Journal of Theoretical Biology 317 (2013) 301–309

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Start/stop codon like trinucleotides extensions in primate alpha satellites

Marija Rosandić, Matko Glunčić *, Vladimir Paar

Faculty of Science, University of Zagreb, 10000 Zagreb, Croatia

HIGHLIGHTS

▸ We find large differences between primates in alpha satellite higher order repeats.
▸ We introduce novel noncoding concept of extended start/stop Codon Like Trinucleotides.
▸ We propose a possible new mechanism of genome regulators acting at distance.

ARTICLE INFO

Article history:
Received 3 January 2012
Received in revised form
7 September 2012
Accepted 19 September 2012
Available online 28 September 2012

Keywords:
Human alpha satellite
Chimpanzee alpha satellite
Orangutan alpha satellite
Macaque alpha satellite
Higher order repeats

ABSTRACT

The centromeres remain “the final frontier” in unexplored segments of genome landscape in primate genomes, characterized by 2–5 Mb arrays of evolutionary rapidly evolving alpha satellite (AS) higher order repeats (HORs). Alpha satellites as specific noncoding sequences may be also significant in light of regulatory role of noncoding sequences. Using the Global Repeat Map (GRM) algorithm we identify in NCBI assemblies of chromosome 5 the species-specific alpha satellite HORs: 13mer in human, 5mer in chimpanzee, 14mer in orangutan and 3mers in macaque. The suprachromosomal family (SF) classification of alpha satellite HORs and surrounding monomeric alpha satellites is performed and specific segmental structure was found for major alpha satellite arrays in chromosome 5 of primates. In the framework of our novel concept of start/stop Codon Like Trinucleotides (CLTs) as a “new DNA language in noncoding sequences”, we find characteristics and differences of these species in CLT extensions, in particular the extensions of stop-TGA CLT. We hypothesize that these are regulators in noncoding sequences, acting at a distance, and that they can amplify or weaken the activity of start/stop codons in coding sequences in protein genesis, increasing the richness of regulatory phenomena.

1. Introduction

Centromere underwent a rapid development on the evolutionary scale (Rudd and Willard, 2004; Alkan et al., 2011). The simplest eukaryote *Saccharomyces cerevisiae*, as a first eukaryote, contains a centromere of only ~120 bp (Clarke and Baum, 1990; Furuyama and Biggins, 2007), while the size of centromeric region in primates grows to several megabases (Rudd and Willard, 2004). The major constituent of primate centromeres is AS and most of it is not yet assembled because of highly repetitive nature. AS consists of tandem repetitions of 171 bp A,T-rich sequence motif, called AS monomer. In humans, a large fraction of AS monomers is organized into chromosome specific HOR structures that are due to more recent evolution and by the process of unequal crossing over enable a rapid evolutionary progress. AS monomers that are not organized into HORs, called monomeric ASs, exhibit 20–40% divergence, while the sequence divergence between the corresponding HOR copies is less than 5%, in most cases less than 2% (Willard, 1985; Tyler-Smith, 1985; Mahtani and Willard, 1990; Warburton and Willard, 1996; Alexandrov et al., 2001; Alkan et al., 2007; Rudd et al., 2006).

The existing and widely used characterization of ASs was usually performed in terms of suprachromosomal family (SF) classification (Alexandrov et al., 2001; Alexandrov et al., 1986; Willard and Waye, 1987; Romanova et al., 1996; Kazakov et al., 2003; Shepelev et al., 2009). The Suprachromosomal Family (SF) classification encompasses 12 SF monomers, organized into 5 SFs: SF1 (monomers J1 and J2), SF2 (monomers D1 and D2), SF3 (monomers W1–W5), SF4(monomer M1) and SF5 (monomers R1 and R2). They are classified into two basic types: A—“old” monomers (J1,D2,W4,W5,M1,R2) containing pJα motif (the binding site for pJα protein), and B—“new monomers” (J2,D1,W1,W2,W3,R1) containing CENP-B box (the binding site for CENP-B protein) (Alexandrov et al., 2001; Romanova et al., 1996). Only A-type satellites are found in lower primates, and both A and B are found in all great apes (Romanova et al., 1996). Previous studies of ASs (without HOR identification) in human chromosome 5 have revealed the presence of SF 4, 5 and 1 monomer classification (Alexandrov et al., 2001; Hulsebos et al., 1988; Puechberty et al., 1999; Baldini et al., 1989). Computational analysis of sequenced data (Build 34.3) identified 13mer HOR in human chromosome 5 (Rosandić et al., 2006).

* Corresponding author. Tel./fax: +385 1 4680321.
E-mail addresses: rosandic@hazu.hr (M. Rosandić), matko@phy.hr (M. Glunčić), paar@hazu.hr (V. Paar).

0022-5193/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2012.09.022

The interest in noncoding sequences is rapidly increasing because of their possible role in gene regulation (Pennachio and Rubin, 2001; Enard et al., 2002; Gilad et al., 2006; Haussler, 2006; King et al., 2007; Wray and Babbitt, 2008; Haygood et al., 2010). Components of regulatory control in human genome include proximal regulatory elements and long-range elements that can act across large genomic distances to influence the spatial and temporal distribution of gene expression (Haussler, 2006; King et al., 2007; Wray and Babbitt, 2008; Haygood et al., 2010). Recent analyses provide evidence of many thousands of distant-acting noncoding sequences, beyond the proximal regulatory sequences, and are beginning to reveal the large-scale regulatory genome architecture (Gilad et al., 2006; Orom et al., 2010; Brawand et al., 2011).

In this study, we investigate computationally AS HORs in human, chimpanzee, orangutan and macaque genomic sequences of chromosome 5 (Build 37.3, 3.1, and 1.2 assemblies) using GRM algorithm (Paar et al., 2011a; Paar et al., 2011b). The resulting consensus sequences are analyzed and compared using suprachromosomal family (SF) consensus sequences from (Alexandrov et al., 2001; Romanova et al., 1996) and novel concept of extended start/stop Codon Like Trinucleotides (CLTs) (see Supplementary Table 1).

2. Results

2.1. HORs in chromosomes 5 of primates

The GRM algorithm (Paar et al., 2011a, 2011b) is applied to Build 37.3, 3.1, and 1.2 DNA assemblies of chromosome 5. We identify AS HORs in human (13mer), chimpanzee (5mer), orangutan (14mer) and macaque (3mers) (Table 1) and determine the corresponding consensus HORs and divergence between monomers within and between HORs.

