Recently transposed Alu repeats result from multiple source genes

A.Gregory Matera, Utha Hellmann, Mary F.Hintz and Carl W.Schmid
Department of Chemistry, University of California, Davis, CA 95616, USA

Received July 2, 1990; Revised and Accepted September 12, 1990

EMBL accession nos X53548 – X53550 (incl.)

ABSTRACT

A human Alu repeat subfamily (the PV subfamily) whose members include insertional polymorphisms is found, as predicted, to differ by five tightly linked mutations relative to another subfamily of recently inserted Alu repeats. Based on these sequence differences some of the small number of polymorphic Alus can be selected from the background of nearly one million member sequences which are fixed in the human genome. Shared patterns of mutations suggest that PV subfamily members are the progeny of several different founder sequences. The additional observation that all members of the PV subfamily end in a stretch of uninterrupted polyadenine residues rather than merely A-rich sequences is evidence for post-transcriptional polyadenylation of the presumptive RNA intermediate. The drift of polyadenine sequences toward tandemly repeated A-rich motifs suggests a biological function that may select for the fixation of dispersed Alu repeats.

INTRODUCTION

The human Alu family of repetitive DNA elements consists of about 0.5 million dispersed members sharing a recognizable 280 nt consensus sequence (Reviews 1–3). Structural evidence strongly suggests that these repeats are retrotransposons dispersed via an RNA intermediate, but the putative transcript has, until recently, been either elusive or controversial (1–4). Moreover, most human Alus are essentially immobile. Many members predate human and ape or even monkey divergence (1–3).

Scrutiny of the many Alu sequences accumulated in data banks have led to essentially identical conclusions by five independent groups (5–9): Alu repeats can be segregated into recognizable subfamilies, each of which is presumably the progeny of a distinct set of founder sequences. The different degrees of divergence exhibited by the members of each subfamily suggests that they appeared at different evolutionary times. Confirming this interpretation, the few human Alus which are known to be polymorphic insertions into the human genome belong to a subfamily which is evolutionarily the most recently inserted, as judged by the sequence homogeneity of its members. Although five independent groups agree on these broad conclusions, there are some minor differences of opinion regarding the exact sequence assignment of each subfamily. For our present purposes, the consensus sequence of the subfamily identified by Deininger

and Slagel (10), which is also identical to the 'precise' subfamily defined by Britten et al. (11), is most informative. This definition can be regarded as a consensus of the several versions proposed for the recent Alu subfamily (5–10). The conclusions of this investigation do not depend on the choice of consensus sequence.

Previous sequence comparisons suggested that two polymorphic human Alu repeats might differ by five shared mutations from the precise subfamily consensus sequence, thereby constituting another discernable subfamily (3,10). This interpretation has been partially confirmed by using an oligonucleotide hybridization probe that incorporates two of these five putative diagnostic differences (4). One Alu repeat selected by library screening under stringent hybridization with this probe is associated with a 300 bp restriction fragment length polymorphism (RFLP). While not yet decisively proven, this RFLP almost certainly results from an Alu insertion. This group of variant Alu sequences has recently expanded in human evolution (i.e. comparing human to chimpanzee and gorilla) and corresponds to one or more transcriptionally active source genes (4).

The five mutations which distinguish these Alu variants from the recently fixed precise subfamily should be tightly linked. Moreover, independently isolated polymorphic Alus should contain these same tightly linked mutations. Here we test and confirm both predictions by sequence analysis. For simplicity, we shall refer to Alus which share these linked mutations as predicted variant (PV) Alus. In predicting this subfamily, we recognize that some PV Alus are fixed in the human genome, possibly predating ape and human divergence, and that not all polymorphic Alus need belong to this group (see Discussion). Rather, these PV Alus result from recognizable source genes that have been active in recent human evolution.

