Volume 15 Number 18 1987 Nucleic Acids Research

Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes

John S.Waye[1] and Huntington F.Willard[1,2]*

[1]Department of Medical Genetics, University of Toronto, Toronto, Ontario M5S 1A8 and [2]Department of Genetics, Research Institute, Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada

Received June 15, 1987; Revised and Accepted August 27, 1987

ABSTRACT

The human alpha satellite DNA family is composed of diverse, tandemly reiterated monomer units of ~171 basepairs localized to the centromeric region of each chromosome. These sequences are organized in a highly chromosome-specific manner with many, if not all human chromosomes being characterized by individually distinct alphoid subsets. Here, we compare the nucleotide sequences of 153 monomer units, representing alphoid components of at least 12 different human chromosomes. Based on the analysis of sequence variation at each position within the 171 basepair monomer, we have derived a consensus sequence for the monomer unit of human alpha satellite DNA which we suggest may reflect the monomer sequence from which different chromosomal subsets have evolved. Sequence heterogeneity is evident at each position within the consensus monomer unit and there are no positions of strict nucleotide sequence conservation, although some regions are more variable than others. A substantial proportion of the overall sequence variation may be accounted for by nucleotide changes which are characteristic of monomer components of individual chromosomal subsets or groups of subsets which have a common evolutionary history.

INTRODUCTION

A general feature of most eukaryotic genomes is the presence of highly repetitive DNA localized to the centric heterochromatin of chromosomes (reviewed in 1,2). These sequences, commonly referred to as satellite DNAs, consist of tandemly repeated non-coding DNA. Although several possible functions have been ascribed to constitutive heterochromatin and satellite DNA (2-4), none has been conclusively demonstrated and the issue of function remains unresolved.

In primate genomes, the centromeric region of chromosomes is dominated by a diverse class of highly repeated DNA, alpha satellite. These sequences, first identified in the genome of

African green monkey (5,6) and subsequently in those of Old World primates (7,8) and man (9,10), are based on tandemly reiterated monomer units of ~171 basepairs (bp). In the human genome, alpha satellite DNA is located at the centromeric region of each chromosome, constituting as much as 5% of the total genomic DNA. Wu and Manuelidis (11,12) first described human alpha satellite as a 340 bp repeat unit consisting of two diverged monomer halves. However, as originally suggested by Manuelidis (13), the dimer configuration is limited to only a subset of human chromosomes and a majority of chromosomal subsets appear to have other repeat organizations (14).

In recent years, there have been numerous studies concerning human alpha satellite repeat units and their sequences (reviewed in 15). Alpha satellite repeat units have isolated from a number of human chromosomes, including chromosomes 1, 3, 6, 7, 8, 11, 13, 16, 17, 18, 20, 21, 22, X, and Y (see ref. 15 and Table 1 of this study). In addition, the isolation and characterization of several unassigned human alphoid repeats have been reported (16-18). Analysis of the above sequences has revealed a high degree of sequence heterogeneity within the alpha satellite DNA family. Sequence variation between any two human alpha satellite monomers is generally on the order of 20-40%.

Despite the extreme heterogeneity of human alphoid sequences, several attempts have been made to derive a human alpha satellite consensus sequence and to identify evolutionarily conserved sequences within the monomer unit. However, in previous studies the analyses have been based on sequences of a single alphoid domain (generally of a single repeated restriction fragment) and therefore may not be representative of alpha satellite in general. To circumvent this bias, we have surveyed the sequences of monomers which originate on at least 12 different human chromosomes. We suggest, given the varied sources of these sequences in the human genome, that their average or consensus sequence may be representative of the DNA family as a whole and may, therefore, reflect more accurately the sequence of an ancestral alpha satellite monomer.

SURVEY AND RESULTS

Sources of the alpha satellite monomer sequences

In this survey, we have compiled the nucleotide sequences of 153 alpha satellite monomer units from the human genome. This includes all complete and independent sequences of cloned human alpha satellite DNA described in the literature or known to the authors as of June 1987. Multiple copies of highly homologous or identical alpha satellite repeats and incomplete and/or ambiguous sequences have not been included. In this way, we have attempted to avoid any undue bias towards a particular monomer sequence.

Table 1 lists the sources of the monomer sequences included in this survey. The monomer sequences have been grouped according to their chromosomal origins and/or their presumed evolutionary histories. The latter consideration is based on monomer-monomer homologies, with evolutionarily related monomers defining distinct homology groupings. An example of such a relationship is among the monomers which define the pentameric alpha satellite subsets of chromosomes 1, 11, 17, and X (19). These monomers have been grouped into five homology groups (designated groups A, B, C, D, and E), with monomers of a given homology group being significantly more homologous to each other than to monomers of other homology groups (19). Similarly, there are numerous alphoid repeat units which are organized as dimeric units consisting of diverged monomer components (designated groups I and II) (12,20). Although the chromosomal distribution of dimeric alpha satellite sequences is rather ill-defined at present, one particular dimeric subset has been shown to be specific for chromosome 7 (20,21), and it is likely that there exist other specific chromosomal subsets based on the dimeric configuration. In addition to the pentamer- and dimer-based alphoid subsets, there are several subsets whose monomer components belong to presently undefined homology groupings. These include alphoid repeats from chromosomes 7 (a subset distinct from the dimeric subset also found on this chromosome, ref. 20), 8, 18, 22, Y, a subset shared by 13 and 21, and several alphoid repeats whose chromosomal distributions have yet to be established (see Table 1 for references).

