Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 6470–6475, June 1996
Evolution

Microsatellite spreading in the human genome: Evolutionary mechanisms and structural implications

(retroposition/3′-extension/homology-driven integration/chromatin organization/scaffold/matrix associated regions)

EYAL NADIR*, HANAH MARGALIT*†, TAMAR GALLILY‡, AND SHMUEL A. BEN-SASSON‡

*Departments of Molecular Genetics and ‡Experimental Medicine and Cancer Research, Hebrew University, Hadassah Medical School, POB 12272 Jerusalem 91120, Israel

Communicated by Howard Green, Harvard Medical School, Boston, MA, March 1, 1996 (received for review November 7, 1995)

ABSTRACT **Microsatellites are tandem repeat sequences abundant in the genomes of higher eukaryotes and hitherto considered as "junk DNA." Analysis of a human genome representative data base (2.84 Mb) reveals a distinct juxtaposition of A-rich microsatellites and retroposons and suggests their coevolution. The analysis implies that most microsatellites were generated by a 3′-extension of retrotranscripts, similar to mRNA polyadenylylation, and that they serve in turn as "retroposition navigators," directing the retroposons via homology-driven integration into defined sites. Thus, they became instrumental in the preservation and extension of primordial genomic patterns. A role is assigned to these reiterating A-rich loci in the higher-order organization of the chromatin. The disease-associated triplet repeats are mostly found in coding regions and do not show an association with retroposons, constituting a unique set within the family of microsatellite sequences.**

Genomic microsatellite (MS) sequences are composed of tandem repeats of various sizes. The repeating unit may be 1–6 bp long, and an MS element may contain up to ≈100 bp. Although MS sequences are widely spread in eukaryotic genomes, their biological role and evolutionary origin are yet unknown, and sometimes they are considered "junk DNA" (1). Still, the length polymorphism of a certain MS among individuals is frequently used as a genetic marker for genome mapping and medical purposes (2). In addition, the abnormal extension and instability of certain MS elements were found to be associated with several genetic diseases (reviewed in ref. 3) and cancer (4), respectively.

Length polymorphism of MSs is thought to be a consequence of elongation by a slippage mechanism (5, 6) and is evident only beyond a certain size [e.g., 20 bp for (CA)n repeats (5)]. Conceivably, a primordial MS sequence of at least such a critical length was first generated by a different mechanism and subsequently served as a substrate for further elongation. Elucidation of the processes that lead to MS formation might shed light on their possible significance and aid in understanding the evolution of eukaryotic genomes. In the current work we take advantage of the tremendous amount of data in the human genome data base to systematically explore this issue. By a detailed analysis of the genomic context of MS sequences, we could show a consistent pattern. Frequently, the most abundant, A-rich MS elements are found to be contiguous to the 3′-end of retroposons. Further dissection of the integration sites of MS-containing retroposons suggests that the mechanism of their insertion involves homology-driven integration through their MS tails. These findings thus assign a unique evolutionary role to A-rich MS sequences in the preservation and amplification of primordial genomic patterns. The evolutionary process that gave rise to the separate group of disease-associated MS sequences is also discussed.

Characterization of MS Sequences in the Human Genome

A computerized analysis of human MS sequences using GenBank as a source must be done with great caution for the following reasons. (*i*) There are many duplicated sequences under different entries. (*ii*) Many sequences were submitted for publication as polymorphic genomic sites and thus cause a bias with respect to MS distribution. (*iii*) Common cDNA sequences do not include introns and therefore are not representative, and poly(A) tracts found in mRNA 3′-ends are not of genomic origin. To overcome these difficulties, we have organized a nonredundant representative data base of long sequences (>10 kb each) of genomic origin, extracted from 122 GenBank entries. The cumulative size of this data base, which amounts to 2.84 Mb, is sufficient to allow a relatively unbiased characterization of MS distribution in the human genome.

A search for MS elements at least 16 nt long revealed 1021 such sequences in our data base. The minimum length of 16 nt was chosen to keep the probability of finding such sequences by chance very low ($<10^{-6}$) and, at the same time, to ensure a reasonable size of the dataset of identified MSs for the analysis. Classification of the MS sequences was done according to their repeating unit prototype, including all permutations on both strands (e.g., AAG represents the following: AAG, AGA, GAA, CTT, CTC, and TTC). Theoretically, 501 types are possible for period lengths from 1 to 6 bp, but only 102 prototypes were identified. The most abundant types are listed in Table 1 with several notable findings. (*i*) Out of 394 polymononucleotides, 393 are poly(A) tracts and only 1 is poly(G). (*ii*) CG is not present at all as a polydinucleotide. Interestingly, AC is more abundant than the AG and AT types. (*iii*) Among the 510 MS sequences with period lengths of 3–6 bp, 312 follow the pattern of A(2–5)N (N = C, G, or T). These results are consistent with a recent compilation by Jurka and Pethiyagoda (7). Table 2 compares the lengths of MS tracts in the different types, classified by their repetitive unit size. As shown, MSs comprised of dinucleotides are the longest found, and the hexanucleotide-based MSs are the shortest. These differences in length may be due either to differences in antiquity or to selective instability of MSs consisting of dinucleotides.

Juxtaposition of MS Sequences and Retroposons

An inspection of MS-flanking sequences immediately points to a close association between a large portion of the MSs and the most frequent short interspersed element (SINE) in the human genome, the Alu element. Many of the MS sequences are located contiguous to the 3′-end of Alu (Table 1). To further

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: MS, microsatellite; SINE, short interspersed element.

