Chromosome Research (2006) 14:735–753
DOI: 10.1007/s10577-006-1078-x

CENP-B box and pJα sequence distribution in human alpha satellite higher-order repeats (HOR)

Marija Rosandić[1], Vladimir Paar[2]*, Ivan Basar[2], Matko Glunčić[2], Nenad Pavin[2] & Ivan Pilaš[3]
[1]*Department of Internal Medicine, University Hospital Rebro, University of Zagreb, 10000 Zagreb, Croatia;* [2]*Faculty of Science, University of Zagreb, 10000 Zagreb, Croatia; Tel: +385-1-4605555; Fax: +385-1-4680336; E-mail: paar@hazu.hr;* [3]*Forest Research Institute, Department of Ecology and Silviculture, 10450 Jastrebarsko, Croatia*
**Correspondence*

Received 23 December 2005. Resubmitted and accepted for publication by Herbert Macgregor 3 June 2006

Key words: alphoid arrays, CENP-B box, higher-order repeat, human alpha satellite, suprachromosomal family

Abstract

Using our Key String Algorithm (KSA) to analyze Build 35.1 assembly we determined consensus alpha satellite higher-order repeats (HOR) and consensus distributions of CENP-B box and pJα motif in human chromosomes 1, 4, 5, 7, 8, 10, 11, 17, 19, and X. We determined new suprachromosomal family (SF) assignments: SF5 for 13mer (2211 bp), SF5 for 13mer (2214 bp), SF2 for 11mer (1869 bp), SF1 for 18mer (3058 bp), SF3 for 12mer (2047 bp), SF3 for 14mer (2379 bp), and SF5 for 17mer (2896 bp) in chromosomes 4, 5, 8, 10, 11, 17, and 19, respectively. In chromosome 5 we identified SF5 13mer without any CENP-B box and pJα motif, highly homologous (96%) to 13mer in chromosome 19. Additionally, in chromosome 19 we identified new SF5 17mer with one CENP-B box and pJα motif, aligned to 13mer by deleting four monomers. In chromosome 11 we identified SF3 12mer, homologous to 12mer in chromosome X. In chromosome 10 we identified new SF1 18mer with eight CENP-B boxes in every other monomer (except one). In chromosome 4 we identified new SF5 13mer with CENP-B box in three consecutive monomers. We found four exceptions to the rule that CENP-B box belongs to type B and pJα motif to type A monomers.

Introduction

The centromere is an essential functional domain for inheritance of eukaryotic chromosomes during cell division. Among the protein components of the centromere, the protein CENP-B, highly conserved in mammalian species, is specifically localized at the centromere on human chromosomes (Earnshaw & Rothfield 1985, Earnshaw *et al.* 1987, Cleveland *et al.* 2003). An abundance of CENP-B boxes was found on all chromosomes except Y on humans, chimpanzee, gorilla and orangutan (Haaf *et al.* 1995). CENP-B protein binds to the 17-bp motif of the CENP-B box sequence that is found in an array of centromere-specific human alpha satellite DNA (Masumoto *et al.* 1989, 1993, Muro *et al.* 1992, Pluta *et al.* 1992).

Alphoid arrays consist of tandem repeats of AT-rich alpha satellite monomer unit. Alpha satellite monomers form chromosome-specific higher-order repeats

Electronic Supplementary Material
Supplementary material is available for this article at http://dx.doi.org/10.1007/s10577-006-1078-x

(HOR) consisting of several monomers, or monomeric organization consisting of diverged monomers (Willard & Waye 1987, Willard *et al.* 1987, Lee *et al.* 1997, Rudd & Willard 2004). Alpha satellite HOR were systematically studied by restriction endonucleases (Maio 1971, Manuelidis & Wu 1978, Willard 1985, Willard & Waye 1987, Wevrick *et al.* 1992, Lee *et al.* 1997). The historical use of restriction enzyme sites has resulted in several different published starts of homologous repeat units. HOR are present in megabase quantities in the centromeric region of all human chromosomes (Willard 1985, Waye & Willard 1986, 1987, Jorgensen *et al.* 1986, Willard & Waye 1987, Wevrick & Willard 1989, Choo *et al.* 1991, Willard 1991, Tyler-Smith & Willard 1993, Warburton & Willard 1996, Lee *et al.* 1997, Alexandrov *et al.* 2001, Rudd & Willard 2004).

In many cases a variety of HOR units can be identified in addition to abundant chromosome-specific HOR. A type of polymorphism found in alphoid arrays are HOR units that differ by an integral number of monomers (monomer addition or deletion), but nonetheless closely related in sequence (Warburton & Willard 1996).

Another approach to identifying HOR is based on computational analysis of genome assembly. However, the sequence of the human genome is not yet complete, and major gaps remain at the centromeric region of each chromosome (Schueler *et al.* 2001, Henikoff 2002, Rudd & Willard 2004). In this way mostly only peripheral HOR are accessible, at the edges of each centromeric region.

The July 2003 assembly of the human genome (Build 34) was analyzed using a combination of BLAST and DOTTER alignment tools (Rudd & Willard 2004). Only the presence of HOR was reported, without detailed HOR structure.

The Key String Algorithm (KSA), our new computational method, was shown to be very effective in identifying and analyzing HOR and their structure from human genome assembly Build 35.1 (Rosandić *et al.* 2003a, 2003b, Paar *et al.* 2005).

In the 17-bp canonical CENP-B box motif 5′-Py TTCGTTGGAAPuCGGGA-3′ (plus strand sequence) only the underlined nucleotides (core recognition sequence) are essential for CENP-B box to bind with CENP-B proteins (Muro *et al.* 1992, Masumoto *et al.* 1993, Ikeno *et al.* 1994, Yoda *et al.* 1996, 1998, Romanova *et al.* 1996, Yoda & Okazaki 1997, Iwahara *et al.* 1998, Tanaka *et al.* 2001, 2004, Masumoto *et al.* 2004, Basu *et al.* 2005). In *de-novo* assembly of human centromeres the role of human centromeres was investigated using various synthetic repetitive sequences; only the combination of both the CENP-B box and HOR provided successful

Table 1. Distribution of CENP-B box and pJα motif (essential parts) in monomers in consensus HOR in human chromosomes 1, 4, 5, 7, 8, 10, 11, 17, 19, and X

Chromosome	*n*mer	c.l. (bp)	Ordinal no. of consensus monomer with CENP-B box	Ordinal no. of consensus monomer with pJα sequence
1	11mer*	1866	6, 8, 10	1, 2, 3
4	13mer*	2211	2, 3, 4, 6	1, 9, 12
5	13mer*	2214	–	–
7	16mer	2734	16	5
8	11mer	1869	5, 9, 11	1, 3, 6, 8, 10
10	18mer*	3058	1, 3, 5, 7, 9, 11, 13, 15	17
11	12mer*	2047	4, 8	2, 12
17	14mer	2379	4, 7, 9, 10, 14	5, 12
19	17mer	2896	15	14
19	13mer*	2214	–	–
X	12mer*	2057	3, 4, 9, 12	1, 6, 11

In each HOR the position of monomer No. 1 is determined by the choice of key string. Column 2: *n*mer identified in Build 35.1 assembly using KSA. An asterisk (*) denotes a plus strand HOR sequence. Otherwise, HOR corresponds to a minus strand sequence. (Note: monomers from Romanova *et al.* (1996) are minus strand sequences.) Dimers are not included in the table. Column 3: HOR consensus length (c.l.) determined using KSA. Column 4: ordinal numbers of monomers containing CENP-B box (essential part). Column 5: ordinal numbers of monomers containing pJα motif (essential part).

binding (Ohzeki *et al*. 2002, Warburton 2004). CENP-B box appears only in alpha satellite HOR (Masumoto *et al*. 1989, Alexandrov *et al*. 2001, Masumoto *et al*. 2004) while no CENP-B boxes were detected in monomeric alpha satellites (Trowell *et al*. 1993, Ikeno *et al*. 1994).

Within the same region of monomeric unit, in some monomers a sequence motif alternative to the CENP-B box was found, recognized by alpha satellite binding protein pJα (Gaff *et al*. 1994, Romanova *et al*. 1996). The 17-bp pJα motif 5′-TTCCTTTTPy CACCPuTAG-3′ (plus strand sequence) reflects some of the nucleotides derived from the alpha satellite monomer which were shown to be effective in binding experiments. A shorter pJα core sequence CCTTTTPyC (Romanova *et al*. 1996), presenting an essential part of the pJα motif, was effective when dimerized, while a number of mutations outside of this core did not abolish binding (Gaff *et al*. 1994, Romanova *et al*. 1996).

Sequence comparison of alpha satellite monomers revealed 12 types of alphoid monomers, which form five suprachromosomal families (SF) (Alexandrov *et al*. 1988, 1991, 2001, Romanova *et al*. 1996, Warburton & Willard 1996). Although each SF has its characteristic types of monomers, they all descend from two basic types, A and B. The differences between A and B types are concentrated in a small region, positions 35 to 51 (for minus strand and base position 1 according to Waye & Willard 1985), which mostly matches functional protein binding sites for pJα in type A and for CENP-B in type B.

In subtypes of alpha satellite DNA consisting of dimers which belong to SF1 and SF2 (-J1J2- and -D1D2-, respectively) (Yoda & Okazaki 1997), the majority of CENP-B boxes are regularly distributed in every other monomer unit leading to the 'every other monomer scheme' (Ikeno *et al*. 1994, 1998, Ohzeki *et al*. 2002). On the other hand, in HOR which belong to SF3, the CENP-B boxes are distributed apparently irregularly and specifically to each chromosome (Masumoto *et al*. 1989, Warburton *et al*. 1993, Yoda & Okazaki 1997). As for pJα motif distribution, no systematic investigation has been reported so far.