Table 1 - Sources of human alpha satellite sequences

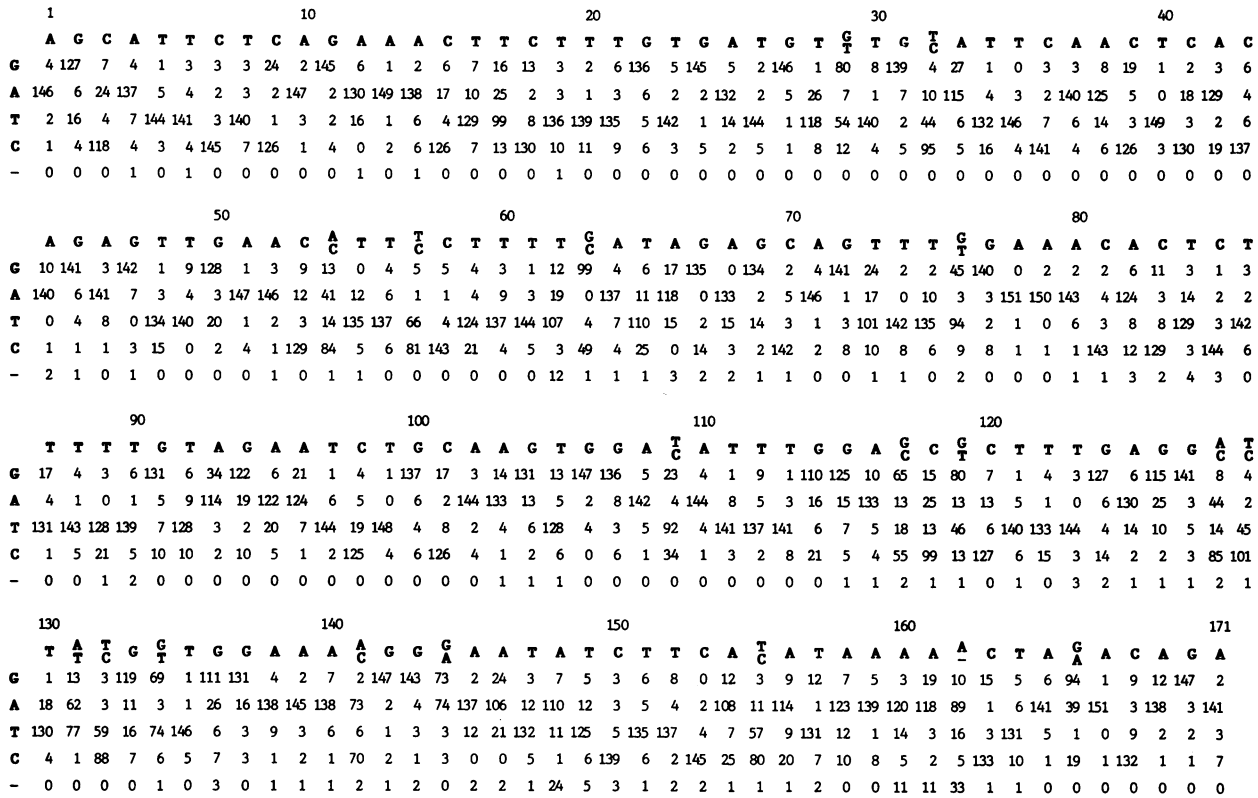
Chromosome[a]	Clone	Number of monomers	Homology groups[c]	References
1	pSD1-1	11	A(3), B(1), C(2), D(2), E(3)	19,46
11	pLC11A	5	A(1), B(1), C(1), D(1), E(1)	19,47
17	p17H8	16	A(5), B(2), C(3), D(3), E(3)	19,48
X	pBamX7	12	A(2), B(2), C(3), D(3), E(2)	19,25,49,50
7	p7d1	2	I(1), II(1)	21
7	p7t1	4	I(2), II(2)	21
-[b]	RI(340)-1′	14	I(14)	20
-[b]	RI(340)-2′	17	II(17)	20
7	pMGB7	16	-	21
8	pJM128	9	-	A. Wyman, unpubl.
13/21	RI(680)21-208	4	-	51
13/21	L1.26	5	-	24
18	L1.84	4	-	24
20	p3-4	8	-	J. Waye, unpubl.
22	p22/1:0.73	2	-	52
22	XI(1020)22-73	4	-	51
22	XI(1020)22-82	4	-	51
Y	cosmid 97	7	-	53
-	pC21	5	-	16
-	pE1	4	-	18

[a] The alphoid subsets of human chromosomes 6 (54), 3 (15), and 16 (15) have also been described. However, since nucleotide sequence data have not been reported, these chromosomal subsets were not included in this survey.

[b] For the dimer sequences of unknown chromosomal origin, more than 20 kb of sequence data have been reported (20). The dimer sequences, RI(340) clones, have been divided into 20 groups based on nucleotide sequence homology (20). For this survey, one member from each group was included where complete monomer sequences were available. This amounted to 14 monomers from the front half of the dimers (1′ or homology group I) and 17 monomers from the back half of the dimers (2′ or homology group II).

[c] The numbers in parentheses indicate the number of monomers belonging to a particular homology group.

Ultimately, one would like to analyze the human alphoid sequences at a chromosomal level and subsequently define monomer homology groups, either chromosome-specific or shared between evolutionarily related subsets of multiple chromosomes. Analysis of nucleotide sequence variation both within and between differ-



	1									10										20										30										40									
	A	**G**	**C**	**A**	**T**	**T**	**C**	**T**	**C**	**A**	**G**	**A**	**A**	**A**	**C**	**T**	**T**	**C**	**T**	**T**	**T**	**G**	**T**	**G**	**A**	**T**	**G**	**T**	**G/T**	**T**	**G**	**T/C**	**A**	**T**	**T**	**C**	**A**	**A**	**C**	**T**	**C**	**A**	**C**
G	4	127	7	4	1	3	3	3	24	2	145	6	1	2	6	7	16	13	3	2	6	136	5	145	5	2	146	1	80	8	139	4	27	1	0	3	3	8	19	1	2	3	6
A	146	6	24	137	5	4	2	3	2	147	2	130	149	138	17	10	25	2	3	1	3	6	2	2	132	2	5	26	7	1	7	10	115	4	3	2	140	125	5	0	18	129	4
T	2	16	4	7	144	141	3	140	1	3	2	16	1	6	4	129	99	8	136	139	135	5	142	1	14	144	1	118	54	140	2	44	6	132	146	7	6	14	3	149	3	2	6
C	1	4	118	4	3	4	145	7	126	1	4	0	2	6	126	7	13	130	10	11	9	6	3	5	2	5	1	8	12	4	5	95	5	16	4	141	4	6	126	3	130	19	137
-	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	50										60										70										80									
	A	**G**	**A**	**G**	**T**	**T**	**G**	**A**	**A**	**C**	**A/C**	**T**	**T**	**T/C**	**C**	**T**	**T**	**T**	**T**	**G/C**	**A**	**T**	**A**	**G**	**A**	**G**	**C**	**A**	**G**	**T**	**T**	**T**	**G/T**	**G**	**A**	**A**	**A**	**C**	**A**	**C**	**T**	**C**	**T**
G	10	141	3	142	1	9	128	1	3	9	13	0	4	5	5	4	3	1	12	99	4	6	17	135	0	134	2	4	141	24	2	2	45	140	0	2	2	2	6	11	3	1	3
A	140	6	141	7	3	4	3	147	146	12	41	12	6	1	1	4	9	3	19	0	137	11	118	0	133	2	5	146	1	17	0	10	3	3	151	150	143	4	124	3	14	2	2
T	0	4	8	0	134	140	20	1	2	3	14	135	137	66	4	124	137	144	107	4	7	110	15	2	15	14	3	1	3	101	142	135	94	2	1	0	6	3	8	8	129	3	142
C	1	1	1	3	15	0	2	4	1	129	84	5	6	81	143	21	4	5	3	49	4	25	0	14	3	2	142	2	8	10	8	6	9	8	1	1	1	143	12	129	3	144	6
-	2	1	0	1	0	0	0	0	1	0	1	1	0	0	0	0	0	0	12	1	1	1	3	2	2	1	1	0	0	1	1	0	2	0	0	0	1	1	3	2	4	3	0

	90										100										110										120									
	T	**T**	**T**	**T**	**G**	**T**	**A**	**G**	**A**	**A**	**T**	**C**	**T**	**G**	**C**	**A**	**A**	**G**	**T**	**G**	**G**	**A**	**T/C**	**A**	**T**	**T**	**T**	**G**	**G**	**A**	**G/C**	**C**	**G/T**	**C**	**T**	**T**	**T**	**G**	**A**	**G**	**G**	**A/C**	**T/C**
G	17	4	3	6	131	6	34	122	6	21	1	4	1	137	17	3	14	131	13	147	136	5	23	4	1	9	1	110	125	10	65	15	80	7	1	4	3	127	6	115	141	8	4
A	4	1	0	1	5	9	114	19	122	124	6	5	0	6	2	144	133	13	5	2	8	142	4	144	8	5	3	16	15	133	13	25	13	13	5	1	0	6	130	25	3	44	2
T	131	143	128	139	7	128	3	2	20	7	144	19	148	4	8	2	4	6	128	4	3	5	92	4	141	137	141	6	7	5	18	13	46	6	140	133	144	4	14	10	5	14	45
C	1	5	21	5	10	10	2	10	5	1	2	125	4	6	126	4	1	2	6	0	6	1	34	1	3	2	8	21	5	4	55	99	13	127	6	15	3	14	2	2	3	85	101
-	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	2	1	1	0	1	0	3	2	1	1	1	2	1