†To whom reprint requests should be addressed. e-mail: hanah@hujivms.huji.ac.il

Table 1. Distribution of abundant microsatellites in the human genome data base (2.84 Mb)

Type of MS monomer	Total	Association with retroposons: 3′ to Alu No.	%	Other Alu association No.	%	3′ to or in L1 No.	%	3′ to other retroposons No.	%	None No.	%
A	393	303	77	31	8	9	2	3	0.8	47	12
AC	78	17	22	3	4	3	4	4	5	51	65
AT	17	4	24	2	12	2	12	0		9	53
AG	22	4	18	2	9	1	5	3	14	12	55
AAC	12	11	92	1	8	0		0		0	
AAT	15	10	67	3	20	1	7	0		1	7
AAG	9	3	33	1	11	0		1	11	4	44
AGG	7	2	29	0		0		2	29	3	43
CAG	7	0		0		0		0		7	100
CGG	7	0		0		0		0		7	100
AAAC	35	28	80	2	6	0		0		5	14
AAAT	57	48	84	3	5	2	4	1	2	3	5
AAAG	36	29	81	5	14	0		1	3	1	3
AAGG	9	6	67	2	22	0		0		1	11
AATC	4	3	75	0		0		0		1	25
AATG	12	3	25	0		0		0		9	75
ACAT	4	1	25	1	25	1	25	0		1	25
AGAT	8	2	25	0		0		0		6	75
AAAAC	26	20	77	0		1	4	0		5	19
AAAAT	25	20	80	0		1	4	0		4	16
AAAAG	17	13	76	1	6	0		0		3	18
AAAAAC	33	27	82	1	3	1	3	0		4	12
AAAAAT	24	17	71	4	17	1	4	0		2	8
AAAAAG	23	18	78	1	4	0		0		4	17
ACCCCC	6	0		0		0		0		6	100
AGAGGG	5	1	20	0		0		0		4	80
Total	891	590	66	63	7	23	3	15	1.7	200	22
Major families											
A	393	303	77	31	8	9	2	3	0.8	47	12
A(2-5)N	312	244	78	22	7	8	2.5	2	0.6	36	11.5
AN	117	25	21	7	6	6	5	7	6	72	62
CRG	14	0		0		0		0		14	100

Included in this table are only those types of microsatellites that appear in at least four different loci in our data base. Three types (ATCC, ACTCC, and ACCATC), which were found more than three times but mostly at the same locus, were excluded.

characterize the association between MS and Alu sequences, all Alu elements in the data base were identified independently by a computer search directed at Alu's right (3′) monomer using a low stringency. More than 2700 candidate sequences were identified. Sequences with low alignment scores, when compared with a consensus Alu right monomer, were eliminated as false positives. This procedure resulted in a collection of 1200 Alu monomers, which account for ≈10% of our 2.84-Mb data base, the majority of which were full-size Alu elements. The identified Alus were in good agreement with the GenBank documentation, suggesting that the Alu selection procedure used by us was reasonably accurate.

Examination of the tight association between the locations of MS and Alu elements revealed that two-thirds (590 of 891) of the more abundant MSs were located immediately downstream to the 3′-end of an Alu element and that at least half (>590 of 1200) of the Alu elements were followed by some type of an MS. Moreover, the MS sequences were preferably oriented with their A-rich, rather than T-rich, tracts, 3′ to the conventional Alu orientation. This phenomenon, which is evident for poly(A) versus poly(T) tracts, encompasses all types of MSs. Fig. 1 shows the distribution of distances between the 3′-end of the Alu elements and the proximal end of their associated MS sequences; such closeness between 590 MSs and Alu elements is highly statistically significant (see legend to Fig. 1).

Table 2. Comparison of MS tract lengths among the different MS types classified by their repetitive unit size

Repetitive unit size	Occurrence	Median MS tract length and interquartile range
1	394	19 (17–22)
2	117	22 (18–32)
3	61	20 (17–26)
4	192	20 (18–24.5)
5	100	19 (17–22.5)
6	157	17 (16–19)
Total	1021	19 (17–23)

Because the length distributions are asymmetrical, median and interquartile range values are reported. The differences between the median of dinucleotide tracts and all other groups were evaluated by a Mann–Whitney test. All differences were statistically significant (ranging from $P < 0.0001$ for the comparison with mononucleotide tracts to $P = 0.0665$ for the comparison with tetranucleotide tracts). Comparison of the dinucleotide tracts with all other MS tracts, taken together, also showed a highly statistically significant difference in lengths ($P < 0.0001$).

Alu's full length is 281 bp, and it is composed of two monomers that show some degree of sequence similarity (8). Occasionally, Alu-associated MSs were not found at its 3′-end but rather between the two Alu monomers, downstream from an Alu's orphan left (5′) monomer, or, rarely, just upstream from the Alu element. One-third of the MS sequences did not show any association with Alu elements. The flanking regions of these MSs were examined for sequence similarity with the entire human genome data base. This comparison, carried out for each MS separately, revealed that ≈20% of them are found in close association with other known repetitive elements of the human genome. The majority of these MS sequences were found in association with the mammalian LINE-L1 (9). The others were associated with less frequent repetitive elements such as MER12, O interspersed repeat, and alpha repetitive DNA (10). Jurka *et al.* (10) have assembled the prototypic sequences of human repetitive DNA and claimed that many of the interspersed repeats contain a simple poly(A) tail. A search for these prototypic sequences in the entire human genome revealed that in addition to poly(A) tails, other MS sequences can be also identified in association with them, as exemplified in Fig. 2.

Repetitive elements are not the only known retroposons in the human genome. Pseudogenes are genomic sequences that result from retroposition of processed mRNA into the genome (11), and in some cases their origin is known in detail. A search through GenBank entries documented as pseudogenes revealed in several instances the presence of MSs at their 3′-ends. Fig. 3 depicts an example of two related pseudogenes with two different 3′-tails, a poly(A) and another MS sequence.

