MOLECULAR AND CELLULAR BIOLOGY, Oct. 1993, p. 6520–6529
0270-7306/93/106520-10$02.00/0

Vol. 13, No. 10

Nonrandom Localization of Recombination Events in Human Alpha Satellite Repeat Unit Variants: Implications for Higher-Order Structural Characteristics within Centromeric Heterochromatin

PETER E. WARBURTON,[1,2] JOHN S. WAYE,[3] AND HUNTINGTON F. WILLARD[1,4]*

Department of Genetics, Stanford University, Stanford, California 94305[1]; Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada M5S 1A8[2]; Department of Pathology, McMaster University, Hamilton, Ontario, Canada L8N 3Z5[3]; and Department of Genetics, Center for Human Genetics, Case Western Reserve University School of Medicine, Cleveland, Ohio 44106[4]

Received 21 January 1993/Returned for modification 25 March 1993/Accepted 23 July 1993

Tandemly repeated DNA families appear to undergo concerted evolution, such that repeat units within a species have a higher degree of sequence similarity than repeat units from even closely related species. While intraspecies homogenization of repeat units can be explained satisfactorily by repeated rounds of genetic exchange processes such as unequal crossing over and/or gene conversion, the parameters controlling these processes remain largely unknown. Alpha satellite DNA is a noncoding tandemly repeated DNA family found at the centromeres of all human and primate chromosomes. We have used sequence analysis to investigate the molecular basis of 13 variant alpha satellite repeat units, allowing comparison of multiple independent recombination events in closely related DNA sequences. The distribution of these events within the 171-bp monomer is nonrandom and clusters in a distinct 20- to 25-bp region, suggesting possible effects of primary sequence and/or chromatin structure. The position of these recombination events may be associated with the location within the higher-order repeat unit of the binding site for the centromere-specific protein CENP-B. These studies have implications for the molecular nature of genetic recombination, mechanisms of concerted evolution, and higher-order structure of centromeric heterochromatin.

The complex genomes of eukaryotes contain large amounts of tandemly repeated DNA, often comprising several percent of an organism's genetic complement. These DNA families contain as many as several thousand repeat units at a given location, arranged in a head-to-tail fashion. Noncoding tandemly repeated DNAs are possible candidates for providing some of the structural requirements for proper chromosome function (14, 20), while multigene families such as ribosomal DNA (rDNA) and immunoglobulin genes are important for the development and biology of complex organisms (16). One interesting but poorly understood property of these DNA families is the high degree of sequence similarity observed among repeat units within a species compared with the relatively low similarity of repeat units between closely related species (5, 16). Models to explain this phenomenon of concerted evolution invoke repeated rounds of homologous genetic exchange processes such as unequal crossing over and conversion, such that mutations arising in individual repeat units can be duplicated and eventually spread throughout the DNA family within a species (24, 30). Genetic exchange processes are thought to be important in shaping complex genomes, resulting in gene duplication, deletion, or fusion (reviewed in reference 15) and contributing to genetic diversity between populations during evolution (8, 32).

Alpha satellite DNA is a tandemly repeated DNA family found in arrays of up to several thousand kilobase pairs at the centromeres of all human and primate chromosomes. The fundamental units of alpha satellite DNA are diverged 171-bp monomers, which are tandemly organized into distinct linear arrangements that form highly homologous chromosome-specific higher-order repeat units (6, 50). These have presumably been homogenized on their particular chromosomes by genetic exchanges between misaligned arrays aligned on the register of the higher-order repeat units (31, 38, 50). Within a particular chromosomal subset, higher-order repeat units that differ in length by integral numbers of monomers but are otherwise closely related in sequence can be detected. These variant repeat units are presumably generated by exchange between misaligned higher-order repeat units aligned on the homologous monomers (43, 44, 50). In these cases, the location of the recombination events can be determined by comparison of the sequence of the variant repeat units to the putative progenitor repeat unit (12, 13, 43, 44).

Two important parameters need to be considered in the examination of these recombination events: the register of (mis)alignment of the two repeat units that are undergoing exchange and the location of the actual recombination event within those sequences. The register of misalignment is in general determined by sequence homology. However, in the case of tandemly repeated DNA families, many such homologous alignments are possible. Data from previous studies suggest that the locations of unequal crossing over and/or conversion events occur preferentially in regions that are among the largest available stretches of perfect sequence identity between the (mis)aligned sequences (19, 29, 36). It is not yet clear, however, what effect secondary characteristics such as chromatin structure may have in the determination of such events. On the basis of a limited number of previously described examples, recombination events in the alpha satellite have been proposed to cluster in a distinct region of

* Corresponding author.

TABLE 1. Alpha satellite repeat unit variants

Name	Chromosome(s)	Source	No. of monomers misaligned[a]	Monomers in crossover	Reference(s)
15-mer	17	PCR	1[b]	12–14	This report
14-mer(AA)	17	PCR	2	12–15	This report
13-mer	17	Clone, PCR	3	9–13	25, 39, 43
12-mer	17	PCR	4	9–14	This report
11-mer	17	PCR	5	9–15	This report
24-mer	17	Cosmid clone	8	13–6	This report
9-mer	17	Cosmid clone	7	9–1	This report
AAA	17	Clone, PCR	—[c]	A–A	44
14-mer(CD)	17	PCR	2	8–11	This report
3kb X	X	PCR	5[d]	7–3	39
3kb X(cl)	X	Lambda clone	5	6–2	This report
CDCD	1, 17, X	Clones	2[e]	D–C	25, 49
EAEA	1, 17	Clones	2[e]	A–E	25, 49

[a] This number is equal to the number of monomers that are deleted or duplicated in the variant repeat unit.
[b] Chromosome 17 repeat unit variants (no. 1 to 9) are described relative to the 16-mer repeat unit (44).
[c] The triple A is an ancestral event in the formation of the present chromosome 17 16-mer and is derived by maximizing the homology between monomers.
[d] Chromosome X repeat unit variants (no. 10 and 11) are described relative to the 2-kb 12-mer repeat unit (42).
[e] These crossovers are derived relative to a canonical pentamer (ABCDE).

the 171-bp monomeric repeat unit rather than being randomly distributed throughout the repeat unit (12, 25, 43, 44). Additionally, this region has been noted (10) to lie in close proximity to the CENP-B box, the 17-bp recognition site for the alpha satellite DNA-binding protein CENP-B (18, 21, 28).

Examination of the molecular basis for multiple recombination events in closely related DNA sequences should provide a better understanding of the biological principles and constraints involved in these processes. To extend previous work, we have determined the molecular basis of a set of 13 variant repeat units from the pentamer subfamily of alpha satellite DNA (suprachromosomal subfamily 3) (1, 2, 49). The positions of recombination events appear to cluster in a window of approximately 20 to 25 bp within the 171-bp monomer. As such, this recombination window may represent an evolutionarily significant beginning of the monomer. Also, since misalignments of higher-order repeat units giving rise to these recombination events invariably involve the juxtaposition of CENP-B boxes, a possible role for the CENP-B protein in genetic exchange processes is suggested.