	130									140										150										160										171
	T	**A/T**	**T/C**	**G**	**G/T**	**T**	**G**	**G**	**A**	**A**	**A**	**A/C**	**G**	**G**	**G/A**	**A**	**A**	**T**	**A**	**T**	**C**	**T**	**T**	**C**	**A**	**T/C**	**A**	**T**	**A**	**A**	**A**	**A**	**A/-**	**C**	**T**	**A**	**G/A**	**A**	**C**	**A**	**G**	**A**
G	1	13	3	119	69	1	111	131	4	2	7	2	147	143	73	2	24	3	7	5	3	6	8	0	12	3	9	12	7	5	3	19	10	15	5	6	94	1	9	12	147	2
A	18	62	3	11	3	1	26	16	138	145	138	73	2	4	74	137	106	12	110	12	3	5	4	2	108	11	114	1	123	139	120	118	89	1	6	141	39	151	3	138	3	141
T	130	77	59	16	74	146	6	3	9	3	6	6	1	3	3	12	21	132	11	125	5	135	137	4	7	57	9	131	12	1	14	3	16	3	131	5	1	0	9	2	2	3
C	4	1	88	7	6	5	7	3	1	2	1	70	2	1	3	0	0	5	1	6	139	6	2	145	25	80	20	7	10	8	5	2	5	133	10	1	19	1	132	1	1	7
-	0	0	0	0	1	0	3	0	1	1	1	2	1	2	0	2	2	1	24	5	3	1	2	2	1	1	1	2	0	0	11	11	33	1	1	0	0	0	0	0	0	0

Figure 1 - A human alpha satellite monomer consensus sequence. The consensus sequence (bold) and the base compositions (G, A, T, C) at each position in the monomer are shown (based on comparison of 153 monomers). (-) indicates gaps introduced in the monomer sequences based on the optimal sequence alignment. As described in the text, a consensus base was assigned if the most abundant base is represented at least three times as often as the next abundant base.

ent homology groups would provide a systematic approach to evaluating the evolutionary history of human alphoid sequences. Because such data are, at present, limited to the pentamer- and dimer-based human subsets, we have included in this survey the sequences of all known chromosomal subsets as well as two unassigned alphoid repeat units whose sequences have been reported (Table 1).

For most published sequences, the beginning and end of the alphoid monomer units have been chosen to coincide with the particular restriction endonuclease cleavage site that defines a repeated fragment. As a result, numerous different frames have been reported for alphoid monomers, complicating the task of comparing different monomeric sequences. For this survey, each monomer sequence has been rearranged to begin at a common point; somewhat arbitrarily, for historical reasons, we have chosen the position which corresponds to the HindIII cleavage site which defines the prototype alpha satellite repeat unit of the African green monkey genome (6). The monomer sequences, as arranged and analyzed for this survey, are available on request from the authors.

Derivation of a consensus human alpha satellite monomer sequence

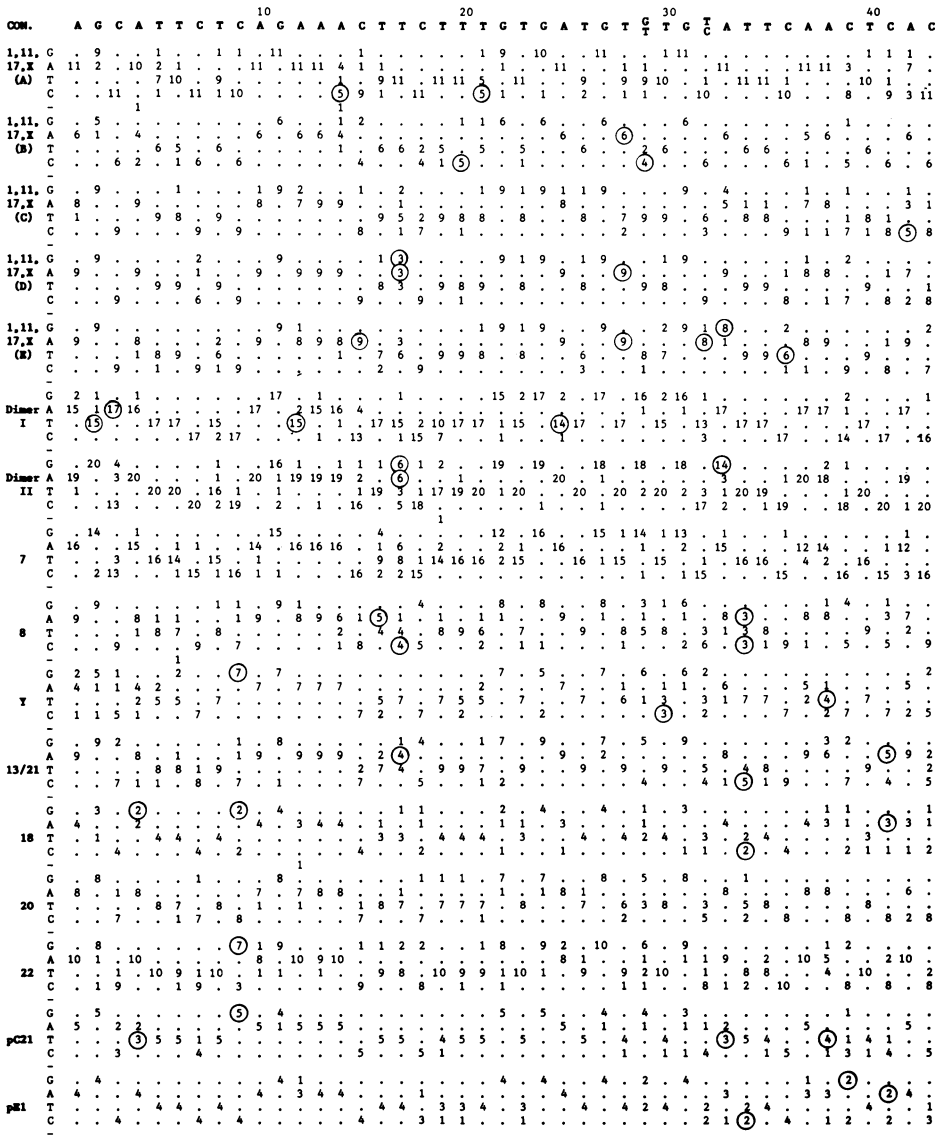
To derive a consensus human alphoid sequence, we have analyzed the nucleotide sequences of 153 monomer units and determined the base distributions at each of the 171 positions (Figure 1). A consensus base has been designated for those positions at which a given base is represented at least three times as often as the next abundant base. Otherwise the two most dominant bases at the position are noted. The overall "strength" of the consensus sequence is measured by the percentage of monomers which are characterized by a consensus base at a given position. At individual positions in the monomer unit, an average of 88% (range 65% to 99%) of the 153 monomers are in agreement with the consensus base designations.