As summarized in Table 1, in many MS types >80% of the sequences are associated with retroposons. Furthermore, there are MS types where all of the sequences are associated with retroposition products. The strong association between MS elements and retroposons can be used for the discovery of the latter. Because MS sequences are easy to identify, they can be used in turn as potential markers for archaic, fading, repetitive retroposons in eukaryotic genomes, such as early SINEs. In the present study, one such (as yet unknown) repetitive sequence has been discovered in the human genome, only by its association with an MS (a detailed report of this will be published elsewhere).

Alu is the most abundant repetitive sequence in the human genome. It has been estimated that altogether it constitutes

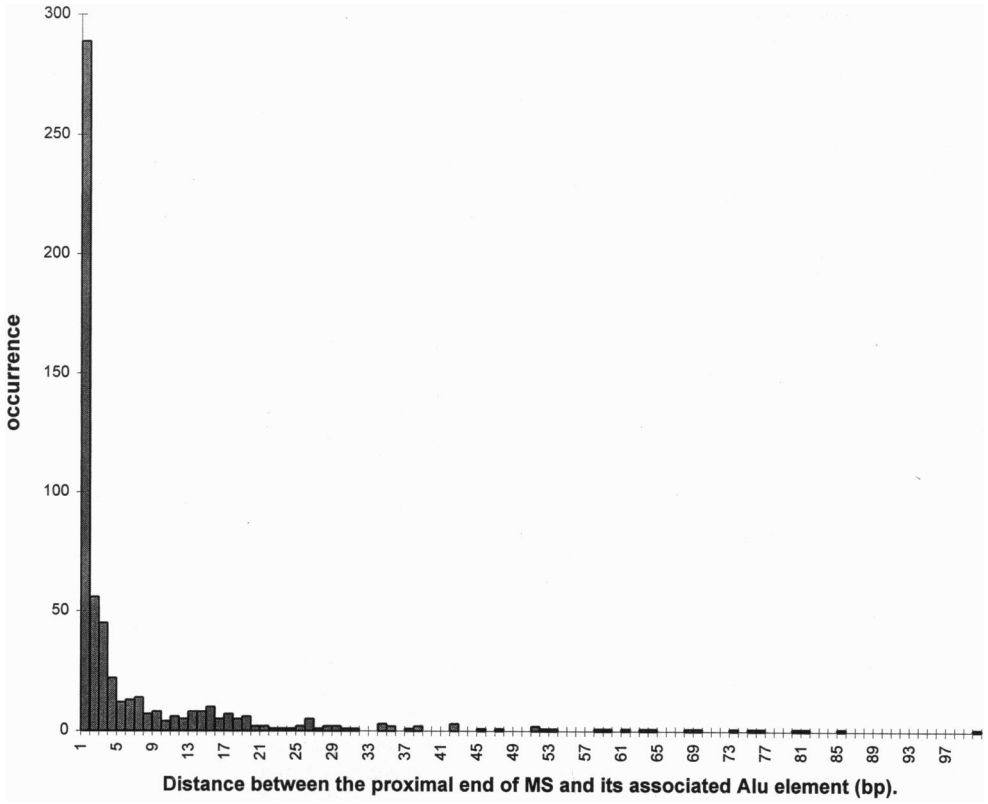

FIG. 1. Histogram of the distances between the 3′-end of an Alu element and the proximal end of the associated MS sequence in the 590 instances where this association existed. Of these instances, in only 7 and 25 cases the distance exceeded 100 and 50 bp, respectively. This association is highly statistically significant; the frequency of the Alu element in our data base implies that the probability of a nucleotide to be a 3′-end of Alu by chance is $1200/2.84 \times 10^6 = 0.0004$. The probability of finding at least one 3′-end of an Alu in a span of 100 nt from an MS is 0.04. Therefore, in our data base of 1021 MSs, the expected number of Alus in close proximity to MSs was ≈41. The observed number of 565 is highly statistically significant ($P < 0.00001$).

≈5–10% of the DNA content (8). It is widely accepted that the extensive spread of this sequence occurred ≈30 million years ago via retroposition of Alu transcripts (12). Therefore, the presence of a poly(A) tract next to the 3′-end of the Alu can be considered as a consequence of the common polyadenylylation process, which takes place at the 3′-end of RNA transcripts (13). Nevertheless, our analysis revealed that many other types of MSs demonstrate an association with Alu as extensively as poly(A) tracts (see below). These relationships, which were noticed in the past in sporadic cases and through computational analyses on smaller data sets (14–16), were demonstrated in the recent comprehensive survey of Jurka and Pethiyagoda (7) and are reinforced here. Especially, our observations emphasize the generality of the phenomenon in the human genome. Several earlier publications have indicated an association between other specific retroposons and MS sequences in eukaryotic genomes. For instance, in the bovine genome, of 44 sequenced Artiodactyl C-A retroposons and their 3′-flanking sequences, 50% are associated with MS sequences, usually (CA)n, and all but one of the 33 (A-A) dimer elements have a (CAG)n (n = 1–9 tails; ref. 17). Taken together, these findings suggest that MS generation and retroposition events probably represent accompanying evolutionary processes, common to a variety of higher eukaryotes.

```
1     MER17   ..CAGTCCTG CaagcTcCAT TtATGGtAAG TACcCTAtAC AGGTGTACCA
    HUMSIRPOCB  ..CAGTCCTG C...tTtCAT TcATGGcAAG TACtCTAaAC AGGTGTACCA

        MER17   Tt
    HUMSIRPOCB  Tattttatna tACACACACA CACACACACA CACACACACA CACACAtttt


2       MER2    .....GTct TGccACcAAT tttCcatgGA TACTgAGGaA cGaCTgtaaa
    HUMVTR14F   .....GTcc TGgaACaAAT actCcataGA TACTaAGGgA tGaCTa...t
     HUMATPSG   .....GTtc TGgaACcAAT ccaCtgcaGA TACTgAGGgA tGgCTatact

        MER2
    HUMVTR14F   atcctactag acttctccAG AGTATACACA CACACACACA CACACACACA
     HUMATPSG   atcttttggt actgatatAG AAAACATAAT GATTAAAAAA CATAATGATT

        MER2
    HUMVTR14F   CACACACACA CACACACACA CACACAcgga catgtaatca gatc
     HUMATPSG   AAAAAAttcc tggaccttcc ca.acatgca atggtttata gatttatgta


3         O             ........CC cAcGATTCAA TTaCcTCCCa CcaGGTCCCT
      HSREP06           ........CC cAtGATTCAA TTgCtTCCCc CtgGGTCCCT
    HUM335MFD           ........CC tAtGATTCAA TTaCtTCCCa CtgGGTCCCT

          O   CCCAtgACAc aTgGGaaTTa TaGGAgcTAc AATTCAgGaT GAGATTTgGg
      HSREP06 CCCAcaACAt gT.GGagTTc TgGGAgaTAt AATTCAaGtT GAGATTT.Ga
    HUM335MFD CCCAccACAc aTgGGgaTTa TgGGAacTAc AATTCAgGaT GAGATTTgGg

          O   tTGGGGACAC AgCCAAACtA TA
      HSREP06 aTGGGGACAC AgCCAAACcA TAtcAAATGA ATATGCTAGA AATGAGGAAA
    HUM335MFD .TGGGGACAC AaCCAAACcA TAtcagatAC ACACACACAC ACACACACAC

          O
      HSREP06 ACAAAAATCA AAAGGTACAA AAAAGGAAAA AAAAAAAAAA AAGAAcagag
    HUM335MFD ACACACACAC ACACACACAC ACATATATGT ACATAGttta aaaaaatatc
```