In this paper our goal was to identify HOR in Build 35.1 assembly, to determine HOR consensus sequences and SF classification, CENP-B box/pJα motif distributions and to discuss them in a broader framework.

Materials and methods

In this study the distribution of CENP-B box and pJα motif was determined in HOR from Build 35.1

Table 2. Positions of HOR from Table 1 in human chromosomes 1, 4, 5, 7, 8, 10, 11, 17, 19, and X

			Position				
Chromosome	*n*mer	c.l. (bp)	Build 35.1	Physical	Arm	Contig	Clone
1	11mer	1866	222205527	120968174	p11.2	NT_077389.3	BX248407.26
4	13mer	2211	1149048	52492332	c-q	NT_022853.14	AC027271.7
5	13mer	2214	135616328	49431760	c-q	NT_006713.14	AC024586.7
7	16mer	2734	3416085	60906906	c-q	NT_023603.5	AC017075.8
8	11mer	1869	111	46948164	c-q	NT_023678.15	AC118650.5
			13341502	–	–	NT_079518.1	AC137085.2
			47675197	43902189	c-p	NT_007995.14	AC127507.4
10	18mer	3058	28836630	41851162	c-q	NT_079540.1	BX322613.6
11	12mer	2047	488728	51426158	c-p	NT_035158.2	AC126345.11
17	14mer	2379	37777322	22158778	c-p	NT_024862.13	AC131274.9
			41000357	–	–	NT_079564.1	AC145160.1
19	17mer	2896	47180181	24385351	p-12	NT_011295.10	AC073541.4
19	13mer	2214	47283624	–	–	NT_078103.1	AC136499.2
X	12mer	2057	6334979	58436530	c-p	NT_011630.14	AL591645.35
			47241021	61455101	c-q	NT_011669.15	BX537339.3

Column 2: *n*mer identified in Build 35.1 assembly using KSA. Column 3: HOR consensus length determined using KSA. Column 4: starting position of HOR in Build 35.1. Column 5: starting physical position of HOR within each chromosome. Column 6: arm side of the chromosome's cen domain containing HOR. Column 7: Contig containing HOR. Column 8: Clone containing HOR.

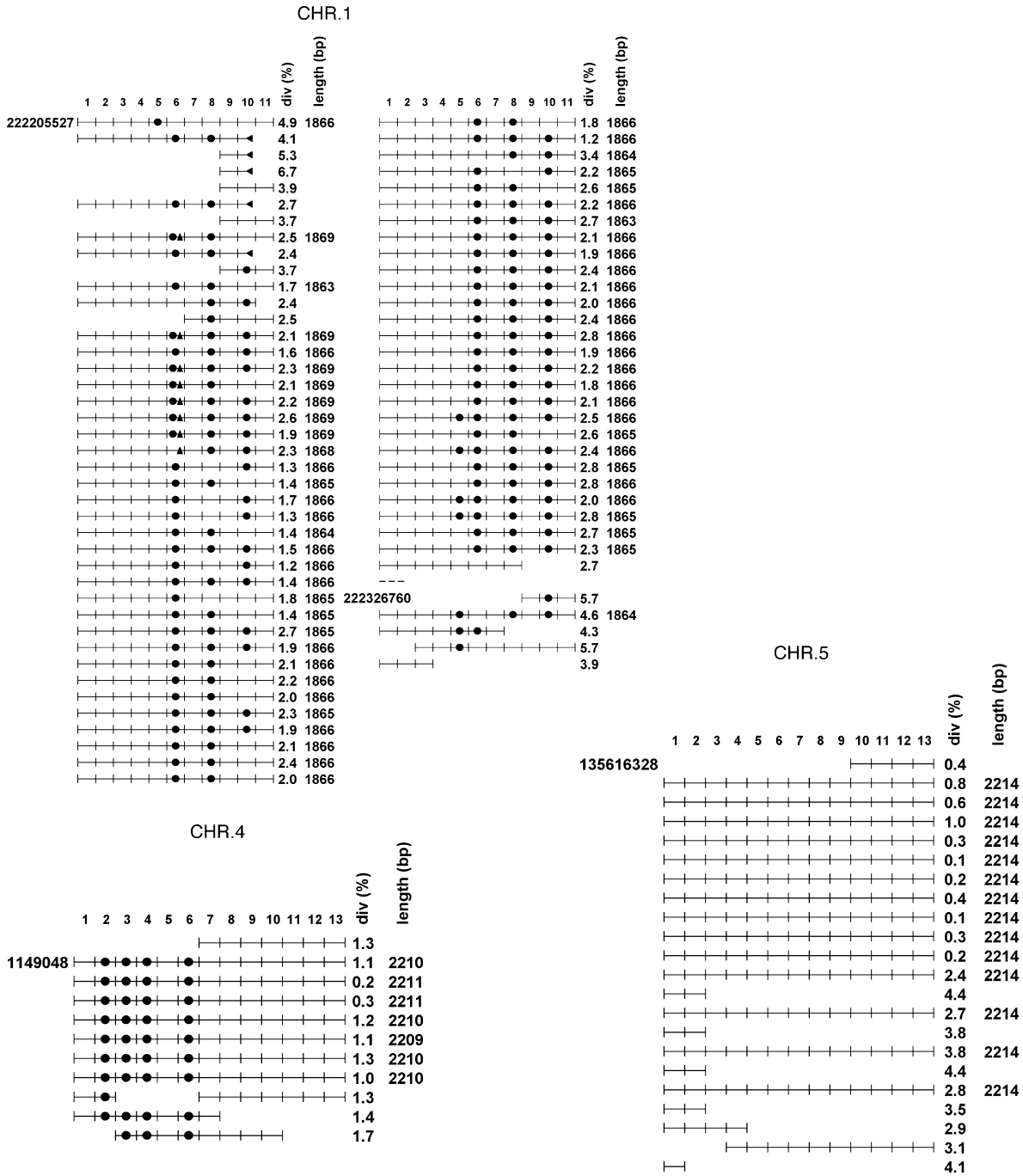

Figure 1. Distribution of CENP-B box (essential part) in all copies within HOR in assembly of human chromosomes 1, 4, 5, 7, 8, 10, 11, 17, 19, and X (Build 35.1). Top row: enumeration of *n* distinct constituent alpha monomers. Monomers within HOR copies are aligned in this scheme. To the left: start position in Build 35.1 of the first HOR copy (or its segment) within chromosome. HOR copies are contained in more than one contig in chromosomes 8 (three contigs), 17 (two contigs) and X (two contigs) (see Table 2). Div (%): divergence of HOR copy (or its segment) to HOR consensus. Length (bp): length of a HOR copy determined by using the KSA method. Closed circle: position of CENP-B box in a monomer. Closed triangle: position of a 3-bp insertion (CTA) in a monomer from chromosome 1. Line with closed triangle at the right end: incomplete monomer (first 97 bases) in chromosome 1.

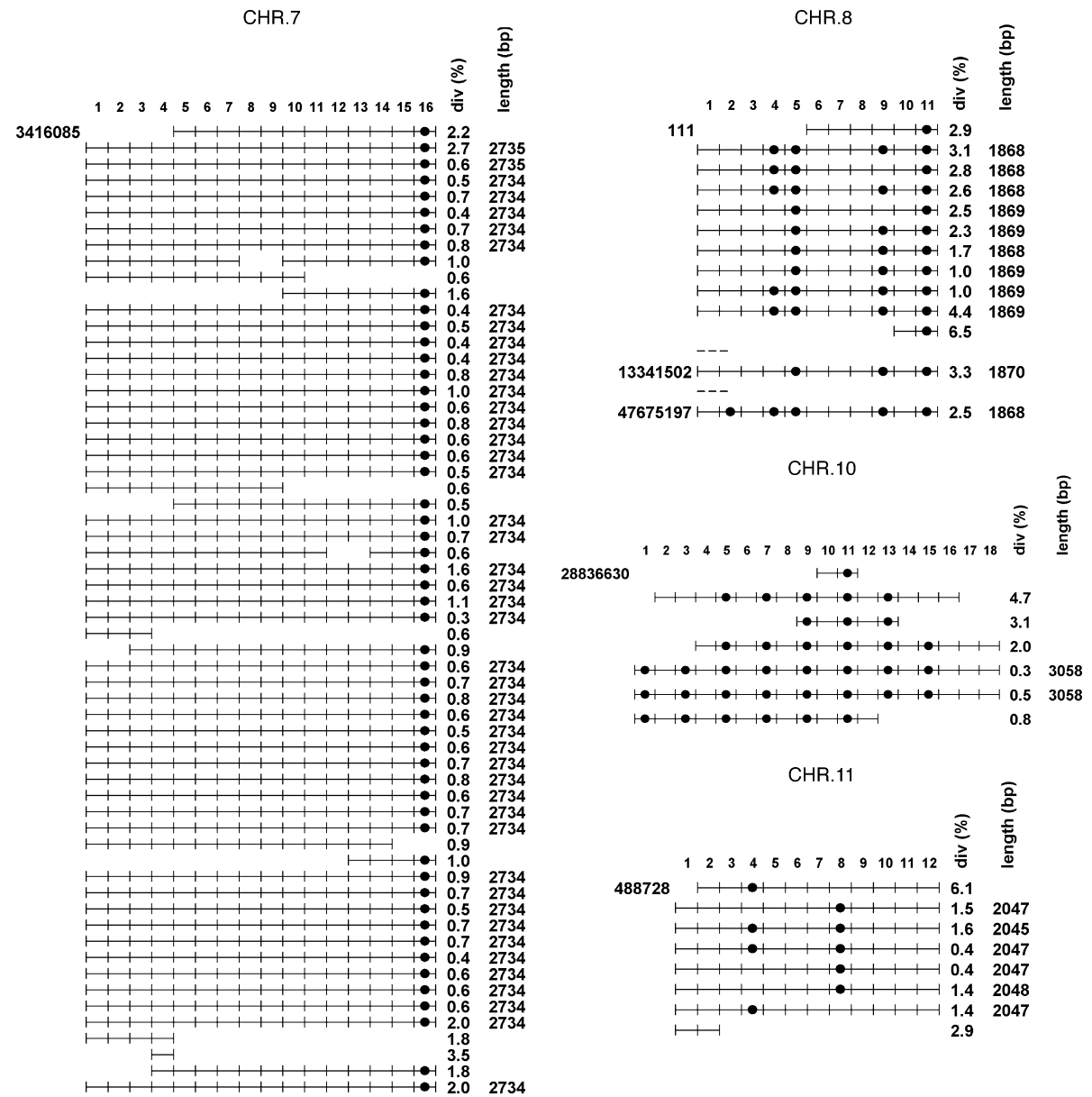

Figure 1. Continued.

