA	AAACATTCGA	ATTCTATATA													
GSE 18	CTCCGATAG	ACATTAAGT	AGACAGCAT	CATTCTTAAT	AAAAAGATC			CAAAAAAGA	ATGTCACG	GGT															
GSE 19	CTCCGATAG	ACATTAAGT	AGACAGCAT	CATTCTTAAT	AAAAAGATC			CAAAAAAGA	ATGTCACG	GGT															
GSE 20	AAAGAAATT	CAACAAAGG	CTCGACACAG	TTCTGACAG	AAATTAATG			CTTAATAATA	ATTTTAATCT	TTGTAAGT	TTCCCAAGGA	CAATCTGTAT													
GSE 32	AAATTAAT	ATTCACTTT	CTAAACACAA	AAATTAATG	CTACTTAAG			TAATGCTATG	TTCTATGAT	TTTCTCTCC	TTGTTATATA	TAGGT													
GSE 36	AAATTAAT	ATTCACTTT	CTAAACACAA	AAATTAATG	CTACTTAAG			TAATGCTATG	TTCTATGAT	TTTCTCTCC	TTGTTATATA	TAGGT													
GSE 40	AAATTAAT	ATTCACTTT	CTAAACACAA	AAATTAATG	CTACTTAAG			TAATGCTATG	TTCTATGAT	TTTCTCTCC	TTGTTATATA	TAGGT													
GSE 41	AAATTAAT	ATTCACTTT	CTAAACACAA	AAATTAATG	CTACTTAAG			TAATGCTATG	TTCTATGAT	TTTCTCTCC	TTGTTATATA	TAGGT													
GSE 42	AAATTAAT	ATTCACTTT	CTAAACACAA	AAATTAATG	CTACTTAAG			TAATGCTATG	TTCTATGAT	TTTCTCTCC	TTGTTATATA	TAGGT													
GSE 43	AAATTAAT	ATTCACTTT	CTAAACACAA	AAATTAATG	CTACTTAAG			TAATGCTATG	TTCTATGAT	TTTCTCTCC	TTGTTATATA	TAGGT													
GSE 44	AAATTAAT	ATTCACTTT	CTAAACACAA	AAATTAATG	CTACTTAAG			TAATGCTATG	TTCTATGAT	TTTCTCTCC	TTGTTATATA	TAGGT													
XAAT	60	64	62	60	64	58	70	66	63	77	79	79	73	62	67	64	66	53	54	64	57	56			

Figure 5. Comparison of sequence and nucleotide composition of regions flanking galago Monomer family members. The nucleotide sequences of 12 Monomer family members were aligned in positive orientation to determine the base composition of their flanking regions as was described in Figure 4. The direct repeats which average 13 nucleotides per Monomer sequence were counted together to determine their overall nucleotide composition. The sequence for each direct repeat is shown in Figure 3A.

Figure 6. Evidence for integration of two related sequences into the central dA-rich region of a galago Type I Alu family sequence. The partial nucleotide sequences of several galago Type II Alu family members are aligned to show homology to the Type II consensus sequence (CONS) from positions 51 to position 150. The sequences are grouped as subfamilies to show similar changes from the consensus sequence. The Type II_A subfamily contains sequences GAL 6,7,16, 21,26,27, and 33 and the Type II_B subfamily has GAL 4,20,35, and 40 as members (13). Similarities to the consensus sequence are indicated by a dot, differences are shown by placing the proper nucleotide at that position. Insertions are placed above an arrowhead at the position in which they appear and deletions are indicated by an X at the position. Asterisks indicate positions which are consistently different for each subfamily when compared to the Type II consensus sequence. Nucleotides in the Type II consensus sequence which are homologous to the Type I Alu family right-half sequence are underlined. Below the Type II subfamily sequence is a schematic of the structure which we propose to have resulted in the formation of the Type II family. Our model requires the integration of a Monomer family member into the central oligo-dA rich region of a galago Type I Alu family member. The resulting composite structure becomes a Type II Alu family after transcription initiates at the Monomer family promoter with subsequent cDNA formation and integration of the Type II sequence (see Figure 1). The Type I Alu family left-half sequence is not carried in transcripts initiated by the intragenic Monomer RNA polymerase III promoter.

integrated with the Monomer sequence. Although we cannot completely rule out some sort of deletion mechanism as having caused the variation at the Type II family junction, the precise location of the deletions at the junction between Monomer and Type I homologous sequences argues in favor of an integration mechanism for Type II subfamily formation. Two similar insertion events occurring at approximately the same location provides further evidence that

integration occurs preferentially in oligo-dA rich regions. In addition, encounters between the Alu Type I and Monomer families may have been enhanced by the high copy numbers of each sequence in the galago genome (Daniels and Deininger, unpublished).