MATERIALS AND METHODS

PCR amplification, isolation, and sequencing of alpha satellite DNA repeat units. Chromosome 17 alpha satellite DNA was polymerase chain reaction (PCR) amplified with primers 17-1A and 17-2A, as described previously (37). Direct sequencing of PCR products, as well as cloning and sequencing of individual repeat units, was performed as described previously (39). Individual variant repeat units were selected by the size of their insert. Up to seven independent 13-monomer (13-mer) repeat units, as well as at least two independent clones of each of the 15-, 14(AA)-, 12-, and 11-mer repeats, from chromosome 17 were analyzed. A single clone of the 14-mer(CD) was obtained and sequenced in its entirety. In each case, the sequences of the variant repeat units were compared with 16-mers from the same chromosome 17, which had been segregated into mouse-human somatic cell hybrids (37, 39). DNA from a mouse-human somatic cell hybrid containing a single chromosome 17 of African pygmy origin gave an amplification product corresponding to the 11-mer repeat unit, while DNA from nonpygmy origins never yielded an 11-mer PCR product (37). Direct sequencing of PCR-amplified DNA was performed for the 13-, 12-, 11-, and 16-mer repeats from each chromosome, as described previously (39).

Comparative sequence analysis was performed with the aid of the DNAstar, Inc. (Madison, Wis.), software package. Sequencing of multiple clones, as well as direct sequencing of PCR products, identified positions of sequence heterogeneity within higher-order repeat units (39). In the present study, any base change between any clones of a particular repeat type was considered to define a heterogeneous position. A match to either of the bases at a heterogeneous position in the comparison of progenitor and variant repeats was not counted as a difference; this results in a conservative and robust estimate of the location and size of the window of recombination (39).

Primers complementary to chromosome 17 alpha satellite DNA used for sequencing have been described previously (39). Additional sequencing primers used in this study and their positions in the published sequence (41) are as follows: 17-7B, 5′-TTG AAC CCT CCT ATG G-3′ (positions 1108 to 1124); 17-11D, 5′-AAA GGA ATT ATC TTC CCG-3′ (positions 1712 to 1730); 17-16E2, 5′-GAA GTT ATT TCC TTT ACT ACG G-3′ (positions 2568 to 2546).

Isolation of variant alpha satellite repeat units from libraries. Isolation and sequence analysis of lambda clones containing the X chromosome alpha satellite have been described previously (9, 42). One such lambda phage contained a single copy of a 3-kb variant, which was subsequently subcloned as a *Bam*HI fragment. This may represent a rare variant, since there was no evidence for its existence on the basis of PCR analysis (39), and it was identified as a single copy within a lambda clone that otherwise contained tandem nonvariant (2-kb) higher-order repeat units (9). Subcloning and sequence analysis were performed with exonuclease digestions, as described previously (44). It is not clear whether this variant [3kb X(cl); Table 1] or the previously described X chromosome PCR product [3kb X(PCR); Table 1] (39) is present constitutively, since they would be indistinguishable by the genomic restriction enzyme analysis that originally showed the presence of this variant (48, 53).

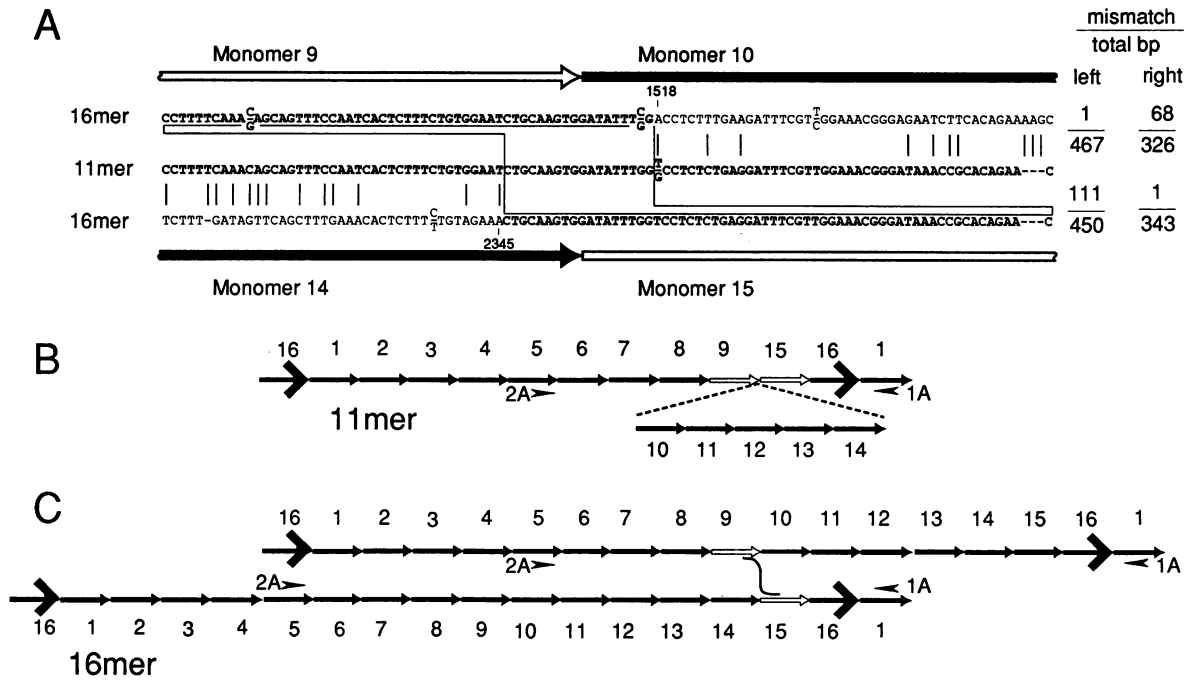


FIG. 1. Analysis of a recombination breakpoint in an alpha satellite repeat unit variant. (A) The abrupt and symmetric breakpoint in sequence similarity at positions 1518 and 2345 allows the recombination event to be localized to a window of 18 bp (boxed region). The sequence similarity is expressed as mismatches per total base pairs for the entire sequenced region of the 11-mer extending to the PCR primers. Heterogeneous positions were determined either by direct sequencing of PCR products or by sequence variation between individual clones. A match to either base at a heterogeneous position was not considered a mismatch. (B) The homology breakpoints allow determination of a deletion of monomers 10 to 14 relative to the 16-mer. (C) The putative register of misalignment between two chromosome 17 16-mers to generate the 11-mer is shown. The 16-mers have misaligned by five monomers. The positions of PCR primers 17-1A and 17-2A in the 16-mer and 11-mer are indicated.

Isolation and characterization of cosmids containing chromosome 17 alpha satellite DNA have been described previously (38). The 9-mer and 24-mer (Table 1) presumably represent low-copy variants because of the fact that they were detected as single copies (not tandemly repeated) in alpha satellite-containing cosmids. Their presence, however, has been confirmed in the hybrid line from which the library was constructed and/or in other rodent-human hybrid lines by hybridization to characteristic unique restriction fragments found in these variants (data not shown).

Computer-modeled statistical significance. The statistical significance of the nonrandomness of the distribution of recombination windows (see Fig. 2B) was tested by using two independent computer algorithms. In general, 13 windows of the length given in Fig. 2B were placed randomly on a 171-unit circle. In the first algorithm, the maximum overlap found at any single point on the 171-unit circle was determined and compared with the value for the actual datum set. Only 28 of 5,000 repetitions ($P = 0.0056$) gave a distribution that was greater than or equal to the maximum overlap for the datum set. In the second algorithm, the number of times that 12 or 13 of the windows overlapped any 25-unit interval within the 171-unit circle (by at least one unit) was determined. This condition reflects the datum set. None of 1,000 repetitions satisfied this condition, giving a value of $P < 0.001$ for the distribution to be seen randomly.

Nucleotide sequence accession numbers. The GenBank accession numbers for the sequences of progenitor and variant repeat units used in the study are L08506 to L08547.