We determine monomer structure of all HOR copies identified in GRM analysis of sequenced data in human, chimpanzee, orangutan and macaque genomic assemblies.The HOR schemes showing composition and alignment of constituent monomers are displayed in Fig. 1. In constituent monomers of each HOR copy we identify the essential part of CENP-B box and pJα motif positions (presented by open circles and squares, respectively). The consensus alpha satellite monomers for basic types A and B have only seven nucleotide differences, five of which are concentrated in a 17 bp region of pJα motif or CENP-B box. Such clustering indicates that these mutations are not random, but are affected by selection (Alexandrov et al., 2001; Romanova et al., 1996; Gaff et al., 1994). Shorter subsegments of pJα motif, GPuAAAAGGAA, and of CENP-B box, TTCG—A—CGGG, presenting the essential parts of the pJα and CENP-B box motifs, were effective when dimerized, while a number of mutations outside these cores did not abolish binding Romanova et al., 1996. The human 13mer HOR consensus array is characterized by duplication of a dimer consisting of Nos. 6 and 7 AS monomers in HOR copies No. 12–15, and by No. 9 AS monomer duplication in HOR copy No. 15. The other HORs in Fig. 1 show a partially riddled monomer structure. Tables 2A and 2B display average divergence between individual AS monomers from consensus HORs, and divergence between consensuses of all monomers from each consensus HOR, respectively. It is interesting that divergence between human and orangutan is slightly smaller than between human and chimpanzee. This might possibly be related to smaller number of sequenced HOR copies in orangutan. On the other hand, according to CLT extensions (discussed in Section 2.3), chimpanzee is closer to human than to orangutan AS HORs.

Divergence between human and great ape's monomers is $\sim$20%, similar to divergence between monomers within HORs for each species. Furthermore, we find substantial differences in CENP-B box and pJα distributions in AS consensus between chromosome 5 in human and other primates (Table 1). Human consensus AS 13mer HOR in chromosome 5 contains neither CENP-B box nor pJα motif. This HOR sequence is the only case of human AS HOR without both CENP-B box and pJα motifs. Only chimpanzee contains CENP-B box (in two AS monomers, No. 1 and No. 4, out of five monomers in consensus HOR). Contrary to humans, the other primates contain pJα motif in chromosome 5: chimpanzee and macaque have one pJα motif per consensus HOR, while orangutan has seven per consensus HOR.

2.2. Start/stop CLTs and extensions

In the next step we analyze the structure of AS HORs in chromosome 5 in primates using a novel approach based on start/stop CLT extensions (Rosandić et al., 2011) . In ASs from human chromosomes we find four pronounced groups of nucleotides in the form of extended cluster organization: although ASs are noncoding sequences, we consider their structure hypothetically as being based on trinucleotides that correspond to the start codon (ATG) and stop codons (TGA, TAG, TAA) in the standard case of coding sequences. In such a broader framework they are referred to as start/stop codon-like trinucleotides (CLTs). What significantly distinguishes AS sequences in human centromeric and pericentromeric regions from non-AS HORs and non-repeat sequences are specificity and the level of extension of start/stop CLTs, where the dominant contribution has the stop-TGA CLT

Table 1
Alpha satellite HORs identified in chromosomes 5 in human, chimpanzee, orangutan and macaque genomes. HORs are identified by GRM algorithm (Brawand et al., 2011; Paar et al., 2011) using Build 37.3 (human), Build 3.1 (chimpanzee) and Build 1.2 (orangutan and macaque) assemblies.

Species	NCBI build	HOR *n*mer[a]	Contig	Start[b]	HOR consensus length (bp)[c]	Orientation	CENP-B box (in mon. No.)[d]	PJ-a (in mon. No.)[e]	No. HOR copies
Human	37.3	13mer	NT_006713.15	243	2214	RC	–	–	16
Chimpanzee	3.1	5mer	NW_003457036.1	150	851	RC	1, 4	5	48
Orangutan	1.2	14mer	NW_002879914.1	1189985	2392	D	–	2, 3, 4, 7, 10, 12, 14	4
Macaque	1.2	3mer (1)	NW_001118154.1	120	515	D	–	3	15
Macaque	1.2	3mer (2)	NW_001118153.1	535827	515	D	–	3	59

[a] Human 13mer HOR was reported in (Baldini et al., 1989), other HORs are identified here.

[b] Start position within contig.

[c] Exact consensus length for human 13mer was determined in (Baldini et al., 1989) and for HORs in chimpanzee, orangutan and macaque are novel. Sequence orientations for human and chimpanzee are reverse complement, and for orangutan and macaque direct.

[d] CENP-B box motifs in consensus HORs. Monomers containing CENP-B box motif within each HOR are listed. Definition of CENP-B box is in accordance with (Romanova et al., 1996).

[e] pJα motifs in consensus HORs. Definition of pJα motif is in accordance with (Romanova et al., 1996).

