Global Repeat Map Method for Higher Order Repeat Alpha Satellites in Human and Chimpanzee Genomes (Build 37.2 Assembly)

Glunčić, Matko; Rosandić, Marija; Jelovina, Denis; Dekanić, Krešimir; Vlahović, Ines; Paar, Vladimir

Source / Izvornik: **Croatica Chemica Acta, 2012, 85, 327 - 351**

Journal article, Published version
Rad u časopisu, Objavljena verzija rada (izdavačev PDF)

https://doi.org/10.5562/cca1987

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:217820

CROATICA CHEMICA ACTA
CCACAA, ISSN 0011-1643, e-ISSN 1334-417X
Croat. Chem. Acta **85** (3) (2012) 327–351.
http://dx.doi.org/10.5562/cca1987

Original Scientific Article

Global Repeat Map Method for Higher Order Repeat Alpha Satellites in Human and Chimpanzee Genomes (Build 37.2 Assembly)

Matko Glunčić, Marija Rosandić, Denis Jelovina, Krešimir Dekanić, Ines Vlahović, and Vladimir Paar*

Faculty of Science, University of Zagreb, Zagreb, Croatia

RECEIVED OCTOBER 12 2011; REVISED SEPTEMBER 20, 2012; ACCEPTED SEPTEMBER 25, 2012

Abstract. Alpha satellites are tandemly repeated sequences found in all human centromeres. In addition to the functional and structural role within centromere they are also a suitable model for evolutionary studies, because of being subject to concerted evolution. The Global Repeat Map (GRM) algorithm is a convenient computational tool to determine consensus repeat units and their exact size within a given genomic sequence, both of monomeric and higher-order (HOR) type. Using GRM, we identify in Build 37.2 assembly fifteen different alpha satellite HORs, three of them novel, not reported previously. In the next step we compute suprachromosomal family classification and CENP-B box / pJα distributions for these HORs. All human alpha satellite sequences originate from one pra-ancestral alpha satellite monomer. For the first time we perform GRM analysis and compare human and chimpanzee alpha satellite HORs for chromosomes 4 and give an evidence that the human and chimpanzee alpha satellites originate from a common ancestor that predated the human-chimpanzee separation. We also compare the codon-like trinucleotide (CLT) extensions of human and chimpanzee chromosome 4. Our results are consistent with the expectation that the alpha satellite HORs in human and chimpanzee have been created after the human-chimpanzee separation. (doi: 10.5562/cca1987)

Keywords: alpha satellites higher order repeats, human chromosome 4, chimpanzee chromosome 4, GRM, trinucleotide extensions

INTRODUCTION

Centromeres in all eukaryotes play an essential role in many of chromosome functions, such as segregation in mitosis and meiosis, recognition and pairing of homologous chromosomes, sister chromatid attachment, and formation of kinetochore structures.[1] They are characterized by highly repetitive DNA regions and bound kinetochore proteins, which are required for the attachment of microtubules to chromosomes during mitosis.

Every human centromere consists of arrays of tandemly repeated 171-bp units, known as alpha satellite DNA that can be several megabases in size;[2] however, among reported chromosome assemblies, the amount and type of alpha satellite varies. These massive arrays are embedded between blocks of pericentromeric heterochromatin containing highly repetitive DNA.[3,4] In situ hybridization with alpha satellite and immunolabeling using antibodies against kinetochore proteins also confirms that centromeres are located in these regions.[5]

Considering the pattern of sequence organization, there are two major types of alpha satellite DNA in human genome: higher-order (HOR) and monomeric.[6–8] Figure 1 schematically shows the overall concept of HORs for an illustrative case of 11mer HOR. Higher-order alpha satellite DNA consists of ≈171-bp monomers organized in the second order arrays of monomeric repeat units that are highly homogenous. After a specific higher-order alpha satellite DNA has been created by amplification, its copies do not just passively accumulate mutations. There is a mechanism that works only

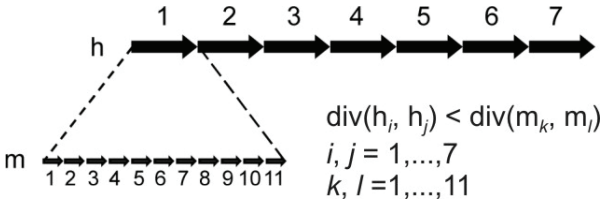

Figure 1. Schematic presentation of seven copies of 11mer alpha satellite HOR. The *i*-th HOR copy is denoted by h_i and the *k*-th constituent alpha satellite monomer by m_k.

* Author to whom correspondence should be addressed. (E-mail: paar@hazu.hr)

within an array to maintain its homogeneity. Owing to that process called "homogenization",[6] all chromosome-specific arrays have the same typical percentage of divergence between HOR copies (1−5 %).[8] By contrast, monomeric alpha satellite DNA lacks detectable higher-order periodicity, and its constituent monomers are far less homogeneous (individual alphoid monomers diverge by 20−40 % from each other).[9]

Higher-order alpha satellite DNA are chromosome specific.[2,7,10−15] every chromosome has its own unique family of higher-order alpha satellite. At least 33 different alphoid subfamilies have been identified so far. Some of these subfamilies are specific for a single chromosome, whereas others are common to a few chromosomes. Certain chromosomes seem to have a single HOR within their centromeres, whereas others contain several different HORs. A type of polymorphism found in alphoid arrays involves higher-order units that differ by an integral number of monomers (monomer insertion or deletion), but nonetheless closely related in sequence.[7,14]

Highly homogeneous arrays of higher-order alpha satellite monomers are relatively recent additions to human genome.[7,8] It was found that the lower primates have only monomeric alpha satellites at their centromere.[16−18] The relatively recent evolution of higher-order alpha satellite DNA and the fact that highly homogeneous arrays of higher-order alpha satellite monomers are allways bordered by more heterogeneous monomeric alpha satellite DNA,[9,19−22] has led to the hypothesis that higher-order alpha satellite DNA evolved from ancestral arrays of monomeric alpha satellites DNA and subsequently transposed to the centromeric regions of all great apes chromosome.[7,8,20,23−25]

In addition to their different sequence organization, higher-order and monomeric alpha satellite DNA also differ in their functionality. On the basic of genomic, biochemical and artificial chromosome analyses it was shown that the centromere function is associated with higher-order and not monomeric alpha satellite DNA in the human genome.[20,26−30] Because of this direct connection with centromere function, aforesaid recent evolution of higher-order alpha satellite DNA raises some intriguing questions.

An explanation for generating higher-order alpha satellite DNA involves unequal crossing over between misaligned HOR units aligned on the register of homologous monomers. Unequal crossing over, restricted to tandem sequences, explains the generation and local homogenization of higher-order units and accounts for large size variation among higher-order alpha satellite DNA on homologous chromosomes.[2,7,25,31−33] By the process of unequal crossing over higher-order alpha satellite DNA enable rapid evolutionary development.

A possible functional role of noncoding sequences and in particular of repeats has been much discussed. Recent studies have indicated a relatively sharp transition between the eucromatin of chromosome arms and the region containing alpha satellites near the centromere,[20,25] raising the possibility that some genes are located close to alpha satellites.[34] Higher-order repeats are in particular interesting since they are, as we mentioned above, due to more recent evolution and by the process of unequal crossing over enable a rapid evolutionary process. Additionally, there is a general concept that the regulatory system of genomes is encoded in the networks of repetitive sequence relationships.[35−38] It was postulated that the chromosomal regions in man, that are gene-poor, harbor gene regulatory elements that have the ability to modulate gene expression even over very long distances.[39] A functional importance of repetitive elements has been considered, suggesting that repetitive components play a major architectonic role in higher order physical structuring. It was argued that a fruitful interpretation of sequence data may result from thinking about genomes as information storage systems with parallels to electronic information storage systems. From this informatics perspective, repetitive DNA is an essential component of genomes; it is required for formatting coding information so that it can be accurately expressed and for formatting DNA molecules for transmission to new generations of cells, and that the cooperative nature of protein-DNA interactions provides another fundamental reason why repeated sequence elements are essential to format genomic DNA.[40] This was accompanied by observation that tandem arrays are often the regions that vary most between related taxa.[40] Considering these facts, it is reasonable, in addition to already known functions, to assume a possible role of different alpha satellite structures as components in gene expression multi-layered regulatory network.

Alpha satellite DNA in great apes were previously studied, for example in Refs. 41−46.

Higher-order and monomeric alpha satellites have been recently studied by computational analysis of the most recent builds of human genome assembly. Despite their obvious functional significance, centromeric regions and their constituent alpha satellite sequences were largely omitted by the Human Genome Project because of their repetitive nature and the expected paucity of genes.[22,25,47] In fact, due to centromere gap, located at the edges of p and q arms,[20] many of higher-order alpha satellite DNA regions are missing in NCBI human genome assembly. Nevertheless, although recent genome assemblies mostly provides alpha satellite content near the centromeric gaps, genomic assemblies of some chromosomes have reached a centromere region and in these cases detailed information on higher-order alpha satellite structure, dynamics and possible new functions can be obtained.