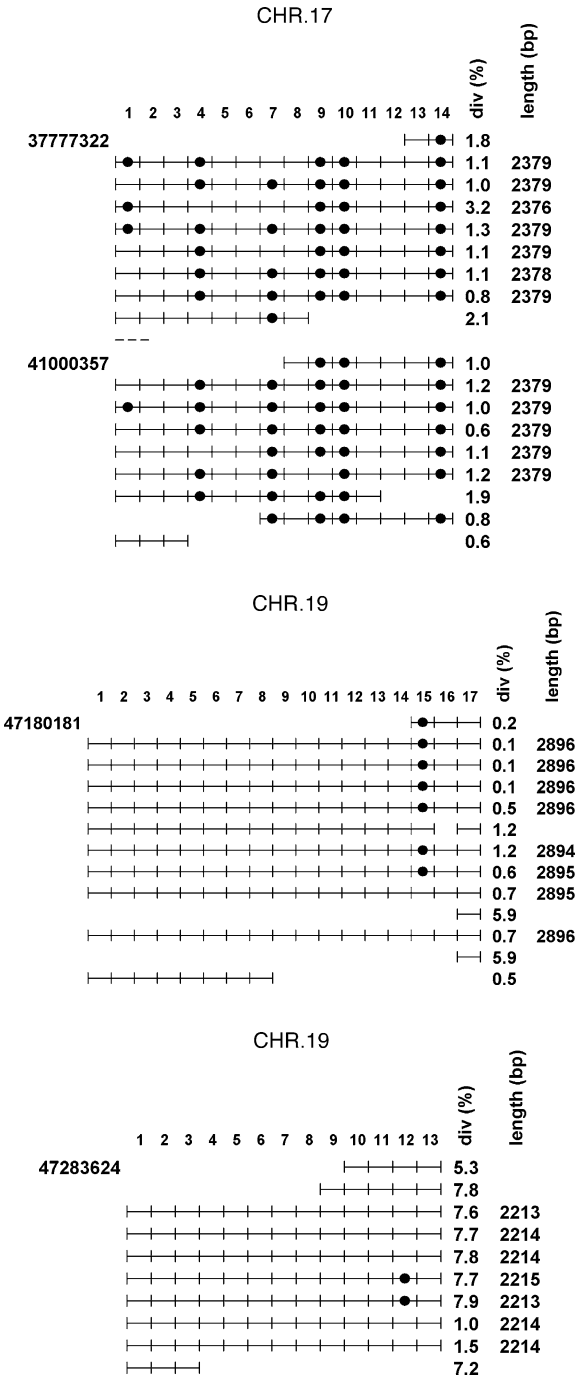

Figure 1. Continued.

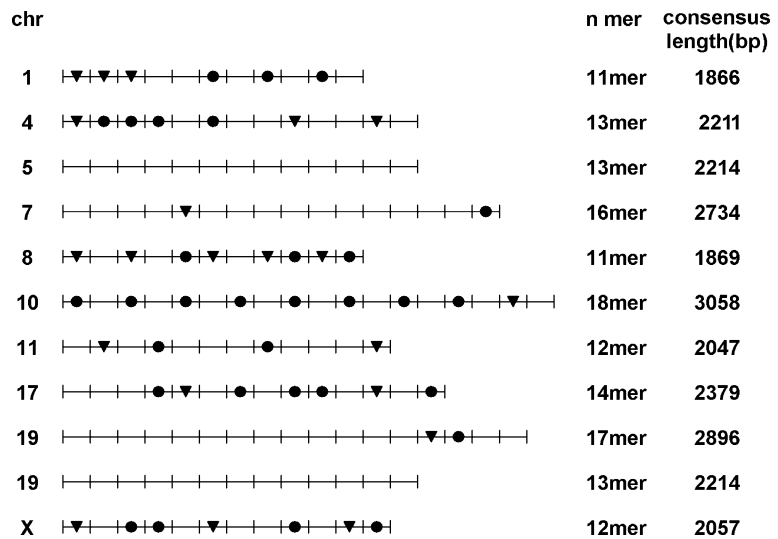

Figure 2. Consensus distribution of CENP-B box and pJα motif in consensus HOR in human chromosomes 1, 4, 5, 7, 8, 10, 11, 17, 19, and X. Closed circle: CENP-B box (essential part) in a constituent monomer. Closed triangle: pJα motif (essential part) in a constituent monomer. To the right of each horizontal line representing consensus HOR: size and length of consensus HOR.

assembly (accessed 24 September 2004) of human chromosomes 1, 4, 5, 7, 8, 10, 11, 17, 19, and X. The Key String Algorithm (KSA) was used to obtain HOR copies and consensus HOR. KSA is a simple and effective computational method to detect and investigate HOR in a given genomic sequence (Rosandić *et al*. 2003a, 2003b, Paar *et al*. 2005).

KSA could be considered as an extended computational analog of the restriction enzyme method. The basic idea behind the KSA is that the computational action of a key string, which is an input to KSA, could be viewed in comparison to the restriction enzyme method. The key string cuts computationally a given single-stranded DNA sequence into fragments (KSA fragments) selectively at the beginning of each key string sequence (KSA site, which could be compared to a restriction site for restriction enzymes). The restriction enzymes cleave double-stranded DNA selectively at specific palindrome sequences, while in the KSA we can use without limitation any sequence as a computational key string. The lengths of generated KSA fragments, forming a distance array, could

Table 3. Comparison of monomers in 13mer consensus HOR in chromosome 5 to SF monomers (divergence (%))

	m01	m02	m03	m04	m05	m06	m07	m08	m09	m10	m11	m12	m13
J1	28	25	21	19	22	21	21	15	21	22	23	22	19
J2	32	29	29	25	24	24	29	23	20	26	22	21	26
D2	25	22	16	18	19	17	17	12	18	18	16	18	15
D1	26	22	22	21	18	19	23	16	14	20	18	16	17
W4	26	22	15	18	18	16	19	11	14	14	16	16	12
W1	32	26	23	25	22	22	27	19	16	24	20	18	21
W5	32	26	22	21	23	21	24	15	20	22	23	21	19
W2	31	26	25	23	22	23	27	22	18	26	21	18	22
W3	30	24	23	25	22	23	27	19	18	25	21	18	21
M1	26	20	16	15	18	16	17	11	14	17	15	15	14
R2	23	17	12	13	15	14	14	9	12	14	13	13	12
R1	24	17	17	17	15	16	18	12	9	18	12	11	14

Underlined: SF monomer having lowest divergence to monomer *mn* (*n*th monomer) in consensus HOR.

Table 4. Comparison of monomers in 18mer consensus HOR in chromosome 10 to SF monomers (divergence (%))

	m01	m02	m03	m04	m05	m06	m07	m08	m09	m10	m11	m12	m13	m14	m15	m16	m17	m18
J1	30	11	26	12	28	10	28	9	25	11	26	10	26	10	29	10	28	8
J2	13	30	11	31	13	31	12	29	11	29	12	30	11	29	14	32	12	28
D2	25	22	21	22	24	22	23	21	21	22	21	22	22	21	25	22	23	18
D1	22	25	19	24	21	26	20	24	18	23	21	25	19	23	21	26	23	23
W4	26	21	21	18	23	20	23	19	21	18	22	20	22	18	23	21	23	18
W1	21	28	19	28	21	30	20	27	16	29	19	28	18	28	21	30	21	28
W5	28	24	25	24	28	23	28	22	25	25	25	23	26	23	29	23	28	21
W2	26	32	22	29	24	30	25	27	21	30	23	30	23	28	26	31	26	28
W3	23	29	20	27	24	30	22	26	20	28	21	28	21	26	25	30	23	27
M1	23	20	19	20	23	21	22	18	20	19	22	19	21	18	24	21	22	17
R2	21	18	17	18	21	18	20	16	18	18	19	17	19	16	22	18	20	14
R1	19	22	15	21	18	22	18	21	15	20	16	22	16	19	19	23	19	19

Underlined: SF monomer having lowest divergence to monomer *mn* in consensus HOR.

be compared to an array of lengths of hypothetical restriction fragments resulting from complete digestion, cutting DNA at a recognition site corresponding to a chosen key-string sequence.