Analysis of nucleotide sequence heterogeneity

In Figure 2, the base compositions have been individually determined for each of the 16 different monomer "classes" (homology groups, chromosomal subsets, and unassigned clones). For a given monomer class, the circled numbers highlight posi-

tions at which a significant proportion of the bases are different from the consensus. This analysis reveals that much of the nucleotide sequence heterogeneity may be accounted for by nucleotide differences (relative to the consensus sequence) that are distributed among the 16 monomer classes in a non-random fashion. For example, at position 122 the vast majority of monomers have a thymine residue (87%), yet 11/11 group A monomers (pentameric) have cytosine at that position. Similarly, at position 2 the dominant base is guanine (83%), yet 15/17 group I monomers (dimeric) have thymine at that position. There are indications that chromosome-specific variation may also be widespread. For example, whereas the majority (82%) of monomers have cytosine at position 9, each of the seven component monomers anaylzed for the Y chromosome subset has guanine at that position. Similarly, the alphoid repeat pC21 (unknown chromosomal distribution) has guanine at position 83 in each of its five monomers, whereas the majority (84%) of monomers from elsewhere in the human genome have cytosine at that position. These and other examples of class-specific sequence variation are summarized in Figure 3.

There are 19 positions in the monomer for which unique consensus bases could not be assigned, based on the criteria we have used in this survey (see Figure 1). Examination of the base compositions for the various monomer classes (Figure 2) indicates that nucleotide heterogeneity is non-random and class-specific at all of these positions. For example, at position 29 (G/T), 50% and 35% of the monomers have guanine or thymine, respectively. Of the monomer classes analyzed, however, only one (pE1) actually has an equal incidence of guanine and thymine at position 29. Another class (homology group B) is dominated by cytosine. The majority of monomer classes are dominated either by guanine (classes I, II, 7, Y, 13/21, 20, 22, and pC21; incidence of G=74, A=4, T=8, C=6) or by thymine (classes A, C, D, E, 8, and 18; incidence of G=4, A=3, T=42, C=2). Base distributions among the different monomer classes are class-specific at the other dimorphic positions also (Figure 2). In this respect, the 19 dimorphic base positions may be viewed as positions characterized by multiple, alternate consensus bases.



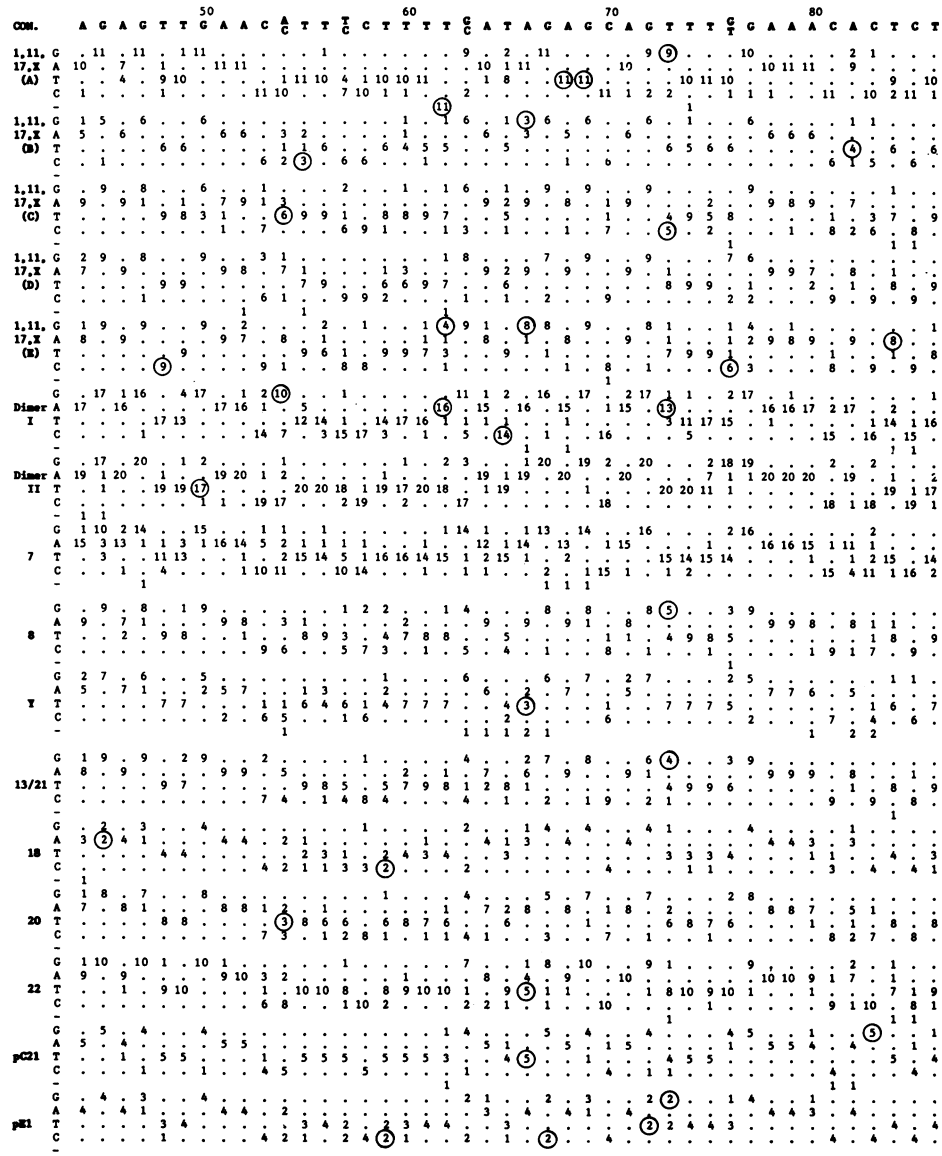


Figure 2 (continued overleaf)