FIG. 2. MS sequences as the 3′-tail of non-Alu SINEs. Each of the consensus sequences of (*i*) MER17, (*ii*) MER2, and (*iii*) O interspersed repetitive DNA (9) is aligned with other GenBank entries that contain MS tails at their 3′-ends. The aligned bases are shown in uppercase letters. The MS or MS-like sequences are in boldface type. Note that the MS tail of Humatpsg, which is aligned with MER2, possesses more than two copies of a hexadecamer (underlined).

Distinct Features Typify Different Subclasses of MS Elements

The MS sequences in the human genome can be characterized according to several criteria: (*i*) frequency, (*ii*) association with various retroposons, (*iii*) evolutionary age, and (*iv*) location with respect to coding regions. When all these criteria are being used in concert, three families of MS elements can be distinguished.

The A-Rich Family. The results shown in Table 1 clearly demonstrate that all 12 A-rich MS types of the general pattern [A(2–5)N]n have the following distinct properties: (*i*) they are the most abundant MSs among their corresponding monomeric size; (*ii*) the vast majority (≈80%) of them are associated with Alu; and (*iii*) the similarity between their associated Alu elements and a consensus Alu is not significantly different from that of poly(A)-associated Alu (Table 3). By all of these

```
HUMLAMB   CTCTTAAGCAGCATGGAAAAATGGTTGATGGAAAATAAACATCAGTTTCT
HSPK2     ------C---A--------T-A----------------...---------
HSLBP32P  ----------A------------------C------------------T-

HUMLAMB
HSPK2     AAAAAAAAAAAAAAAATTCCTTCCTTTTAGGC
HSLBP32P  CAAATAAATAAATAAATAAGTAAAATTACATTGGCATGTATT
```

FIG. 3. Different MS sequences at the 3′-ends of related pseudogenes. Alignment of the 3′-ends of the mRNA of human colon carcinoma laminin-binding protein (HUMLAMB), human cHD4 pseudogene K2 (HSPK2), and human HLBP32 pseudogene (HSLBP32P). The two pseudogenes are related to the same gene as the mRNA. HSPK2 has a poly(A) (boldface type) next to its 3′-end, and HSLBP32P has (AAAT)n MS (boldface type) instead of a poly(A). Polyadenylylation signals are italic and underlined.

criteria, the A-rich MS is indistinguishable from the pure (A)n MS (see Table 1). In other words, poly(A) tracts are authentic MS sequences that, together with the [A(2–5)N]n, form the largest family of MS elements, which spread in the human genome together with the relatively recent primate-specific SINE Alu. Indeed, the distribution of pure (A)n MSs among the various retroposons is almost identical to that of A-rich MSs, as evident from Table 1 ($P > 90\%$ by a χ^2 test).

The Dinucleotide MS Family. MS elements with AC, AT, and AG repeating units share some unique features that justify their separate classification. (*i*) They are the most abundant among the non-A-rich MSs; (AC)n is the second most frequent MS after (A)n. (*ii*) The dinucleotide MS tracts are longer than the other types, including poly(A) (Table 2). (*iii*) Their association with retroposons shows a different pattern; much fewer are adjacent to Alu and more are associated with other retroposons, *vis-à-vis* the A-rich MSs (bottom of Table 1). (*iv*) The Alu sequences with which the dinucleotide MSs are associated seem to belong to the more archaic representatives of this SINE, as they are further remote as a group from the consensus Alu sequence (Table 3). Taken together, these findings suggest that the dinucleotide MSs preceded the A-rich MSs in their spreading, although the mechanism by which both families evolved might be the same.

The (CRG)n Family. As already demonstrated by Stallings (21), (CRG)n MS elements are mainly located within or very close to coding sequences, unlike the two other MS families that appear almost exclusively in introns or between genes. This observation is also reflected in (CRG)n MS involvement in human diseases (3). None of the representatives of this family was found to be associated with retroposons (Table 1). All of the above suggest that (CRG)n sequences constitute a completely separate family of MS elements, not only with respect to composition, but also by their mechanism of formation. For these reasons they will be discussed separately.