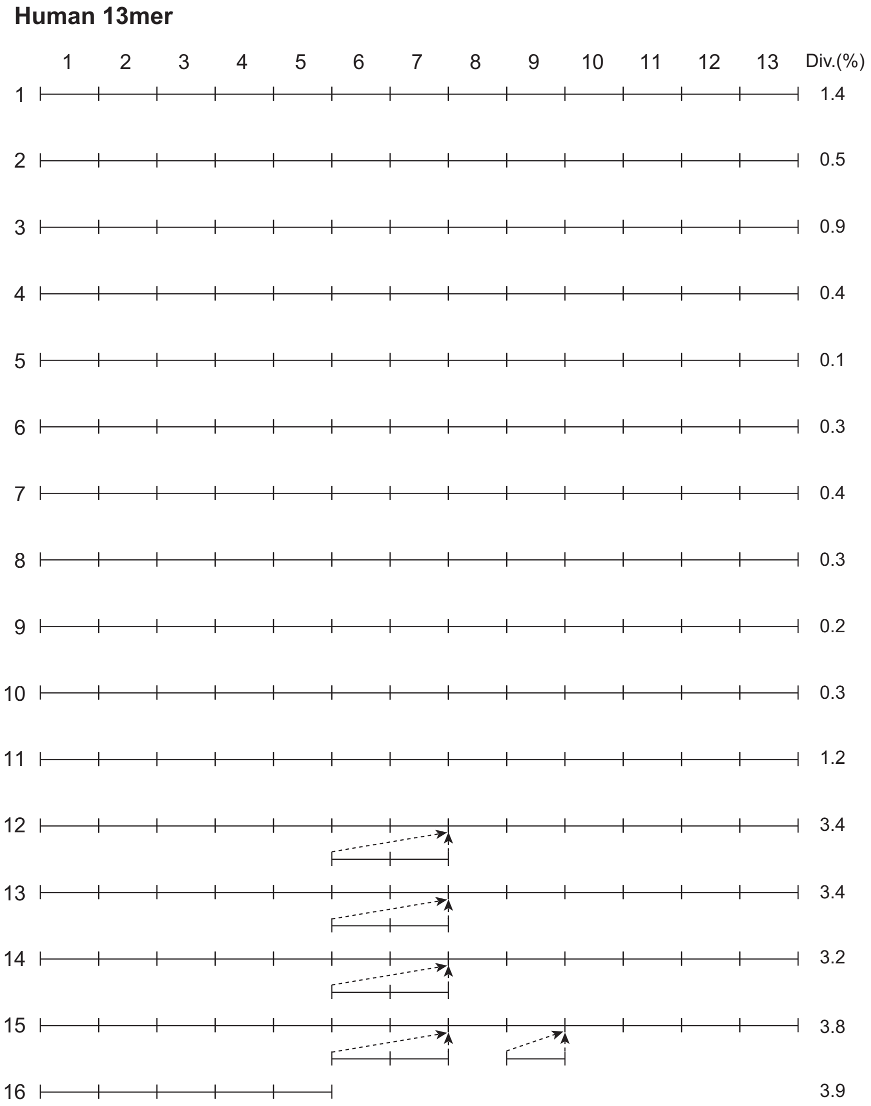

Fig. 1. Schematic presentation of aligned monomer structure of alpha satellite HORs in human (Build 37.3) chromosome 5 and distribution of CENP-B box and pJα motif. Top row: enumeration of *n* columns corresponding to *n* distinct constituent monomers (enumerated Nos. 1 to *n*) from the corresponding consensus *n*mer HOR. Each monomer copy is presented by a horizontal bar placed in the corresponding column numerated at the top. Each horizontal line formed of bars (i.e., array of monomers) presents a HOR copy. Monomers from different HOR copies corresponding to the same monomer from consensus HOR are presented by bars in the column corresponding to its enumeration at the top. For example, in the first HOR copy the second monomer corresponds to the monomer No. 2 from consensus HOR and is presented in the first horizontal line by a bar placed at the 2nd column. Open circle: CENP-B box (essential part) in a monomer. Open square: pJα motif (essential part) in a monomer. Average divergence between HOR copies and consensus HOR is 1.5%. In HOR copy No. 12 the constituent monomer Nos. are 1-5,6,7,6,7,8–13, i.e., the dimer of monomers No. 6, 7 is duplicated. We note that the monomer No. 7 between the two monomers No. 6 has a sizably enhanced divergence with respect to consensus. Similar pattern appears for HOR copies Nos. 13 and 14. In HOR copies No. 15 there is an additional duplication of monomer No. 9, i.e., the monomer composition is 1-5,6,7,6,7,8,9,9,10–13; the first monomer No. 9 has a sizably enhanced divergence with respect to consensus too. To the left of each HOR copy: average divergence between monomers in HOR copy with respect to consensus.

Rosandić et al., 2011. The frequency of nonextended and extended stop-TGA CLTs, i.e., the size of the corresponding cluster, guides us to the concept of codon like start/stop trinucleotides in noncoding sequences. In most cases we find in stop-CLTs within monomers in HORs more pronounced successive multiplications of T and less pronounced of A nucleotides. This results in poly-T–G-poly-A extended motifs, like for example TTTTGAA. Arrays dominated by T multiplications, are referred to as poly-T class. Such type of CLT extensions exhausts about 30% of all T and A nucleotides in ASs, and represents an extended form of stop-TGA CLT. In fewer cases we find AS arrays with reversed situation, having more A multiplications than T. Such arrays, dominated by A multiplications, are referred to as poly-A class. They are mostly present in monomers outside HORs. Poly-T and poly-A classes represent a new AS classification. Human HOR consensus ASs belong to the poly-T class, while the monomeric ASs can belong either to poly-A or to poly-T class.

We find in each AS monomer from consensus HOR of human chromosome 5 two-to-four start-CLTs and ten-to-fourteen stop-CLTs. The highest frequency is associated with stop-TGA CLT. This result is in accordance with our findings for other human chromosomes (Rosandić et al., 2011). Here we obtain a similar start/stop CLT-pattern in great apes (chimpanzee and orangutan) with small individual differences (see Fig. 2). In most of chromosomes, the ratio of extended to non-extended CLTs, referred to as *r*-factor, is significantly higher in human than in chimpanzee and orangutan ASs (Rosandić et al., 2011). The exception is the chromosome 5, studied here, in which case the human *r*-factor is similar to chimpanzee, while the *r*-factor for orangutan is smaller (Table 3A). Thus, the case of chromosome 5 is anomalous in the sense that the dominant stop-TGA CLT extensions in human and chimpanzee are similar, even slightly larger in chimpanzee. On the other hand, in macaque we find significantly different distribution of start/stop CLTs and substantial CLT difference between the two 3mer HORs (see Fig. 2).

On the average, the *r*-factor in human HOR region is ~9, while in the monomeric region it is ~8, but this difference is still smaller than standard deviation.

In consensus 5mer HOR in chromosome 5 of chimpanzee all AS monomers belong to the poly-T class, while among the monomeric ASs some belong to poly-A and some to poly-T class, similarly as in human genome. In consensus AS HOR in chromosome 5 of orangutan we also find the poly-T class assignment; only one of fourteen AS monomers is of poly-A class. Contrary to the poly-T class pattern of human, chimpanzee and orangutan, in macaque the consensus sequences of both 3mer HORs in chromosome 5 belong to the poly-A class. We also find that in both 3mer HORs in macaque the stop-TGA CLT extension is substantially smaller and the stop-TAG CLT extension substantially larger than in human and great apes.

Furthermore, we find that great apes are characterized by low-TA dinucleotide pattern, similarly as found in human ASs (Rosandić et al., 2011). On the contrary, macaque shows no low-TA pattern at all. Thus, the ASs in both HORs in chromosome 5 in macaque resemble, for example, to non-repeat regions in human chromosomes 5, 7 and 21 and to intra-gene HOR with 2.4 kb repeat unit in human chromosome 1, where the similarity refers to the dinucleotide pattern in the satellite (Rosandić et al., 2011). On this basis, one could argue that ASs in macaque genome were evolutionary slowed down at a lower level of development and HORs seem to be still in the process of creation. With respect to extensions, poly-A pattern and the absence of TA-low behavior, HORs in macaque resemble to the non-repeat regions.

In great apes and especially in humans the interior structure of ASs is organized with pronounced robust segments. Within ASs there are only some smaller qualitative differences in distribution of start/stop CLTs, which are species-specific (Fig. 2). Quantitative differences are mostly present on the level of extension of start/stop CLTs, in such a way that in human ASs the stop-TGA CLT extension is most pronounced, while the extension of stop-TAG CLT is absent or small.

Table 2A
Average divergence (%) between monomers from consensus HORs in human, chimpanzee, orangutan and macaque AS arrays. H13mer—human 13mer, C5mer—chimpanzee 5mer, O14mer—orangutan 14mer, M3meR1—macaque 3mer HOR(1), M3mer2—macaque 3mer HOR(2).

	H13mer	C5mer	O14mer	M3mer1	M3mer2
H13mer	21	23	19	31	28
C5mer		24	21	33	29
O14mer			16	31	27
M3mer1				22	23
M3mer2					20

Table 2B
Divergence (%) between consensus of all monomers from each consensus HOR. For example, consensus monomer corresponding to human 13mer HOR is determined as consensus of all 13 monomers from consensus human 13mer HOR.