RESULTS

Analysis of a chromosome 17 alpha satellite repeat unit variant. The predominant alpha satellite repeat unit on chromosome 17 is a 2.7-kb 16-mer higher-order repeat unit (16 171-bp monomers) (44). Chromosome 17 alpha satellite DNA is further characterized by several polymorphic repeat unit variants, ranging in size from 11- to 15-mer repeat units (51), which form the basis for haplotypic analysis of this subset (47). 11-mers are found only in an isolated population of African pygmies and thus define a rare pygmy-specific variant (47). A PCR assay has been developed to simultaneously amplify all of these variant repeat units in a haplotype-specific manner by using primers 17-1A and 17-2A (37). The 11-mer PCR product, as well as 16-mer PCR products from the same pygmy-derived chromosome 17 hybrid, was excised from an agarose gel and directly sequenced. Additionally, two independent repeats of each type from the PCR were cloned and sequenced (39).

By aligning the sequences of the 11-mer and the 16-mer repeat units (nucleating at either the 5′ or 3′ end, as shown), we identified the position of the putative recombination event between two 16-mers which gave rise to the 11-mer (Fig. 1). When aligned on PCR primer 17-2A, the sequences were more than 99% identical (one change of 467 bp sequenced) until position 1518 at the beginning of monomer 10 (numbering as in reference 44), after which identity dropped to less than 80% (68 changes of 326 bp). Conversely, when aligned on PCR primer 17-1A, the sequences were less than 76% identical (111 changes of 450 bp) until position 2345 at

the end of monomer 14, beyond which identity was more than 99% (one change of 343 bp). This abrupt and symmetric breakpoint in sequence similarity at positions 1518 and 2345 localizes the recombination event to within an 18-bp window, between the end of monomers 9 and 14 or the beginning of monomers 10 and 15 (Fig. 1A). Thus, monomers 10 through 14 are deleted in the 11-mer relative to the 16-mer (Fig. 1B). This deletion could result from two 16-mers being misaligned by a register of five monomers (Fig. 1C), with a recombination event within the 18-bp window.

Isolation and analysis of multiple repeat unit variants. In order to further investigate recombination in alpha satellite DNA, a total of 13 variant repeat units from the pentamer subfamily were isolated and analyzed (Table 1). PCR products corresponding to the prominent chromosome 17 repeat unit variants (11- to 15-mers), including the previously published PCR-amplified 13-mer (39), were analyzed and compared. Two different 14-mer repeat units, 14-mer(AA) and 14-mer(CD), were cloned from the PCR-amplified DNA. Two additional low-copy chromosome 17 alpha satellite variants, a 9-mer and a 24-mer, have been isolated from a cosmid library. An additional 3-kb X chromosome repeat unit with an origin apparently different from that shown previously by PCR (39) was isolated from a lambda library (see Materials and Methods).

Further examples of repeat unit variation were obtained by examination of the pentamer subfamily of alpha satellite found on chromosomes 1, 11, 17, and X, which is based on a canonical pentameric repeat unit (ABCDE) in which each monomer can be unambiguously assigned to one of five homology groups (44, 49). The different higher-order repeat units on these chromosomes are postulated to have been derived from several crossover events between canonical pentamers (2, 40). The present higher-order repeat units contain three variations from the canonical pentamer (pentameric deviations) consisting of the monomeric pattern ...ABCDCDE... (42, 44), the monomeric pattern ...ABCD EAEAB... (44, 49), and a triplication of highly homologous A monomers resulting in an ...AAA... pattern (44). The 16-mer from chromosome 17 includes all three of these pentameric deviations (44).

Location of windows of recombination. A total of 13 variant repeat units (Table 1) were analyzed to determine whether recombination events occur randomly or within a particular region of the 171-bp monomer. The location of recombination windows was determined (as for Fig. 1B) by the abrupt and symmetric breaks in homology of the variant repeat units to the presumed progenitor repeat in both registers (Fig. 2A).

For each variant repeat, the size of the window to which the recombination can be localized depends on several factors, including the degree of identity of the monomers at the recombination breakpoint (i.e., monomers 9 and 14 and monomers 10 and 15; Fig. 2B). For example, since the recombination event for the 24-mer (Fig. 2A) falls within highly homologous A group monomers from the pentamer family (49), the window is large (43 bp). Similarly, the windows of recombination derived for the 15-mer and 14-mer, which both fall within an alignment of highly homologous A group monomers, are too large (~100 bp each) to be included in Fig. 2A.

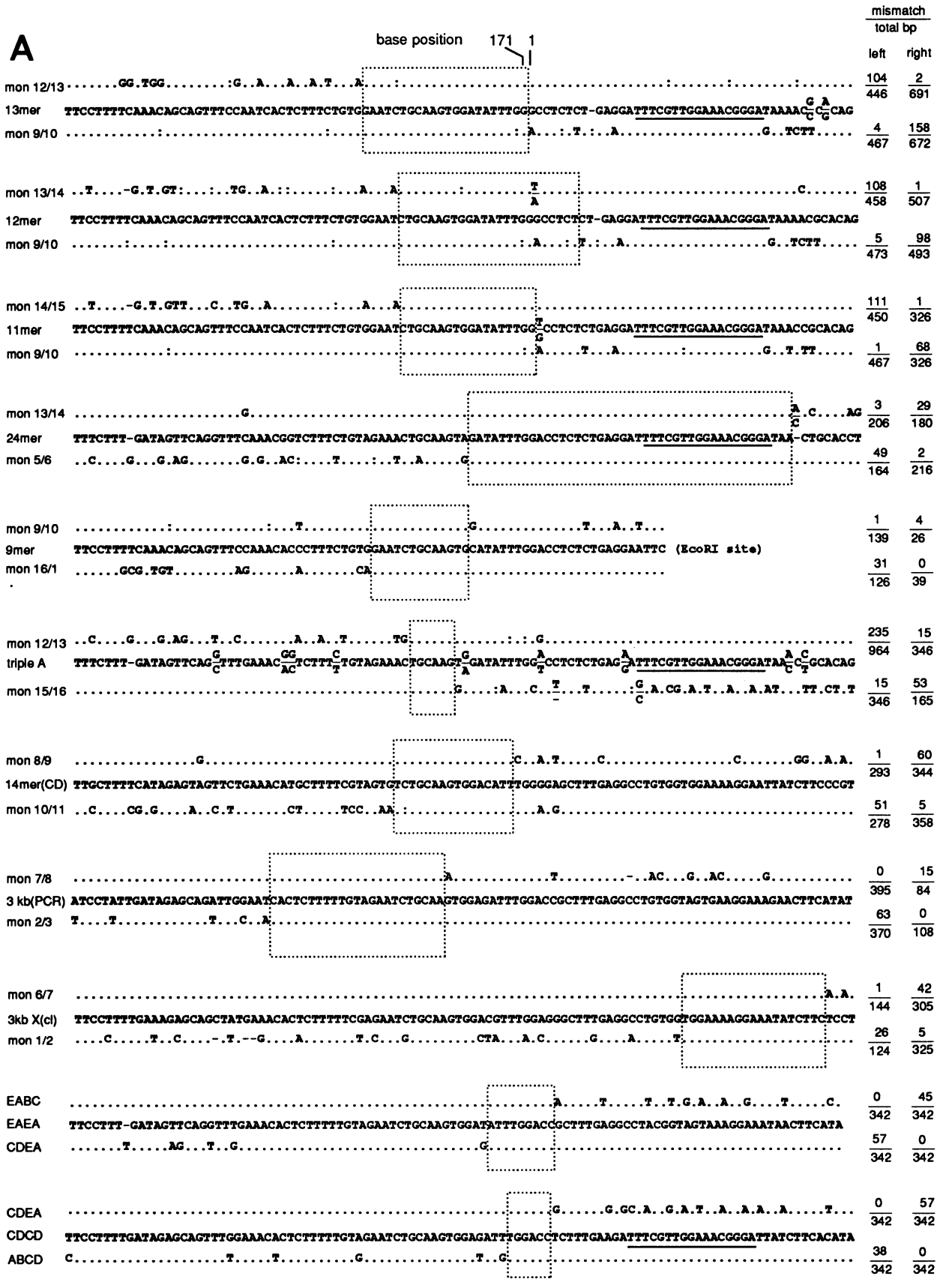
For each of the pentameric deviations (ABCDCDE and CDEAEABC), the monomer register that maximizes homology between duplicated monomers was determined (49). The location of the transition between areas of different levels of sequence identity represents the putative position of recombination, which is analogous to similar analyses of duplicated gene families (11, 19). Independent analysis of the chromosomal higher-order repeat units in which the variants occur (see Table 1) gave similar locations for these windows of recombination, shown as a consensus in Fig. 2A (25, 49). The window size in these cases is established by the number of base pairs that the register can be shifted without decreasing the maximized identity. In a similar fashion, analysis of the -AAA- repeat is derived by maximizing the identity of the three A monomers within the sequence of the 16-mer from chromosome 17 and determining the 171-bp register that maximizes this identity (44).