MS Sequences Were Generated by 3′-Extension of the Retrotranscript

Typically, retroposons are bounded by direct repeats—i.e., target site duplication of 7–21 bp generated during the integration into the genome via a staggered nick (22). Thus, identification of the direct repeats enables a precise delineation of the integrated element. An investigation of a representative sample of ≈200 Alu elements and available pseudogenes was conducted in an attempt to trace the MS origin. A few examples of this search are shown in Fig. 4. An analysis of the context of the direct repeats, which can be readily detected in >90% of the Alu elements studied, demonstrates that at the 3′-end, the direct repeat is consistently located downstream from the MS sequence. In other words, the MS sequence was a contiguous part, indistinguishable from the retroposon itself, at the moment of the integration event. This finding supports the proposition of Beckmann and Weber (14) that the MS sequences were first formed as a 3′-tail of the retroposon before its joining into the genome.

Table 3. Average similarity scores for the Alu elements associated with the three different groups of MS sequences when compared with a consensus Alu sequence

Group	*N*	Average similarity score	SE
(A)n	299	0.6829	0.0064
[A(2–5)N]n	235	0.6677	0.0068
(AN)n	25	0.606	0.0240

Scores are expressed as the ratio score given by the GCG BESTFIT program (18). The scores of each family are distributed normally, as shown by χ^2 test for goodness of fit ($P > 0.25$, 0.025, and 0.5, respectively). The variances are homogenous as shown by Bartlett's test (19) ($P > 0.25$). The comparison between Alu scores was performed by Dunnett's test (20) with the poly(A) family taken as a control group. (AN)n scores were found to be significantly lower than those of the poly(A) ($P < 0.01$), whereas the [A(2–5)N]n were not ($P > 0.05$).

The process of 3′ extension of RNA transcripts, starting with polyadenylylation of mRNA, is very common in all eukaryotes (13). It is tempting to speculate that the mechanism of 3′-extension, which gave rise to MS sequences, employed a U-rich guide RNA as a template for the editing of the primary transcript. If, in addition to the pure U 5′-tail, some guide RNAs encoded U(2–5)N or even UGUG 5′-tails, they could give rise to the (A(2–5)N)n and (AC)n MS sequences, respectively. This putative guide RNA might also serve as a primer for the reverse transcription of the complete element similar to the role played by tRNA in the life cycle of retroviruses (23). The eukaryotic telomerase complex that utilizes a template guide-RNA for the complete replication of chromosomes' free ends (24) can be regarded as a prototype of this mechanism; by multiple rounds of the process, the telomerase forms the telomere, which is in essence a pure MS sequence.

Homology-Driven Integration of MS Sequences

An inspection of the context of the direct repeats at the 5′-ends of retroposons consistently reveals an A-rich box of ≈5 bases upstream from the direct repeat, as already noted by several investigators for a variety of primates' retroposons, including pseudogenes (ref. 25 and references cited therein). A closer look shows that there is an outstanding similarity, and in many cases even complete identity, between this box and the MS sequence upstream from the 3′-end direct repeat. Namely, a pure A box is just a special case of a more general rule, as depicted in Fig. 4. Due to the overwhelming abundance of the A-rich MS sequences and to their recent appearance, this phenomenon can be readily observed within this family. As for the dinucleotides, the limited sample of Alu-associated poly dinucleotides and the difficulty in determining the direct repeats due to their antiquity made the identification of appropriate examples more troublesome; yet the examples of the (AC)n and (AT)n MS sequences given in Fig. 4 suggest that the same rule also applies to the dinucleotide MSs. Thus, the conjecture of Deininger (8) that integration of SINEs is not random but directed at A-rich sites becomes more general. It might be concluded that retroposons integrated into the early genome at specific sites where adequate similarity existed between retroposonal MS and a corresponding sequence at the genomic locus. Consequently, a significant expansion of the A-rich target box to an A-rich MS was achieved at that particular locus. While delineation of the molecular events associated with such homology-driven integration is beyond the scope of this paper, a model of a possible mechanism is schematically outlined in Fig. 5.

1) Humhbb 45069
GTTGATGCTGGATAGAGGTGTTTATTTCTATTCT**AAAAAGTGATGAAATT** [GGCCGGGCGCGGTGGCTCAC..
..ACAAAGCGAGACTCCATCTC] (A)$_{20}$ (AAAG)$_4$ AAAA**AAAAAGTGATGAAATT**GTGTATTCAA

2) Humserg 9036
AGGTTTTTCTACAATGCCACATTGTCTTTTCCAT**AAAAAGTACTTTC** [GGCCGGGTGTGGTGGCTCAC..
..ACAGAGCGAGACTTCATCTC] AAAAAAAAAAAA**AAAAAGTACTTTT**CTTCATTTGGTTAGTATTCTCTTATGAGTTGATGCCTTGTAATTTA

3) Hsacpp1 2742 (C.S.)
TTTTAAATTAAAGAAGAAAAGTAACAGCAAGTTTATTTT**GAAAAATTTTA** [GGCTGGGCGTGGTGGCTCAT..
..ACAGAGCAAGACTCCATCTC] AAAACACAGCA(AAAC)$_3$ AACAACAAAAAACAAACAAAA**GAAAAATTTTA**GTCACACCTTAAGACAATTGGAC

4) Humserg 314 (C.S.)
GGTGCAGATTTCTGGAGGACTAAGGCTCCAT**ATTAAAGTTATACAATAAA** [GGCCAGGTGCTGTGGTCCACAC..
..ACAGAGAGAGACTCTATTTC] AAAAAACTAAAAATAATAAAATAAAATAA**AATAAAGTTATACAATAAA**TACAAGCCTATCATACATCCT

5) Hsaladg 8955 (C.S.)
CAAAAAGAGAAAAATCCCCTGCAATTCTACCTCT**AGAGAAAAGTACT**GTT [GGCCGGGCATGGCGGCTCAC..
..ACAGAGCAAGACTTCATTCC] AAAAAAAAAAAGAAAGAAAGAAAA**AGAGAAAAGTACT**TGGGACATGGATTTTTTGTTAGGTACAA