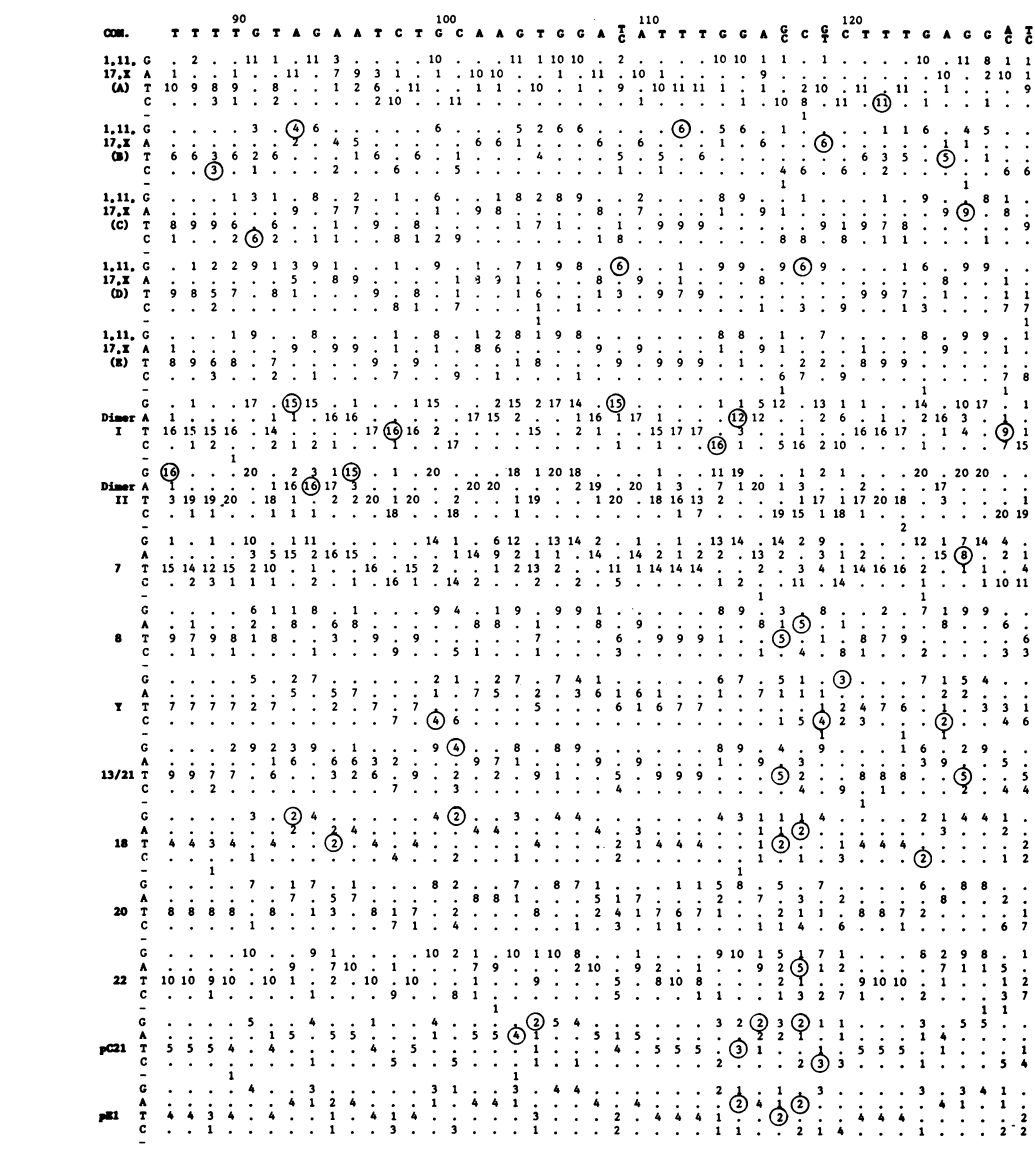
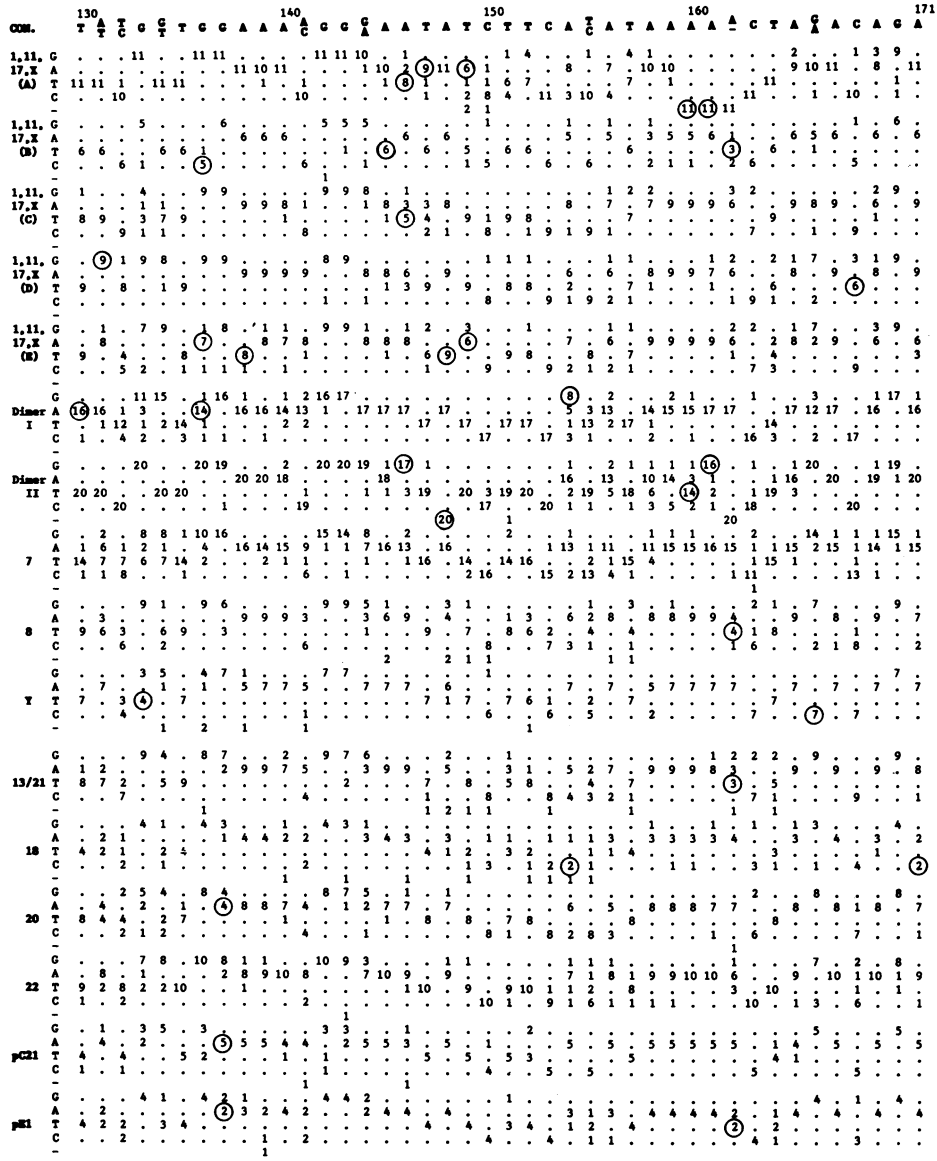


Figure 2 - Analysis of nucleotide sequence heterogeneity in the human alpha satellite DNA family. The base compositions are shown for 16 different classes (homology groups, chromosomal subsets, unassigned clones) of human alpha satellite monomers. The numbers indicate the number of monomers within a class which have a given base [G, A, T, C, or a gap (-)] at a particular



position. For ease of alignment, periods have been substituted for zeros in the G, A, T, and C columns. The circled numbers highlight non-consensus bases in individual classes. These are positions at which the most abundant base differs from the designated consensus base and has an incidence of two or more monomers/class.