The sequences in Fig. 2A are aligned on the basis of the homology between monomers, such that the relative location of each window can be seen within the 171-bp repeat unit. This is shown schematically in Fig. 2B, including the size and location of the windows for the 14-mer and 15-mer. The windows of recombination are not randomly distributed throughout this 171-bp region; instead, 12 of the windows overlap in a 20- to 25-bp region. The cloned copy of the X chromosome 3-kb variant falls slightly outside the window of the other pentameric family members. Thus, the windows of recombination cluster in a distinct region of the alpha satellite monomer as opposed to being randomly distributed throughout the 171-bp repeat unit. This clustering is highly significant statistically, as two independent algorithms to test the randomness of this distribution gave P values of 0.0056 and <0.001 (see Materials and Methods).

Registers of misalignments of repeat units. The location of windows of recombination enables analysis of the molecular basis of each of the variant repeat units, as demonstrated explicitly for the 11-mer in Fig. 1. The eight repeat unit variants from chromosome 17 are shown for convenience as deletions (or an insertion) relative to the constitutive 16-mer (Fig. 3). The putative register of misalignment between two 16-mers that would generate these variants (i.e., a misalignment by one monomer gives the 15-mer, one by two monomers gives the 14-mer, etc.) is implied by these deletions (Table 1). Additional misalignments and recombination events between repeat units other than the 16-mer can also be postulated in some cases to give rise to repeat unit variants. However, given the abundance and constitutive nature of the 16-mer in all haplotypes (47), we feel that it is reasonable (and parsimonious) to consider the events implied by Fig. 3.

Each of the two 3-kb variant repeat units from the X chromosome would result from the same misalignment of 2-kb repeat units, with the recombination breakpoint in different locations (39). Both the -CDCD- and -AEAE- pentameric deviations could result from similarly misaligned canonical pentamers, with recombination in different locations (2, 40). The chromosome 17 14-mer(CD) variant involves a deletion (relative to the 16-mer) in the duplicated pairs of CD monomers which in fact restores a canonical pentamer. This variant may represent an ancestral repeat that preceded the -CDCD- duplication (Fig. 3) and therefore may not be an independent event.

Crossover and sequence conversion events have been found to occur preferentially in regions that are among the largest available stretches of sequence identity between two (mis)aligned sequences (19, 29, 36). Therefore, the length and position of regions of sequence identity between misaligned 16-mers from chromosome 17 (in each register) were examined relative to the recombination breakpoints. In three of the eight chromosome 17 variants, in which two or more highly homologous A group monomers are aligned [14-(AA),


A
base position 171 1
mismatch
total bp
left right
mon 12/13
13mer TTCCTTTTCAAACAGCAGTTTCCAATCACTCTTTCTGTGGAATCTGCAAGTGGATATTTGGGCCTCTCT-GAGGATTTCGTTGGAAACGGGATAAAACGCCAG
mon 9/10
104/446 2/691
4/467 158/672
mon 13/14
12mer TTCCTTTTCAAACAGCAGTTTCCAATCACTCTTTCTGTGGAATCTGCAAGTGGATATTTGGGCCTCTCT-GAGGATTTCGTTGGAAACGGGATAAAACGCACAG
mon 9/10
108/458 1/507
5/473 98/493
mon 14/15
11mer TTCCTTTTCAAACAGCAGTTTCCAATCACTCTTTCTGTGGAATCTGCAAGTGGATATTTGGCCTCTCTGAGGATTTCGTTGGAAACGGGATAAACCGCACAG
mon 9/10
111/450 1/326
1/467 68/326
mon 13/14
24mer TTTCTTT-GATAGTTCAGGTTTCAAACGGTCTTTCTGTAGAAACTGCAAGTAGATATTTGGACCTCTCTGAGGATTTCGTTGGAAACGGGATAA-CTGCACCT
mon 5/6
3/206 29/180
49/164 2/216
mon 9/10
9mer TTCCTTTTCAAACAGCAGTTTCCAAACACCCTTTCTGTGGAATCTGCAAGTGCATATTTGGACCTCTCTGAGGAATTC (EcoRI site)
mon 16/1
1/139 4/26
31/126 0/39
mon 12/13
triple A TTTCTTT-GATAGTTCAGTTTGAAACTCTTTTGTAGAAACTGCAAGTGATATTTGGATCCTCTCTGAGATTTCGTTGGAAACGGGATAACTGCACAG
mon 15/16
235/964 15/346
15/346 53/165
mon 8/9
14mer(CD) TTGCTTTTCATAGAGTAGTTCTGAAACATGCTTTTCGTAGTGTCTGCAAGTGGACATTTGGGGAGCTTTGAGGCCTGTGGTGGAAAAGGAATTATCTTCCCGT
mon 10/11
1/293 60/344
51/278 5/358
mon 7/8
3 kb(PCR) ATCCTATTGATAGAGCAGATTGGAATCACTCTTTTTGTAGAATCTGCAAGTGGAGATTTGGACCGCTTTGAGGCCTGTGGTAGTGAAGGAAAGAACTTCATAT
mon 2/3
0/395 15/84
63/370 0/108
mon 6/7
3kb X(cl) TTCCTTTTGAAAGAGCAGCTATGAAACACTCTTTTTCGAGAATCTGCAAGTGGACGTTTGGAGGGCTTTGAGGCCTGTGGTGGAAAAGGAAATATCTTCTCCT
mon 1/2
1/144 42/305
26/124 5/325
EABC
EAEA TTCCTTT-GATAGTTCAGGTTTGAAACACTCTTTTTGTAGAATCTGCAAGTGGATATTTGGACCGCTTTGAGGCCTACGGTAGTAAAGGAAATAACTTCATA
CDEA
0/342 45/342
57/342 0/342
CDEA
CDCD TTCCTTTTGATAGAGCAGTTTGGAAACACTCTTTTTGTAGAATCTGCAAGTGGAGATTTGGACCTCTTTGAAGATTTCGTTGGAAACGGGATTATCTTCACATA
ABCD
0/342 57/342
38/342 0/342