Analyzing KSA distance arrays, we identify and determine the detailed structure of HOR, including all substitutions, deletions and insertions. While the choice of restriction sites for restriction enzymes is severely limited, the choice of key string, i.e. KS sites in the KSA method, is not restricted. In particular, a HOR-specific key string segments the sequence into HOR and monomers. KSA is a very robust method, effective even in the case of large deletions, insertions and substitutions. This method enables determination of detailed HOR annotation and structure in a given genomic sequence. KSA enables a straightforward ordering of KSA fragments. The size of KSA fragments is not limited, regardless of whether one deals with small fragments of a few nucleotides or as many as hundreds of kilobases.

Using computed HOR copies and consensus HOR, we determined the SF classification and CENP-B box/pJα motif distributions.

Results

Using the KSA computational method, HOR were identified in Build 35.1 assembly of human chromosomes 1 (11mer), 4 (13mer), 5 (13mer), 7 (16mer), 8 (11mer), 10 (18mer), 11 (12mer), 17 (14mer), 19 (17mer and 13mer), and X (12mer). Previously we identified HOR and derived exact consensus lengths (Paar *et al*. 2005) for most of the HOR from Table 1. The 13mer in chromosome 4 is determined here for the first time. Positions of HOR in genomic sequences of chromosomes are displayed in Table 2.

Here we derived HOR consensus sequences and determined the monomer structure of all HOR copies and divergence of each copy with respect to consensus HOR. Alignment of constituent monomers of HOR are shown in Figure 1. This pattern is in accordance with the fact that structural variants of HOR usually differ in length as a result of the presence or absence of an integral number of monomers. In constituent monomers of each HOR copy we identified the reduced CENP-B box positions (closed circles in Figure 1).

In the next step the monomer structure of consensus HOR and the corresponding consensus distri-

Table 5. Suprachromosomal family (SF) classification of HOR from Table 1. Base position 1 within monomers was assigned in a standard way (Waye & Willard 1985, Romanova *et al*. 1996).

	HOR																				
Chromosome	*nmer*	c.l. (bp)	SF consensus monomer structure of HOR																		SF
1	11mer*	1866	W4	W3	W4	W3	W2	W1	W5	W1	W5	W1	W5								3
			p2	p4	p1	–	–	C0	–	C0	–	C1	–								
4	13mer*	2211	R2	R1	R2	R2	R2	R1	R1	R2	R2	R2	R2	R2	R2						5
			p1	C3	C3	C2	–	C3	–	–	p0	–	–	p0	–						
5	13mer*	2214	R2	R2/R1	R2	R2	R2/R1	R2	R2	R2	R1	R2/W4	R1	R1	R2						5
			–	–	–	–	–	–	–	–	–	–	–	–	–						
7	16mer	2734	R1	R1	R1	R2	R2	R1	R2	R2	R2	R2	R1	R2	R1	R2	R2	R1			5
			–	–	–	–	p2	–	–	–	–	–	–	–	–	–	–	C0			
8	11mer	1869	D2	D1	D2	D1	D1	D2	D1	D2	D1	D2	D1								2
			p3	–	p1	–	C0	p2	–	p0	C0	p1	C0								
10	18mer*	3058	J2	J1	J2	J1	J2	J1	J2	J1	J2	J1	J2	J1	J2	J1	J2	J1	J2	J1	1
			C0	–	C0	–	C1	–	C0	–	C0	–	C0	–	C0	–	C2	–	p4	–	
11	12mer*	2047	W3	W4	W3	R1	W1	W5	W4	W3	W2	W1	W5	W4							3
			–	p1	–	C2	–	–	–	C1	–	–	–	p5							
17	14mer	2379	W2	W3	W4	W3	W4	W5	W1	W5	W1	W2	W3	W4	W5	W1					3
			–	–	–	C0	p1	–	C0	–	C2	C1	–	p2	–	C0					
19	17mer	2896	R2	R2	R2/R1	R2	R2	R2/R1	R2	R2	R1	R1	R2/W4	R1	R2	R2	R1	R1	R2		5
			–	–	–	–	–	–	–	–	–	–	–	–	–	p0	C0	–	–		
19	13mer*	2214	R2	R2	R2/R1	R2	R2	R2/R1	R2	R2	R2	R1	R2/W4	R1	R1						5
			–	–	–	–	–	–	–	–	–	–	–	–	–						
X	12mer*	2057	W4	W3	R1	W1	W5	W4	W3	W2	W1	W5	W4	W3							3
			p3	–	C2	C0	–	p1	–	–	C0	–	p2	C0							

Column 2: the number of monomers in HOR. An asterisk (*) denotes that monomers in HOR correspond to a plus strand sequence; otherwise, monomers in HOR correspond to minus strand sequence. Column 3: HOR consensus length. Column 4: SF classification of monomers in consensus HOR. Below each sequence of SF consensus monomers is the corresponding CENP-B box (C*m*) or pJα motif (p*m*). C*m* denotes a CENP-B box with *m* base differences from the canonical CENP-B box outside of its essential part; p*m* denotes a pJα motif with *p* base differences from the canonical pJα motif outside of its essential part. It was suggested that a monomer can be assumed to be box-positive if, in addition to agreement at essential sites, the number of mismatches to the canonical CENP-B consensus sequence is ≤ 3 (Kouprina *et al*. 2003). Here, this criterion is satisfied for all CENP-B box sequences. Column 5: suprachromosomal family (SF).

Table 6. Comparison of present HOR/SF results of KSA analysis for Build 35.1 assembly to previous compilations (Lee *et al*. 1997, Alexandrov *et al*. 2001)

	Present		Alexandrov *et al.*				Lee *et al.*		
Chr.	*n*mer/length (bp)	SF	Locus	*n*mer/length (bp)	SF	Ref.	Locus	*n*mer/length (bp)	Ref.
1	11mer/1866	3	D1Z5	11mer/1900	3	a	D1Z5	11mer/1900	i
4	13mer/2211	5	–	–		–	D4Z1	15mer/2600	j
							D4Z1	4mer/600	j
5	13mer/2214	5	D5Z1	–	5	b	D5Z1	13mer/2250	b
			D5Z12	–	5	c			
7	16mer/2734	5	D7Z2	16mer/2700	5	d	D7Z2	16mer/2700	d,k
8	11mer/1869	2	D8Z2	15mer/2550	2	e	D8Z2	15mer/2500	e,l
10	18mer/3058	1	D10Z1	6mer/1020	1	f	D10Z1	8mer/1350	k,m
				8mer/1360		f			
11	12mer/2047	3	D11Z1	5mer/850	3	d	D11Z1	5mer/850	d,k
17	14mer/2379	3	D17Z1	16mer/2712	3	g	D17Z1	16mer/2700	g
							–	13mer/2200	n
19	17mer/2896	5	–	–	–	–			
19	13mer/2214	5	–	–	5	b	–	13mer/2250	b
			–	–	5	c			
X	12mer/2057	3	DXZ1	12mer/2000	3	h	DXZ1	12mer/2000	o

To each HOR/SF determination (Columns 2, 3) the corresponding HOR/SF data from Alexandrov *et al*. are assigned (Columns 3–7). For each locus (Column 4) the corresponding data from Lee *et al*. 1997 are displayed (Columns 8, 9). (a) Willard & Waye 1987, (b) Hulsebos *et al*. 1988, (c) Puechberty *et al*. 1999, (d) Waye *et al*. 1987a, (e) Ge *et al*.1992, (f) Looijenga *et al*. 1992, (g) Waye & Willard 1986, (h) Waye & Willard 1985, (i) Waye *et al*. 1987c, (j) Mashkova *et al*. 1994, (k) Wevrick &Willard 1989, (l) Donlon *et al*. 1987, (m) Devilee *et al*. 1988, (n) Choo *et al*. 1987, (o) Yang *et al*. 1982, Mahtani & Willard 1990.

Table 7. Frequency of CENP-B box C*m* and pJα motif p*m* in consensus HOR from Table 5; for description see Table 5

No. of monomers of type A with pJα motif p*m*

p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
4	7	5	2	2	1	–	–	–	–

No. of monomers of type A with CENP-B box motif C*m*

C0	C1	C2	C3	C4	C5	C6	C7	C8
–	–	1	1	–	–	–	–	–

No. of monomers of type B with CENP-B box motif C*m*

C0	C1	C2	C3	C4	C5	C6	C7	C8
19	4	5	3	–	–	–	–	–

No. of monomers of type B with pJα motif p*m*

p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
–	–	–	–	2	–	–	–	–	–

butions of CENP-B box and pJα motif (essential parts) are given (Figure 2, Table 1).

Monomers from each of our consensus HOR were aligned to SF monomers from Romanova *et al*. 1996 to maximize monomer similarity. (Base position 1 was assigned according to Waye & Willard 1985, as used in Romanova *et al*. 1996.) Values of divergence were calculated for pairwise comparison of all monomers from consensus HOR and SF monomers. To each monomer from consensus HOR we assign the corresponding SF monomer with the lowest mutual divergence. As an example of determination of SF classification we present the divergence matrix for 13mer consensus HOR in chromosome 5, revealing the SF5 classification (Table 3) and for 18mer consensus HOR in chromosome 10, revealing the SF1 classification (Table 4).