MATERIALS AND METHODS

Library construction and screening

A human λ DASH genomic library (Stratagene) was screened using an Alu oligonucleotide hybridization probe GM-002 at high stringency (67°C and 5×SSPE) (4). This stringency corresponds to the melting temperature of an exactly paired duplex with GM-002. GM-002, which is shown below, incorporates two of the five diagnostic mutations of the proposed PV subfamily. Sequences of the resulting clones PV71, PV83, and PV93, are reported in the Results section. Poly A plus RNA from a human hydatidiform mole was isolated using the 'Fast Track' mRNA isolation kit (Invitrogen). A *Not I*/*Eco* RI-linked 1 GT10 cDNA

10	20	30	40	50	60	70	
GGCCGGGCGC	GGTGGCTCAC	GCCTGTAATC	CCAGCACTTT	GGGAGGCCGA	GGCGGGCGGA	TCACGAGGTC	Precise
.T.....							MLvi
.....							PV 92
.....					T.....		APO
.....							TPA
.....							PV 71
Truncated ...							PV 83
.....							PV 6
GGCCGGGCGC	GGTGGCTCAC	GCCTGTAATC	CCAGCACTTT	GGGAGGCCGA	GGCGGGCGGA	TCACGAGGTC	PV Consensus
.....							C1 INH
80	90	100	110	120	130	140	
AGGAGATCGA	GACCATCCTG	GCTAACACGG	TGAAACCCCG	TCTCTACTAA	AAATA-CAAAA	AATTAGCCGG	Precise
.....	C.....	A.....	-.....		A.....		MLvi
.....	C.....	A...C	T.....		-.....		PV 92
.....	C.....	A.....			-.....		APO
.....	C.....	A.....			C-.....	=.....	TPA
.....	C.....	A.....			C-.....	=.....	PV 71
.....	C.....	A.....			-.....		PV 83
.....	C.....	A.....			-.....		PV 6
AGGAGATCGA	GACCATCCCG	GCTAAAAACGG	TGAAACCCCG	TCTCTACTAA	AAATA-CAAAA	AATTAGCCGG	PV Consensus
.....	T.....	C.....		[A ₁₀].....	-.....		C1 INH
150	160	170	180	190	200	210	
GCGTGGTGGC	GGGCGCCTGT	AGTCCCAGCT	ACTCGGGAGG	CTGAGGCAGG	AGAATGGCGT	GAACCCGGGA	Precise
-...A..A..	..-...=...		...T.....				MLvi
...A.....	=...		...T.....				PV 92
...A.....	=...		...T.....				APO
...A.....		...TG...	...T.....		...A.....		TPA
...A..=...		...T.....	...T.....				PV 71
.....=...			...T.....			...T.....	PV 83
...A..=...	...T.....		...T.....				PV 6
GCGTAGTGGC	GGGCGCCTGT	AGTCCCAGCT	ACTTGGGAGG	CTGAGGCAGG	AGAATGGCGT	GAACCCGGGA	PV Consensus
...G..A..			...C.....		G.....		C1 INH
220	230	240	250	260	270	280	
GGCGGAGCTT	GCAGTGAGCC	GAGATCGCGC	CACTGCACTC	CAGCCTGGGC	GACAGAGCGA	GACTCCGTCT C	Precise
.....--...		C...					MLvi
.....		CT..					PV 92
.....		C...					APO
.....		C...			A.....		TPA
.....		C...		T.....		...A...	PV 71
.....		C...					PV 83
.....		C...					PV 6
GGCGGAGCTT	GCAGTGAGCC	GAGATCCCGC	CACTGCACTC	CAGCCTGGGC	GACAGAGCGA	GACTCCGTCT C	PV Consensus
.....CG..		G...				-	C1 INH

Figure 1. PV subfamily Alu sequences are compared to the consensus of the precise subfamily (7,9). These Alus share 5 diagnostic mutations relative to the precise subfamily consensus sequence. The sequences of the MLvi, TPA and PV92 Alus were previously published (4,14,15). The APO sequence was determined using a subclone of the region flanking the human apolipoprotein A1 gene (12). The other PV subfamily Alus were isolated as described in Methods. Sequence positions that are underlined identify mutations from the PV consensus that are shared by other PV family members. The sequence of the C1 INH Alu (19) is compared to the PV subfamily consensus. The location of the hybridization probe GM2 (position 76 to 97) is underlined.

library (insert size 500nt and larger) was constructed and amplified from this RNA by Invitrogen, Inc. The library was screened using standard procedures (4) with end-labeled oligonucleotide GM-002 at 60°C in 5×SSPE. Clone PV 6 was isolated from this library and subcloned as described below.