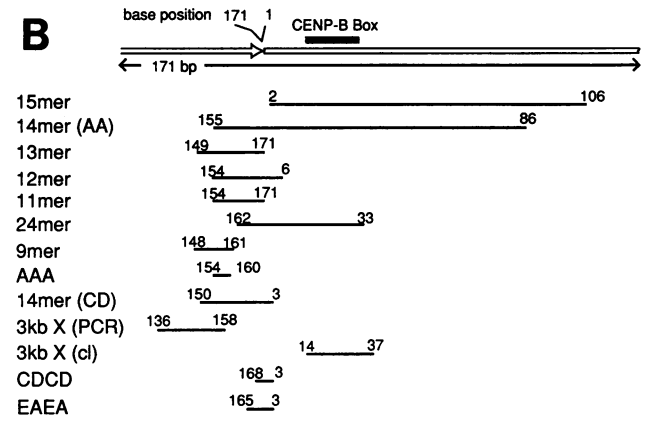


FIG. 2. Location of recombination windows in variant repeat units. (A) Approximately 100 bp of the sequence of each repeat unit around the position of the breakpoint in sequence similarity are shown. The sequences are aligned such that the relative location of each window within the homologous monomers can be seen. In each example, the variant repeat is shown in the middle, with the hypothesized progenitor repeat unit in either register shown above and below for comparison. A dot represents a match to the variant sequence; a colon represents a heterogeneous position that matched the variant repeat unit (see the legend to Fig. 1). Each recombination event can be located to a window of several base pairs (boxed region). The mismatch per total base pairs for the entire available sequence is shown on the right. The locations of CENP-B boxes are underlined. The 15-mer and 14-mer(AA) from chromosome 17 are not shown (see text). The 9-mer is cloned into the *Eco*RI site as indicated; restriction mapping of the cosmid from which this variant originated showed continuation into monomer 1 of a typical chromosome 17 16-mer. (B) Schematic representation of the location of the windows of recombination in panel A relative to the complete 171-bp monomer. The large windows from the 15-mer and 14-mer(AA) are included here. As opposed to being randomly distributed, these windows appear to cluster within a distinct region within the 171-bp monomer.

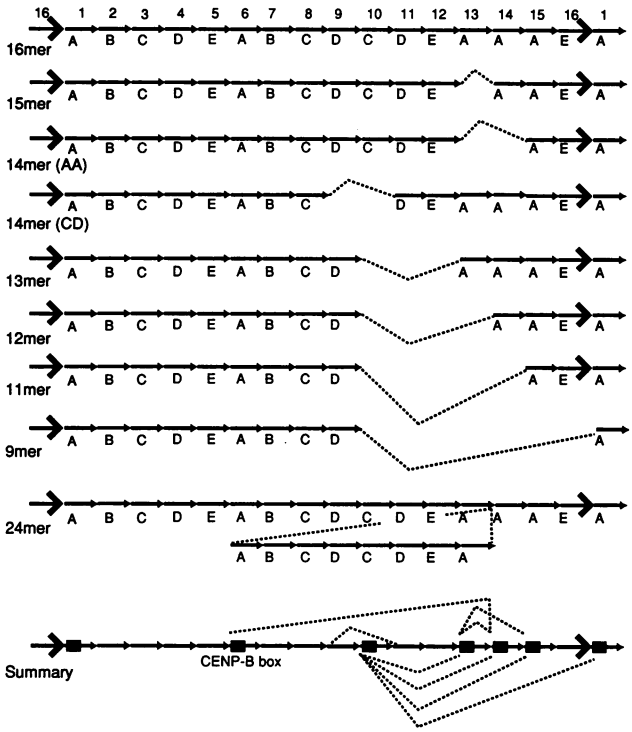


FIG. 3. Structure of variant repeat units from chromosome 17 alpha satellite DNA. The variant repeat units are shown as deletions (and an insertion) relative to the 16-mer, which is the most abundant constitutive chromosome 17 repeat unit. The pentameric homology group (A to E) for each monomer is indicated. At the bottom is a summary of these recombination events, indicating their apparent interrelatedness. The locations of the CENP-B boxes in the 16-mer are shown. In each case, the register of misalignment resulted in the juxtaposition of two CENP-B boxes in close proximity to the position of recombination.

15-, and 24-mer], the breakpoint is found in the longest region of sequence identity available in that register of misalignment. The other recombination events (with the exception of the relatively ancient pentameric deviations, which may have accumulated mutations) occur in windows of sequence identity of 12 to 24 bp. While these regions are among the largest regions of sequence identity in each register of misalignment, there are many other regions of sequence identity of greater or equal length found throughout the repeat units in all registers of misalignment. For example, there are an average of 21 regions of sequence identity of 15 bp or more in each hypothesized register of misalignment (data not shown).

We therefore examined the misaligned 16-mers further to identify any additional features associated with preferential occurrence of the recombination events in these locations. A search for regions of identity that were aligned between misaligned repeats revealed a total of five juxtaposed regions of 15 bp or more in each register of misalignment of chromosome 17 16-mers (Table 2). Of these five regions, region I seems particularly notable, since it contains the recognition site for the centromere-specific alpha satellite DNA-binding protein CENP-B (18, 21, 28) and occurs within several base pairs of seven of the eight recombination events in the chromosome 17 variant repeats (Fig. 3). Examination of these sequences in the derived consensus monomer for the pentamer subfamily (49) reveals that region I has a high degree of sequence heterogeneity among the five monomer groups. In contrast, the other four regions are quite conserved, and in fact region II is found in the consensus of all five monomer groups and in a large number of monomers from chromosomes 17 and X (Table 2). Misalignment involving such conserved regions would be predicted to result in a much larger number of variants, not just the set observed here. Further, regions IV and V would not be juxtaposed in the predicted misalignment of the X chromosome repeat units (Table 2). Thus, the limited pattern of predicted misalignments best correlates with the presence of juxtaposed pairs of region I.

In support of this correlation, we have not observed any evidence for misalignment not involving juxtaposed copies of region I (e.g., misalignment of two 16-mers by six monomers, resulting in a 10-mer or reciprocally a 6-mer). Neither of these variant repeat units has been isolated, nor are they apparent by PCR or on even highly overexposed Southern blots with at least three appropriate restriction enzymes (data not shown). This, of course, does not formally rule out their presence in low copy number or their isolation in an untested population. Finally, the predicted recombination points for both the PCR and cloned X chromosome 3-kb variants (Table 1) are one and two monomers away, respectively, from juxtaposed copies of region I (39) (Table 2), while the recombination points of the two pentameric devi-

TABLE 2. Juxtaposed sequence motifs in misaligned higher-order repeat units

Region	Position[a] and sequence	Chromosome and monomer containing sequence	No. of times the no. of misaligned monomers[b] is juxtaposed in misalignments on:							
			X chromosome	Chromosome 17						
			5	1	2	3	4	5	7	8	6[b]
I	15–30; TTCGTTGGAAACGGGA	17: 1, 6, 10, 13, 14, 15 X: 1, 4, 9	1	2	1	2	3	2	3	2	0
II	54–69; GAAGCATTCTCAGAA	17: 4, 5, 6, 8, 10, 11, 13, 14 X: 1, 2, 3, 5, 7, 10, 11	4	4	4	4	3	3	4	4	4
III	86–101; TGCATTCAACTCACAG	17: 7, 9, 10, 13, 14, 15 X: 12, 5	1	3	2	2	2	2	1	1	2
IV	132–146; TGAAACACTCTTTTT	17: 1, 3, 6, 10, 13, 14, 15 X: 6, 2	0	2	2	3	3	2	3	2	1
V	157–171; CAAGTGGATATTTGG	17: 1, 5, 6, 9, 12, 14 X: 9, 12	0	1	1	3	2	3	2	3	1

[a] Position in pentamer family consensus monomer (49).
[b] Number of monomers misaligned (see Table 1).
[c] A 6-monomer misalignment of chromosome 17 16-mers was not seen (see text).

ations are each one monomer away from aligned monomers which have region I (including the CENP-B box) in their consensus sequences.