	C5mer	O14mer	M3mer1	M3mer2
H13mer	2	1	23	16
C5mer		3	23	17
O14mer			24	17
M3mer1				14

2.3. Suprachromosomal family classification and CLT extension

We determine SF classification of consensus HORs in human, chimpanzee, orangutan and macaque AS HORs in chromosome 5 in a standard way by aligning primate monomers to SF consensus monomers from (Romanova et al., 1996) and finding for each primate monomer the closest SF consensus monomer (Table 4, last column and Supplementary Table 1). As seen, the human, chimpanzee and orangutan HORs are derived from SF5 sequences R2 and R1, representing the AB type ASs. In general, SF5 is formed by two types of monomers, R1 and R2, alternating irregularly. R2 is similar to M1 (class A), and R1 represents the first appearance of novel class B monomers, which bind CENP-B protein and presumably have invaded the A-arrays before the great ape divergence (Alexandrov et al., 2001; Romanova et al., 1996; Shepelev et al., 2009).

Average divergence of primate monomers of R1-type, R2-type and M1-type from the R1, R2 and M1 SF consensus monomers from (Romanova et al., 1996), respectively, is shown (Table 5). We see that the average divergence for human AS R1-type monomers is lower than for R2-type AS monomers. On the contrary, for chimpanzee and orangutan average divergence is lower for R2-type AS monomers than for R1-type. Divergence of R2-type monomers with respect to SF consensus is lower for orangutan than for human and chimpanzee monomers from HORs. The macaque sequences belong to ancient alpha AS M1 which existed before the appearance of the B-type monomers R1 in the human lineage group, with the A-only type AS (Shepelev et al., 2009). The next larger divergence of macaque monomers from 3mer HORs is with respect to R2 (on the average, R2 divergence is by only 1%

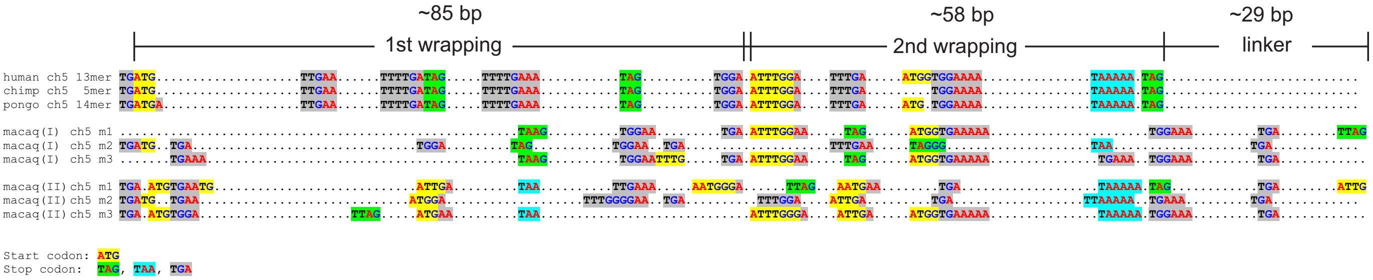

Fig. 2. CLTs and extensions in consensus alpha satellites in chromosome 5 of human, chimpanzee, orangutan and macaque genomes. Alignment of start/stop CLTs and of the corresponding extended cluster structure in consensus of AS monomers from consensus HORs in chromosome 5 in human, chimpanzee, orangutan, and macaque genome. Top: Segments corresponding to the first wrapping (~85 bp), to the second wrapping (~57 bp) and to the linker (~29 bp). Two vertical bars encompassing 85th nucleotide T position between the two wrappings is assigned to the first wrapping. This position, in the middle of AS monomer, is labeled by a dot. Rows 1–3, human, chimpanzee, and orangutan, respectively. Monomer sequence in each row is consensus of AS monomers from consensus HOR. Rows 4–6, the first array of AS 3mer HOR (containing 15 HOR copies) in macaque. Because of sizeable divergence among three monomers in consensus HORs, the rows correspond to the first, second and third monomer from consensus HOR, respectively. Rows 7–9, the second array of AS 3mer HOR (containing 59 HOR copies) in macaque. The rows correspond to the first, second and third monomer from consensus HOR, respectively.

Table 3
Stop-TGA CLTs in primate chromosomes 5. (A) Extended and non-extended stop-TGA CLTs and *r*-factor in chromosome 5. *E*(nt), percentage of nucleotides in extended stop-TGA CLTs; *NE*(nt), percentage of nucleotides in non-extended stop-TGA CLTs in genomic sequence; *r*, ratio of *E* and *NE*, *r*=*E*(nt)/*NE*(nt). (B) Eight copies of dominant stop-TGA CLTs appearing in consensus AS monomer of human chromosome 5. (C) Example of possible impact of a single-point mutation on the subsequence TTTTGAA, an extended stop-TGA CLT. Underlined, positions of mutated nucleotide C substituting one of nucleotides from TTTTGAA. A consequence of such single mutation may be destroying the stop-TGA effect (right column) or reducing/without effect (left column).

(A)

		TGA CLT (nt)		
Species	**HOR *n*mer**	***E*(nt)**	***NE*(nt)**	***r***
Human	13 mer	21.9	1.8	12.4
Chimpanzee	5 mer	22.7	1.8	12.9
Orangutan	14 mer	21.6	2.1	10.1
Macaque	3 mer (1)	14.6	4.7	3.1
Macaque	3 mer (2)	16.9	5.2	3.2

(B)

Stop TGA CLTs
TGA
TTGAA
TTTTGA
TTTGAAA
TGGA
TTTGGA
TTTGA
TGGAAAA

(C)

Effect of point mutation within TTTTGAA CLT

Without effect or reducing	**Destroying**
$\underline{\mathrm{C}}$TTTGAA	TTT$\underline{\mathrm{C}}$GAA
T$\underline{\mathrm{C}}$TTGAA	TTTT$\underline{\mathrm{C}}$AA
TT$\underline{\mathrm{C}}$TGAA	TTTTG$\underline{\mathrm{C}}$A
TTTTGA$\underline{\mathrm{C}}$	

larger than the dominant M1 divergence for HOR 1, and by 5% for HOR 2). Most of AB-type satellites belong to poly-T group, while within the A-only ASs both are significantly present, with poly-A group prevailing.

The predominant SF structure of major AS arrays in human, chimpanzee, orangutan and macaque chromosome 5 is displayed in Table 6. Each of arrays in human, chimpanzee, and orangutan is characterized by central position of R-type AS monomers R1 and R2, organized in HORs, surrounded by monomeric predominantly M1-type AS monomers. On the contrary, in macaque the central segment consists of predominantly M1-type AS monomers. In human genome the monomeric R-type segment consists of 60% R2 and 33% R1, and the 13mer HOR R-type segment of 52% R2 and 46% R1. In chimpanzee R2 and R1 contributions are comparable and in orangutan the contribution of R2-type monomers is much larger than of R1-type monomers.

Considering contributions to the TT/AA ratio separately for SF groups of R1- and R2-type AS monomers (Table 7) we find that the average value of the TT/AA dinucleotide ratio is larger for R1 than for R2 groups, i.e., that the B-type monomers contribute more to nucleotide clusterization than the A-type monomers, although the individual values have sizeable fluctuations. A similar relation is obtained for the trinucleotide ratio TTT/AAA. Both the B-class SF R1-type percentage and the CLT extensions are slowly increasing from macaque over great apes to human AS HORs.