Plasmid construction and DNA sequencing

Plasmid subclones of Alu family members PV 83 and PV 71 were prepared as follows: a 2.0kb *Bst* YI and a 1.3kb *Hinc* II/*Bst* YI fragment from PV 83, a 1.7kb *Dra* I/*Bst* YI and a 0.5kb *Dra* I/*Bst* YI fragment from PV 71. PV 6 was subcloned into pUC

TATGAGATTA <u>GAACACTAC</u>	[ALU] (A) ₂₂	<u>GAACACTAC</u>	ACATTACTGA	TPA
AGGCTGAGAT <u>GAAAATGT</u>	[ALU] (A) ₃₁	<u>GAAAATGT</u>	CATCTGTCCT	MLvi
GGCTACAGAG <u>GAAAGAA</u>	[ALU] (A) ₂₈	<u>GAAAGAA</u>	TTCCCTCTCT	PV 92
CTTCNCAAAA <u>GAAGACATTTATG</u>	[ALU] (A) ₂₉	<u>GAAGACATTTATG</u>	CAGCCAAAAA	PV 71
GCCTTCTCAA <u>GAAGGGCCTTCT</u>	[ALU] (A) ₂₀	<u>GAAGGGCCTTCT</u>	AAAACCTCTC	APO
GAGGCTTAAA <u>GAAGGCTTAAA</u> TTGG	*[ALU] (A) ₁₅	<u>GAAGGCTTAAA</u>	AAAATACTGG	PV 83
<u>GAARR (N) x</u>	[ALU] (A) _n	<u>GAARR (N) x</u>		Consensus

Figure 2. The sequences flanking PV Alus. The sequences of the short direct repeats are underlined. Note that unlike most Alu repeats, the A-rich tails are composed only of adenine residues. All direct repeats begin with the dinucleotide GA, with a further preference for 5'-GAARR(N)_x-3'.

19 from a λ GT10 cDNA clone as a 650bp *Eco* RI insert. A 2.4kb *Hind* III subclone of the region flanking the human Apolipoprotein A1-C3-A4 gene cluster was a generous gift of Dr. S. Karathanasis (12,13). A 0.6 kb *Sma* I fragment was subcloned from the original *Hind* III subclone to obtain sequence overlaps. Sequencing was performed on pUC dsDNA with Sequenase 2.0 (USB) using forward, reverse and three different Alu oligonucleotide primers (GM-002, 5'-ATCGAGACCAT-CCCGGCTAAAA-3'; GM-003, 5'-GGTTTCACCGTTTGA-GCCG-3'; GM-004, 5'-GCAGTGAGCCGAGATCC-3' [Synthecell, Inc]). Alignment of sequences was accomplished using the MicroGenie data analysis software (Beckman).

RESULTS

PV Alus share diagnostic sequence features

The base sequences of polymorphic Alus mapping within the MLvi-2 and TPA loci, as well as that of clone PV 92, have been previously published (4,14,15) (Fig. 1). As mentioned, these Alus share five differences with respect to the precise subfamily sequence. Several investigators have described an Alu insertion-deletion polymorphism mapping 5' to the human apolipoprotein A1 gene (12,13,16). The base sequence of a subclone of this region identifies the presence of an Alu repeat that includes all five diagnostic mutations (APO, Fig. 1). A comparison of the restriction sites within this Alu repeat to those used to map the length polymorphism confirms that the observed polymorphism results from an insertion of this Alu repeat.