DISCUSSION

In this report, we describe the molecular basis for 13 variant higher-order repeat units from the pentamer subfamily of alpha satellite DNA (1, 49). These analyses allow us to examine multiple recombination events in closely related sequences in an attempt to identify possible underlying mechanisms that may affect this process. Several mechanisms have been proposed for the recombination of DNA (for a review, see reference 26). A simple explanation for the generation of these variant repeats involves unequal crossing over between misaligned higher-order repeat units aligned on the register of homologous monomers (Fig. 1C). Unequal crossing over, which unlike other mechanisms is restricted to tandem DNA sequences, satisfactorily explains the generation and local homogenization of higher-order repeat units (30, 31, 50) and accounts for the large size variation observed among alpha satellite arrays on homologous chromosomes (45). Nonetheless, the lack of observed variant repeat units representing reciprocal crossover products may argue for nonreciprocal processes such as sequence conversion and double-strand gap repair (26) in the generation of the variant repeat units. However, given the low probability and stochastic nature of the generation and subsequent fixation of any particular variant repeat unit, it may not be surprising that pairs of reciprocal unequal crossover events are not observed. This might be addressed by further screening of clonal libraries for less abundant variants, such as the 9-mer and 24-mer described here (Fig. 3).

Sequence conversion between misaligned repeat units has been shown to account for repeat unit homogeneity, especially if a conversion bias (lack of parity in heteroduplex repair) is considered, but may not account for changes in array size (23). The paucity of heterologies (due to mismatch repair) around the abrupt and symmetric recombination breakpoints described in this report (Fig. 2A) could suggest a lack of extensive heteroduplex formation. However, a conversion bias in heteroduplex repair in favor of the progenitor repeat unit could also account for the cleanliness of these breakpoints. It seems plausible to us that the existing abundant repeat units may have had the opportunity to accumulate such a bias during their widespread homogenization and fixation. In this light, the relatively few heterogeneous positions detected in multiple clones (see, for example, Fig. 1A and 2A) may represent positions with a greater parity in the direction of mismatch repair. Other models that do not require extensive heteroduplex formation, such as double-strand break repair processes, could also account for formation of these variant repeats if one or both ends of a break misaligned on a monomer register, with subsequent repair by ligation. Since a single recombinational mechanism need not be solely responsible for the homogeneity of higher-order repeat units, differences in array length, and/or the generation of variant repeat units, it seems plausible that the observed patterns of variation in tandemly repeated DNA reflect the complex interplay of several recombinational processes (8).

The beginning and end of a tandem repeat. A general paradigm of tandemly repeated DNA is that repeat units of *n* bp, arranged in a head-to-tail fashion, have an arbitrary beginning or end that can be described as starting at any of *n* registers (31). Any of the 171 positions within tandem monomers reveals the underlying homology, as, for example, any of the 16 171 positions would reveal the underlying homology among tandem 16-mers (50). The historical use of the location of restriction enzyme sites has resulted in many different published starts of homologous repeat units which do not necessarily have any biological or evolutionary significance. For the 13 variant repeat units from the pentamer subfamily described here, the window of recombination was clustered in a distinct region of approximately 20 to 25 bp of the 171-bp monomer (Fig. 2B), allowing us to reject the a priori hypothesis of random distribution throughout the entire 171-bp region. This recombination window delineates a biologically and evolutionarily significant beginning of the 171-bp monomeric repeat (Fig. 2B) since such recombination events are a major force in the generation and homogenization of tandemly repeated DNA (30, 44, 50). One predicted effect of every recombination event occurring within a particular region on the monomer is the evolution of distinct types of 171-bp units. If the location of each event were variable, then many novel monomer types would arise and the distinct monomeric homology groups characteristic of the pentamer family would likely not be apparent (49).

Interestingly, this window of recombination corresponds to the spacer region for the most abundant nucleosome phase determined for alpha satellite DNA (22, 52, 55). The delineation of a preferred location of recombination provides supporting evidence for models in which recombination events occur in the linker region between phased nucleosomes on repeat units. Such models have been previously proposed to explain the evolution of tandemly repeated mammalian DNA (17, 43) as well as the chromatin structure of *Tetrahymena* rDNA genes from different species (27). A limited number of additional windows of recombination have been described for the other suprachromosomal subfamilies (12, 13, 35, 41), raising the interesting possibility that different suprachromosomal subfamilies and chromosome-specific subsets are characterized by different monomeric registers of recombination and, possibly, by different nucleosomal phases. Indeed, several additional minor nucleosome phases were reported for alpha satellite DNA in a study that did not distinguish among different chromosome-specific subsets (52). Additional analysis of the chromatin structure of human alpha satellite DNA from the pentamer and other subfamilies, performed with the present knowledge and utility of chromosome specificity, would be extremely informative.

Registers of misalignment. Data for the 13 variant repeats analyzed here suggest that recombination events in alpha satellite DNA do occur in relatively long regions of sequence identity, which is consistent with the conclusion of Metzenberg et al. (19) based on an analysis of unequal crossover events at the human β-globin locus. However, the large number of such regions in each examined register of misalignment suggests that factors other than (or in addition to) sequence homology affect the position of recombination. In all variant repeat units examined, the proposed misalignment between higher-order repeat units results in the juxtaposition of several sequences of 15 or more base pairs (Table 2), including the binding site for the centromere-specific protein CENP-B. This protein has been reported to form a complex that contains dimerized CENP-B protein and two DNA molecules (21, 54). The close proximity of the recombination junctions to these CENP-B dimer-alpha satellite DNA complexes suggests that they may play a role in these events, for example, by providing a barrier for strand migration or boundaries for double-strand breaks. Significant homology between CENP-B and a protein encoded by the Pogo DNA transposable elements of *Drosophila melanogaster* may further implicate CENP-B in DNA recombination (34).

Alignment of arrays of alpha satellite DNA on the register of a complete higher-order repeat unit may be stabilized by multiple CENP-B dimer-alpha satellite DNA complexes that would be formed along the length of the array (Fig. 4A), which is consistent with models of centromeric higher-order chromatin (21, 54). Upon misalignments of higher-order repeats (Fig. 1C), monomeric registers that retain some CENP-B dimer-alpha satellite DNA complexes (Fig. 4B) might be stabilized over those that do not. Thus, in this model, registers of misalignment (and hence variant repeat units) would be determined in part by the position and number of available pairs of juxtaposed CENP-B boxes. The observed conservation of the 17-bp CENP-B box in otherwise highly diverged centromeric DNA from distantly related species (for example, mouse) (18, 46) would be consistent with this model because identical sequences that are juxtaposed in misalignments would be maintained despite the evolution of surrounding sequences (Fig. 4B). In contrast, other registers of misalignment (i.e., those not favoring a recombination event) could result in the eventual loss of sequences important for recombination over evolutionary time (i.e., thousands of such crossover events) by stochastic processes (7, 30).