AS monomers in consensus 13mer HOR in human chromosome 5 belong to SF R1 and R2 which are of A-type and B-type, respectively, but they do not have either CENP-B or pJα binding sites. Because they contain monomers which class as AB type, they do have what could be called A and B boxes from (Romanova et al., 1996), as opposed to functional binding sites. One could also argue that the alignment of SF5 monomers from HORs to SF5 consensus monomers R1 and R2 gives divergence of 7–15%. Assuming random distribution of mutations along each monomer, the CENP-B box could lose essential core. Deviations of HOR SF monomers from SF consensus monomers R1 and R2 involves nucleotide differences which can hit the region of CENP-B box. By allowing relaxation of essential CENP-B box and pJα core by appropriately changing only one nucleotide, the binding site boxes can arise in ten out of thirteen monomers from 13mer consensus HOR: CENP-B box in monomers Nos. 1,3 and pJα in monomers Nos. 6–13. Thus, one-nucleotide substitution and subsequent multiplication could lead to the observed CENP-B box-free and pJα-free 13mer HOR.

It was found that most of active human centromeres are made of great ape-specific ASs organized in nearly identical HORs belonging to new SFs 1, 2 and 3, surrounded by less homogeneous HOR-free monomeric ASs of SF 4 and SF 5 types, often disrupted by transposon insertions (Alexandrov, 2001; Romanova et al., 1996; Kazakov et al., 2003; Shepelev et al., 2009). The SF5 HOR in human chromosome 5, investigated here, is a low copy number HOR with no ortholog in the sequenced chimpanzee genomic sequence and it could represent a secondary recent amplification of a segment in dead SF5 centromere that belonged to a common ancestor of orangutan and humans, functionally different from the new family HORs of SFs 1, 2 and 3. On the other hand, orangutans do not have the new SFs and their current centromeres are formed by SF5 (Shepelev et al., 2009).

Specification of copy number and homogeneity for primate HORs is seen in the schemes of HORs (see Fig. 1). A low-copy number human 13mer HOR, free of CENP-B box and pJα binding sites, could be perceived as small scale amplifications in the dead layers and the other HORs may be part of the homogeneous cores of functional centromeres. From this figure we can see increased homogeneity within HOR region, which decreases towards the edges of HOR regions.

In Table 6 we display the major AS arrays in primate chromosome 5 and their predominant SF classification. In human AS array we find two R2/R1 segments, separated by a 3 Mb unsequenced gap, and surrounded by monomeric M1 segments. One of the two R2/R1 segments is largely organized into CENP-B box-free and pJα-free 13mer HOR, bordering with the neighboring M1 segment, while the other R2/R1 segment is monomeric. In chimpanzee AS array we find R2/R1 segment containing 5mer HORs and in orangutan R2/R1 segment containing 14mer HORs. In macaque AS we find no R2/R1 segment, but there is a M1 segment containing 3mer HOR. In human chromosome 5 the monomeric segments I and III, with similar SF assignments, have different polyT/polyA composition. Considering different copies of similar mononucleotide repeats, their polyT/polyA pattern can widely differ. Thus we conclude that similar non-HOR sequences do not fall into the same polyT/polyA class. This is consistent with different biological content of SF and polyT/polyA classifications. Similar results are obtained for other primates considered here. The general conclusion is in accordance with previous assertion that the human HOR ASs belong mostly to poly-T class, while the monomeric ASs belong either to poly-T or poly-A class.

Table 4
SF classification and TT/AA, TTT/AAA ratios for alpha satellite monomers in primate consensus HORs. The fourth and fifth column show the ratio of number of TT and AA dinucleotides within consensus monomers, and of TTT and AAA trinucleotides, respectively. The last column denotes SF classification computed by aligning to the consensus SF monomers from Romanova et al., 1996. Two different 3mer HORs, identified in macaque, are denoted as 3mer(1) and 3mer(2).

Species	HOR	Monomer	TT/AA	TTT/AAA	SF class
Human	13mer	m01	2.9	1	R1
		m02	1.6	1.9	R2
		m03	1.2	1	R1
		m04	1.2	1.1	R1
		m05	1.5	1.8	R2
		m06	1.2	1.4	R1
		m07	1.3	2.5	R1
		m08	1.1	0.9	R2
		m09	0.8	0.6	R2
		m10	1.4	1.2	R1
		m11	1.2	0.9	R2
		m12	1.1	1.2	R2
		m13	1.2	1.3	R2
Chimpanzee	5mer	m01	1.4	1.8	R1
		m02	1.1	1	R2
		m03	2.2	5.5	R1,R2
		m04	1.5	2	R1
		m05	1.2	1.3	R2
Orangutan	14mer	m01	1.5	2.2	R1,R2
		m02	1.1	1.2	R2
		m03	0.9	1	R1,R2
		m04	1.2	1.1	R1,R2
		m05	0.9	0.8	R2
		m06	1.1	1.1	R2
		m07	1.4	1.4	R2
		m08	1	1.3	R2
		m09	1	1	R2
		m10	1.1	1.1	R2
		m11	1.1	1.1	R2
		m12	1.4	1.5	R2
		m13	1.2	1.8	R2
		m14	1.2	1.3	R2
Macaque	3mer(1)	m01	0.6	0.3	M1
		m02	1.1	1.4	M1/*R2*
		m03	0.8	0.7	M1
	3mer(2)	m01	1	1.1	M1
		m02	1	0.9	M1
		m03	0.7	0.4	M1

2.4. *Wrapping of human AS in nucleosome*

Up to now, it was known that the structure of ~171 bp AS monomers is characterized by (A+T)-rich content (Warburton and Willard, 1996; Romanova et al., 1996), and that the TA dinucleotide is broadly underrepresented genome-wide, intrinsically less stable energetically than the other dinucleotides (Nussinov, 1984; Karlin and Ladunga, 1994; Karlin and Mrazek, 1997). We find here that the internal structure of ASs is characterized by extended clusters of trinucleotides. A dominant role in this respect is due to stop-TGA CLT, as the most pronounced extensions involving multiplications of T and A nucleotides, as already mentioned. Just the level of extension of start/stop CLTs distinguishes most of human chromosomes from other primates.

Distribution of extended and non-extended CLTs in AS monomers does not appear randomly, but forms a well organized structure. The TGATG segment represents a junction of neighboring AS copies forming a tandem. In our opinion, each monomer copy ends with stop-TGA CLT and the next copy starts with start-ATG CLT, forming a very robust TGATG subsequence, with nucleotide A overlap in the middle. In this way, we segment the structure of AS into three parts: two wrapping sequences, the first of ~85 bp with non-extended start-ATG CLT and the second of ~57 bp with extended start-ATG CLT (ATTTGG) at the beginning (Rosandić, 2011). There are no other ATG CLTs in overall human consensus ASs. The third segment represents a linker of ~29 bp, ending with TG which forms the stop/start CLT TGATG as already mentioned (see Fig. 3). Outside of TGATG there are no start/stop CLTs in the linker, in contrast to the CLT-rich structure of wrapping sequences. Thus, a robust CLT alignment of human HOR ASs enables detailed reconstruction of AS DNA wrapping in nucleosomes and of the linker.