Various computational tools have been developed for computational analyses of repetitions in a given genomic sequence, with a goal to achieve a compromise between efficiency and sensitivity requirements. However, there still remain challenges in the case of large scale and/or significantly distorted repetitions. In particular, for higher-order alpha satellites the difficulties are largely due to imperfect patterns containing substitutions, insertions and deletions.

Analysis of the NCBI assembly was performed recently using two different computational approaches. Rudd and Willard[22] have used standard computational tools. Monomers of alpha satellites were extracted using Repeat-Masker and characterized as monomeric or higher-order using dot matrix program DOTTER. Percent identity among monomeric alpha satellite monomers and among higher-order alpha satellites was examined using CLUSTALW. BLAST alignments of all known HORs reported in the literature versus all alpha satellite in the July 2003 assembly was performed in Ref. 22 revealing that many of higher-order alpha satellites reported in the literature were missing in the genome assembly.

Having in mind possibly important information regarding the evolutionary and functional role of human higher-order alpha satellite DNA and a demanding task of studying bioinformatically this higher-order units, we perform here an extensive study applying novel robust bioinformatics tools Global Repeat Map (GRM)[48–54] (see Methods). We investigate the major alpha satellite higher-order repeats from Build 37.2 assembly of all human chromosomes and determine detailed monomer scheme and consensus sequences, finding three novel higher-order alpha satellite structures, not reported previously. Furthermore, we identify and analyze alpha satellite HOR from chimpanzee chromosome 4 centromere and analyze higher-order, monomer, and base-to-base divergences in human and chimpanzee homologous chromosomes. We find that the human and chimpanzee HORs are widely different, both in size and composition of HOR units and in the constituting monomer structure. To analyze differences in possible regulatory elements in human and chimpanzee higher-order alpha satellite consensuses we apply here our new method of stop/start codon like trinucleotide extensions.[55]

METHODS

Key String Algorithm (KSA)

In spite of powerful standard computational tools in bioinformatics, there are still difficulties to identify and analyze long repeat units. For example, the Tandem Repeat Finder can identify tandem repeat units up to 2 kb.[56,57] Here we use a new approach useful in particular for investigations of very long and/or complex repeats.

The KSA framework[48,49,52,53] is based on the use of a short sequence of nucleotides, referred to as key string, which cuts a given genomic sequence at each location of the key string appearing within the sequence. Going along genomic sequence, the lengths of ensuing KSA fragments form KSA length array. The length array could be compared to an array of lengths of restriction fragments resulting from hypothetical complete digestion cutting genomic sequence at recognition sites corresponding to KSA key string. While restriction enzymes cleave double stranded DNA selectively at specific palindrome sequences, in KSA we have no limitations on the choice of computational key string cutting a given genomic sequence. Periodicities appearing in KSA length array enable identification and location of repeats in genomic sequence. Analysis of repeat sequences at positions of any periodicity in the KSA length array provides consensus repeat unit and divergence of repeat copies with respect to consensus. A presence of higher order periodicity in KSA length array reveals the presence of HOR and enables determination of consensus HOR repeat unit (secondary repeat unit) and divergence of HOR copies with respect to consensus.

Similarly, with a proper choice of key string, the KSA fragments a given tandem repeat into monomers, as for example cutting Alu sequence at two identical positions providing identification of Alu sequences, cuts a palindrome providing identification of large palindrome sequences and their substructure, and so on. KSA provides a straightforward ordering of KSA fragments, regardless of their size (from small fragments of a few bp to as large as tens of kilobasepairs). KSA provides high degree of robustness and requires only a modest scope of computations using a PC. Due to its robustness, KSA is effective even in cases of significant deletions, insertions and substitutions, providing detailed HOR annotation and structure, consensus sequence and exact consensus length in a given genomic sequence even if it is highly distorted, intertwined and riddled (segmentally fuzzy repeats). Using HOR consensus sequence, in the next step KSA computes finer characteristics, as for example the suprachromosomal family (SF) classification and CENP-B box / pJα distributions.

Global Repeat Map (GRM)

The GRM program is an extension of KSA framework, executed as follows.

Step 1. GRM-Total module: Computes the frequency vs. fragment length distribution for a given genomic sequence by superposing results of consecutive KSA segmentations computed for ensemble of all 8-bp key strings (4^8 = 65536 key strings).[49] Figures 2a and 2b

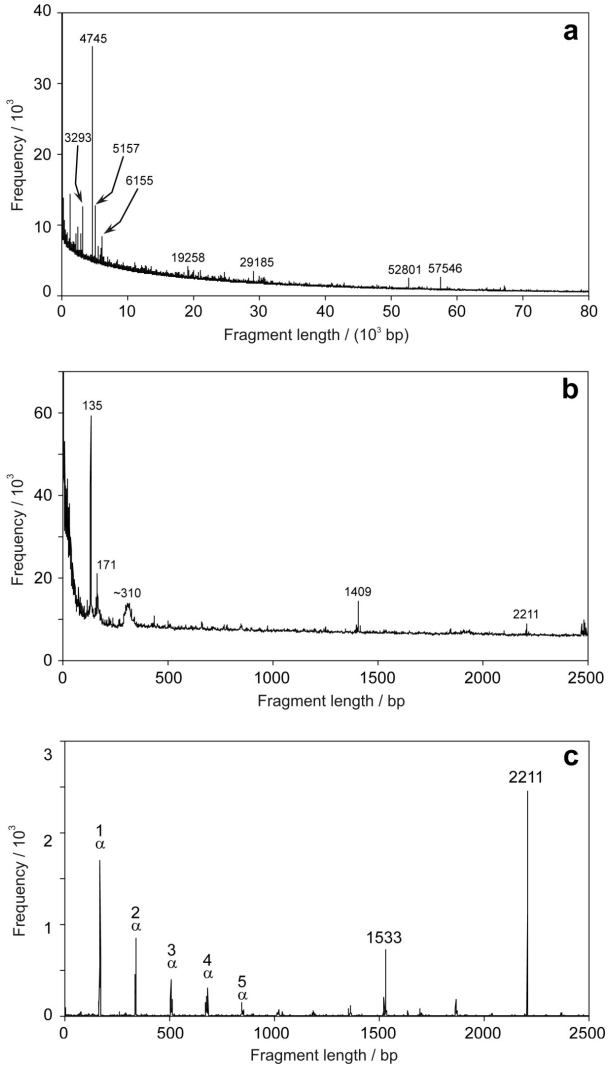

Figure 2. GRM diagram for Build 37.2 genomic assembly of human chromosome 4 for intervals of fragment lengths: **a** 0–80000 bp. Pronounced peaks above 2 kb are denoted by the corresponding fragment lengths. The most pronounced peaks above 2 kb are at approximately 3293, 4745, 5157, 6155, 19258, 29185, 52801 and 57546 bp. **b** 0–2500 bp. The most pronounced peaks are at approximately 135, 171, 310, 1409, and 2211 bp. **c** contig NT_022853.15 containing alphoid HOR in chromosome 4. There is a pronounced tandem array with alphoid repeat units of 171 bp. The peaks at multiples of alphoid monomer repeat unit 171 bp, i.e., n·171 bp, are denoted by nα. For description of peaks see the text.

show GRM diagrams for genomic sequence of human chromosome 4 (NCBI Build 37.2). In a GRM diagram each pronounced peak corresponds to one or more repeats at that length, tandem or dispersed.

Step 2. GRM-Dom module: Determines dominant key string corresponding to fragment length for each peak in the GRM diagram from step 1. An 8-bp key string (or a group of 8-bp key strings) that gives the largest frequency for a fragment length under consideration is referred to as a dominant key string.

Step 3. GRM-Seg module: Performs segmentation of a given genomic sequence into KSA fragments using dominant key string from the step 2. Any periodic segment within the KSA length array reveals the location of repeats and provides genomic sequences of the corresponding repeat copies.

Step 4. GRM-Cons module: Aligns all sequences of repeat copies from step 3 and constructs consensus sequence.

Step 5. NW module: Computes divergence between each repeat copy from step 3 and consensus sequence from step 4 using Needleman-Wunsch[58] algorithm.

Code for GRM modules is available upon request to the authors.

Regarding the 8-bp choice of the key string size: using an ensemble of all r-bp key strings the average length of KSA fragments is $\approx 4^r$. With increasing length of key strings the overall frequency of large fragment lengths increases. For an ensemble of all 8-bp key strings, from computed GRM diagrams we can identify the primary and secondary repeat units as large as hundred kilobases.