In Table 5 we present our SF classification of monomers in consensus HOR. We used this SF classification as a basis for further discussion of CENP-B box and pJα motif distributions. In Table 6 we compare our results for consensus HOR and their

Table 8. Alignment of monomers in 13mer from chromosomes 5 and 19 determined by calculating divergence matrix; for description see text

Chromosome 5	13mer*	*m*01	*m*02	*m*03	*m*04	*m*05	*m*06	*m*07	*m*08	*m*09	*m*10	*m*11	*m*12	*m*13
Chromosome 19	13mer*	*m*02	*m*03	*m*04	*m*05	*m*06	*m*07	*m*08	*m*09	*m*10	*m*11	*m*12	*m*13	*m*01

Table 9. Alignment of monomers in 17mer and 13mer HOR in chromosome 19 determined by calculating divergence matrix; for description see text

Chromosome 19	17mer	*r*01	*r*02	*r*03	*r*04	*r*05	*r*06	*r*07	*r*08	*r*09	*r*10	*r*11	*r*12	*r*13	*r*14	*r*15	*r*16	*r*17
Chromosome 19	13mer*	*m*08	*m*07	*m*06	*m*05	*m*04	*m*03	*m*02	*m*01	*m*13	*m*12	*m*11	*m*10	–	–	–	–	*m*09

Table 10. Alignment of SF monomers in SF classification for 16mer from Alexandrov *et al.* (2001) and for our 14mer from Table 5 (chromosome 17)

Chromosome 17, 16mer:	W2	W3	W4	W3	W4	W5	W1	W1	W1	W5	W1	W2	W3	W4	W5	W1
Chromosome 17, 14mer:	W2	W3	W4	W3	W4	W5	W1	–	–	W5	W1	W2	W3	W4	W5	W1

SF classification to the corresponding previous results reviewed in Alexandrov *et al*. 2001 and Lee *et al*. 1997. In Table 7 we display the frequency distribution of CENP-B box and pJα motif (essential parts) in SF monomers of types A and B within HOR, and differences with respect to canonical CENP-B box/pJα motif (out of essential part). In Table 8 we align monomers in 13mers from chromosomes 5 and 19. In Table 9 monomers in 17mer and 13mer from chromosome 19 are aligned.

Let us illustrate our enumeration of monomers in HOR for the case of chromosome 19. Monomers in 17mer from chromosome 19 are minus strand sequences (i.e. corresponding to the strand in Romanova *et al*. 1996). The sequence of minus strand monomers *rn* in 17mer from chromosome 19 is denoted by:

$$17mer,\ \{r\} : r01,\ r02,\ r03, \ldots,\ r17$$

The sequence of plus strand monomers *sn* in 13mer from chromosome 19 is denoted by:

$$13mer,\ \{s\} : s01,\ s02,\ s03, \ldots,\ s13$$

To compare 13mer to 17mer, we construct the reverse complement of each monomer *sn*, and denote it by *mn*:

$$13mer^*,\ \{m\} : m01,\ m02,\ m03, \ldots,\ m13$$

Here, $m01$ = reverse complement of $s01$, $m02$ = reverse complement of $s02$, etc.

Aligning the arrays $\{r\}$ and $\{m\}$ we align monomers in 17mer and 13mer (Table 9).

According to this procedure the ordering of monomers in 13mer* is reversed to the ordering in 17mer. As shown in Table 9, $r01$ in 17mer is aligned with $m08$ in 13mer*, r02 in 17mer is aligned with $m07$ in 13mer* etc. Here the start monomer in each HOR is determined by the choice of key string.

In Table 10 for chromosome 17 we align monomers in SF classification of 16mer from Alexandrov *et al*. 2001 to our 14mer from Table 5.

Discussion

SF5 assignments of HOR in Build 35.1 assembly

Out of 11 HOR in Table 5, HOR in chromosomes 4, 5, 7 and 19 belong to SF5, HOR in chromosomes 1, 11, 17, and X to SF3, HOR in chromosome 10 to SF1 and in chromosome 8 to SF2. Five HOR from Table 1 correspond to SF5, while the contribution of SF5, in the total human genome assembly was estimated to only 5% (Kazakov *et al*. 2003). This indicates that the SF5 type HOR are more clustered toward the edges of the centromeric region.

13mers in chromosomes 4, 5 and 19, the 16mer in chromosome 7, and 17mer in chromosome 19 are assigned here as SF5. The average homology of monomers from consensus HOR to the corresponding SF monomers is 82–87%, in comparison to the overall homology to all SF monomers of 75–81%. The two monomeric classes R1 (type B) and R2 (type A) are alternating irregularly, as a characteristic feature of SF5.

SF5 13mer in chromosome 5 lacking CENP-B box and pJα motif

In chromosome 5 we determined consensus HOR (2214 bp) for highly homogeneous 13mer (copies more than 95% identical), belonging to SF5. This HOR lies within the q arm contig NT_006713.14 (Table 2), adjacent to the centromere gap. In previous computational analyses of Build 34 assembly no HOR domain was found (Rudd & Willard 2004).

Human chromosome 5 contains at its centromere an alphoid array, detectable by the probe pG-A16. This alphoid array was shown to be common to chromosomes 5 (D5Z1) and 19 (D19Z2). According to Alexandrov *et al*. (2001) a HOR which belongs to SF5 is contained in D5Z1, while a 13mer/2250 bp HOR was reported earlier for D5Z1 without SF determination (Hulsebos *et al*. 1988, Lee *et al*. 1997).

The array D5Z1 represents the very end of alphoid domain on the q side of chromosome 5, as an array of 2.25-kb HOR (13mer) (Hulsebos *et al*. 1988, Puechberty *et al*. 1999). It was possible to distinguish the arrays D5Z1 and D19Z2 from each other despite their high sequence homology (Puechberty *et al*. 1999).

Another alphoid array, composed of 340-bp dimers that contain CENP-B boxes, is common to chromosomes 1 (D1Z7), 5 (D5Z2), and 19 (D19Z3) (Baldini *et al*. 1989, Archidiacono *et al*. 1995, Schindelhauer & Schwarz 2002) and shown to be physically distinct from D5Z1 on chromosome 5 (5p > D5Z2 > D5Z1 > 5q) (Finelli *et al*. 1996). One

new alphoid array (D5Z12) was detected and partially characterized on the p side of 5cen (Puechberty *et al*. 1999). D5Z2 was reported to belong to SF1 (Alexandrov *et al*. 2001).

We showed that the 13mer consensus HOR does not contain any CENP-B box or pJα motif. To our knowledge this is the first case of a completely CENP-B box-free and pJα motif-free HOR in the human genome. The only HOR in the human genome reported so far to have no CENP-B box were known in chromosome Y (Pluta *et al*. 1990), but they contain pJα motifs. However, an important difference is the complete lack of CENP-B boxes and CENP-B protein from the Y chromosome, while the centromere domains in chromosomes 5 and 19, containing CENP-B box-free 13mer, have other alphoid arrays (for example, dimers and 17mers) that contain CENP-B boxes.

High sequence homology of SF5 13mers in chromosomes 5 and 19

Homology of 13mers in chromosomes 5 and 19 is displayed in Table 8. (The start monomer in each HOR was determined by the choice of key string.) Monomers 1–13 in 13mer from chromosome 5 are aligned to monomers 2–13 and 1 in 13mer from chromosome 19, respectively. The average divergence between the two 13mers is 4%, in comparison to the average pairwise divergence of 22% between all pairs of monomers.

High sequence homology of new 17mer and 13mer in chromosome 19

In chromosome 19 the 13mer HOR have been reported (Hulsebos *et al*. 1988, Warburton & Willard 1996, Lee *et al*. 1997), but the corresponding SF assignment was not available. In addition to the 13mer, here we identified a new 17mer HOR (consensus length 2896 bp) (Table1). This HOR lies within the p arm contig NT_011295.10 (Table 2), adjacent to the centromere gap. In previous computational analyses of Build 34 assembly a HOR domain was reported within the p arm (Rudd & Willard 2004).

Monomers in 13mer are plus strand sequences, i.e. reverse complement with respect to monomers in 17mer. Taking this into account, the 13mer is within 5–9% divergence identical to 13 out of 17 monomers in the 17mer. We conclude that the 13mer in chromosome 19 arises by deletion of four monomers from 17mer.

On the other hand, monomers in dimers identified in chromosome 19 diverge from monomers in 17mer by more than 25%.

We calculated homology by pairwise alignment of monomers in 17mer to reverse complement of monomers from 13mer (17mer is a minus strand and 13mer plus strand sequence), and determined pairs of monomers with smallest divergence (Table 9). Average divergence is 11%, in comparison to the average pairwise divergence of 23% between all monomers from 17mer and 13mer. We align monomers *r*09–*r*17 and *r*01–*r*08 from 17mer to the monomers *m*13–*m*01 from 13mer, respectively. Both 17mer and 13mer in chromosome 19 belong to SF5 (Table 5).

17mer in chromosome 19 contains one CENP-B box and one pJα motif, positioned in two neighboring monomers (Table 5). These are just two out of four monomers deleted in aligning 17mer to 13mer (Table 9).

New SF5 13mer in chromosome 4

In chromosome 4 we identified a new 13mer HOR (consensus length 2211 bp), which belongs to SF5. This is the first time that a SF5 HOR was found in chromosome 4. This HOR lies within the q arm contig NT_022853.14 (Table 2), adjacent to the centromere gap.

Previous investigations of chromosome 4 have determined a 3.2-kb HOR of SF2, which consists either of a single 3.2-kb fragment or of 2.6 and 0.6 fragments (Mashkova *et al*. 1994), while a HOR belonging to SF5 was not identified (Alexandrov *et al*. 2001). Furthermore, a 340-bp dimer was reported which belongs to SF2 (Hulsebos *et al*. 1988, D'Aiuto *et al*. 1993). In previous computational analyses of Build 34 assembly, a HOR domain was reported within the q arm (Rudd & Willard 2004).

13mer in chromosome 4 contains four CENP-B boxes and three pJα motifs (Figure 2), with three CENP-B boxes occurring in three consecutive monomers. This is a completely different pattern for CENP-B box and pJα motif distribution than in the 13mer in chromosome 19 (13mer in chromosome 19 contains neither CENP-B box nor pJα motif). The 13mer in

chromosome 4 diverges from 13mer in chromosome 19 by ≈20%. Evidently, 13mer in chromosome 4 is not homologous to the 13mer in chromosomes 5 and 19, although all three belong to SF5.