Implications for the transposition mechanism.

The presence of dA-rich sequences in the direct repeats of Alu family members has previously been observed (4,6,10,11,13) and a correlation with the oligo-dA region at the 3' end of the Alu family has suggested that this region may play some role in the integration mechanism (10,11). Our finding that the dA-richness resides almost exclusively at the 5' end of the direct repeats makes this putative role even more plausible. As shown by the schematic diagram in Figure 1, our data on SINE family integration preferences suggests that the interaction between DNA molecules often takes place as a direct hybridization between the 5' end of the direct repeat and the 5' end of a single-stranded cDNA. This hybridization could then stabilize the interaction while repair processes link the DNA molecules. A similar mechanism has recently been proposed for the integration of RNA polymerase II transcribed, processed pseudogenes (26,27). Thus, direct binding of DNA species by hybridization may be a common mechanism for the integration of RNA-mediated transpositions. In further support of such a direct interaction, we note that the goat C family does not have a dA-rich 3' end and the flanking direct repeats are not dA-rich (28). An alternative possibility for Alu family integration might be that the T-rich 3' end of a staggered DNA break (complementary strand to the dA-rich end) could prime reverse transcription directly on the Alu family RNA. The proposed ability of Alu family RNA to self-prime (see Figure 1) and the observation of self-priming in U3 pseudogenes (29), makes this less likely to be the major mechanism.

There is a distinct difference between our data on Alu family integration and the data on U2 pseudogenes which lack a 3' dA-rich region (12). All six U2 pseudogenes had an intact 5' end, but variable deletions were seen at the 3' end of the U2 sequence. These observations suggested that the linkage at the 5' end of the transposing cDNA occurred first, and that variable deletions occurred at the downstream end either during cDNA formation or as it subsequently integrated. Moreover, the same authors suggested that short regions of homology between the U2 cDNA and the downstream end of the direct repeat may occasionally limit the 3' deletions seen in these U2 pseudogenes (12). Our data are not inconsistent with this model, as an upstream linkage may still be the first event to occur, although the nature of that linkage

must remain highly speculative. Our findings do suggest the importance of homology at the downstream end of the Alu family member with the direct repeat. The 3' dA-rich tail of Alu family members and most other high copy number SINES may insure consistent hybridization to the downstream direct repeat. This dA-rich region may not only limit deletions, but may also explain the high efficiency of Alu family transposition relative to the U2 pseudogenes (in excess of 100,000 copies versus 1000 copies or less, respectively).

ACKNOWLEDGEMENTS

This work was supported by USPHS grant GM29848. We would like to thank Kathleen Atkinson for her technical assistance and Harvey Bradshaw and Vicki Traina-Dorge for use of their data before publication.

REFERENCES

1. Jelinek, W.R. and Schmid, C.W. (1982) *Annu. Rev. Biochem.* 51, 813-844.
2. Schmid, C.W. and Jelinek, W.R. (1982) *Science* 216, 1065-1070.
3. Singer, M.F. (1982) *Int. Rev. Cytol.* 76, 67-112.
4. Rogers, J. (1985) *Int. Rev. Cytol.* 93, 185-200.
5. Rinehart, F.P., Ritch, T.G., Deininger, P.L., and Schmid, C.W. (1981) *Biochemistry* 20, 3003-3010.
6. Schmid, C.W. and Shen, C.-K.J. (1986) in *Molecular Evolutionary Genetics*, MacIntyre, R.J., Ed., Plenum Press, New York.
7. Duncan, C.H., Jagadeeswaran, P., Wang, R.R.C., and Weissman, S.M. (1981) *Gene* 13, 185-196.
8. Elder, J.T., Pan, J., Duncan, C.H., and Weissman, S.M. (1981) *Nucleic Acids Res.* 9, 1171-1189.
9. Fuhrman, S.A., Deininger, P.L., LaPorte, P., Friedmann, T., and Geiduschek, E.P. (1981) *Nucleic Acids Res.* 9, 6439-6456.
10. Jagadeeswaran, P., Forget, B.G., and Weissman, S.M. (1981) *Cell* 26, 141-142.
11. Van Arsdell, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Manser, T., and Gesteland, R.F. (1981) *Cell* 26, 11-17.
12. Van Arsdell, S.W. and Weiner, A.M. (1984) *Nucleic Acids Res.* 12, 1463-1471.
13. Daniels, G.R. and Deininger, P.L. (1983) *Nucleic Acids Res.* 11, 7595-7610.
14. Messing, J., Crea, R., and Seeburg, P.H. (1981) *Nucleic Acids Res.* 9, 309-321.
15. Benton, W.D. and Davis, R.W. (1977) *Science* 196, 180-182.
16. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A., and Roe, B. (1980) *J. Mol. Biol.* 143, 161-178.
17. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 75, 5463-5467.
18. Biggin, M.D., Gibson, T.J., and Hong, G.F. (1983) *Proc. Natl. Acad. Sci. USA* 80, 3963-3965.
19. Daniels, G.R. and Deininger, P.L. (1985) *Nature* (in press).
20. Daniels, G.R., Fox, G.M., Loewensteiner, D., Schmid, C.W., and Deininger, P.L. (1983) *Nucleic Acids Res.* 11, 7579-7594.