Between the end of the first wrapping sequence, ending with TGGA CLT, and at the start of the second wrapping sequence, starting with ATTTGG CLT, one T nucleotide is positioned in the

Table 5
Average divergence of primate R1-type, R2-type and M1-type monomers from consensus HORs with respect to R1, R2 and M1 SF consensus monomers, respectively.

	Average divergence (%)		
	R1-type vs. R1 SF	**R2-type vs. R2 SF**	**M1-type vs. M1 SF**
Human (13mer)	11.5	12.1	–
Chimpanzee (5mer)	14.6	11.5	–
Orangutan (14mer)	12.7	10.2	–
Macaque (3mer 1)	–	–	25.8
Macaque (3mer 2)	–	–	20.5

Table 6
SF structure of major AS arrays containing HORs in human, chimpanzee, orangutan and macaque chromosome 5. Human AS array: Human monomeric segment HI contains 1955 AS monomers, with dominant classification M1. In this segment the non-AS insertions amount to 86 kb; the three largest are of 22 kb, 10 kb and 8 kb. Direct (denoted d) and reverse complement (denoted r) subsegments alternate seven times; the longest subsegment d is of 169 kb and the longest subsegment r 45 kb. Percentage of AS monomers containing essential part of CENP-B box is listed in column CBB and percentage of monomers containing pJα motif is listed in column pJα. Human monomeric segment HIIa contains 362 AS monomer copies of predominantly R2-type SF, including also a sizeable number of R1-type monomers. There are seven poly-R2 arrays longer than 5R2 (the longest contain 23R2 and 11R2), but the longest poly-R1 is only 5R1. One third of monomers contain essential part of pJα motif and $\sim$5% contain essential part of CENP-B box. Human monomeric segment HIIc contains 209 AS monomer copies of predominantly R2 and R1 SF classification (109 R2, 97 R1), organized into $\sim$16 HOR copies. The consensus HOR contains neither pJα nor CENP B Box essential motifs. Human monomeric segment HIII contains 506 AS monomers of predominantly M1 SF. All monomers in this segment are reversed complement (r). Almost half of AS monomers contain scattered essential pJα motifs and none contains CENP-B Box motif.

Chimpanzee AS array: Chimpanzee monomeric segment CI contains predominantly M1 SF monomers. Towards its end several poly-R2 subsequences are inserted. Several non-AS insertions are present, the largest is 4.4 kb long. Around one third of ASs contains scattered essential parts of pJα motifs and none contains CENP B Box motif. Chimpanzee HOR segment CII contains AS monomers of predominantly R2 SF classification, with a significant R1-type contribution. (831 R2, 552 R1). Around 30% of AS monomers in HOR copies contain essential part of pJα motif and 5% contain CENP-B Box. Chimpanzee monomeric segment CIII contains predominantly M1 SF monomers. Around one third of AS contain scattered essential parts of pJα motifs and in most of this segment there is no CENP B Box motif. Only near the end of this segment there appear CENP-B Box motifs and moreover at high density, even more frequently than pJα.

Orangutan AS array: In the monomeric segments OI and OIII the dominant SF classification is M1. In the 14mer HOR segment OIIb and the pre-HOR monomeric segment OIIa AS monomers are predominantly R2, with a sizeable contribution of R1. In the post-HOR monomeric segment OIIc there is equal number of R1 and R2-type AS monomers. In the monomeric segments OI, OIIa, OIIc, and OIII around one third of ASs contains scattered essential parts of pJα motifs. In the 14mer HOR segment two third of AS contain pJα motifs and none contains CENP B box motif. Macaque AS array: Among monomeric segments of predominantly M1 type we find two segments MIIa and MIIc with different 3mer HORs, both of M1-type and without *R*1-type contributions. In HOR and inter-HOR segments MIIa, MIIb, MIIc the number of AS monomers with essential parts of pJα motifs is significantly higher in comparison to the other segments.

Species	Segment	Largest SFs	Length (kb)	Orientation	CBB (%)	pJα (%)	AS organization
Human							
	HI.	95%M1, 3%R2	420	7(d,r)	0	32	monomeric
	HIIa	60%R2, 33%R1	77	1d,1r	5	29	monomeric
	HIIb	gap	3000				
	HIIc	52%R2, 46%R1	36	all r	0	0	13merHOR
	HIII	89%M1, 11%R2	118	all r	0	53	monomeric
Chimpanzee							
	CI	74%M1, 25%R2	128	all d	0	45	monomeric
	CII	58%R2, 38%R1	312	2(d,r)	6	27	5merHOR
	CIII	89%M1, 11%R2	270	3(d,r)	0	32	monomeric
Orangutan							
	OI	94%M1,5%R2	591	5(d,r)	0	35	monomeric
	OIIa	75%R2, 20%R1	147	all r	0	44	monomeric
	OIIb	71%R2, 25%R1	14	all d	0	71	14merHOR
	OIIc	49%R2, 49%R1	79	all r	0	30	monomeric
	OIII	64%M1, 25%R2	312	all r	0	37	monomeric
Macaque							
	MIa	83%M1, 12%R2	48	1d, 1r	0	2	monomeric
	MIb	86%M1, 9%R2	80	1d, 1r	0	0	monomeric
	MIc	68%M1, 19%R2	17	all d	0	5	monomeric
	MIIa	89%M1, 3%R2	41	all d	0	18	3merHOR(2)
	MIIb	89%M1, 8%R2	53	2 d, 3 r	0	33	monomeric
	MIIc	88%M1, 8%R2	36	all d	0	23	3merHOR(1)
	MIII	81%M1, 15%R2	28	all r	0	1	monomeric

Table 7
Average values of dinucleotide TT/AA ratios separately for R1-type and R2-type alpha satellite monomers in human, chimpanzee and orangutan. In cases of monomers with equal divergence with respect to both R1 and R2 consensus SF monomer groups, the same value of ratio is assigned to both R1 and R2 groups.

	TT/AA	
	R1	R2
Human (13mer)	1.5	1.2
Chimpanzee (5mer)	1.3	1.1
Orangutan (14mer)	1.2	1.1

middle of consensus AS monomer (see Fig. 2). The second half of AS contains the second wrapping sequence and the linker. The second ATG CLT is positioned just after the first half of the monomer. The three non-extended stop-TAG CLTs are at special positions in consensus AS wrapping sequences. The first stop-TAG CLT is positioned in the middle of the first wrapping sequence, and the second stop-TAG CLT in the middle of a combined sequence composed of the first and second wrapping sequences. The third stop-TAG CLT is positioned at the end of the second wrapping sequence.