SF3 11mer in chromosome 1

In chromosome 1 we identified 11mer which belongs to SF3. The average homology of monomers from our consensus HOR to the corresponding SF3 monomers is 87%, while the average homology to all SF monomers is 78%.

The sequence of SF monomers in SF classification of 11mer in chromosome 1 (Table 5) is derived from our consensus HOR using reverse complement sequence of monomers $\{m\}$, as defined in the previous section. Our definition leads to reversed ordering of SF monomers in plus strand HOR, when compared to SF monomer sequence from Alexandrov *et al.* (2001). For the plus strand HOR (denoted by an asterisk in Table 1) we are using this definition of reverse complement sequence throughout this paper.

11mer HOR in chromosome 1, identified from Build 35.1 assembly, lies within the p arm contig NT_077389.3 (Table 2), adjacent to the centromere gap. In previous computational analyses of Build 34 assembly, a HOR domain in chromosome 1 was not reported (Rudd & Willard 2004).

The length of 11mer in chromosome 1 was previously reported as 1.9 kb (Willard 1985, Waye *et al.* 1987c, Willard & Waye 1987, Wevrick & Willard 1989, Warburton & Willard 1996). A nucleotide sequence of 1861 bp was determined for pSD1-1 in chromosome 1 (Willard & Waye 1987). The consensus length of 11mers in chromosome 1, derived from Build 35.1 assembly using KSA, is 1866 (Paar *et al.* 2005). We note that in some 11mers there is a systematic 3-bp insertion (CTA), increasing the HOR length to 1869 (Figure 1).

In chromosome 1 the array D1Z7 (2mers) belongs to SF1, while D1Z5 (11mers) belongs to SF3; the ordering of these domains was determined as 1p → D1Z5 → D1Z7 → D1Z5 → 1q (Archidiacono *et al.* 1995, Finelli *et al.* 1996). The 11mer HOR unit is present in at least 100 copies (Waye *et al.* 1987c). The length of individual centromeric arrays was reported to range from 440 to 1510 kb (Wevrick & Willard 1989).

Analyzing the Build 35.1 array in chromosome 1 we found two close-lying 11mer arrays containing 59 complete and 14 incomplete HOR of total length 130087 bp.

The 11mer in chromosome 1 contains three CENP-B boxes and three pJα motifs (Table 1). A CENP-B box appears in type B monomers W1 in subsequence W1W5W1W5W1W5 within the SF classification of 11mer (Table 5). On the other hand, the pJα motif occurs in three consecutive monomers, where the middle one is associated with W3 (of type B). This is an exception to the rule that the pJα motif occurs in type A monomers only.

CENP-B box/pJα motif – poor SF5 16mer in chromosome 7

In chromosome 7 the 16mer belongs to SF5 (Table 5). Consensus HOR contains only one CENP-B box (monomer No. 16) and one pJα motif (monomer No. 5).

This HOR lies within the q arm contig NT_023603.5 (Table 2), adjacent to the centromere gap. In previous computational analyses of Build 34 assembly a HOR domain was reported within the q arm (Rudd & Willard 2004).

Analyzing Build 35.1 array for chromosome 7 we found an alphoid array containing 46 complete and 14 incomplete 16mers (total length of 148147 bp).

The centromeric region of chromosome 7 contains two distinct arrays of alpha satellites, D7Z1 (1–3 Mb) and D7Z2 (100–500 kb). D7Z1 is composed of 6mer HOR, and D7Z2 of 16mer HOR (Waye *et al.* 1987a, Wevrick & Willard 1989, 1991, Wevrick *et al.* 1992, Archidiacono *et al.* 1995, Finelli *et al.* 1996, Alexandrov *et al.* 2001). In addition, 2mers were reported (Jorgensen *et al.* 1986).

Previously it was shown that the D7Z1 alphoid array is associated with 6mer which belongs to SF1, while the D7Z2 array, containing 16mer, belongs to SF5 (Alexandrov *et al.* 2001). The alphoid array D7Z1 is CENP-B box-rich, while the array D7Z2 is CENP-B box-poor (Haaf & Ward 1994, Ikeno *et al.* 1994, Choo 1997). Only the CENP-B box-poor alphoid array D7Z2 is present in Build 35.1 assembly.

SF3 14mer in chromosome 17

In chromosome 17 we identified 14mer HOR (consensus length 2379 bp) belonging to SF3, with five CENP-B boxes and two pJα motifs (Table 1).

This HOR lies within the p arm contig NT_024862.13 (Table 2), adjacent to the centromere gap. In previous

computational analyses of Build 34 assembly a HOR domain was reported within the p arm (Rudd & Willard 2004).

The predominant HOR on chromosome 17 is a 2.7-kb 16mer (Waye & Willard 1986). Chromosome 17 is further characterized by several polymorphic HOR variants, 11mer, 12mer, 13mer, 14mer, and 15mer arising from the 16mer by deletion of some of its monomers (Willard *et al.* 1986, 1987, Choo *et al.* 1987, Warburton *et al.* 1993, Lee *et al.* 1997, Alexandrov *et al.* 2001).

The 16mer in chromosome 17, found previously, belongs to SF3 (Alexandrov *et al.* 2001). It is characterized by a cluster of three W1 monomers (Warburton *et al.* 1993, Alexandrov *et al.* 2001). By deleting two out of three W1 monomers in the W1W1W1 triplet in 16mer, alignment with our calculated SF structure of 14mer was achieved (Table 10).

The 16mer in chromosome 17 has six CENP-B boxes, three of them associated with triplet W1W1W1 (Warburton *et al.* 1993). By deleting two W1 monomers from the triplet in 16mer, in the resulting 14mer the CENP-B boxes are positioned in accordance with the result of our calculation for 14mer (Table 10). However, in 14mer calculated from Build 35.1 assembly we obtain an additional CENP-B box in monomer No. 1, which was not reported in 16mer from Warburton *et al.* 1993.

New SF1 18mer in chromosome 10

18mer in chromosome 10 belongs to SF1, with regularly alternating dimers -J2J1-. The average homology of monomers from consensus HOR to the corresponding SF1 monomers is 89%, while the average homology to all SF monomers is 78%. For comparison, we note that figures for homology to SF1 monomers, obtained here, are similar to figures previously identified, as for example for SF2 10mers and 8mers in chromosome 18 (Alexandrov *et al.* 1991).

The 18mer HOR in chromosome 10 lies in the q arm contig NT_079540.1 (Table 2), adjacent to the centromere gap. In previous computational analyses of Build 34 assembly a HOR domain was reported within the q arm (Rudd & Willard 2004).

Previous investigations of chromosome 10 have reported an alphoid array D10Z1, containing 6mer and 8mer belonging to SF1 (Devilee *et al.* 1988, Wevrick & Willard 1989, Wu & Kidd 1990, Looijenga *et al.* 1992, Alexandrov *et al.* 2001). The 18mer HOR unit in chromosome 10 was not reported in earlier investigations (Warburton & Willard 1996, Alexandrov *et al.* 2001).

In alternating dimers -J2J1-, each type B monomer J2 (except the last one) contains a CENP-B box, while the type A monomers J1 are CENP-B box free. Thus, there is a sequence of eight CENP-B boxes in every other monomer. Only in the last J2 monomer a pJα motif occurs instead of the CENP-B box. This is an exception to the rule that the pJα motif corresponds to type A monomers only.

New SF2 11mer in chromosome 8

11mer in chromosome 8 belongs to SF2, with regularly alternating dimers -D2D1-. In addition, one D1 monomer insertion occurs in the alternating structure. The average homology of monomers from consensus HOR to the corresponding SF2 monomer is 86%, while the average homology to all SF monomers is 76%. This HOR contains three CENP-B boxes and five pJα motifs. (For HOR having monomers with plus strand sequence, our ordering of SF monomers is reverse with respect to Alexandrov *et al.* 2001.)

In our calculation the 11mer in chromosome 8 lies in two distinct domains, in q arm contig NT_023678.15 and in p arm contig NT_007995.14 (Table 2), adjacent to the centromere gap. In previous computational analyses of Build 34 assembly the HOR domains were reported both within q and p arms (Rudd & Willard 2004).

Previous investigations of chromosome 8 have reported an alphoid array D8Z2, containing SF2 15mers (Ge *et al.* 1992, Alexandrov *et al.* 2001).

Homology between new SF3 12mer in chromosomes 11 and 12mer in chromosome X

For the first time we determined the SF3 classification for 12mer in chromosome 11. The average homology of monomers from consensus HOR to the corresponding SF monomers is 89%, while the average homology to all SF consensus monomers is 79%.

Previously, a 5mer HOR was identified in chromosome 11 (Waye *et al.* 1987b, Wevrick & Willard 1989) and the pentamer SF3 classification W1W2 W3W4W5 was assigned (Alexandrov *et al.* 2001). The corresponding pentamer in this paper is reversed (as discussed previously).

Here we identify 12mer HOR in chromosome 11 with suprachromosomal classification W3W4W3R1 W1W5W4W3W2W1W5W4. A subsequence consisting of the last seven monomers in this array could have evolved from a simple ancestral pentamer W5W4 W3W2W1. In a subsequence consisting of the first five monomers, W3W4W3R1W1, two monomers differ from the ancestral pentamer, the first W3 and R1. With respect to R1 assignment, the fourth monomer in consensus HOR showed the highest homology to R1 monomer (14.0% divergence), but homology to the W2 monomer was only slightly worse (16.4% divergence). Regarding the assignment of W3 monomer to the first position, divergence for W5 of 25% is much larger than 14% for W3.