The GRM method is a straightforward method to provide a global repeat map in a single diagram, identifying all pronounced repeats in a given sequence, without any prior knowledge of the sequence structure. Once the size of a repeat is determined, GRM provides in a straightforward way location of the corresponding repeat arrays and their precise analysis. GRM is particularly useful for precise sequence analysis since the method does not involve averaging procedure. It is also useful that the method is robust with respect to sizeable substitutions and indels. Once the consensus repeat unit is determined using GRM, in the next step it could be well combined with BLAST for search of dispersed units or their fragments. For very large repeat units Tandem Repeat Finder has limitations, why GRM has no such size limitations. On the other hand, Tandem Repeat Finder may be more effective for short sequences.

RESULTS AND DISCUSSION

Global Repeat Map for Human Chromosome 4

As an illustration of GRM study of higher-order and monomeric alpha satellite arrays in genomic sequence, we compute here the GRM diagram for genomic sequence of chromosome 4 (Figures 2a and 2b). The most pronounced peaks in this diagram correspond to the following tandem repeats in chromosome 4: alpha satellite repeats (GRM peaks at multiples of the ≈171 bp repeat unit), GRM peaks at 135 bp, 166 bp, and ≈310 bp which are signature of Alu sequences, GRM peak at

Table 1. Alpha satellite monomer repeat structure of 13mer HORs in Contig NT_022853.15 in human chromosome 4. Monomers are denoted by m01, m02, ..., m13. Divergence is expressed with respect to the corresponding monomer sequences in consensus 13mer HOR

HOR Copy No.	Position / bp	Divergence / %	Composition
1	2591	1.0	m01, ..., m13
2	4802	0.7	m01, ..., m13
3	7013	0.5	m01, ..., m13
4	9223	0.7	m01, ..., m13
5	11433	0.4	m01, ..., m13
6	13642	0.4	m01, ..., m13
7	15852	0.2	m01, ..., m13
8	17385	0.8	m01, ..., m06,m11, m12, m13
9	19429	0.9	m01, ..., m06,m11, m13
10	20440	0.4	m08,...m13
11	21628	8.0	m01,m02,m07, ...,m11
12	22979	6.9	m07, ..., m10,m11,m11,m12,m13

1409 bp and GRM peaks at 2210 bp (also multiple of ≈171 bp repeat unit and possible higher-order alpha satellite). In addition, there are eight pronounced GRM peaks at repeat lengths above 2500 bp.

Higher-order Alpha Satellite Repeats in Human Chromosome 4

In the next step we perform detailed study for alpha satellite HORs. Analyzing partial contributions to GRM diagram of chromosome 4 from individual contigs we find that the largest frequencies contributing to alpha satellite peaks are arising from contig NT_022853.15. The relevant GRM interval of fragment lengths for genomic sequence NT_022853.15 is shown in Figure 2c. Peaks at approximate multiples of basic alpha satellite repeat length ≈171 bp are decreasing with increasing multiple orders. That is a natural trend for tandem repeats. On top of that multiple pattern there is a strong peak at 2211 bp corresponding to the consensus HOR length. Actually, the peak at 2210-bp reveals higher-order structure of alpha satellite organization: thirteen (2211 bp / 171 bp ≈ 13) tandemly arranged alpha satellite monomers, which mutually diverge by 20−40 %, are arranged into more homogenous second order units. The high homogeneity of second order units, as well as relatively high heterogeneity of primary repeat units, are reflected in GRM diagram (Figure 2c) with a characteristic pattern of HOR-signature. At the end of decreasing array of primary repeat peaks there is a pronounced peak which corresponds to higher periodicity. Furthermore, in GRM diagram of contig NT_022853.15 there is one pronounced peak at the fragment length 1553 bp (≈ 9α) which disturbs HOR-signature pattern. This peak arises due to deletion of four monomers (m07, m08, m09, and m10) in HOR copy No 8 (Tables 1 and 2).

Next, we determine computationally a dominant key string, TTTG, which maximally segments the NT_022853.15 sequence into ≈171-bp fragments. Performing KSA segmentation using this dominant key string we obtain an array of ≈171-bp fragments. Mutual alignment of all, in this way obtained, ≈171-bp monomers (see heatmap, Figure 3) approved above deduction; thirteen different monomers are constituent blocks of higher-order structure (13mer alpha satellite HOR). The corresponding basic consensus monomers are denoted as m01, …, m13 (consensus sequences in Table 3).

We also compute GRM diagram for each of thirteen alpha satellite basic monomers and we find no pronounced peak. This reflects their monomeric structure, *i.e.*, the absence of internal repeat structure. In the next step, divergences between each repeat copy and HOR consensus monomers are computed, revealing internal structure of each alpha satellite HOR copy (Table 2). Detailed monomer structure of alpha satellite HOR copies in contig NT_022853.15 is summarized in Table 1.

Position of alpha satellite HOR copies in chromosome 4 (see ideogram on Figure 4) and heatmap in Figure 3 reveal that the 13mer HOR, in fact, seems to be a truncated tail of a major HOR block positioned in unsequenced domain in front of the contig NT_022853.15.

From results in Tables 1−3 and Figure 3 it is obvious that homogenization works better near the center of higher-order repeat arrays, and less well at the array of HOR edges, bordering some non-related sequences.[8] Our results (Table 1) are in accordance with publica-

Table 2. Divergence (%) between each repeat copy and monomer sequences in human consensus 13mer HOR. Divergences less then 10 % are colored

Contih Position / bp	44	215	386	551	722	894	1065	1236	1407	1578	1747	1917	2086	2254	2425	2591	2761	2932	3104	3275	3446	3617	3788	3957	4127	4297	4465	4636	4802	4972	5143	5315	5486	5657	5828	5999	6168	6338	6508	6676	6847	7013	7183	7354	7526
m01	23	17	29	1	15	12	19	12	20	24	23	27	26	23	29	17	0	15	12	19	12	19	23	22	26	26	24	29	17	0	15	12	19	11	19	24	22	26	26	23	27	18	1	15	12
m02	25	15	32	14	1	18	21	20	24	27	25	33	27	26	32	15	14	1	18	21	19	23	27	25	33	27	26	32	15	14	1	18	21	19	23	27	25	33	27	25	30	16	15	1	18
m03	27	20	32	12	17	0	21	19	25	27	29	30	31	27	32	20	12	19	0	21	19	25	27	28	30	31	27	32	20	12	19	0	21	19	25	27	28	30	31	27	30	22	13	19	0
m04	26	21	33	19	22	21	0	25	17	27	29	31	29	26	33	21	19	21	21	0	25	16	26	28	31	27	27	32	21	19	21	21	0	25	16	27	28	31	27	26	31	22	19	21	21
m05	26	20	30	12	19	19	25	1	24	26	26	31	29	27	31	20	11	19	19	25	1	24	25	26	30	29	27	31	20	11	19	19	25	0	24	26	26	30	29	26	30	21	12	19	19
m06	24	23	30	19	24	24	17	24	1	24	27	30	26	25	30	23	19	24	24	17	24	2	24	26	30	25	25	30	23	19	24	24	17	24	2	24	26	30	25	24	28	23	20	24	24
m07	7	28	15	24	27	27	26	25	24	1	15	21	14	8	15	28	23	27	27	26	25	25	0	14	20	13	8	15	28	23	27	27	26	25	25	1	14	20	13	7	13	29	24	27	27
m08	13	27	18	22	25	28	27	27	27	13	3	22	19	14	19	27	22	25	28	27	27	26	12	1	21	18	15	20	27	22	25	28	27	26	26	13	1	21	18	13	18	28	22	25	28
m09	20	30	23	26	33	29	30	29	30	20	22	2	23	21	22	30	26	33	29	30	30	30	20	22	1	21	21	22	30	26	33	29	30	30	30	20	22	1	21	20	20	31	26	33	29
m10	13	30	19	26	28	31	27	29	26	14	20	22	4	14	18	29	26	28	31	27	29	26	13	19	21	0	15	19	29	26	28	31	27	29	26	14	19	21	0	13	16	30	27	28	31
m11	0	27	13	23	25	27	26	26	25	8	16	21	13	1	13	27	23	25	27	26	27	25	7	15	21	13	1	14	27	23	25	27	26	26	25	8	15	21	13	0	12	28	23	25	27
m12	12	30	3	27	30	30	31	29	29	13	18	22	19	12	2	30	27	30	30	31	30	29	13	19	21	16	13	2	30	27	30	30	31	30	29	13	19	21	16	12	0	32	27	30	30
m13	27	1	32	16	15	20	21	21	23	27	28	30	28	27	32	0	17	15	20	21	21	23	28	28	31	29	26	32	0	17	15	20	21	20	23	28	28	31	29	27	30	1	16	15	20
Min Div / %	0	1	3	1	1	0	0	1	1	1	3	2	4	1	2	0	0	1	0	0	1	2	0	1	1	0	1	2	0	0	1	0	0	0	2	1	1	1	0	0	0	1	1	1	0
Min Div Monomer	11	13	12	1	2	3	4	5	6	7	8	9	10	11	12	13	1	2	3	4	5	6	7	8	9	10	11	12	13	1	2	3	4	5	6	7	8	9	10	11	12	13	1	2	3