It is remarkable that the dominant stop-TGA CLTs appear eight times in the AS monomer consensus and in seven out of eight cases in extended form (Table 3B). The only exception of non-extended form is TGATG at the start of the first wrapping sequence. At the start of second wrapping sequence the extended TGA appears in front and behind the extended start-ATG CLT, in TGGA · ATTTGGA. The remaining five extended stop-TGA CLTs, together with the only extended stop-TAA CLT at the end of the second wrapping sequence can be considered as three pairs of CLTs. Such pronounced dominance of stop-TGA CLT appears as a signature of a subtle AS dynamics.

3. Discussion

Previously, some information on ASs in chimpanzee and orangutan genomes was obtained (Haaf et al., 1997; Haaf and Willard, 1998). It was suggested that the subset homologous to the clone PPY2–5 is organized in distinct HOR structures consisting of 18 adjacent monomers. However, this subset resides on three PPY

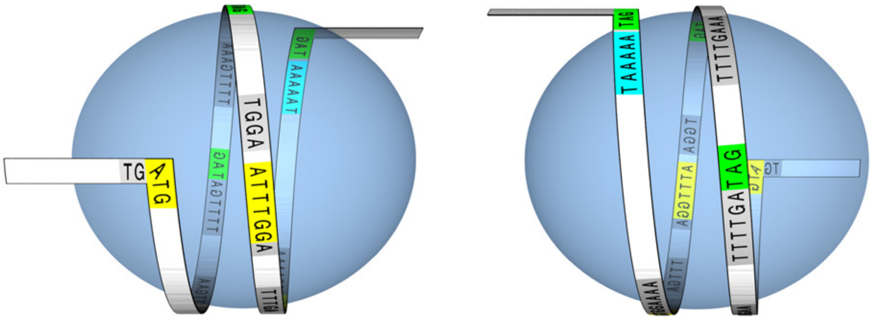

Fig. 3. Schematic presentation of human consensus AS monomer wrapping around the nucleosome. The histone core is represented as ellipsoid and DNA as twisted ribbon. (a) Front view; (b) Back view. Start/stop CLTs are colored according to Fig. 2.

chromosomes but the target chromosomes were not identified. It was not known on which chimpanzee chromosomes the PTR subset resides. Some digests revealed more prominent bands at the 6mer and 12mer position, but the bulk of the delineated subsets remained undigested. It was noted that this may indicate the presence of an evolutionary very young HOR structure that has not yet been extensively amplified in the orangutan genome. The chimpanzee subset recognized by PPY3–5 showed a prominent band at the 16mer position(Haaf et al., 1998). Clones $p\alpha$ PTR4N and $p\alpha$ PTR4H hybridized only to the centromeric region of the chimp chromosome 3 (in old notation assigned as chromosome 4) (Haaf et al., 1997). Hybridization patterns were produced by both clones. In several digests a prominent band at the pentamer position (850 bp, 5mer) was found (Haaf et al., 1997), but without chromosome assignment.

As more AS subsets from great ape chromosomes were cloned and characterized (Haaf and Willard, 1998; Jorgensen et al., 1992; Warburton et al., 1996), it became evident that the evolutionary relationship within and between human and higher primate species should rely more on DNA sequence comparisons (Haaf et al., 1998). Our analysis of chromosome 5 in chimpanzee and orangutan identifies 5mer and 14mer HORs, respectively. The macaque centromere was previously characterized as 2mer (Alves et al., 1998), while our analysis of macaque Build 1.2 assembly identifies two distinct 3mer HOR arrays (consisting of 15 and 59 HOR copies, respectively).

We hypothesize that a regulatory system including start/stop CLTs can act locally, as for example TGATG as the end and start signal at a junction of neighboring AS monomers in tandem. We note a possibility of long range interaction arising so that different multiplication patterns of A, T and G nucleotides in extensions of start/stop CLTs in ASs might influence expression of codons in coding sequences. For example, the stop-TGA CLT, which is most strongly extended within the AS, takes the extended form TTTTGAA. We hypothesize that in this way it might exert the stop function on some segments of the genome, possibly on trinucleotides underlying the extended stop-TGA CLTs (i.e., in this case TTT, TTT, TTG, TGA, GAA).

In this connection we note possible consequences of point mutations within start/stop CLTs. In the same case of TTTTGAA as the extended stop-TGA CLT, single point mutations are shown (Table 3C). Point mutations that are not destroying the stop-TGA CLT pattern can at most reduce its possible stop-effect by decreasing the number of multiplications of T and A nucleotides. On the contrary, the effect of point mutations that destroy the TGA motif might lead to destroying of the stop effect in its zone of action if the AS is involved in the regulator function. Thus, some locations of point mutations in extended CLTs could result in their reduction or destruction. This might lead to irregular expansion, as for example of GAA trinucleotides, with possible consequences in protein genesis (Gatchel et al., 2005; Kozlowski et al., 2010).

There is a kind of complementarity between extended start/stop CLTs and existing SF approaches. On one hand, SF classification concerns the classification of the structure of ASs; in particular, the differences betwen A and B types of SF classes are concentrated in a small region which matches functional protein binding sites, for pJα in type A and for CENP-B in type B (Romanova et al., 1996). On the other hand, extended start/stop CLTs mechanism is present in all noncoding sequences, and ASs provide an example with pronounced dominance of extended stop-TGA CLTs. For difference with respect to noncoding sequences, in coding sequences there is no such pronounced scope of CLT extensions. The start/stop CLTs appear in clusters in non-extended and extended forms, that we consider as sophisticated structure of strongly specific ASs in HORs in primates, most pronounced in human genome. We hypothesize that these extended CLTs are regulators in noncoding sequences, acting at a distance. This opens a new avenue to follow the evolutionary development of centromere. On this evolutionary path we could follow how they enrich in start/stop CLTs and how their extensions proceed by multi-T and/or multi-A multiplications.

Author contributions

All authors contributed to investigations, discussions, and the writing of the manuscript. M.G. performed the computations.

Acknowledgments

We thank C. Tyler-Smith for stimulating our interest in the structure of ASs in primates and to T. Haaf for useful information on experimental studies of chimpanzee and orangutan ASs. We wish to thank anonymous referees for very helpful comments and recommendations to improve the manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2012.09.022.

References

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

Alexandrov, I.A., Yurov, Y.B., Mitkevich, S.P., Gindilis, V.M., 1986. Chromosome organization of human alphoid DNA. Dokl. Akad. Nauk 288, 242–245.

Alkan, C., Ventura, M., Archidiacono, N., Rocchi, M., Sahinalp, S.C., Eichler, E.E., 2007. Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. PLoS Comput. Biol. 3, e18.

Alkan, C., Cardone, M.F., Catacchio, C.R., Antonacci, F., OBrien, S.J., Ryder, O.A., Purgato, S., Zoli, M., Della Valle, G., Eichler, E.E., Ventura, M., 2011. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res. 21, 137–145.

Alves, G., Seuanez, H.N., Fanning, T., 1998. A clade of new world primates with distinctive alphoid satellite DNAs. Mol. Phylogenet. Evol. 9, 220–224.