12mer in chromosome X corresponds to SF3. The average homology of monomers from consensus HOR to the corresponding SF monomers is 89%, while the homology to all SF monomers is 80%. Our SF classification for 12mer consensus HOR in chromosome X is in accordance with the previous classification (Alexandrov *et al.* 2001), except for R1 classification at the third position in Table 5. We note that divergence of 13.5% for SF monomer R1 is very close to divergence of 14.0% for the W2 monomer, so this R1 deviation from SF3 is not very significant.

As shown in Table 5, the SF classifications of 12mers in chromosomes 11 and X are identical (with a one-monomer shift). The calculated divergence between these two consensus HOR after alignment is 12%, in comparison to 24% of average pairwise divergence between all monomers, showing a moderate homology of 12mers in chromosomes 11 and X. However, there are differences regarding the CENP-B box/pJα motif: 12mers in chromosome 11 have two CENP-B boxes and two pJα motifs, while 12mers in chromosome X contain four CENP-B boxes and three pJα motifs.

Dimers in Build 35.1 assembly

In Build 35.1 sequence of chromosome 1, besides 11mers we found dimers. In chromosome 7 dimers were found in addition to 16mers (Rosandić *et al.* 2003a,b, Paar *et al.* 2005). In chromosomes 2, 10 and 19 some clones included in contigs in Build 35.1 contain dimers. Each sequence of dimers from Build 35.1 assembly contains a CENP-B box in almost every other alpha monomer. However, divergence between dimers is more pronounced than between HOR investigated here. We note a tendency of intermittent onset of HOR pattern along genomic sequence, resembling a phenomenon of intermittency associated with nonlinear systems (Berge *et al.* 1984). Going toward the region with regular CENP-B box distribution, the frequency of appearance of CENP-B boxes intermittently increases and becomes more and more regular. Such a situation occurs, for example, in dimers in clone AC010517.3 of chromosome 19.

Monomeric alpha satellites not organized into HOR are almost CENP-B box-free, e.g. in alphoid arrays AC104789.4 and AC092585.2 in chromosome 7.

The Build 35.1 assembly for chromosome 21 contains no HOR. However, we found that the clone AC001464.1 contains an incomplete pattern of HOR organization, showing a transitional HOR-region with random distribution of CENP-B boxes.

Exceptions to the rule assigning CENP-B box to SF type B and pJα motif to SF type A monomers

From previous analyses of nucleotide frequencies in positions 35–51 (for minus strand) of a sample of monomers of types A and B it was calculated that a functional CENP-B box (essential positions only) would occur three times in a million of type A monomers and in 55% of type B monomers (Romanova *et al.* 1996). In a sample of monomers from Romanova *et al.* (1996) the frequency of CENP-B boxes was 60% in type B and 0% in type A. The pJα core sequence would occur in 12% of type A and 0.02% of type B monomers. These estimates indicate that the probability of a functional CENP-B box randomly occurring in a type A alpha satellite is negligibly small.

Contrary to that, our results of analysis of 11 HOR in 10 human chromosomes show that two out of 31 CENP-B boxes in consensus HOR, i.e. 6% of CENP-B box core sequences, occur in type A monomers. These exceptions occur in 13mer in chromosome 4: R2 monomers at positions 3 and 4 contain a CENP-B box. They contain two and three bases outside of the essential core, respectively, differing from the canonical sequence.

Regarding the pJα core sequence, two out of 21 pJα core sequences in consensus HOR, i.e. 10%, occur in the type B monomers. Such cases are the

pJα motif in B-type monomers W3 in 11mer in chromosome 1 and the J2 B-type monomer in 18mer in chromosome 10.

CENP-B box/pJα motif distribution

In monomers of consensus HOR we determined the distribution of CENP-B box and pJα motif. Our results illustrate a broad spectrum of CENP-B box pattern: HOR without any CENP-B box (for example, 13mers in chromosomes 5 and 19), or HOR with only one CENP-B box per copy (for example, 16mer in chromosome 7, and 17mer in chromosome 19), or HOR with a CENP-B box occurring in consecutive monomers (for example, 13mer in chromosome 4), or HOR with a CENP-B box occurring in almost every other monomer (for example, 18mer in chromosome 10).

The number of pJα motifs in consensus HOR studied in this paper is mostly comparable to the number of CENP-B box motifs (eight from 11 cases). In two cases the number of CENP-B box motifs is sizeably higher than the number of pJα motifs (for example, the SF1 18mer in chromosome 10), and in one case (SF2 11mer in chromosome 8) it is smaller. Totally, in consensus HOR studied in 10 chromosomes we found 31 CENP-B box and 21 pJα motif sequences.

In four of the investigated consensus HOR, we found that eight or more monomers occur per CENP-B box/pJα motif (low CENP-B box density). Such cases are, for example: 16mers in chromosome 7 and 17mers in chromosome 19. The 13mers in chromosome 5 and 19 are CENP-B box-free and pJα-motif-free in consensus HOR. All four HOR belong to SF5, which is associated with low CENP-B box density.

Seven consensus HOR contain about two monomers per CENP-B box/pJα motif (high CENP-B box density). The average number of monomers per CENP-B box/pJα motif in these HOR is 1.9, resembling a generalized 'every other monomer scheme.' In some HOR this scheme is rather well preserved (for example, 18mer in chromosome 10), while in some HOR more pronounced clustering and/or deletions occur (for example, 11mer in chromosome 8). In general the density of CENP-B box/pJα motif sequences in investigated cases is much lower in HOR belonging to SF5 than in HOR belonging to SF1, SF2, and SF3.

References

Alexandrov IA, Kazakov A, Tumeneva I, Shepelev V, Yurov Y (2001) Alpha-satellite DNA of primates: old and new families. *Chromosoma* **110**: 253–266.

Alexandrov IA, Mashkova TD, Akopian TA *et al*. (1991) Chromosome-specific alpha satellites: two distinct families on human chromosome 18. *Genomics* **11**: 15–23.

Alexandrov IA, Mitkevich SP, Yurov YB (1988) The phylogeny of human chromosome-specific alpha satellites. *Chromosoma (Berlin)* **110**: 253–266.

Archidiacono N, Antonacci R, Marzella R, Finelli P, Lonoce A, Rocchi M (1995) Comparative mapping of human alphoid sequences in great apes using fluorescence *in situ* hybridization. *Genomics* **25**: 477–484.

Baldini A, Smith DI, Rocchi M, Miller OJ, Miller DA (1989) A human alphoid DNA clone from the EcoRI dimeric family: genomic and internal organization and chromosomal assignment. *Genomics* **5**: 822–828.

Basu J, Stromberg G, Compitello G, Willard HF, Van Bokkelen G (2005) Rapid creation of BAC-based human artificial chromosome vectors by transposition with synthetic alpha-satellite arrays. *Nucleic Acids Res* **33**: 587–596.

Berge P, Pomeau Y, Vidal C (1984) *Order Within Chaos*. New York: Wiley.

Choo KHA (1997) *The Centromere*. Oxford: Oxford University Press.

Choo KH, Brown R, Webb G, Craig IW, Filby RG (1987) Genomic organization of human centromeric alpha satellite DNA: characterization of a chromosome 17 alpha satellite sequence. *DNA* **6**: 297–305.

Choo KH, Vissel B, Nagy A, Kalitsis P (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.

Cleveland DW, Mao Y, Sullivan KF (2003) Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell* **112**: 407–421.

D'Aiuto L, Antonacci R, Marzella R, Archidiacono N, Rocchi M (1993) Cloning and comparative mapping of a human chromosome 4-specific alpha satellite DNA sequence. *Genomics* **18**: 230–235.

Devilee P, Kievits T, Waye JS, Pearson PL, Willard HF (1988) Chromosome-specific alpha satellite DNA: isolation and mapping of a polymorphic alphoid repeat from human chromosome 10. *Genomics* **3**: 1–7.

Donlon TA, Burns GA, Latt SA, Mulholland J, Wyman AR (1987) A chromosome 8-enriched alphoid repeat. *Cytogenet Cell Genet* **46**: 607.

Earnshaw WC, Rothfield N (1985) Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. *Chromosoma* **91**: 313–321.

Earnshaw WC, Sullivan KF, Machlin PS *et al*. (1987) Molecular cloning of cDNA for CENP-B, the major human centromere autoantigen. *J Cell Biol* **104**: 817–829.

Finelli P, Antonacci R, Marzella R, Lonoce A, Archidiacono N, Rocchi M (1996) Structural organization of multiple alphoid arrays coexisting on human chromosomes 1, 4, 5, 7, 9, 15, 18, and 19. *Genomics* **38**: 325–330.

Gaff C, du Sart D, Kalitsis P, Iannello R, Nagy A, Choo KH (1994) A novel nuclear protein binds centromeric alpha satellite DNA. *Hum Mol Genet* **3**: 711–716.

Ge Y, Wagner MJ, Siciliano M, Wells DE (1992) Sequence, higher order repeat structure, and long-range organization of alpha satellite DNA specific to human chromosome 8. *Genomics* **13**: 585–593.

Haaf T, Ward DC (1994) Structural analysis of alpha-satellite DNA and centromere proteins using extended chromatin and chromosomes. *Hum Mol Genet* **3**: 697–709.

Haaf T, Mater AG, Wienberg J, Ward DC (1995) Presence and abundance of CENP-B box sequences in great ape arrays of primate-specific alpha-satellite DNA. *J Mol Evol* **41**: 487–491.

Henikoff S (2002) Near the edge of a chromosomes 'black hole'. *Trends Genet* **18**: 165–167.