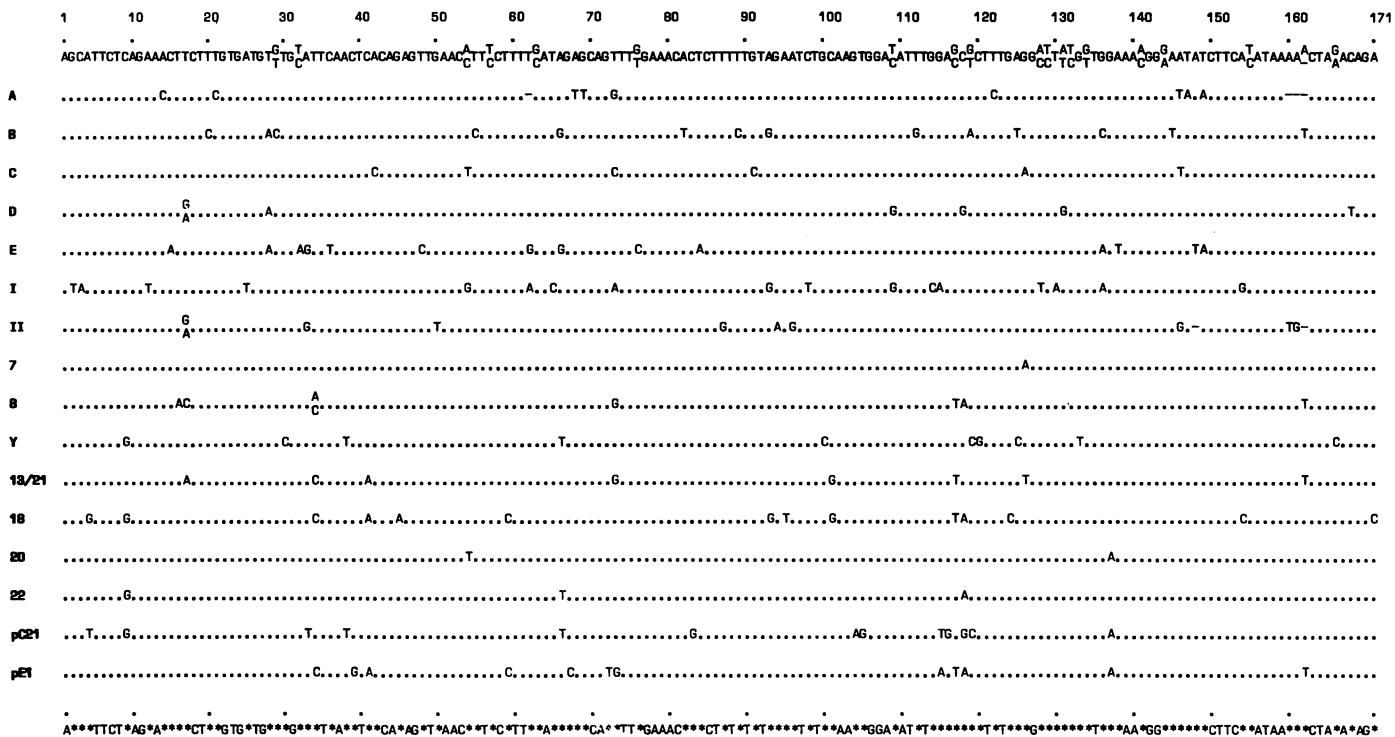


Figure 3 - Distribution of variable base positions in the human alpha satellite monomers. For each of the 16 monomer classes, the positions of non-consensus bases (as defined in Figure 2) are shown. Periods indicate conformity with the consensus. The bottom line indicates the distribution of consensus positions which are not interrupted by non-consensus or dimorphic bases in the 16 monomer classes.

163	CTA$^{G}_{A}$ACA	GAAG	CAT	TCT		8
9	CA	GAAA	C T	TCTTT	GTGATGT$^{G}_{T}$TG$^{T}_{C}$ATTCAAC	39
40	TCACAGAGTT	GAAC	$^{A}_{C}$TT	$^{T}_{C}$CTTTT	$^{G}_{C}$ATAGAG	69
70	CAGTTT$^{G}_{T}$	GAAA	CAC	TCTTTTT	GTAGAAT	97
98	CTGCAAGTG	GA$^{T}_{C}$A	TTTGGA$^{G}_{C}$C	$^{G}_{T}$CTTT	GAGG$^{AT}_{CC}$T$^{A}_{T}$	131
132	$^{T}_{C}$G$^{G}_{T}$TGGAAA$^{A}_{C}$G	G$^{G}_{A}$AA	TA	TCTT	CA$^{T}_{C}$ATAAAA$^{A}_{-}$	162
	Motif	GAAA	C$^{A}_{T}$T	TCT$_{n}$		

Figure 4 - Subrepeats within the human alpha satellite monomer unit. The sequences of four putative subrepeats and two more distantly related repeats are shown. The numbers indicate the positions of each repeat within the monomer consensus sequence.

The distribution of non-consensus and dimorphic base positions (as defined for Figures 1 and 2) has been used to assess the degree to which particular bases have been "conserved" at various positions in the monomer unit. As summarized in Figure 3, it appears that these positions are more or less uniformly distributed throughout the monomer unit, although the apparent concentration of dimorphic positions in the region 109-166 may be notable. There are 78 positions at which the overall consensus base is the consensus base for each of the 16 classes of monomers analyzed (Figure 3, bottom). There are no extended regions of striking sequence conservation within the monomer; the longest stretch of sequence uninterrupted by any non-consensus or dimorphic base position is only 5 bp (extending from positions 77-81 of the monomer unit) (Figure 3). Within this region are found two bases (positions 78 and 79) with the least variation of all monomer positions (~99% homogeneity in each case).

Internal sequence redundancy of the 171 bp alpha satellite monomer unit

It has been proposed that the tandem expansion of short oligonucleotide sequences can lead to the formation of longer

repeat lengths which eventually come to represent the fundamental unit of tandem repitition for a given satellite DNA (2,22). With respect to alpha satellite, there are conflicting views as to whether the 171 bp unit has evolved from the tandem reiteration of a shorter sequence. Some reports have noted the presence of short direct repeats within the monomer unit (23,24) whereas others have found no compelling evidence for internal sequence redundancy (6,12). Considering the substantial degree of sequence heterogeneity among members of the human alpha satellite DNA family (see above), it is unlikely that the analysis of individual monomer sequences could adequately address the possibility of internal sequence redundancy. Instead, analysis of the consensus monomer sequence may provide a more meaningful approach towards resolution of this issue.

We have searched the consensus monomer sequence for direct, internally repeated oligonucleotide sequences. Four diverged repeats were found, each related by one or more mutations to the sequence motif GAAAC(A/T)TTCT$_n$ (Figure 4). Two additional regions in the consensus sequence (positions 107-123 and 137-155) also appear to be loosely related to this motif and to each other (Figure 4). Based on the existence of these six repeats throughout the monomer unit, we propose that the 171 bp monomer may have evolved from a primordial oligonucleotide sequence, similar to the identified motif. Models based on consecutive rounds of unequal crossing-over between tandemly arranged short repeats to explain the generation of a longer repeat have been presented (22,23).

DISCUSSION

Significance of a consensus human alpha satellite sequence

Alphoid DNA, defined as tandemly repeated ~171 bp units which bear appreciable homology to the prototype human (12) or African green monkey (6) consensus monomer sequences, has been detected in the genomes of most primates (7,8). Assuming that the alphoid sequences of all primate species have a common evolutionary beginning and given the likely chromosomal phylogenies of various primate karyotypes, it follows that the monomer unit originated early in the course of primate evolution and that

the various human alpha satellite DNA subsets considered here originated from multiple ancestral repeats which were already distributed among different chromosomal domains before the evolutionary divergence of individual primate species (and karyotypes) (15). Given these considerations, it is not unreasonable to propose the existence of a monomer sequence from which all contemporary primate (and thus human) alphoid sequences have evolved. It is our contention that a consensus sequence derived from representative human alphoid sequences may accurately reflect the sequence of such an evolutionary precursor.

In this report, we have analyzed the sequences of human alpha satellite monomers from at least 12 different chromosomes and derived a consensus human alpha satellite sequence. Approximately one half (81/153) of the monomers surveyed belong to one of seven different homology groups (A, B, C, D, E, I, and II). The remaining monomers have yet to be categorized in this fashion, yet ultimately may define other homology groups. The survey is, therefore, somewhat biased toward the pentamer- and dimer- based human alphoid sequences. However, the pentameric subsets are found on four human chromosomes (19) and dimeric sequences are thought to reside on at least eight different chromosomes (13, our unpublished results). Given the relative abundance of these subsets, the bias in this survey may, in fact, be justified as reflecting the representation of particular sequences in the genome. The consensus is apt to change as new sequences are reported; such changes will likely involve the designation of additional dimorphic positions or the clarification of existing ones.