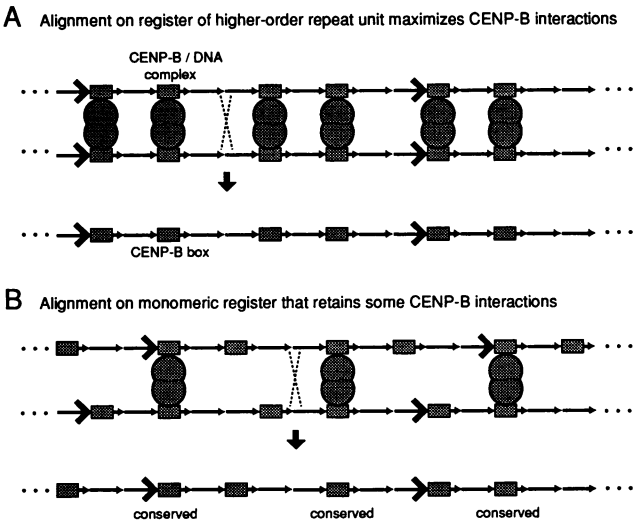


FIG. 4. Model for the role of CENP-B in the misalignment and evolution of alpha satellite DNA. Hypothetical higher-order repeat units each containing four CENP-B boxes (boxed) are shown. (A) Misalignment on the register of the higher-order repeat unit will be stabilized because of the formation of large numbers of CENP-B dimer-alpha satellite complexes. The number and position of CENP-B boxes are maintained. (B) Misalignments on monomeric registers that retain some of these complexes will also be stabilized, as seen for every variant described in this report. Recombination in such misaligned repeat units will result in the conservation of juxtaposed CENP-B boxes.

Tandemly repeated DNA at centromeres. In effect, we have used the description of multiple unequal recombination events to probe indirectly the higher-order structural characteristics of alpha satellite DNA at centromeres. A unique chromatin structure at centromeres would likely be formed by phasing nucleosomes on alpha satellite DNA, because of the spacer regions between nucleosomes being of identical length, as opposed to other regions of the genome that contain phased nucleosomes but do not have similar-length linker regions (3). Such a regular repeating structure over megabase-sized arrays of alpha satellite DNA (45) might lead to formation of large highly compact higher-order chromatin structures, which is consistent with the observation of heterochromatin at centromeres (33). The nucleosomal phase suggested by the window of recombination observed here would place the CENP-B box directly adjacent to the proposed linker region (see Fig. 2B). Thus, it is possible that CENP-B would displace the nucleosome in monomers that contain a CENP-B box, perhaps providing a boundary for the phasing of the nucleosomes along the remainder of the higher-order repeat unit. Although alternative interpretations are possible (28), the number and/or spacing of CENP-B dimer-alpha satellite complexes within arrays (4, 14, 54) may be important for permitting correct interactions between neighboring structures in the formation of functional centromeric heterochromatin.

ACKNOWLEDGMENTS

We thank Rebecca Bernat, Gillian Greig, Rachel Wevrick, and Melanie Mahtani for helpful discussions and comments on the

manuscript. We are especially indebted to Rebecca Bernat for pointing out the proximity of CENP-B boxes to the previously published alpha satellite crossover events. We gratefully acknowledge Melanie Mahtani, David Hull, and the Stanford University statistical consulting services for development of computer programs.

This work was supported by grant HG00107 from the National Institutes of Health.

REFERENCES

1. **Alexandrov, I. A., T. D. Mashkova, T. A. Akopian, L. I. Medvedev, L. L. Kisselev, S. P. Mitkevich, and Y. B. Yurov.** 1991. Chromosome-specific alpha satellites: two distinct families on human chromosome 18. Genomics **11**:15–23.
2. **Alexandrov, I. A., S. P. Mitkevich, and Y. B. Yurov.** 1988. The phylogeny of human chromosome specific alpha satellites. Chromosoma **96**:443–453.
3. **Benezra, R., C. R. Cantor, and R. Axel.** 1986. Nucleosomes are phased along the mouse beta-major globin gene in erythroid and nonerythroid cells. Cell **44**:697–704.
4. **Bernat, R. L., M. R. Delannoy, N. F. Rothfield, and W. C. Earnshaw.** 1991. Disruption of centromere assembly during interphase inhibits kinetochore morphogenesis and function in mitosis. Cell **66**:1229–1238.
5. **Brown, D. D., P. C. Wensink, and E. Jordan.** 1971. A comparison of the ribosomal DNAs of Xenopus laevis and Xenopus mulleri: the evolution of tandem genes. J. Mol. Biol. **63**:57–73.
6. **Choo, K. H., B. Vissel, A. Nagy, E. Earle, and P. Kalitsis.** 1991. A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. Nucleic Acids Res. **19**:1179–1182.
7. **Dover, G. A.** 1982. Molecular drive: a cohesive mode of species evolution. Nature (London) **299**:111–117.
8. **Dover, G. A.** 1986. Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. Trends Genet. **2**:159–165.
9. **Durfy, S. J., and H. F. Willard.** 1989. Patterns of intra- and interarray sequence variation in alpha satellite from the human X chromosome: evidence for short-range homogenization of tandemly repeated DNA sequences. Genomics **5**:810–821.
10. **Earnshaw, W. C., and J. E. Tomkiel.** 1992. Centromere and kinetochore structure. Curr. Opin. Cell Biol. **4**:86–93.
11. **Ellison, J. W., and L. E. Hood.** 1983. Human antibody genes: evolutionary and molecular genetic perspectives. Adv. Hum. Genet. **13**:113–147.
12. **Ge, Y., M. J. Wagner, M. Siciliano, and D. E. Wells.** 1992. Sequence, higher order repeat structure, and long-range organization of alpha satellite DNA specific to human chromosome 8. Genomics **13**:585–593.
13. **Greig, G. M., P. E. Warburton, and H. F. Willard.** The organization and evolution of an alpha satellite subset shared by chromosomes 13 and 21. J. Mol. Evol., in press.
14. **Haaf, T., P. E. Warburton, and H. F. Willard.** 1992. Integration of human alpha satellite DNA into simian chromosomes: centromere protein binding and disruption of proper chromosomal segregation. Cell **70**:681–696.
15. **Hood, L., J. H. Campbell, and S. C. R. Elgin.** 1975. The organization, expression, and evolution of antibody genes and other multigene families. Annu. Rev. Genet. **9**:305–353.
16. **Maeda, N., and O. Smithies.** 1986. The evolution of multigene families: human haptoglobin genes. Annu. Rev. Genet. **20**:81–108.
17. **Maio, J. J., F. L. Brown, and P. R. Musich.** 1977. Subunit structure of chromatin and the organization of eukaryotic highly repetitive DNA: recurrent periodicities and models for the evolutionary origins of repetitive DNA. J. Mol. Biol. **117**:637–655.
18. **Masumoto, H., H. Masukata, Y. Muro, N. Nozaki, and T. Okazaki.** 1989. A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J. Cell Biol. **109**:1963–1973.
19. **Metzenberg, A., G. Wurzer, T. H. J. Huisman, and O. Smithies.** 1991. Homology requirements for unequal crossing over in humans. Genetics **128**:143–161.
20. **Miklos, G. L. G.** 1985. Localized highly repetitive DNA sequences in vertebrate and invertebrate genomes, p. 241–321. *In* J. R. Macintyre (ed.), Molecular evolutionary genetics. Plenum Publishing Corp., New York.
21. **Muro, Y., H. Masumoto, K. Yoda, N. Nozaki, M. Ohashi, and T. Okazaki.** 1992. Centromere protein B assembles human centromeric alpha satellite DNA at the 17-bp sequence, CENP-B box. J. Cell Biol. **116**:585–596.
22. **Musich, P. R., F. L. Brown, and J. J. Maio.** 1982. Nucleosome phasing and micrococcal nuclease cleavage of African green monkey component alpha DNA. Proc. Natl. Acad. Sci. USA **79**:118–122.
23. **Nagylaki, T., and T. D. Petes.** 1982. Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. Genetics **100**:315–337.
24. **Ohta, T.** 1980. Evolution and variation of multigene families. Lect. Notes Biomath. **37**:1–106.
25. **Okumura, K., R. Kiyama, and M. Oishi.** 1987. Sequence analysis of extrachromosomal Sau3A and related family DNA: analysis of recombination in the excision event. Nucleic Acids Res. **15**:7477–7489.
26. **Orr-Weaver, T. L., and J. W. Szostak.** 1985. Fungal recombination. Microbiol. Rev. **49**:33–58.
27. **Palen, T. E., and T. R. Cech.** 1984. Chromatin structure at the replication origins and transcription-initiation regions of the ribosomal RNA genes of Tetrahymena. Cell **36**:933–942.
28. **Pluta, A. F., N. Saitoh, I. Goldberg, and W. C. Earnshaw.** 1992. Identification of a subdomain of CENP-B that is necessary and sufficient for localization to the human centromere. J. Cell Biol. **116**:1081–1093.
29. **Rubnitz, J., and S. Subramani.** 1984. The minimum amount of homology required for homologous recombination in mammalian cells. Mol. Cell. Biol. **4**:2253–2258.
30. **Smith, G. P.** 1976. Evolution of repeated DNA sequences by unequal crossover. Science **191**:528–535.
31. **Southern, E. M.** 1975. Long range periodicities in mouse satellite DNA. J. Mol. Biol. **94**:51–69.
32. **Strachan, T., D. Webb, and G. A. Dover.** 1985. Transition stages of molecular drive in multi-copy DNA families in Drosophila. EMBO J. **4**:1701–1708.
33. **Sumner, A. T.** 1991. Scanning electron microscopy of mammalian chromosomes from prophase to telophase. Chromosoma **100**:410–418.
34. **Tudor, M., M. Lobocka, M. Goodell, J. Pittitt, and K. O'Hare.** 1992. The pogo transposable element family of *Drosophila melanogaster*. Mol. Gen. Genet. **232**:126–134.
35. **Tyler-Smith, C., and W. R. A. Brown.** 1987. Structure of the major block of alphoid satellite DNA on the human Y chromosome. J. Mol. Biol. **195**:457–470.
36. **Waldman, A. S., and R. M. Liskay.** 1988. Dependence of intrachromsomal recombination in mammalian cells on uninterrupted homology. Mol. Cell. Biol. **8**:5350–5357.
37. **Warburton, P. E., G. M. Greig, T. Haaf, and H. F. Willard.** 1991. PCR amplification of chromosome-specific alpha satellite DNA: definition of centromeric STS markers and polymorphic analysis. Genomics **11**:324–333.
38. **Warburton, P. E., and H. F. Willard.** 1990. Genomic analysis of sequence variation in tandemly repeated DNA: evidence for localized homogeneous sequence domains within arrays of alpha satellite DNA. J. Mol. Biol. **216**:3–16.
39. **Warburton, P. E., and H. F. Willard.** 1992. PCR amplification of tandemly repeated DNA: analysis of intra- and interchromosomal sequence variation and homologous unequal crossing-over in human alpha satellite DNA. Nucleic Acids Res. **20**: 6033–6042.
40. **Waye, J. S., S. J. Durfy, D. Pinkel, S. Kenwrick, M. Patterson, K. E. Davies, and H. F. Willard.** 1987. Chromosome-specific alpha satellite DNA from human chromosome 1: hierarchical structure and genomic organization of a polymorphic domain spanning several hundred kilobase pairs of centromeric DNA. Genomics **1**:43–51.
41. **Waye, J. S., S. B. England, and H. F. Willard.** 1987. Genomic