6) Humserg 1267
AGAGAGCCAATGATCCAATGCAAGAAACTGAGTTT**AAGAAGAAACTGTT** [GCTGGGCACGGTGGCTCAC..
..ACAGAGTGAGACTCCATCTC] (A)$_{27}$ GAAGAAAAAGAAGAAG**AAGAAGAAACTGTT**CATCTGAAATCCGA

7) Humadag 13734
AGCAGAGCTGGGAGGTGGGCTCTTGAATTGGCATT**TAAATACTTTTGCCA** [GGCAGGGTAAGGCAGCTCAC..
..ACAGAGCCAGACTCCATCTC] AAAAAAAAAAA(AAC)$_4$ (AAAT)$_3$ GAA**TAAATACTTTAGCC**AGAAGTAGCCATGCAGACC

8) Hsatpcp2 9429
ACTTTATGATGATGCAAAAATGATATGCATTCATT**AGAAACCATACTTCG** [GGCTGGGCGTGGTGGCTCAT..
..ATACAGCAAGACTCCGTCTC] AAAAAAAAA(AAAG)$_3$ **AAAGAAACCATACTTCG**AATTAGGAATCTTTATCTTTTCCCAAGCTAGCA

9) Hummmdbc 58868 (C.S.)
CCACCTCAGTGGTACAGCCCCCAGCCTCAAAACCTG**TAAATACCATACTT** [GGCCGGGCGCGGTGGCTCAC..
..ACAGAGCGAGACCCTGTCTC] AAAAAG(AT)$_{18}$ A**TATATAC.ATATTTTTTTTTTC**TTTACTGTTTTTAG

10) Humtpa 7230 (C.S.)
ACATAGAAAATATTTTTTGAATATATAATATGCA**AAATATCACTAAGAGG** [GACTGGGCACAGTGGCTCTT..
..ATGGAGTCAGACCTTGTCTC] TAAAT(TA)$_{12}$ (CA)$_{14}$ TACAAAC**ACATATTGCTAAGAGA**ATGTGCTGTCTT

11) Hscalm - Human CaMII-psi2 calmodulin retropseudogene. (38)
AATCCTAGGGCCCTG**AAGAATGACACTAAATC**AGTCCAAGTGGAGCAAGTGA [GCGAGCAGAGTGGTTGTGTGGTT..
..TTTCC<u>AATTAAA</u>AAAATTACAATTA] AAAA**AAGAATTATACTAAATC**TACTCTGCCTACCACAC

12) Humfolp21 + Humfolp23 - Human dihydrofolate reductase (DHFR) psi-2 pseudogene. (39)
AGTTTCCAAGC**AAAAACCTTCCGGCC** [AATCTTGGCCGCGCGTCCTGGTGT..
..ACTCTATCAT<u>AATAAA</u>AAGTCACACTTGTTCAT] AGA(AAAC)$_7$ AA**AAAAACCTTCCTGCC**CAGCCCATCTCATTCTCCCTGA

13) Hsgapdp - Human x-linked GAPD pseudogene (glyceraldehyde-3-phosphate dehydrogenase). (40)
GTGGTTAAGGCCAAAGACAGGTAAGTC**AGAAAAATCTTGGCT** [CGGCTCTCTGCTCCTCCTGTTCTACAGT..
..TCATATACCATC<u>AATAAA</u>GTACCCTGTGCTCAGCC] AAAAAAAAAAAAAAGAGAG**AGAAAAATCTTGGCT**TACATTTTGACTTTGTGAAAATTACTTTACC

FIG. 4. Comparison between sequences at the 3′-ends of retroposon-associated MSs and the boxes upstream from the 5′-direct repeats. Representative examples of direct repeats flanking 10 Alu elements and three pseudogenes are displayed. The entry in GenBank and the position of the A-rich MS within this sequence are indicated at the top of each comparison. The 5′- and 3′-ends of the retroposon are shown in brackets. The direct repeats are in boldface type and the A-rich boxes at the 5′-end of the direct repeats are boldface and magnified. The erosion of the direct repeat in the presumably more archaic (AC)n-containing Alu element (no. 10), is noteworthy. C.S., complementary strand.

Furthermore, homology-driven integration can readily explain the tendency of Alu elements to cluster; it is not uncommon to find genomic loci where several Alus are positioned head-to-tail or are even integrated into one another. Integration of one element implies the generation of two new genomic A-rich tracts (one is built-in between Alu's two monomers and the other is the MS, which forms its 3′-end). These tracts, in turn, might become favorable integration sites for new retroposons, which will further enrich and intensify the presence of A-rich tracts in that particular locus. This model for the dynamics of Alu clustering was already proposed by Daniels and Deininger (25) and implies that in one cluster all the elements will be oriented in the same direction. Hence, it is consistent with the results of Moyzis *et al.* (26), who noted that the number of oppositely oriented neighboring Alus is disproportionately low.

Implications of MS-Driven Retroposition

We have assigned a role for retroposons as "A-rich MS generators" and for A-rich MS sequences as "retroposition navigators," which together result in a site-specific generation of regular A-rich sequences. The work of Gasser and Laemmli (27) has demonstrated the presence of A and T boxes in the sequences of scaffold/matrix associated regions. Typically, these boxes, which control chromatin organization, have consensus sequences of AATAAA(T/C)AAA and TT(A/T)T(T/A)TT(T/A)TT, respectively; motifs that are shared by many of the MSs. Hence, it is conceivable that such MS-derived boxes are bona fide scaffold/matrix associated regions, which originated by retroposition events and contribute to the higher-order organization of eukaryotic nuclei. It is noteworthy that the association between individual A and T boxes and the nuclear matrix is relatively weak (28). In other words, the reiteration of MS-like sequences at certain locations enhances significantly the affinity of these sites to the immobilized scaffold of the nucleus. It is therefore possible that MSs play a global role in chromatin organization via matrix attachment. At the biochemical level, the existence of a protein that binds specifically to a pure (A)n MS has been demonstrated (29, 30). Also, the high mobility group of nuclear proteins has a higher affinity to A-rich sequences (31, 32). Yet another type of MS sequence that has a proven regulatory role is AGAT, better known as GATA (recently reviewed in ref. 33). Fig. 6 shows a sequence from the human CD4 locus in which four Alu elements are co-integrated with this prominent MS.