Consensus alpha satellite sequences have been reported previously for human (12,19,25,26) as well as other primate genomes (6,23,27-29). However, many of these were based on cloned monomers from single chromosomes (25) or related subsets (19) or were based on the sequences of uncloned restriction fragments (e.g., a purified 343 bp BamHI fragment of the baboon genome, ref. 23) and, as such, may not be representative of the total alphoid sequences in those genomes. Indeed, a number of studies have indicated that the alphoid sequences of non-human primate

genomes are also heterogeneous and may be organized as distinct chromosomal subsets (7,30,31). Considering the diverse nature of human alpha satellite and the probable complexity of these sequences in other primate genomes, one would not expect comparisons between the sequence of a single repeated fragment (of unknown genomic distribution or chromosomal location) from one species and a consensus sequence from another species to reveal regions of striking sequence conservation or divergence not evident from the derivation of the consensus sequence itself. Indeed, comparisons of the human consensus sequence derived here with the limited primate sequences available (see Pike et al., ref. 27), including the prototype AGM monomer sequence (6), were unremarkable.

In contrast to the above, the analysis of representative chromosomal subsets from various primate genomes may be more meaningful. Not only would this provide information on the evolution of specific ancestral chromosomal subsets, it would ultimately provide a data base necessary for the derivation of true consensus sequences for different primate alphoid DNAs. In theory, comparisons among the derived consensus sequences of homologous alphoid subsets from each primate genome could then be used to generate an overall consensus primate sequence for that alpha satellite subset(s).

In addition to the proposed evolutionary significance of the alphoid consensus sequence, the consensus may be of practical use for the identification, characterization and alignment of alphoid DNAs (15). Previously, human alphoid sequences have been designated as such based on "significant" nucleotide sequence homology with the dimer consensus sequence of Wu and Manuelidis (12). In some cases, low levels of sequence identity (70-75%) with the dimer consensus have been taken as an indication that a particular repetitive DNA is non-alphoid. Consequently, "new" repetitive DNA families have emerged in the literature (e.g. the "Sau3A family", ref. 26). We suggest that use of the consensus sequence presented here may prevent further ambiguities pertaining to the identification and designation of alphoid repetitive DNAs.

A basis for alpha satellite nucleotide sequence heterogeneity

In this survey, we have attempted to determine the basis for

the high degree of nucleotide sequence variation observed within the human alpha satellite DNA family. We have derived a consensus human alpha satellite sequence and demonstrated that variation from this sequence is non-random. The data are consistent with a model based on the largely independent recent evolution of alpha satellite DNA on heterologous chromosomes. This notion is supported by the observation that many of the non-consensus bases are specific for individual chromosomal subsets (see Figures 2 and 3). The pentamer-based subsets of chromosomes 1, 11, 17, and X may represent variation upon the theme of chromosome independence, whereby an ancestral pentameric repeat has become dispersed among these chromosomes and subsequently evolved in a chromosome-independent manner (see further discussion in refs. 15 and 19). Similarly, it seems reasonable to propose that the various dimeric subsets also share a common origin and have evolved in a chromosome-specific manner after dispersal to multiple human chromosomes.

Evolutionary origins of the alpha satellite monomer unit

Our analysis of the consensus human alpha satellite sequence indicates that it may have evolved from a shorter oligonucleotide sequence which became tandemly arranged in ~171 bp blocks. We note the presence of at least four diverged internal repeats, each related to the sequence motif GAAAC(A/T)TTCT$_n$ (Figure 4). Although there is a high degree of sequence variability among these shorter repeat units, this may merely reflect the length of time involved in the formation of the "original" primate alpha satellite monomer unit from the shorter repeat units (12,23). There are several other examples which illustrate a concept of repeat construction beginning with simple oligonucleotides and building towards complex higher-order repeat units. Most notably, the ~240 bp satellite repeat of mouse (22) and the 1.715 and 1.706 g/cm^3 bovine satellites (32,33) are thought to have evolved from the progressive expansion (amplification) of shorter sequences. During the expansion process, the shorter sequences would be expected to accumulate base changes and, in time, to lose their sequence identity. The larger repeat units (~171 bp monomers and multimeric higher-order repeat units in the case of alphoid DNA) represent more recent units of amplification/

fixation and hence are the most recognizable units of tandem repetition.

Nucleotide sequence conservation

One motivation for the comparison of different alphoid sequences is the identification of "conserved" sequences which may be of functional significance. In particular, previous studies (17,24,25,27) have focused on the sequences which correspond to three protein-binding domains within the African green monkey (AGM) monomer unit (34). These studies have indicated conservation of some or all of these domains in primate species. However, as discussed above, the sequences used to address this issue may only represent a portion of the alphoid subsets and thereby understate the degree of sequence divergence associated with the alphoid sequences of any given species.

Alpha-protein is an HMG-like nuclear protein isolated from cultured AGM cells that has been shown to bind to three regions of the AGM monomer unit (34) and, more generally, to any AT stretches of at least 6 bp length (35). The regions of the consensus human sequence extending from positions 30-39, 106-115, and 143-152 correspond to the AGM alpha-protein binding domains I, II, and III, respectively. Binding region I, represented by the sequence TTAATTCATC in AGM, is substantially different in the consensus human monomer sequence [TG(T/C)ATTCAAC]. Analysis of the 153 monomer sequences reveals that the AT stretch has not been conserved at this position. In fact, only 2/153 monomers have a 6 bp AT stretch in this region. Similarly, regions which correspond to binding domains II and III (GGA(T/C)ATTTGG and G(G/A)AATATCTT, respectively) have not been strictly maintained. However, each does contain a 6 bp AT stretch in a significant proportion of the monomers surveyed (40% and 39%, for domains II and III, respectively). At present, the significance of these observations is unclear. Assuming that there exists a human alpha-protein with binding specificity similar to that of the AGM protein, one might predict that it would not bind to the same regions of human monomers as previously demonstrated for AGM monomers, particularly in the case of region I. Certainly, it would be of interest to determine if such a protein exists and what, if any, its binding specificity is with respect to differ-

ent subsets of human alpha satellite DNA.

The centromeric location of alphoid sequences on each human chromosome has prompted speculation that these sequences (and, by analogy, the satellite sequences found at the centromeres in most higher eukaryotes) may be components of functional centromeres. Limited, yet statistically significant sequence homology has been reported between various satellite DNAs (including the prototype human alpha dimer, ref. 12) and the sequences of functional centromeres in the yeast Saccharomyces cerevisiae (36). Recently, it has been reported that p82H (37), a cloned alphoid sequence which corresponds to a low-copy domain localized to human chromosome 14 (38), exhibits a degree of sequence relatedness with the centromeric sequences of S. cerevisiae (39). However, the regions of homology are not extensive and their significance may be questionable given the interchromosomal variability of S. cerevisiae centromeric sequences (40-43) and the observation that the functional centromeres of S. cerevisiae differ considerably in their sequence and complexity from those of the fission yeast Schizosaccharomyces pombe (44,45). Indeed, the region of the consensus previously (39) implicated as being similar to the yeast centromere DNA element I (positions 39-46) does not appear to be particularly conserved among the monomers examined in this survey.

Lastly, there have been numerous observations of inverted repeats in alphoid DNAs (6,17,23,24). Although no specific functional properties have been associated with such regions, their incidence appears to be a general feature of alphoid DNAs. In the consensus sequence, the most compelling region of symmetry is an imperfect inverted repeat that extends from positions 18-47. Other, less extensive regions of symmetry exist throughout the monomer. It should be noted, however, that these regions of dyad symmetry appear not to have been "conserved" in the different alphoid subsets, i.e. chromosome- or group-specific changes have not occured in a compensatory manner (see Figure 2, positions 18-47). Like other features of alpha satellite monomer sequences, these observations may be more relevant to the evolution of these sequences rather than their possible function(s).