Baldini, A., Smith, D.I., Rocchi, M., Miller, O.J., Miller, D.A., 1989. A human alphoid DNA clone from the EcoRI dimeric family: genomic and internal organization and chromosomal assignment. Genomics 5, 822–828.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F.W., Zeller, U., Khaitovich, P., Grutzner, F., Bergmann, S., Nielsen, R., Paabo, S., Kaessmann, H., 2011. The evolution of gene expression level in mammalian organs. Nature 478, 343–348.

Clarke, L., Baum, M.P., 1990. Functional analysis of a centromere from fission yeast—a role for centromere-specific repeated DNA sequences. Mol. Cell. Biol. 10, 1863–1872.

Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S.L., Wiebe, V., Kitano, T., Monaco, A.P., Paabo, S., 2002. Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418, 869–872.

Furuyama, S., Biggins, S., 2007. Centromere identity is specified by a single centromeric nucleosome in budding yeast. PNAS 104, 14706–14711.

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

Gatchel, J.R., Zoghbi, H.Y., 2005. Diseases of unstable repeat expansion: mechanisms and common principles. Nat. Rev. Genet. 6, 743–755.

Gilad, Y., Oshlack, A., Smyth, G.K., Speed, T.P., White, K.P, 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature 440, 242–245.

Haaf, T., Willard, H.F., 1997. Chromosome-specific alpha satellite DNA from the centromere of chimpanzee chromosome 4. Chromosoma 106, 226–232.

Haaf, T., Willard, H.F., 1998. Orangutan alpha satellite monomers are closely related to the human consensus sequence. Mamm. Genome 9, 440–447.

Haussler, D., 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441, 87–90.

Haygood, R., Babbitt, C.C., Fedrigo, O., Wray, G.A., 2010. Contrasts between adaptive coding and noncoding changes during human evolution. PNAS 107, 7853–7857.

Hulsebos, T., Schonk, D., van Dalen, I., Coerwinkel-Driessen, M., Schepens, J., Ropers, H.H., Wieringa, B., 1988. Isolation and characterization of alphoid DNA sequences specific for the pericentric regions of chromosomes 4, 5, 9, and 19. Cytogenet. Cell Genet. 47, 144–148.

Jorgensen, A.L., Laursen, H.B., Jones, C., Bak, A.L., 1992. Evolutionarily different alphoid repeat DNA on homologous chromosomes in human and chimpanzee. PNAS 89, 3310–3314.

Karlin, S., Ladunga, I., 1994. Comparisons of eukaryotic genomic sequences. PNAS 91, 12832–12836.

Karlin, S., Mrazek, J., 1997. Compositional differences within and between eukaryotic genomes. PNAS 94, 10227–10232.

Kazakov, A.E., Shepelev, V.A., Tumeneva, I.G., Alexandrov, A.A., Yurov, Y.B., Alexandrov, I.A., 2003. Interspersed repeats are found predominantly in the old alpha satellite families. Genomics 82, 619–627.

King, D.C., Taylor, J., Zhang, Y., Cheng, Y., Lawson, H.A., Martin, J., Chiaromonte, F., Miller, W., Hardison, R.C., 2007. Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. Genome Res. 17, 775–786.

Kozlowski, P., de Mezer, M., Krzyzosiak, W.J., 2010. Trinucleotide repeats in human genome and exome. Nucleic Acids Res. 38, 4027–4039.

Mahtani, M.M., Willard, H.F., 1990. Pulsed-field gel analysis of alpha-satellite DNA at the human X chromosome centromere: high-frequency polymorphisms and array size estimate. Genomics 7, 607–613.

Nussinov, R., 1984. Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res. 12, 1749–1763.

Orom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q.H., Guigo, R., Shiekhattar, R., 2010. Long noncoding RNAs with enhancer-like function in human cells. Cell 143, 46–58.

Paar, V., Glunčić, M., Basar, I., Rosandić, M., Paar, P., Cvitković, M., 2011a. Large tandem, higher order repeats and regularly dispersed repeat units contribute substantially to divergence between human and chimpanzee Y chromosomes. J. Mol. Evol. 72, 34–55.

Paar, V., Glunčić, M., Rosandić, M., Basar, I., Vlahović, I., 2011b. Intragene higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees. Mol. Biol. Evol. 28, 1877–1892.

Pennacchio, L.A., Rubin, E.M., 2001. Genomic strategies to identify mammalian regulatory sequences. Nat. Rev. Genet. 2, 100–109.

Puechberty, I., Laurent, A.M., Gimenez, S., Billault, A., Brun-Laurent, M.E., Calenda, A., Marcais, B., Prades, C., Ioannou, P., Yurov, Y., Roizes, G., 1999. Genetic and physical analyses of the centromeric and pericentromeric regions of human chromosome 5: recombination across 5cen. Genomics 56, 274–287.

Romanova, L.Y., Deriagin, G.V., Mashkova, T.D., Tumeneva, I.G., Mushegian, A.R., Kisselev, L.L., Alexandrov, I.A., 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., Glunčić, M., Pavin, N., Pilaš, I., 2006. CENP-B box and pJα sequence distribution in human alpha satellite higher-order repeats (HOR). Chromosome Res. 14, 735–753.

Rosandić, M., Glunčić, M., Paar, V., 2011. Start/stop codon-like trinucleotides (CLTs) and extended clusters as new language of DNA. Croat. Chem. Acta 84, 331–341.

Rudd, M.K., Willard, H.F., 2004. Analysis of the centromeric regions of the human genome assembly. Trends Genet. 20, 529–533.

Rudd, M.K., Wray, G.A., Willard, H.F, 2006. The evolutionary dynamics of alpha satellite. Genome Res. 16, 88–96.

Shepelev, V.A., Alexandrov, A.A., Yurov, Y.B., Alexandrov, I.A., 2009. The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. PLoS Genet. 5, e1000641.

Tyler-Smith, C., 1985. Structure of repeated sequences in the centromeric region of the human Y chromosome. Development 101, 93–100.

Warburton, P., Haaf, T., Gosden, J., Lawson, D., Willard, H.F., 1996. Characterization of a chromosome-specific chimpanzee alpha satellite subset: evolutionary relationship to subsets on human chromosomes. Genomics 33, 220–228.

Warburton, P.E., Willard, H.F., 1996a. Evolution of centromeric alpha satellite DNA: molecular organization within and between human and primate chromosomes. In: Jackson, M., Strachan, T., Dover, G. (Eds.), Human Genome Evolution. BIOS Scientific, Oxford, pp. 121–145.

Willard, H.F., 1985. Chromosome-specific organization of human alpha satellite DNA. Am. J. Hum. Genet. 37, 524–532.

Willard, H.F., Waye, J.S., 1987. Hierarchical order in chromosome-specific human alpha satellite DNA. Trends Genet. 3, 192–198.

Wray, G.A., Babbitt, C.C., 2008. Enhancing gene regulation. Science 321, 1300–1301.