The involvement of MSs in the higher organization of chromatin is also suggested by the (CRG)n family, which constitutes a separate group of repetitive elements. This is evident by the defect in the packaging of the chromatin during mitosis in the fragile X syndrome, as its name implies, due to a corresponding (CRG)n expansion (34, 35). Moreover, a recent report demonstrates the implications of the triplet repeat extension in myotonic dystrophy in the alteration of the adjacent chromatin structure (36).

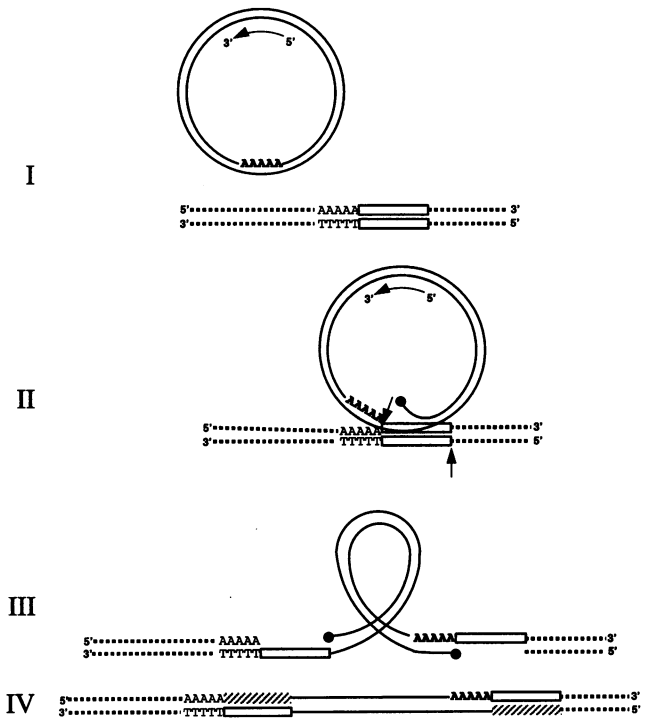

FIG. 5. A model for homology-driven integration of a circular, MS-containing retroposon into a target DNA. The retroposon is drawn in boldface, and the target DNA is indicated by a broken line. Arrows indicate the positions where staggered nicks occur. The direct repeats that are generated by the integration process are indicated by open boxes. I, alignment; II, staggered nicks and integration of one strand; III, integration of the second strand; and IV, gap filling and ligation.

Disease-Associated MSs Were Generated by a Different Mechanism

Two polytrinucleotides MS elements, (CGG)n and (CAG)n, formally considered as MS elements, are associated with a variety of human diseases and syndromes including fragile X, myotonic dystrophy, and Huntington disease (3). In our survey, none of the 14 MS sequences of the (CRG)n type was associated with a retroposon (Table 1). As already illustrated by Stallings, (CAG)n MSs are excluded from intronic regions in a strand-specific fashion (21). A thorough analysis that has been published recently by Green and Wang (37) ascribed the extension of CAG and CAG-like MS sequences to the process of codon reiteration in the evolution of proteins. These authors have demonstrated that this process is widespread along the evolution and provided strong arguments in favor of a precise, targeted mechanism of codon reiteration for the evolution and development of coding regions. The generation of (CGG)n MS sequences has occurred presumably by a mechanism similar to that of (CAG)n (ref. 37 and unpublished data). A comparison of the amino acids that show a tendency to reiterate in human versus yeast proteins shows that the major differences reside in poly-proline and poly-glycine tracts that are present in human proteins and are completely absent from yeast proteins (data not shown). These two amino acids are encoded by CG-rich codons, particularly CCG for proline and GGC for glycine.

In other words, the mechanism of generation of (CRG)n MS elements is different from that of other MS sequences. Nevertheless, from a broader evolutionary perspective they complement each other. While (CRG)n enhance preservation and expansion of certain motifs associated with coding regions, the A-rich and other MS elements play a similar role with respect to intergenic and intronic domains.

```
7500 TGTGGAGAGGAATGGGGCACTCAGAAGGTGTGGGGGCTTGGTCCTCTTAC
     GGTGGTTTAGGAGAAGATAAGAAAAGTGACTCCATCTTTGTTAGCTGCCT
     CCAGTTACTCCATCTGCCTGGAAGAGGAGGTAGCCAGGGCTTAATACGTG
     GTAGGTGCTTGAGATAGGTAGGTAGATAGATGGATGGACAGATAGATAGA
     TAGATAAAATAGATTAGATAGACAGATTAGATAGATTAGATGATAGATTG
     ATAGATTAGGTAAATAGATTAGATAAACATACAGACAGATTAGATAGATG
     ATGAATAAGATAGATAAAATAGATTAGGTTAGACAGAAAGATAGATTAGA
     TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGAGACAGA
     CATTTAG [Alu] AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
     CAGACATACAGACAGATAGATTAGATAAATAGATGATAGATAAGATAAAC
     AGATAAAATAGATGAATTAGATTAGACAGTCAGACAGACAAGATAGATAT
     ATAA [Alu] AAAAAAAAAAAAGATAGATATAT [Alu] AAAAAAAAAAAAA
     GATAGATAGATCAATAGATAAGATAGATACATAAAGATAGATAGACAGAT
     ATAGATAGATTAGATAGATAGGTGATAGATAGATAGATAGATAGATAGAT
     AGATAGATAGATAGATAGATATAGATAGATAGAACCGGGT [Alu] AAAAA
     AAAAAAAAAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAG
     ATAGATAGATGATAGATAGATACAATTTTGTTTAAACCGGGGAAGTAGGA
     AGAAAAGAGGTCAGGAAGGAGGTTTCTGAAAAAGGAGGAAGGGAGACTAG
     CCACCAGTGATGGTGAAGTGTTTACAGGTTAGCATGCAGATGGGAGG 5501
```