Four Alu repeats designated 'PV' in Figure 1 were selected by library screening under stringent hybridization conditions with an Alu oligonucleotide (GM-002, Materials and Methods). PV 92 is a probable insertional polymorphism in the human genome and is absent from chimpanzee DNA, as judged by restriction fragment length polymorphisms (4). PV 83 Alu is truncated, having an seventeen nucleotide deletion on its 5' end, but is bounded by direct repeats. Thus, PV 83 presumably results from the insertion of an incomplete cDNA. Interestingly, primer extension of Alu RNA was also found to prematurely terminate at this position (4). This Alu is fixed in a panel of eight humans (4). PV 6 is an incomplete cDNA clone in which the Alu is inverted with respect to the sense of the original transcript. The end of the cDNA occurs at position 270 so that ten terminal nucleotides are missing, none of which are diagnostic (Fig. 1).

Six of the seven predicted variant Alus share all five diagnostic mutations (Fig. 1). PV 83 has four of the diagnostic mutations; the only exception, G at position 145, matches the precise subfamily consensus. While these Alus share the five diagnostic mutations, there are clear differences between members (Fig. 1).

The average pairwise divergence of these Alus is approximately seven nucleotides; they differ by an average of four mutations from the PV consensus. Based on estimates of the fidelity of reverse transcriptase and the mutational rate of non-selected DNA, it is unlikely that these sequences acquired all of their observed differences post-transcriptionally (17,18). Some of the differences among these Alus might result from sequencing artifacts; however, the diversity of these newly inserted Alus is consistent with the possibility of several distinct source genes. The non-random pattern of mutations is a stronger indication that discernable subgroups were encoded by distinct source genes. As the best example, TPA and PV 71 Alus share three differences: C at position 123, a deletion at position 134 and a T at position 166, relative to the other PV subfamily members (Fig. 1). Two other possible subgroupings are MLvi and PV 92 which share a deletion at position 157, and PV 71, PV 83, and PV 6 which share a deletion at position 148 (Fig. 1). However, a single deletion, particularly one occurring near a GC rich run, is not as convincing as the three mutations linking the TPA and PV 71 subgroup. The suggestion that several distinct source genes generated newly inserted Alu repeats is strengthened by the existence of an Alu (C1 INH) that clearly belongs to the precise rather than the PV subfamily (Fig. 1, 19).

The 3' ends of PV Alus are polyadenylated

The 3' ends of Alu repeats are typically A-rich suggesting that they, like the 3' ends of other retrotransposons, result from post-transcriptional polyadenylation of an RNA intermediate (1-3). However, most Alu repeats do not have the usual recognizable polyadenylation signals, and their 3' ends are not exclusively adenine but include other nucleotides often arranged in simple repeating structures (1, 2).

Figure 2 compares the ends of PV Alus. These heterogeneous-length sequences are composed exclusively of adenine; exactly the type of products expected for polyadenylation of RNA intermediates. The C1 INH Alu insertional polymorphism which does not belong to the PV subfamily (Discussion) also ends in a run of forty-two A's (19). Thus, the simple repeats in the A-rich 3' ends of other Alu members must result from post-insertional events and are not encoded by the Alu source gene(s). These findings complement the observation of a high degree of polymorphism associated with the poly A tail of fixed Alu repeats (20).

PV Alus insert at a defined target sequence

The direct repeats surrounding Alu repeats are typically also A-rich and often include similarity to the 5' end of the Alu insert (1). However, because of their mutational divergence and

similarity to the simple repeating structures within the A-rich 3' ends, the exact structure of the direct repeats is often ambiguous (1). All direct repeats surrounding PV Alus include the sequence 5'-GAN_x-3' (Fig. 2). At present, we do not know the molecular basis for this striking target site specificity.