ACKNOWLEDGEMENTS

We thank A. Wyman for providing unpublished sequences. This work was supported by grants from the Medical Research Council of Canada and the March of Dimes Birth Defects Foundation. HFW is a research scholar of the Medical Research Council of Canada.

*To whom correspondence should be addressed

REFERENCES

1. Singer, M.F. (1982) Int. Rev. Cytol. **76**,67-112.
2. Miklos, G.L.G. (1985) In McIntyre, J.R. (ed), Molecular Evolutionary Genetics, Plenum Press, New York, pp. 241-321.
3. John, B. and Miklos, G.L.G. (1979) Int. Rev. Cytol. **58**,1-114.
4. Miklos, G.L.G. and John, B. (1979) Am. J. Hum. Genet. **31**,264-280.
5. Maio, J.J. (1971) J. Mol. Biol. **56**,579-595.
6. Rosenberg, H., Singer, M., and Rosenberg, M. (1978) Science **200**,394-402.
7. Musich, P.R., Brown, F.L., and Maio, J.J. (1980) Chromosoma **80**,331-348.
8. Donehower, L. and Gillespie, D. (1979) J. Mol. Biol. **134**, 805-834.
9. Manuelidis, L. (1978) Chromosoma **66**,1-21.
10. Darling, S.M., Crampton, J.M., and Williamson, R. (1982) J. Mol. Biol. **154**,51-63.
11. Manuelidis, L. and Wu, J.C. (1978) Nature **276**,92-94.
12. Wu, J.C. and Manuelidis, L. (1980) J. Mol. Biol. **142**, 363-386.
13. Manuelidis, L. (1978) Chromosoma **66**,23-32.
14. Willard, H.F. (1985) Am. J. Hum. Genet. **37**,524-532.
15. Willard, H.F. and Waye, J.S. (1987) Trends in Genet. **3**, 192-198.
16. Jones, R.S. and Potter, S.S. (1985) Nucleic Acids Res. **13**, 1027-1042.
17. Shmookler Reis, R.J., Srivastava, A., Beranek, D.T., and Goldstein, S. (1985) J. Mol. Biol. **186**,31-41.
18. Gray, K.M., White, J.W., Costanzi, C., Gillespie, D., Schroeder, W.T., Calabretta, B., and Saunders, G.F. (1985) Nucleic Acids Res. **13**,521-535.
19. Willard, H.F. and Waye, J.S. (1987) J. Mol. Evol. (in press)
20. Jorgensen, A.L., Bostock, C.J., and Bak, A.L. (1986) J. Mol. Biol. **187**,185-196.
21. Waye, J.S., England, S.B., and Willard, H.F. (1987) Mol. Cell. Biol. **7**,347-356.
22. Southern, E.M. (1975) J. Mol. Biol. **94**,51-69.
23. Donehower, L., Furlong, C., Gillespie, D., and Kurnit, D. (1980) Proc. Natl. Acad. Sci. USA **77**,2129-2133.
24. Devilee, P., Slagboom, P., Cornelisse, C.J., and Pearson, P.L. (1986) Nucleic Acids Res. **14**,2059-2073.
25. Waye, J.S. and Willard, H.F. (1985) Nucleic Acids Res. **13**,2731-2743.
26. Kiyama, R., Matsui, H., and Oishi, M. (1986) Proc. Natl. Acad. Sci. USA **83**,4665-4669.
27. Pike, L.M., Carlisle, A., Newell, C., Hong, S.-B., and Musich, P.R. (1986) J. Mol. Evol. **23**,127-137.

28. Prassolov, V.S., Kuchino, Y., Nemoto, K., and Nishimura, S. (1986) J. Mol. Evol. **23**,200-204.
29. Rubin, C.M., Houck, C.M., Deininger, P.L., and Schmid, C.W. (1980) J. Mol. Biol. **136**,151-167.
30. Maio, J.J., Brown, F.L., and Musich, P.R. (1981) Chromosoma **83**,103-125.
31. Lee, T.H.N. and Singer, M.F. (1982) J. Mol. Biol. **161**, 323-342.
32. Pech, M., Streeck, R.E., and Zachau, H.G. (1979) Cell **18**, 883-893.
33. Roizes, G.P. and Pages, M. (1982) In Dover, G.A. and Flavell, R.B. (eds), Genome Evolution, Academic Press, London, pp. 95-112.
34. Strauss, F. and Varshavsky, A. (1984) Cell **37**,889-901.
35. Solomon, M.J., Strauss, F., and Varshavsky, A. (1986) Proc. Natl. Acad. Sci. USA **83**,1276-1280.
36. Fitzgerald-Hayes, M., Clarke, L., and Carbon, J. (1982) Cell **29**,235-244.
37. Mitchell, A.R., Gosden, J.R., and Miller, D.A. (1985) Chromosoma **92**,369-377.
38. Waye, J.S., Mitchell, A.R., and Willard, H.F. (1987) Hum. Genet. (in press)
39. Aleixandre, C., Miller, D.A., Mitchell, A.R., Warburton, D.A., Gersen, S.L., Disteche, C., and Miller, O.J. (1987) Hum. Genet. (in press)
40. Panzeri, L. and Philippsen, P. (1982) EMBO J. **1**,1605-1611.
41. Hieter, P., Pridmore, D., Hegemann, J.H., Thomas, M., Davis, R.W., and Philippsen, P. (1985) Cell **42**,913-921.
42. Mann, C. and Davis, R.W. (1986) Mol. Cell. Biol. **6**,241-245.
43. Neitz, M. and Carbon, J. (1985) Mol. Cell. Biol. **5**,2887-2893.
44. Clarke, L., Amstutz, H., Fishel, B., and Carbon, J. (1986) Proc. Natl. Acad. Sci. USA **83**,8253-8257.
45. Nakaseko, Y., Adachi, Y., Funahashi, S., Niwa, O., and Yanagida, M. (1986) EMBO J. **5**,1011-1021.
46. Waye, J.S., Durfy, S.J., Pinkel, D., Kenwrick, S., Patterson, M., Davies, K.E., and Willard, H.F. (1987) Genomics (in press)
47. Waye, J.S., Creeper, L.A., and Willard, H.F. (1987) Chromosoma **95**,182-188.
48. Waye, J.S. and Willard, H.F. (1986) Mol. Cell. Biol. **6**, 3156-3165.
49. Willard, H.F., Smith, K.D., and Sutherland, J. (1983) Nucleic Acids Res. **11**,2017-2033.
50. Yang, T.P., Hansen, S.K., Oishi, K.K., Ryder, O.A., and Hamkalo, B.A. (1982) Proc. Natl. Acad. Sci. USA **79**, 6593-6597.
51. Jorgensen, A.L., Bostock, C.J., and Bak, A.L. (1987) Proc. Natl. Acad. Sci. USA **84**,1075-1079.
52. McDermid, H.E., Duncan, A., Higgins, M.J., Hamerton, J.L., Rector, E., Brasch, K., and White, B.N. (1986) Chromosoma **94**,228-234.
53. Wolfe, J., Darling, S.M., Erickson, R.P., Craig, I., Buckle, V.J., Rigby, P., Willard, H.F., and Goodfellow, P.N. (1985) J. Mol. Biol. **182**,477-485.
54. Jabs, E.W., Wolf, S.F., and Migeon, B.R. (1984) Proc. Natl. Acad. Sci. USA **81**,4884-4888.