Contih Position / bp	7697	7868	8039	8210	8379	8549	8719	8887	9058	9223	9393	9564	9736	9907	10078	10249	10419	10588	10758	10928	11096	11267	11433	11603	11774	11946	12117	12288	12458	12628	12797	12967	13137	13305	13476	13642	13812	13983	14155	14326	14497	14668	14838	15007	15177
m01	20	11	19	23	21	26	25	23	27	19	1	14	12	19	11	19	23	22	26	26	23	28	17	0	14	12	19	12	19	24	22	25	25	23	27	17	1	14	12	19	11	19	23	22	26
m02	22	19	24	27	24	32	26	26	30	16	15	0	18	21	19	23	27	25	32	27	25	30	14	14	0	18	21	20	23	27	25	31	26	25	30	15	14	0	18	21	19	23	27	25	32
m03	22	19	24	27	27	29	30	27	30	22	13	18	1	21	19	24	27	28	29	31	27	30	20	12	18	0	21	19	24	27	28	28	30	27	30	21	13	18	0	21	19	24	27	28	29
m04	1	25	16	26	27	30	26	27	32	23	20	21	21	0	25	17	26	27	30	27	26	31	21	19	21	21	0	25	17	27	27	30	26	26	31	22	20	21	21	0	25	17	26	27	30
m05	25	0	24	25	26	30	29	26	30	22	12	19	19	25	0	23	25	26	30	29	26	30	20	11	19	19	25	1	23	26	26	30	29	26	30	21	11	19	19	25	0	23	25	26	30
m06	17	24	1	24	26	29	24	25	28	23	20	24	24	17	24	2	24	26	29	25	24	29	23	19	24	24	17	24	2	24	26	29	24	24	28	23	20	24	24	17	24	2	24	26	29
m07	26	25	24	0	12	20	12	8	14	30	24	26	27	26	25	23	0	12	20	13	7	13	27	23	26	27	26	26	23	1	12	19	12	7	13	28	24	26	27	26	25	23	0	12	20
m08	27	26	26	12	1	21	16	14	18	29	22	25	27	27	26	26	12	0	21	18	13	18	27	22	25	28	27	27	26	13	0	20	16	13	18	28	22	25	28	27	26	26	12	0	21
m09	30	30	29	20	20	1	21	20	22	32	26	32	28	30	30	29	20	21	0	21	20	21	29	26	32	29	30	30	29	20	21	1	21	20	20	31	26	32	29	30	30	29	20	21	0
m10	27	29	25	13	17	22	1	14	18	31	27	27	31	27	29	25	13	18	21	0	13	17	28	26	27	31	27	30	25	14	18	21	1	13	16	29	27	27	31	27	29	25	13	18	21
m11	25	26	25	7	13	20	12	1	13	29	23	25	27	26	26	24	7	13	20	13	0	12	26	23	25	27	26	27	24	8	13	20	12	0	12	27	23	25	27	26	26	24	7	13	20
m12	30	30	29	13	17	22	16	12	1	32	27	30	30	31	30	28	13	18	20	16	12	1	30	27	30	30	31	30	28	13	18	21	16	12	0	30	27	30	30	31	30	28	13	18	20
m13	21	20	23	28	27	31	28	27	31	2	16	15	20	21	20	22	28	27	30	29	27	31	1	17	15	20	21	20	22	28	27	30	28	27	30	1	17	15	20	21	20	22	28	27	30
Min Div / %	1	0	1	0	1	1	1	1	1	2	1	0	1	0	0	2	0	0	0	0	0	1	1	0	0	0	0	1	2	1	0	1	1	0	0	1	1	0	0	0	0	2	0	0	0
Min Div Monomer	4	5	6	7	8	9	10	11	12	13	1	2	3	4	5	6	7	8	9	10	11	12	13	1	2	3	4	5	6	7	8	9	10	11	12	13	1	2	3	4	5	6	7	8	9

Contih Position / bp	15347	15515	15686	15852	16022	16193	16365	16536	16707	16878	17049	17219	17385	17555	17726	17898	18069	18240	18411	18581	18750	18920	19090	19258	19429	19595	19765	19935	20103	20274	20440	20610	20781	20953	21121	21290	21460	21628	21797	21968	22137	22307	22475	22644	22813	22979
m01	26	23	27	17	1	14	12	19	11	19	25	28	17	0	14	12	19	11	19	23	23	26	26	23	26	22	25	26	23	27	19	0	18	26	24	24	28	24	23	24	24	25	23	23	24	18
m02	27	25	30	15	14	0	18	21	19	24	27	30	15	14	0	18	21	19	23	27	25	32	27	25	28	25	31	27	25	30	16	14	13	28	27	29	30	26	26	27	29	27	27	24	26	19
m03	31	27	30	21	13	18	0	21	19	26	28	30	21	12	18	0	21	19	24	27	29	29	31	27	29	28	28	31	27	30	22	12	23	29	28	27	31	27	27	30	26	28	26	26	26	23
m04	27	26	31	22	20	21	21	0	25	17	28	31	22	19	21	21	0	25	17	26	29	30	27	26	29	27	30	27	26	31	23	19	23	25	29	29	29	25	25	30	28	27	24	26	26	21
m05	29	26	30	21	11	19	19	25	0	24	26	30	21	11	19	19	25	0	23	25	26	30	29	26	27	26	30	29	26	30	22	11	22	26	28	28	30	28	26	28	28	29	27	26	29	23
m06	25	24	28	23	20	24	24	17	24	2	25	29	23	19	24	24	17	24	2	24	26	29	25	24	25	26	29	25	24	28	24	19	26	26	28	29	28	25	24	30	28	26	24	24	25	23
m07	13	7	13	28	24	26	27	26	25	26	4	13	28	23	26	27	26	25	23	0	14	20	13	7	8	12	19	13	7	13	29	23	29	12	9	17	15	7	6	14	16	13	8	8	12	30
m08	18	13	18	28	22	25	28	27	26	26	15	18	28	22	25	28	27	26	26	12	2	21	18	13	15	0	20	18	13	18	28	22	28	14	8	18	19	13	12	7	17	17	14	12	16	30
m09	21	20	20	31	26	32	29	30	30	30	21	21	31	26	32	29	30	30	29	20	22	0	21	20	23	21	1	21	20	20	32	26	27	21	19	6	20	19	18	20	6	19	20	18	18	32
m10	0	13	16	29	27	27	31	27	29	26	15	17	29	26	27	31	27	29	25	13	19	21	0	13	15	18	21	0	13	16	30	26	29	14	18	20	10	12	13	19	19	8	13	10	16	31
m11	13	0	12	27	23	25	27	26	26	26	4	12	27	23	25	27	26	26	24	7	15	20	13	0	8	13	20	13	0	12	28	23	27	14	12	17	15	9	5	15	17	13	10	7	13	30
m12	16	12	0	30	27	30	30	31	30	30	13	1	30	27	30	30	31	30	28	13	18	20	16	12	7	18	21	16	12	0	32	27	28	20	14	19	17	15	10	17	18	15	16	14	7	33
m13	29	27	30	1	17	15	20	21	20	23	28	31	1	17	15	20	21	20	22	28	27	30	29	27	30	27	30	29	27	30	2	17	17	28	28	29	27	27	28	28	29	26	27	28	28	6
Min Div / %	0	0	0	1	1	0	0	0	0	2	4	1	1	0	0	0	0	0	2	0	2	0	0	0	7	0	1	0	0	0	2	0	13	12	8	6	10	7	5	7	6	8	8	7	7	6
Min Div Monomer	10	11	12	13	1	2	3	4	5	6	11	12	13	1	2	3	4	5	6	7	8	9	10	11	12	8	9	10	11	12	13	1	2	7	8	9	10	11	7	8	9	10	7	11	12	13