DISCUSSION

Sequence comparisons have compelled the conclusion that Alu repeats can be segregated into subfamilies resulting from distinct source genes (5–9). Further, the shared sequence features of two previously known polymorphic Alus strongly indicates that the currently active source gene(s) will have five distinct base substitutions compared to the most recently fixed subfamily (3,4,7,9,10). This hypothesis is experimentally tested and confirmed by two complementary approaches. A polymorphic Alu mapping near the apolipoprotein gene cluster is found to have the five linked base substitutions. Diagnostic polymorphic restriction cleavage sites reported for this insertional polymorphism (12,13,16) originally drew our attention to the possibility that a PV Alu was responsible for the polymorphism. Conversely, using an oligonucleotide that incorporates two of the five diagnostic mutations to isolate other PV Alus, we find that all five mutations are tightly linked. As reported by Matera et al. (1990), at least one of the four PV Alus described here is associated with a human DNA length polymorphism (4). Thus, we can selectively enrich for a small number of newly inserted Alus against a background of nearly one million Alu repeats which are fixed in the human genome.

When compared to the sequence diversity of the entire human Alu family, these new Alus obviously result from a select source gene or genes. Although sequence differences between PV subfamily members might reflect multiple source genes, these differences might also result from an error prone insertion mechanism, random drift or simple sequencing errors. The present finding that PV Alus can be further resolved into recognizable sub-subfamilies strongly supports the existence of multiple but closely related source genes. This interpretation is confirmed by the existence of a polymorphic Alu that closely matches the precise subfamily consensus sequence and does not share any of the tightly-linked PV Alu mutations (19). It is notable that 4.5S RNA, an Alu homolog in rodents is encoded by multiple genes that are arranged in tandem (21). However, there is but a single locus for another Alu homolog, the human 7SL RNA gene (22).

Rosenberg et al. (23) recognized that a consensus sequence for a repeat sequence family is merely an average, which need not exactly match any individual member sequence. Remarkably, the APO Alu differs by only a single nucleotide from the PV consensus sequence. The PV consensus sequence identified here is likely an exact match to one of the source genes encoding new members of this subfamily.

A-rich 3' ends of Alu repeats are often highly structured, consisting of tandem repeats of simple A-rich elements (1, 2). This observation and the absence of conventional polyadenylation signals prompted the suggestion that the A-rich ends might be encoded by Alu genes (2). In contrast, 3' ends of predicted variant Alus are composed of variable-length tracts of pure adenine, which suggests their addition via a post-transcriptional mechanism. Tandemly repeated elements which are present in the 3' A-rich ends of many Alu repeats must form subsequently to their post-transcriptional polyadenylation. Tandem A-rich elements in many instances resemble the short direct repeats

flanking the Alu repeats and, apparently, 'diffuse' into the A-rich tract from the 3' end (1). This interpretation is confirmed by the additional observation that many Alu tails begin with a 5' run of pure A's and end with a 3' region consisting of tandemly repeated A-rich elements (20). A high frequency of length polymorphism is associated with the A-rich 3' ends of individual Alu repeats but these length polymorphisms are stably inherited (20). Evidently, the polyadenine tracts are stabilized by acquiring other nucleotides through some common mechanism.

Another issue is whether the dispersed Alu repeats have a function. We speculate that the A-rich tails of some Alu repeats may serve as transcription terminators. A-rich regions are known *Pol* II transcription terminators; complementary T rich regions are thought to be both *Pol* II and *Pol* III terminators (24,25). Human 5S genes are interrupted by an inverted Alu repeat providing a probable *Pol* III transcription terminator (26). Similarly, the A-rich transcription terminator of sea urchin histone genes closely resembles the A-rich 3' ends of numerous Alu repeats (27). The observation that an interspersed Alu blocks transcriptional interference is also consistent with this speculation (28). One problem in reconciling short interspersed repeats (Sines) with a biological function has been that entirely different Sine families populate divergent mammalian genomes (1,2). Presumably, the exact sequence of the dispersed repeat is not essential to whatever, if any, function it serves. Our present speculation also circumvents this difficulty. Selection for A-rich as well as complementary T-rich transcription termination signals between transcription units might account for the ubiquity of mammalian Sines. Human Alus merely provide one selected pathway to disperse such A-rich elements.

ACKNOWLEDGEMENTS

We thank Dr. S. Karathanasis for helpful discussions and for the apolipoprotein gene subclone and Dr. J. Gatewood for RNA isolation. This research was supported by USPHS grant GM 21346.