21. Thiery, J.-P., Macaya, G., and Bernardi, G. (1976) *J. Mol. Biol.* 108, 219-235.
22. Hess, J.F., Fox, M., Schmid, C., and Shen, C.-K.J. (1983) *Proc. Natl. Acad. Sci. USA* 80, 5970-5974.
23. Degen, S.J.F., MacGillivray, R.T.A., and Davie, E.W. (1983) *Biochemistry* 22, 2087-2097.
24. Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T. and Schmid, C.W. (1981) *J. Mol. Biol.* 151, 17-33.
25. Hammarstrom, K., Westin, G., Bark, C., Zabielski, J., and Petterson, U. (1984) *J. Mol. Biol.* 179, 157-169.
26. Moos, M. and Gallwitz, D. (1983) *E.M.B.O.J.* 2, 757-761.
27. Vanin, E.F. (1984) *Biochim. Biophys. Acta* 782, 231-241.
28. Spence, S.E., Young, R.M., Garner, K.J., and Lingrel, J.B. (1985) *Nucleic Acids Res.* 13, 2171-2186.
29. Bernstein, L.B., Mount, S.M., and Weiner, A.M. (1983) *Cell* 32, 461-472.
30. Tsukada, T., Watanabe, Y., Nakai, Y., Imura, H., Nakanishi, S., and Numa, S. (1982) *Nucleic Acids Res.* 10, 1471-1479.
31. Sawada, I., Beal, M., Shen, C.-K.J., Chapman, B., Wilson, A., and Schmid, C. (1983) *Nucleic Acids Res.* 11, 8087-8102.
32. Shen, C.-K.J. and Maniatis, T. (1982) *J. Mol. Appl. Genet.* 1, 343-360.
33. Poncz, M., Schwartz, E., Ballantine, M., and Surrey, S. (1983) *J. Biol. Chem.* 258, 11599-11609.
34. Maeda, N., Bliska, J.B., and Smithies, O. (1983) *Proc. Natl. Acad. Sci. USA* 80, 5012-5016.
35. Baralle, F.E., Shoulders, C.C., Goodbourn, S., Jeffreys, A., and Proudfoot, N. (1980) *Nucleic Acids Res.* 8, 4393-4404.
36. DiSegni, G., Carrara, G., Tocchini-Valentini, G.R., Shoulders, C.C., and Baralle, F.E. (1981) *Nucleic Acids Res.* 9, 1151-1170.
37. Shen, S., Slightom, J.L., and Smithies, O. (1981) *Cell* 26, 191-203.
38. Bell, G.I., Pictet, R., and Rutter, W.J. (1980) *Nucleic Acids Res.* 8, 4091-4109.
39. Pan, J., Elder, J.T., Duncan, C.H., and Weissman, S.M. (1981) *Nucleic Acids Res.* 9, 1151-1170.
40. Flemington, E., Bradshaw, H., Traina-Dorge, V., Slagel, V., and Deininger, P.L. (manuscript submitted).
41. Lee, M.G.-S., Loomis, C. and Cowan, N.J. (1984) *Nucleic Acids Res.* 12, 5823-5836.