Hulsebos T, Schonk D, van Dalen I *et al*. (1988) Isolation and characterization of alphoid DNA sequences for the pericentric regions of chromosomes 4, 5, 9, and 19. *Cytogenet Cell Genet* **47**: 144–148.

Ikeno M, Grimes B, Okazaki T *et al*. (1998) Construction of YAC-based mammalian artificial chromosomes. *Nat Biotechnol* **16**: 431–439.

Ikeno M, Masumoto H, Okazaki T (1994) Distribution of CENP-B boxes reflected in CREST centromere antigenic sites on long-range alpha-satellite DNA arrays of human chromosome 21. *Hum Mol Genet* **3**: 1245–1257.

Iwahara J, Kigawa T, Kitagawa K, Masumoto H, Okazaki T, Yokoyama S (1998) A helix-turn-helix structure unit in human centromere protein B (CENP-B). *EMBO J* **17**: 827–837.

Jorgensen AL, Bostock CJ, Bak AL (1986) Chromosome-specific subfamilies within human alphoid repetitive DNA. *J Mol Biol* **187**: 185–196.

Kazakov AE, Shepelov VA, Tumeneva IG, Alexandrov AA, Yurov YB, Alexandrov IA (2003) Interspersed repeats are found predominantly in the 'old' alpha satellite families. *Genomics* **82**: 619–627.

Kouprina N, Ebersole T, Koriabine M *et al*. (2003) Cloning of human centromeres by transformation-associated recombination in yeast and generation of functional human artificial chromosomes. *Nucleic Acids Res* **31**: 922–934.

Lee C, Wevrick R, Fisher RB, Ferguson-Smith MA, Lin CC (1997) Human centromeric DNAs. *Hum Genet* **100**: 291–304.

Looijenga LH, Oosterhuis JW, Smit VT, Wessels JW, Mollevanger P, Devilee P (1992) Alpha satellite DNAs on chromosomes 10 and 12 are both members of the dimeric suprachromosomal subfamily, but display little identity at the nucleotide sequence level. *Genomics* **13**: 1125–1132.

Mahtani MM, Willard HF (1990) Pulsed-field gel analysis of a satellite DNA at human X chromosome centromere: high-frequency polymorphisms and array size estimate. *Genomics* **7**: 607–613.

Maio JJ (1971) DNA strand reassociation and polyribonucleotide binding in the African green monkey, *Cercopithecus aethiops*. *J Mol Biol* **56**: 579–595.

Manuelidis L, Wu JC (1978) Homology between human and simian repeated DNA. *Nature* **276**: 92–94.

Mashkova TD, Akopian TA, Romanova LY *et al*. (1994) Genomic organization, sequence and polymorphism of the human chromosome 4 specific alpha satellite DNA. *Gene* **140**: 211–217.

Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T (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.

Masumoto H, Nakano M, Ohzeki J (2004) The role of CENP-B and alpha-satellite DNA: *de novo* assembly and epigenetic maintenance of human centromeres. *Chromosome Res* **12**: 543–556.

Masumoto H, Yoda K, Ikeno M, Kitagawa K, Muro Y, Okazaki T (1993) Properties of CENP-B and its target sequence in a satellite DNA. In *Chromosome Segregation and Aneuploidy*. Berlin: Springer-Verlag, pp. 31–43.

Muro Y, Masumoto H, Yoda K, Nozaki N, Ohashi M, Okazaki T (1992) Centromere protein B assembles human centromeric alpha satellite DNA at 17-bp sequence, CENP-B box. *J Cell Biol* **116**: 585–596.

Ohzeki J, Nakano M, Okada T, Matsumoto H (2002) CENP-B box is required for the novo centromere chromatin assembly on human alphoid DNA. *J Cell Biol* **159**: 765–775.

Paar V, Pavin N, Rosandić M *et al*. (2005) ColorHOR – novel graphical algorithm for fast scan of alpha satellite higher-order repeats and HOR annotation for GenBank sequence of human genome. *Bioinformatics* **21**: 846–852.

Pluta AF, Cooke CA, Earnshaw WC (1990) Structure of the human centromere at metaphase. *Trends Biochem Sci* **15**: 181–185.

Pluta AF, Saitoh N, Goldberg I, Earnshaw WC (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.

Puechberty J, Laurent AM, Gimenez S *et al*. (1999) Genetic and physical analyses of the centromeric and pericentromeric regions of human chromosome 5: recombination across 5cen. *Genomics* **56**: 274–287.

Romanova LY, Deriagin GV, Mashkova TD *et al*. (1996) Evidence for selection of alpha satellite DNA: the central role of CENP-B/pJα binding region. *J Mol Biol* **261**: 334–340.

Rosandić M, Paar V, Basar I (2003a) Key-string segmentation algorithm and higher-order repeat 16mer (54 copies) in human alpha satellite DNA in chromosome 7. *J Theor Biol* **221**: 29–37.

Rosandić M, Paar V, Glunčić M, Basar I, Pavin N (2003b) Key-string algorithm – novel approach to computational analysis of repetitive sequences in human centromeric DNA. *Croat Med J* **44**: 386–406.

Rudd MK, Willard HF (2004) Analysis of the centromeric regions of the human genome assembly. *Trends Genet* **20**: 529–533.

Schindelhauer D, Schwarz T (2002) Evidence for a fast, intra-chromosomal conversion mechanism from mapping of nucleotide variants within a homogeneous alpha-satellite DNA array. *Genome Res* **12**: 1815–1826.

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF (2001) Genomic and genetic definition of a functional human centromere. *Science* **294**: 109–115.

Tanaka Y, Nureki O, Kurumizaka H *et al*. (2001) Crystal structure of the CENP-B protein–DNA complex: the DNA-binding domains CENP-B induce kinks in the CENP-B box DNA. *EMBO J* **20**: 6612–6618.

Tanaka Y, Kurumizaka H, Yokoyama S (2004) CpG methylation of the CENP-B box reduces human CENP-B binding. *FEBS J* **272**: 282–289.

Trowell HE, Nagy A, Vissel B, Choo KH (1993) Long-range analyses of the centromeric regions of human chromosomes 13, 14 and 21: identification of a narrow domain containing two key centromeric DNA elements. *Hum Mol Genet* **2**: 1639–1649.

Tyler-Smith C, Willard HF (1993) Mammalian chromosome structure. *Curr Opin Genet Dev* **1993**: 390–397.

Warburton PE (2004) Chromosomal dynamics of human neocentromere formation. *Chromosome Res* **12**: 617–626.

Warburton PE, Willard HF (1996) Evolution of centromeric alpha satellite DNA: molecular organization within and between human and primate chromosomes. In *Human Genome Evolution*. Oxford: BIOS Scientific, pp. 121–145.

Warburton PE, Waye JS, Willard HF (1993) Nonrandom localization of recombination events in human alpha satellite repeat unit variants: implications for higher-order structural characteristics within centromeric heterochromatin. *Mol Cell Biol* **13**: 6520–6529.

Waye JS, Willard HF (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.

Waye JS, Willard HF (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.

Waye JS, Willard HF (1987) Nucleotide sequence heterogeneity of alpha satellite DNA: a survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res* **15**: 7549–7569.

Waye JS, Creeper LA, Willard HF (1987b) Organization and evolution of alpha satellite DNA from human chromosome 11. *Chromosoma* **95**: 182–188.

Waye JS, Durfy SJ, Pinkel D *et al*. (1987c) 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.

Waye JS, England SB, Willard HF (1987a) 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.

Wevrick R, Willard HF (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.

Wevrick R, Willard HF (1991) Physical map of the centromeric region of human chromosome 7: relationship between two distinct alpha satellite arrays. *Nucleic Acids Res* **19**: 2295–2301.

Wevrick R, Willard VP, Willard HF (1992) Structure of DNA near long tandem arrays of alpha satellite DNA at the centromere of human chromosome 7. *Genomics* **14**: 912–923.

Willard HF (1985) Chromosome-specific organization of human alpha satellite DNA. *Am J Hum Genet* **37**: 524–532.

Willard HF (1991) Evolution of alpha satellite. *Curr Opin Genet Dev* **1**: 509–514.

Willard HF, Waye JS (1987) Chromosome-specific arrays of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal arrays and evidence for an ancestral pentameric repeat. *J Mol Evol* **25**: 207–214.

Willard HF, Greig GM, Powers VE, Waye JS (1987) Molecular organization and haplotype analysis of centromeric DNA from human chromosome 17: implications for linkage in neurofibromatosis. *Genomics* **1**: 368–373.

Willard HF, Waye JS, Skolnick MH, Schwartz CE, Powers VE, England SB (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.

Wu JS, Kidd KK (1990) Extensive sequence polymorphisms associated with chromosome 10 alpha satellite DNA and its close linkage to markers from the pericentromeric region. *Hum Genet* **84**: 279–282.

Yang TP, Hansen SK, Oishi KK, Ryder OA, Hamkalo BA (1982) Characterization of a cloned repetitive DNA sequence concentrated on the human X chromosome. *Proc Natl Acad Sci USA* **79**: 6593–6597.

Yoda K, Okazaki T (1997) Site-specific base deletions in human alpha-satellite monomer DNAs are associated with regularly distributed CENP-B boxes. *Chromosome Res* **5**: 207–211.

Yoda K, Ando S, Okuda A, Kikuchi A, Okazaki T (1998) *In vitro* assembly of the CENP-B/alpha satellite DNA/core histone complex: CENP-B causes nucleosome positioning. *Genes Cells* **3**: 533–548.

Yoda K, Nakamura T, Masumoto H *et al.* (1996) Centromere protein B of African green monkey cells: gene structure, cellular expression and centromeric localization. *Mol Cell Biol* **16**: 5169–5177.