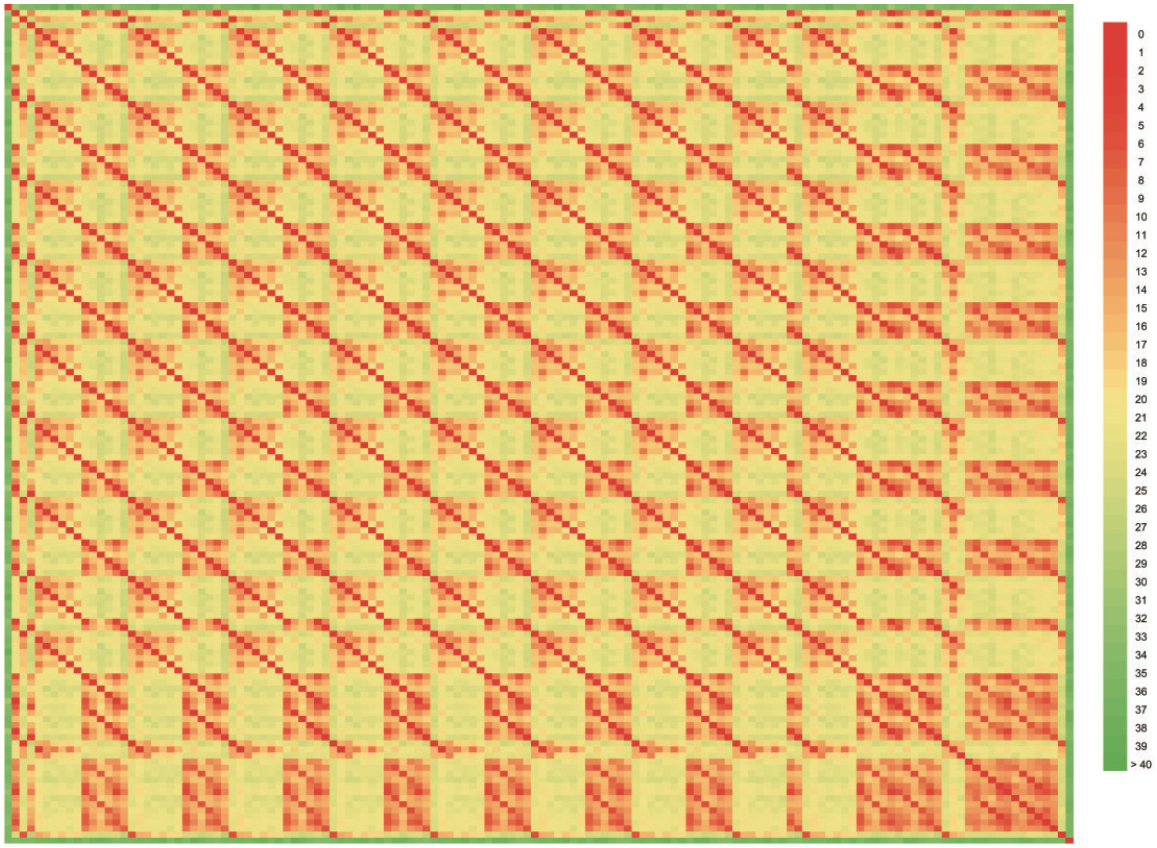

Figure 3. Percent divergence scores for base to base comparison of alpha satellite monomers from contig NT_022853.15. Percent divergence scores are colored according to the color scale shown on the right.

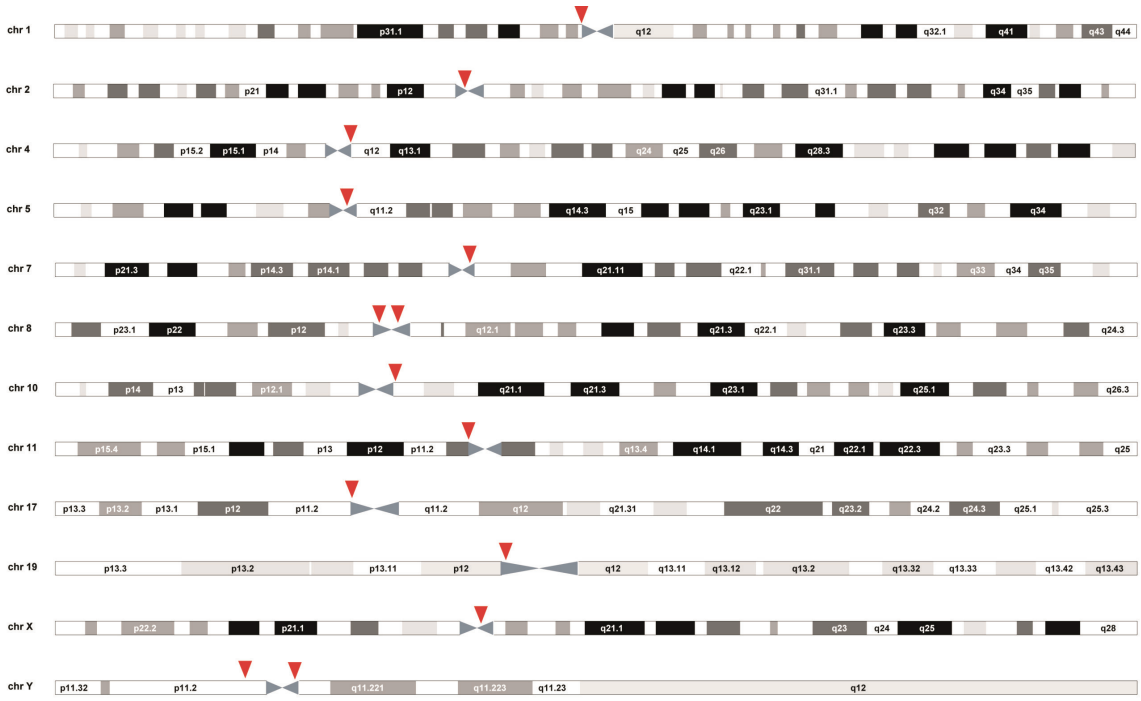

Figure 4. Human chromosomes ideogram with denoted positions of alpha satellite HORs investigated in this paper.

Table 3. Consensus sequences of human alpha satellite HORs in chromosomes 2, 4, and 9 determined by the GRM analysis of Build 37.2 assembly. Constituent alpha satellite monomers are denoted by m01, m02, ...

chromosome 2 / 10mer D satellite HOR

```
m01 171 TCAGAAACCTCTTTGGGATGTTAGCGTTCGAGTCACAGAGTTGATCACTCCCTTTCATAGAGCAGGTTTGAAGCACTCTTTTTGTAGTATCTGGAAGTGGACGTTTTGATCGCTTTGAGGCGTAAGGTGAAAAAGGAAATATCTTGCCACAAAAACTACACAGAAGCATTC
m02 171 TCAGAAACTACGTTGTGATGTGTTTACTCAACTAACAGAGTTGAACCTTTCTTTTGATAGAGCAGTTTTGAAACACTCTTTTTGGAGAATCTGCAGGTGGATATTTGGATAGCTTTGAGGATTTCCTTGGAAAAGGAAATATCTTCATATAAAATCTAGACAGAAGCATTC
m03 171 GCAGAAACACCTTTGTGATGTTTGCATTGAAGTCAGAGAGTTGTACATTCCCTTTCATAGAGCAGCTTTGAAACACTCTTTTTGTAGTATCTGGAGATGGACATTTAGATCGCTTTGAGGCCTATGGTGAAATAGGAAATATCTTCGCATAAAAACTAGACGGAAGCAGTC
m04 171 TCCAAAACTTGCTTGGAATGTGTGTACTCAACTAACAGAGTTGAATCTTTCTTTTGATAGAGCAGTTTTGAAACACTCTTTTTGTAGAGTCTGCAAGTGGATATTTGGATAGCTTAGAGGATTTCGTTGGAAACGGGAATATGTCCATACAAAACCTAGACAGAAGCATTC
m05 171 TCAGAAAAATCTCTGTGAGGATTGCATTCAAGTCCCAGTGTTGAACATTCCCTTTCATAAAGCAGGTGTGAACACAAGATTTTGTAGTATATGGAACTGGACATTTGGAGTGCTTTGTGACCTATTGTGAAAAAGGAAATATCTTCCCATACAAAACCTAGACAGAAGCATTC
m06 170 TCAGAAACCAGTTTGTGATGTGTGACTCAACTAACAGGGTTGAACCTTTCTTTTGAGAGAGCACTCTTGAAACACTCTTTTTGTAGACTCTGCAAGGGGATATTTGGACAGCTTTGAGGATTTCGTTGGAAACGGGATATCTTCATATAAAATCTCGACAGAAGCATCC
m07 171 TCAGAAACATCTTTGGGATGTTTGCATTCAAGTCACAGAGTTGAACATTCCCTTTCATGGAGCAGGTTTGAAACACTCTTTTTGTGGAATCTGGAAGTGGACATTTGGATCGCATTGAGGCCTACGGTGAAAAAGGGAATATCTTCGAATAAAAACTAGACAGAAGCATTC
m08 171 TCATAAACTAGTTTGTGATGTGTGCTTAACTAACAGAGCTGAACCTTTCTTTTCATAGAGCGGTTTTGAAACACTCTTTTTGTAGAATCTGCATGTGGATATTTGGAAAGCTTTGAGGATTTCGTTGGAAACGGGAATATCTTCACATAAAATCTAGACAGAAGCATTC
m09 170 TCAGAAACGTCTTTGGGGTGTTAGCATTCAAGTCACAGAGTTGAACGTTCCTTTTCATAGAGCAGGTTTGAAACACTCTTTTTGTGGAATCTGGAAGTGGACATTTGGATCGCTTTGAGGCCTGCGGTGAAAAAGGTATATCTTCGCATAAAAACTAGACAGAAGTATTC
m10 171 TCATGAACTAGTTTGTGATGTGTGTGCTCAACTACCAGAGTTGAACCTTTCTTTTGATAGAGCAGTTTTGAAACACTCTTTTTGTAGAATTTGCATGTGGATATTTGGACAGCTTTGAGGATTTCGTTGGAAACGGGAAAATCTTCATATGAAATCGAGACACAAGCATTC
```