FIG. 6. Alu elements integrated within a (GATA)n-rich region, retrieved from the complementary strand of GenBank entry GB_pR:Humcd4. The four full-size Alu elements were omitted and only the tails are depicted. The GATA and GATA-like MS sequences are underlined.

We thank Shmuel Pietrokovski & Philipp Bucher for providing software application for direct accession to GenBank sequences. We are grateful to Prof. Amos Oppenheim for his helpful comments. This work has been partially supported by a grant from the Israeli National Science Foundation to H.M.

1. Nowak, R. (1994) *Science* **263,** 608–610.
2. Sutherland, G. R. & Richards, R. I. (1994) *N. Engl. J. Med.* **331,** 191–193.
3. Sutherland, G. R. & Richards, R. I. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 3636–3641.
4. Eshleman, J. R. & Markowitz, S. D. (1995) *Curr. Opin. Oncol.* **7,** 83–89.
5. Weber, J. L. (1990) *Genomics* **7,** 524–530.
6. Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E. & Inouye, M. (1966) *Cold Spring Harbor Symp. Quant. Biol.* **31,** 77–84.
7. Jurka, J. & Pethiyagoda, C. (1995) *J. Mol. Evol.* **40,** 120–126.
8. Deininger, P. L. (1989) in *Mobile DNA*, eds. Berg, D. E. & Howe, M. M. (Am. Soc. Microbiol., Washington, DC), pp. 619–636.
9. Singer, M. F. & Skowronski, J. (1985) *Trends Biochem. Sci.* **10,** 119–122.
10. Jurka, J., Walichiewicz, J. & Milosavljevic, M. (1992) *J. Mol. Evol.* **35,** 286–291.
11. Vanin, E. F. (1985) *Annu. Rev. Genet.* **19,** 253–272.
12. Britten, R. J. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 6148–6150.
13. Wahle, E. (1995) *Biochim. Biophys. Acta* **1261,** 183–194.
14. Beckmann, J. S. & Weber, J. L. (1992) *Genomics* **12,** 627–631.
15. Economou, E. P., Bergen, A. W., Warren, A. C. & Antonarakis S. E. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2951–2954.
16. Zuliani, G. & Hobbs, H. H. (1990) *Am. J. Hum. Genet.* **46,** 963–969.
17. Kaukinen, J. & Varvio, S. (1992) *Nucleic Acids Res.* **20,** 2955–2958.
18. Genetics Computer Group (1994) *Program Manual for the Wisconsin Package* (Genetics Computer Group, Madison, WI), Version 8.
19. Bartlett, M. S. (1937) *J. R. Stat. Soc. Suppl.* **4,** 137–170.
20. Dunnett, C. W. (1955) *J. Am. Stat. Assoc.* **50,** 1096–1121.
21. Stallings, R. L. (1994) *Genomics* **21,** 116–121.
22. Weiner, A. M., Deininger, P. L. & Efstratiadis, A. (1986) *Annu. Rev. Biochem.* **55,** 631–661.
23. Taylor, J. M. (1977) *Biochim. Biophys. Acta.* **473,** 57–71.
24. Singer, M. S. & Gottschling, D. E. (1994) *Science* **266,** 404–409.
25. Daniels, G. R. & Deininger, P. L. (1985) *Nucleic Acids Res.* **13,** 8939–8954.
26. Moyzis, R. K., Torney, D. C., Meyne, J., Buckingham, J. M., Wu, J.-R., Burks, C., Sirotkin, K. M. & Goad, W. B. (1989) *Genomics* **4,** 273–289.
27. Gasser, S. M. & Laemmli, U. K. (1987) *Trends Genet.* **3,** 16–22.
28. Mielke, C., Kohwi, Y., Kohwi-Shigematsu, T. & Bode, J. (1990) *Biochemistry* **29,** 7475–7485.
29. Winter, E. & Varshavsky, A. (1989) *EMBO J.* **8,** 1867–1877.
30. Reardon, B. J., Winters, R. S., Gordon, D. & Winter, E. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 11327–11331.
31. Strauss, F. & Varshavsky, A. (1984) *Cell* **37,** 889–901.
32. Churchill, M. E. A. & Travers, A. A. (1991) *Trends Biochem. Sci.* **16,** 92–97.
33. Weiss, M. J. & Orkin, S. H. (1995) *Exp. Hematol.* **23,** 99–107.
34. Verkerk, A. J. M. H., Pieretti, M., Sutcliffe, J. S. Fu, Y.-H., Kuhl, D. P. A., *et al.* (1991) *Cell* **65,** 905–914.
35. Kremer, E. J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S. T., Schlessinger, D., Sutherland, G. R. & Richards, R. I. (1991) *Science* **252,** 1711–1714.
36. Otten, A. D. & Tapscott, S. J. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 5465–5469.
37. Green, H. & Wang, N. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 4298–4302.
38. Koller, M., Baumer A. & Strehler, E. E. (1991) *Gene* **97,** 245–251.
39. Chen, M. J., Shimada, T., Moulton, A. D., Harrison, M. & Nienhuis, A. W. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 7435–7439.
40. Hanauer, A. & Mandel, J. L. (1984) *EMBO J.* **3,** 2627–2633.