organization of alpha satellite DNA on human chromosome 7: evidence for two distinct alphoid domains on a single chromosome. Mol. Cell. Biol. **7**:349–356.

42. **Waye, J. S., and H. F. Willard.** 1985. Chromosome-specific alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X chromosome. Nucleic Acids Res. **13**:2731–2743.
43. **Waye, J. S., and H. F. Willard.** 1986. Molecular analysis of a deletion polymorphism in alpha satellite of human chromosome 17: evidence for homologous unequal crossing-over and subsequent fixation. Nucleic Acids Res. **14**:6915–6927.
44. **Waye, J. S., and H. F. Willard.** 1986. Structure, organization, and sequence of alpha satellite DNA from human chromosome 17: evidence for evolution by unequal crossing-over and an ancestral pentamer repeat shared with the human X chromosome. Mol. Cell. Biol. **6**:3156–3165.
45. **Wevrick, R., and H. F. Willard.** 1989. Long-range organization of tandem arrays of alpha satellite DNA at the centromeres of human chromosomes: high frequency array-length polymorphism and meiotic stability. Proc. Natl. Acad. Sci. USA **86**: 9394–9398.
46. **Willard, H. F.** 1990. Centromeres of mammalian chromosomes. Trends Genet. **6**:410–416.
47. **Willard, H. F., G. M. Greig, V. E. Powers, and J. S. Waye.** 1987. Molecular organization and haplotype analysis of centromeric DNA from human chromosome 17: implications for linkage in neurofibromatosis. Genomics **1**:368–373.
48. **Willard, H. F., K. D. Smith, and J. Sutherland.** 1983. Isolation and characterization of a major tandem repeat family from the human X chromosome. Nucleic Acids Res. **11**:2017–2033.
49. **Willard, H. F., and J. S. Waye.** 1987. Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. J. Mol. Evol. **25**:207–214.
50. **Willard, H. F., and J. S. Waye.** 1987. Hierarchical order in chromosome-specific human alpha satellite DNA. Trends Genet. **3**:192–198.
51. **Willard, H. F., J. S. Waye, M. H. Skolnick, C. E. Schwartz, V. E. Powers, and S. B. England.** 1986. Detection of restriction fragment polymorphisms at the centromeres of human chromosomes by using chromosome-specific alpha satellite DNA probes: implications for development of centromere-based genetic linkage maps. Proc. Natl. Acad. Sci. USA **83**:5611–5615.
52. **Wu, K. C., F. Strauss, and A. Varshavsky.** 1983. Nucleosome arrangement in green monkey alpha satellite chromatin: superimposition of non-random and apparently random patterns. J. Mol. Biol. **170**:93–117.
53. **Yang, T. P., S. K. Hansen, K. K. Oishi, O. A. Ryder, and B. A. Hamkalo.** 1982. Characterization of a cloned repetitive DNA sequence concentrated on the human X chromosome. Proc. Natl. Acad. Sci. USA **79**:6593–6597.
54. **Yoda, K., K. Kitagawa, H. Masumoto, Y. Muro, and T. Okazaki.** A human centromeric protein, CENP-B, has a new DNA binding motif with partial homology to homeodomain motif at the N-terminus and a dimerization domain at the C-terminus. J. Cell Biol., in press.
55. **Zhang, X. Y., F. Fittler, and W. Horz.** 1983. Eight different highly specific nucleosome phases on alpha satellite DNA in the African green monkey. Nucleic Acids Res. **11**:4287–4306.