chromosome 4 / 13mer D satellite HOR

```
m01 171 TCAGAAACTGCTTTGGGATGTGTGCATTCAACTCACAGAGTTGAACGCTTTCTTTGATACAGCAGTTTAGAAGCATTCTTTTTGTAGAATCTGCAAGTGTTCATTTGGAGCGCTTTGAAGCCTACGGTGGAAAAGGAAATATCTTCACATAAAAGCTAGACAGAAGTATTC
m02 172 TCAGAAACTGCTTTGTGATGTGTCCATTGAACTCACAGAGTCCAACCTTTCTTTTGATAGAGCAGATTTGAGGCATTCTCTGTGCATAATCTGCAAGTGTTCCTTTGGAGCGCTTTGAAGGCTTATGGTGGAAAGCAAAATATCTTCACATAAAAACTAAACAGAAGCATTC
m03 171 TCAGAAACTGCTCTGTGTTGTGTTCATTGAAGTCAAAGGGTTGAACTCCGTCTTTGATAGAGCAGTTTTGAAACATTCTTTTTGTAGAATCTGCAAGTGTTCATTTGGAGCGCTTTGAAGCCTAGGGTGCAAAAGGAAATATCTTCACAGAAAAACCGGACAGAAGCATTC
m04 171 TCAGACAGTTCCTTGTGATGTGCGCATTGAACTCAAAGAGTTGAACCTTCCTTTTGAGAAAGCAGTTTTGAAACAGTGTTTTCTTACGATCTACAAGTGGATATTTGGAGCGATTTGAGGCCTATGGTGGAAAAGGAAATATCTTCGGATAAAAACTACACAGAAGCATTC
m05 171 TCAGAAACTGCTTTGGGATGTGTGCATTCAACTCACAGAGGTGAACGCTTTCTTTGATAGAGCAGTTGTGAAACATTCTTGTTGCAGTATCTGCAAGTGTTCTTTTGGAGCGCTTCGAAACCTGCGGTGGAAATGGTAATATCGTCACATAAAAACTAGACGGAAGTTTGC
m06 171 TCAGAAACTTTTTTGTGATGTCTGCATTCAACTCACAGAGTTGAACCTTCCTTTGGGAGAGCAGCTCTGAAACAGTGTTTTTGCAGTATCACCCAAGTGGATATTTGGAGCGATATGAGGCCTATGGTGGAAAAGGAGATACCTTCACATGGAAGCTACACAGAAGCATTC
m07 169 TCAGAAACTTTTTTGGGATGTGTGGATTCAACTCACAGAGTTCACCCTTCCTTTTGATAGAGCAGTCTTGAAATACTGCTTTTGTAGAATCTGCTTGAGGATATTTGGAGCTCTTGAGGAATTCGTTGTAAACGGGGTATCTGCACATACAAACTGGACGGAAGCATTC
m08 170 TCAGAGACTTTTTTGTGATGTGTGAACTCAACTCACAGAGTTCAACCTTTCTTTTGATAGAGCAGTTTTGCAGCACTGCTTTTGTAGAATCTGCCTGAGGACATTTGGAGCTCTTGAGGAATTCGTTGTAAACGGGATGTCCGCACATACAAACTAGACACAACTATTC
m09 170 TCTCAGAAAATTTTGTGTTGTGTGCATTCAAGTCACAGAGTTGAACGTCTTTTTGGATAGATCAGTTTTGAAACCATCCTTTTCTAGAATCAGCTTGAGGATATTTGTAGGTCTTTGAGGAATTCGTTGTAAACGGGATATCTGCACATACAAGGTAGACAGAAACATGG
m10 168 CCAGAAACTTTTTCGTGATGTGTGGATTCAACTCACAGAATTCAACCTTCCTTTTGATAGAGCATTCCTGAAACCCTGCTTGGGTAGAATCTGCTTGAGGATATTTGGAGCTCTTGGAGGAATTCGTGGCAAACGAGATATCTCCACATACAATCTAGACAGAAGCAT
m11 171 TCAGAAACTTTTTTTGGGATGTGTGGATTCAACTCACACAGTTCAACCTTCCTTTTGATAGAGCAGTCTTGAAATACTGCTTTTGAAGAATCTGCTTCAGGATATTTGGAGCTCTTGAGGAATTCGTTGCAATCGGGATATCTGCACATACAAATTAGACAGAAGCATTC
m12 166 CTCAGAAACATTTAGGGATGTGTGAACTCAACTCACAGAGTTCAACTTTCCTTTTGATAGAGCTGTCTTGAAATCCTGCTTTTGCAGAATCTGCTTGAGGATATTTGGAGCTCTTTCAGGAATTCGTTGCCAACGGGATATCTGCACAAAATTGACAGAAGCATTC
m13 170 TCAGACAGTTCTTTGTGATATGTCCATTCAACTCACAGAGTTGAATCTTTCCTTTGATAGAGAAGATTGGAAACATTCTTTCTGCATACTCTGCAAGTGTTCATTTGGAGCGCTTAGAGGCTCATGGTGGAAAGTGAAATATCTTCACATATAAACTAGACAGAAGCATT
```

chromosome 9 / 11mer D satellite HOR

```
m01 171 TGAGAAACTACTTTGTGATGTGTGCATTCATCTCACAGAGTTGAACCTTTCTTTTGATTGAGCAGTTTTGAAACACTCTTTTTGTTGTATATGCAAGTGGATATTTGGAGCGATTTGTGGTCTATGGTGGAAAAGGAAATATCTTCACGTAAAAACTAGACAGAAGCACTC
m02 171 TGAGAAACTTCTTTGTGATGTGTGCATTCATCTCACCAAGCGGAACCATTCTTTTGATGGAGCTGTTTTGAAATACTCTTTTTGTAGAATCTGCAAGTGAATATTTGGAGTGCTTTCAGGCCTGTGGTGGAAAAGGAAATATCTTCACATAAAAACTAGACAGAAGCATTC
m03 170 GGAGAAACTTCTTTTTAGTGTGTGCATTCATCTCACAGTGTTGAAACTTTCTTTTGATTGAGGGTTTTGAAACAGTCTTTTTGATAAATCTGCAAGTGGATATTTGGAGCGAATTGTGGCCTATGGTTTAAAAGGAAATATCTTCACATAAAAGCTAGACAGAAGCTTTC
m04 169 TGAGAAACTTCTTTGTGATATGTGCGTTCATCTCACCGGGTTGAATCTTTCCTTTCATCGAGCAATATTGAAACACCCTTTTTTTTGAATCTGAAATAGATATTTGGAGCGATTGTGTCCTATGGTAGTAAAGGCAATATTTTCACAGAAAAACTAGACAGAAGCATTC
m05 171 TGCAAAACTTCCTCGTGATGTGTTCATTTATCTCACCAAATTGAACCATTCTTTTCCTTGAGCAGATTTGATACACTCTTTTTGTAGAATGTGCAAGTGAATATTTGGAACGCTTTGATGAGTTTGACGGAAAAGGAAATACCTTCACATATAAACGAGACAGAAGCATTC
m06 170 TGAGAAACTTCTTTCTGATATGTGCATTCAACTCACAGAGTTGAACCTTTCTTTTGATTCAGCAGTTTTGAAACACACTTTTTGAAGGATCTGTAAGTGGATATTTGGAGTGCTTAGGGGTCGATGCTAGAAAGGAAATATCTCCACATAAAAACTTGATGGAAGCATTC
m07 171 TGAGAAACTTCTTTGTGATGTGTGCATTCATCAGAGAGAGTTTAACATTTCTTTTGACTGAGCAGTTTTGAAACTCTCTTTTTGTAGAATCTGCAAGTGGACATTTGGAGCCACTTGAGGCCTATTGTGGAAAAGGGAATAAATTCACATAAAAGCTACACGGAAGCATTC
m08 169 TGACAAACTTCTTTGTGATGTGCACATTCATCTCACAGATTGAAAATTTCTTTTGATTGAGCAGTTTTGAAATGCTCTTTTCGGAGAATCAGCCAGTGGATATTTGGAGCACTTTGAGGACTATGGTGGAAAAGGAAATATCTTCACATAAAAACTAGAGGAAACATTC
m09 170 TGAGAAACTTATTTGTGATGTGTGCATTCCTCTCACAGAGTTGAACATTTCTTTTGATTAAGCAGTTTTGAAACACTCTTTTGTGGAATCTGCTAGAGGATATTTGGAGCGCTTTGAGGCCTATGGAGGAAAAGGAAATATCTTCACTTAAAAACTAGACAGAAGCATTC
m10 171 TGAGAAACTTCTTTGTGATGTTTGCATTCATCTCACAGGGTTGAAATTTTCTTTTGATTGAGAAGTTTTGAGACACTCTTTTTGTAGAATCTGCCTGTGGATAATAGGAGCGCTTTGGGGCATATTTTGGAAAAGGAAATACCTTCACATAAATAGTAGACAGAAGCATTC
m11 169 TGAGAAACTTCTTTGTGACGTGTGCATATATATCACAGAGTTGAACCTTTCTTTTCATTTAGCCTTTTGAAACACTCTTTTTCTAGAATCTGCAAGTGAATATTTGGAGCGCTTTGCGGCCTATGGTGGAAAATAAATATCTTCACATAAAAACTAGACAGAAGCAATC
```