REFERENCES

- Schmid, C.W. and Shen, C.-K. J. (1985). In MacIntyre, R.J. (ed.) *Molecular Evolutionary Genetics*. Plenum, New York, p. 323–358.
- Weiner, A.M., Deininger, P.L. and Efstratiadis, A. (1986). In *Annual Review of Biochemistry*. Annual Reviews, Inc., Palo Alto, Vol. 55, p. 631–661.
- Schmid, C.W., Deka, N. and Matera, A.G. (1989). In Adolph, K. W. (ed.) *Chromosomes: Eukaryotic, Prokaryotic and Viral*. CRC Press, Inc., Boca Raton, Vol. 1, p. 3–29.
- Matera, A.G., Hellmann, U. and Schmid, C.W. (1990) *Mol. Cell Biol.*, accepted.
- Willard, C.W., Nguyen, H.T. and Schmid, C.W. (1987) *J. Mol. Evol.*, **26**, 180–186.
- Quentin, Y. (1988) *J. Mol. Evol.*, **27**, 194–202.
- Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H., Deininger, P. (1987) *Mol. Biol. Evol.*, **4**, 19–29.
- Jurka, J. and Smith, T. (1988) *Proc. Nat. Acad. Sci. USA*, **85**, 4775–4778.
- Britten, R.J., Baron, W.F., Stout, D.B. and Davidson, E.H. (1988) *Proc. Nat. Acad. Sci. USA*, **85**, 4770–4774.
- Deininger, P.L. and Slagel, V.K. (1988) *Mol. Cell Biol.*, **8**, 4566–4569.
- Britten, R.J., Stout, D.B. and Davidson, E.H. (1989) *Proc. Nat. Acad. Sci. USA*, **86**, 3718–3722.
- Antonarakis, S.E., Chakravanti, A., Halloran, S.L., Hudson, R.R., Feissee, L. and Karathanasis, S.K. (1988) *Hum. Genet.* **80**, 265–273.
- Mietus-Snyder, M., Charnley, P., Korf, B., Ladias, J.A.A., Gatti, R.A. and Karathanasis, S.K. (1990), manuscript submitted.
- Economou-Pachnis, A. and Tsiachlis, P.N. (1985) *Nucleic Acids Res.*, **13**, 8379–8387.
- Degen, S.J.F., Rajput, B. and Reich, E. (1986) *J. Biol. Chem.*, **261**, 6972–6985.

16. Lim, D., Coleman, R. T., Assmann, G. and Frossard, P.M. (1986) *Am. J. Hum. Genet.*, **39**, abstr. 621.
17. Ricchetti, M. and Bue, H. (1990) *EMBO J.*, **9**, 1583–1593.
18. Nei, M. (1987). In *Molecular Evolutionary Genetics*. Columbia Univ. Press, New York, p. 267.
19. Stoppa-Lyonnet, D., Carter, P.E., Meo, T. and Tosi, M. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 1551–1555.
20. Economou, E.P., Bergen, A.W., Warren A.C. and Antonarakis, S.E. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 2951–2954.
21. Schoeniger, L.O. and Jelinek, W.R. (1986) *Mol. Cell. Biol.*, **6**, 1508–1519.
22. Ullu, E. and Weiner, A.M. (1985) *Nature*, **318**, 371–374.
23. Rosenberg, H., Singer, M. and Rosenberg, M. (1978) *Science*, **200**, 394–402.
24. Proudfoot, N.J. (1989) *TIBS*, **14**, 105–110 .
25. Bogenhagen, D.F. and Brown, D.D. (1981) *Cell*, **24**, 261–270.
26. Little, R.D. and Braaten, A.C. (1989) *Genomics*, **4**, 376–383.
27. Briggs, D., Jackson, D., Whitelaw, E. and Proudfoot, N.J. (1989) *Nucl. Acids Res.*, **17**, 8061–8071 .
28. Wu, J., Grindlay, J., Bushell, P., Mendelsohn, L. and Allan, M. (1990) *Mol. Cell Biol.*, **10**, 1209–1216.