chromosome 9 / 7mer D satellite HOR

```
m01 170 TCAGAAACTTCCTTGTGATGTGTGTACTCAAGTAACAGAGTTGAACCTTACTTTTGACAGAGCCGTTTTGAAACAGTCTTTTTGTAGAATCTGGAAGTAGATATTTGGATACCTTTGAGGATTTCTTTGGAAACGGGATATCTTCATATAAAATCTAGACAGAAGCATTC
m02 171 TCAGAAACTTCTCTGTGATGTTTGCATTCAACTCATAGAGTTGAACACTTCCCTTCATAGAGCAGGTTTGAAACACTCTTTTTGTAATATTTGGAAGTGGACATTTGCAGCGCTTTGAGGCCTATGTTGAAAAAGGAAATATCTTCTCCTAAAAACCAGACAGAAGCATTC
m03 171 TCAGAAACTTATTTGCGATGTGTGTTCTCAACTAACAGAGTTGAACCTTTGTTTTGATACAGCATTTTGGAAACACTCTTTTTGTAGAATCTGCAGGTGGATATTTGGATAGCTTTGAAGGTTTCGTTGGAAACGGGAATATCTTCATATAAAATCAAGACAGAAGCATTC
m04 171 TCAGGAACTTCTTTGTGATGTTTGCATTCAAGTCACAGAACTGAAACATTCCCTTTCATAGAGCATGTTTGAAACACTCTTTCTGTAGTATCTGCAAGCGGACGTTTCAAGCGCTTTCAGGCCTATGGTGAGAAAGGAAATATCTTCAAATAAAAACTAGACAGAAGCATTC
m05 167 TCAGAAACTTGTTTGTGATGTGTGTATTCAACTAACAGAGATGAACCTTTCTTTTTACAGAGCAGTTTTGAAACACTCTTTTTGTGGAATCTGAAAGTGGATATTTGGATAGCTTTGAGGATTTCGTTGGAAACGGGATTACATATAAAATCTAGAGAGAAGCATTC
m06 171 TCAGAAACTGCTTTGTGATGTTTTCATTCAAGTCACAGAGTAGAATGTTCCCTTTTATAGAGCAGGTTTGAAACACTCTTTCTGCACTACCTGGAAGTGGACATTTGGAGCGCTTTGAGGCCTATGTTGAAAAAGGAAATATCTTTCCCATAAAAACTAGACAGAAGCATTC
m07 171 TCAGAAACTTCTTTGTGCTGTATGTCCTCAATTAACAGAGTTGAACCTTTGTGTGGATACAGCATTTTGGAAACATTCCTTTAGTAGAATCTGCAAGTTGATATTTAGATAGCTAGGAAGATTTCCTTGGAAACGGGAATATCTTCATATAAAATCTAGACGGAAGCATTC
```

The consensus sequences of HORs in chromosomes 1, 5, 7, 8, 10, 11, 17, 19, X and Y are the same as given in Ref. 55. Consensus sequence for chromosome 9 (7mer) is reproduced here because in published sequence in Ref. 55 the last two nucleotides in each row were missing.

tions where it was shown that structural variants of HORs usually differ in length as a result of the presence or absence of an integral number of monomers. Warburton *et al.* in Ref. 59 have already described duplications of one monomer, as happens here for instance in HOR copy No 12, or deletions and duplications of a number of monomers within a HOR, as for instance in HOR copies No 8, 9, 10, 11 (Table 1). Generation of all these structural variants can be satisfactory explained by unequal crossover between two misaligned wild-type HORs or by non-reciprocal processes such as gene conversion or double strand gap repair.[8,59,60]

Global Repeat Maps and Higher-order Alpha Satellite Repeats for All Human Chromosomes

Using GRM algorithm we identify and analyze higher-order and/or monomeric alpha satellite units in all human chromosomes (Build 37.2 assembly). In the first step, we compute GRM diagrams for all human chromosomes for two relevant intervals of fragment lengths (Figures 5−8). (The same diagrams with the corresponding magnification ability are presented in Supplementary Figures 1 and 2).

We perform detailed study of alpha satellite HORs in every human chromosome in the same way as for chromosome 4 in the previous chapter. Summary of all human alpha satellite HORs corresponding to Build 37.2 assembly and positions of HOR blocks are given in Table 1 and Figure 4.

We have determined consensus HORs for chromosomes 1, 2, 4, 5, 7, 8, 10, 11, 17, 19, X, and Y (Build 37.2 assembly). Aligned monomers in consensus *n*mer HOR are denoted *m*01, *m*02, ... Arrays correspond to consensus HOR if monomer sequences correspond to the convention of[61] (referred to as direct (D) monomers). This is the case for 10mer in chromosome 2, 16mer in chromosome 7, 11mer in chromosome 8, 11mer in chromosome 9, 14mer in chromosome 17, and 17mer in chromosome 19. If the consensus HOR contains alpha monomers which are reverse complement to the convention of[61] (referred to as revers-complement (RC) monomers), then the array *m*01, *m*02, ... is reverse complement to consensus HOR; this is the case for 11mer in chromosome 1, 13mer in chromosome 4, 13mer in chromosome 5, 7mer in chromosome 9, 18mer in chromosome 10, 12mer in chromosome 11, 13mer in chromosome 19, 12mer in chromosome X and 45mer in chromosome Y. HOR consensus sequences are presented in Table 3 (chromosomes 2/10mer, 4/13mer, 9/11mer, and 9/7mer) and in Ref. 55 (chromosomes 1/11mer, 5/13mer, 7/16mer, 8/11mer, 10/18mer, 11/12mer, 17/14mer, 19/13mer, 19/17mer, X/12mer, and Y/45mer). For convenience, the consensus for the second group of HORs is also given in Supplementary Table 1.

Only two chromosome (8 and Y) assemblies have arrays of highly homogenous higher-order alpha satellite DNA both on p and q arms (Figure 4). Because all chromosomes are known to contain higher-order alpha satellites at centromeres,[7,8] the fact that only the chromosomes 8 and Y have this level of success indicates that most current assemblies probably terminate at some distance from functional centromere. In two cases with higher-order alpha satellite DNA both on p and q arms, the alpha satellite tandem repeats are oriented in the same direction on both arms (see Table 4: D-D for chromosome 8 and RC-RC for chromosome Y), consistent with both being part of the same homogeneous tandem array. By contrast, within the heterogeneous monomeric arrays, the orientation of alpha satellite DNA typically switches several times within each arm contig.[8,20] On the other hand, we have found[51] two 30mer HOR arrays in chimpanzee chromosome Y, positioned one after the other (with a gap of 599 bp in between). The first HOR, truncated at the start of the contig was referred to as direct. The second HOR which is reverse complement and highly identical to the first HOR array, was referred to as reverse complement. We conclude that the direct and reverse complement HOR arrays are positioned on the opposite arms of a palindrome and also are a part of the same homogeneous tandem array.

Suprachromosomal Family Assignment

Sequence comparison of alpha satellite monomers in human chromosomes revealed 12 types of alpha satellite monomers, which form five suprachromosomal families (SFs). They all descend from two basic types of monomers, A and B. To the subset A belong the SF types J1, D2, W4, W5, M1, and R1, and to the subset B belong J2, D1, W1, W2, W3, and R2.[7,8,61] Subtypes of alpha satellites are: SF1 (dimeric structure -J1_J2-), SF2 (dimeric structure -D1_D2-), SF3 (pentameric structure -W1_W2_W3_W4_W5-), SF4 (monomeric structure –M1-), and SF5 (dimeric structure -R1_R2-).[8] We calculate divergence for pairwise comparison of every monomer from consensus HOR and SF monomers. To each monomer constituting alphoid HOR the corresponding SF monomer from[61] with the lowest mutual divergence is assigned. Results are summarised in Table 5. In this way we find that, out of fifteen alpha satellite consensus HORs, ten could be clearly assigned to one of suprachromosomal families, while five are a combination of different types of SF monomers. However, even within this limited dataset, a significant number of consensus alpha satellites within HORs shows a sizable divergence with respect to previously described families[61] (row bellow family assignment in Table 5), suggesting that the complete suprachromosomal family classification has yet to be determined.

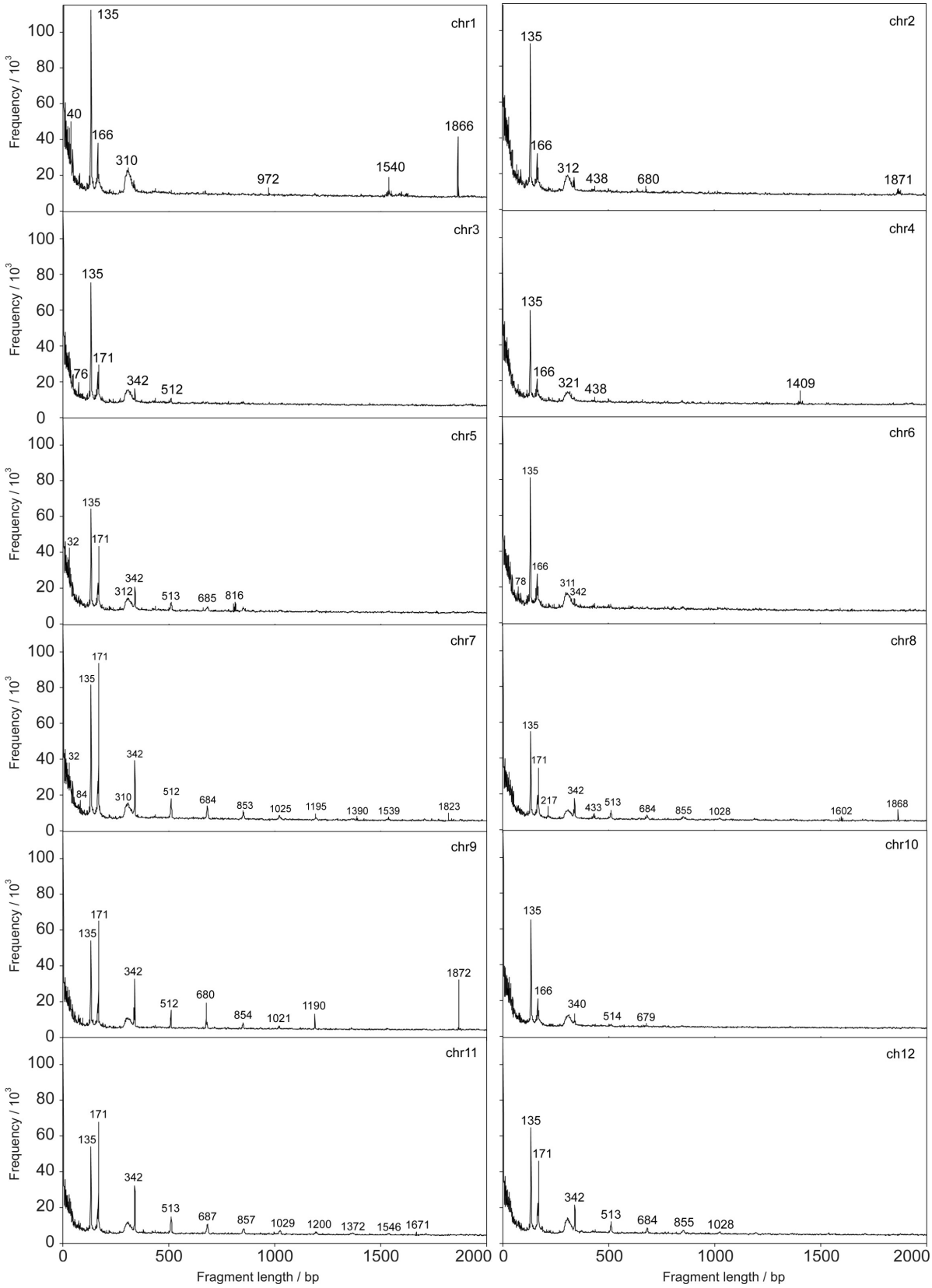

Figure 5. GRM diagrams for Build 37.2 ansembly of human chromosomes 1 to 12 in the interval of fragment lengths 0 - 2000 bp. Pronounced peaks are denoted by fragment lengths (repeat unit lengths).

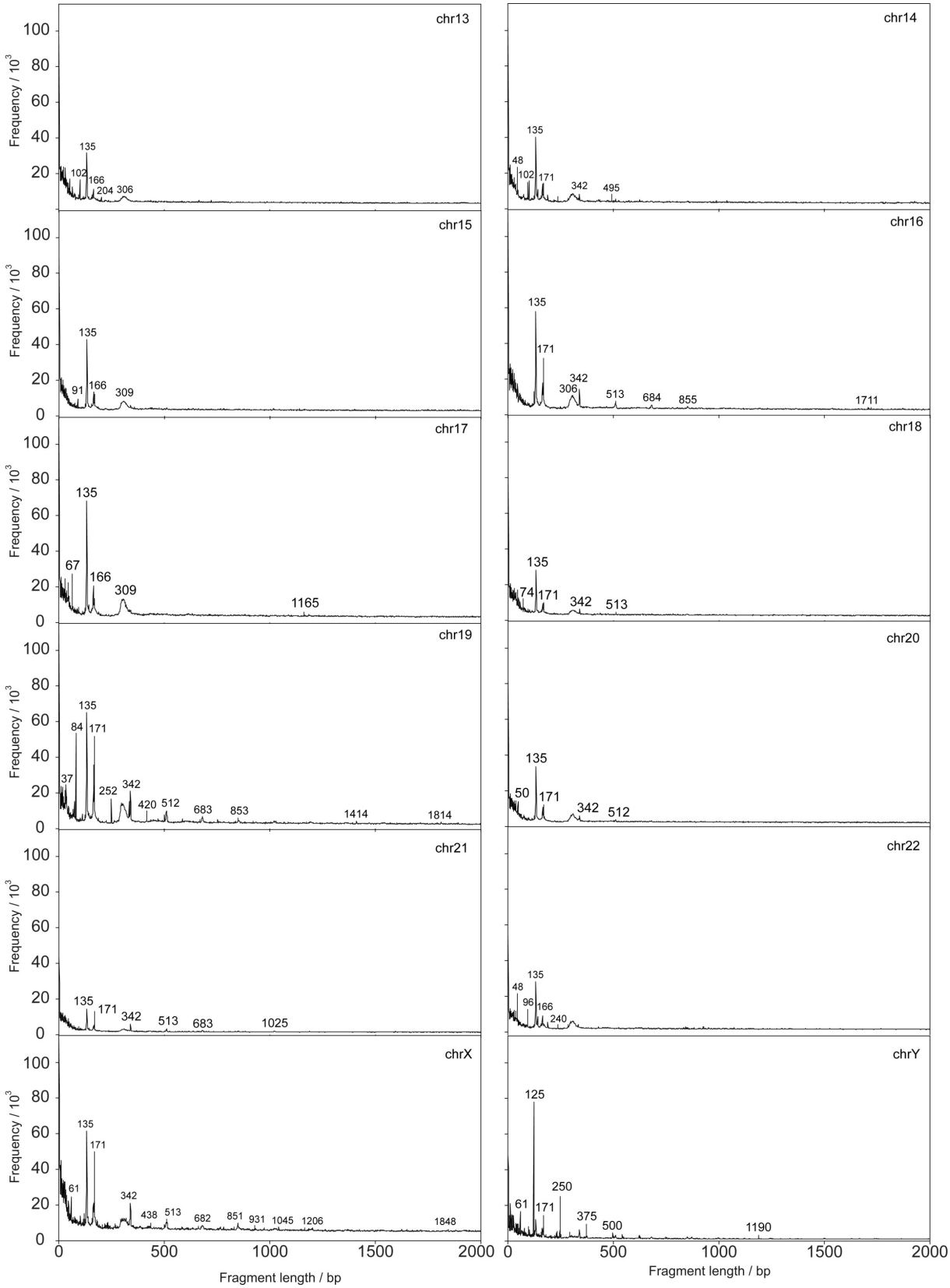

Figure 6. GRM diagrams for Build 37.2 assembly of human chromosomes 13 to Y in the interval of fragment lengths 0−2000 bp.

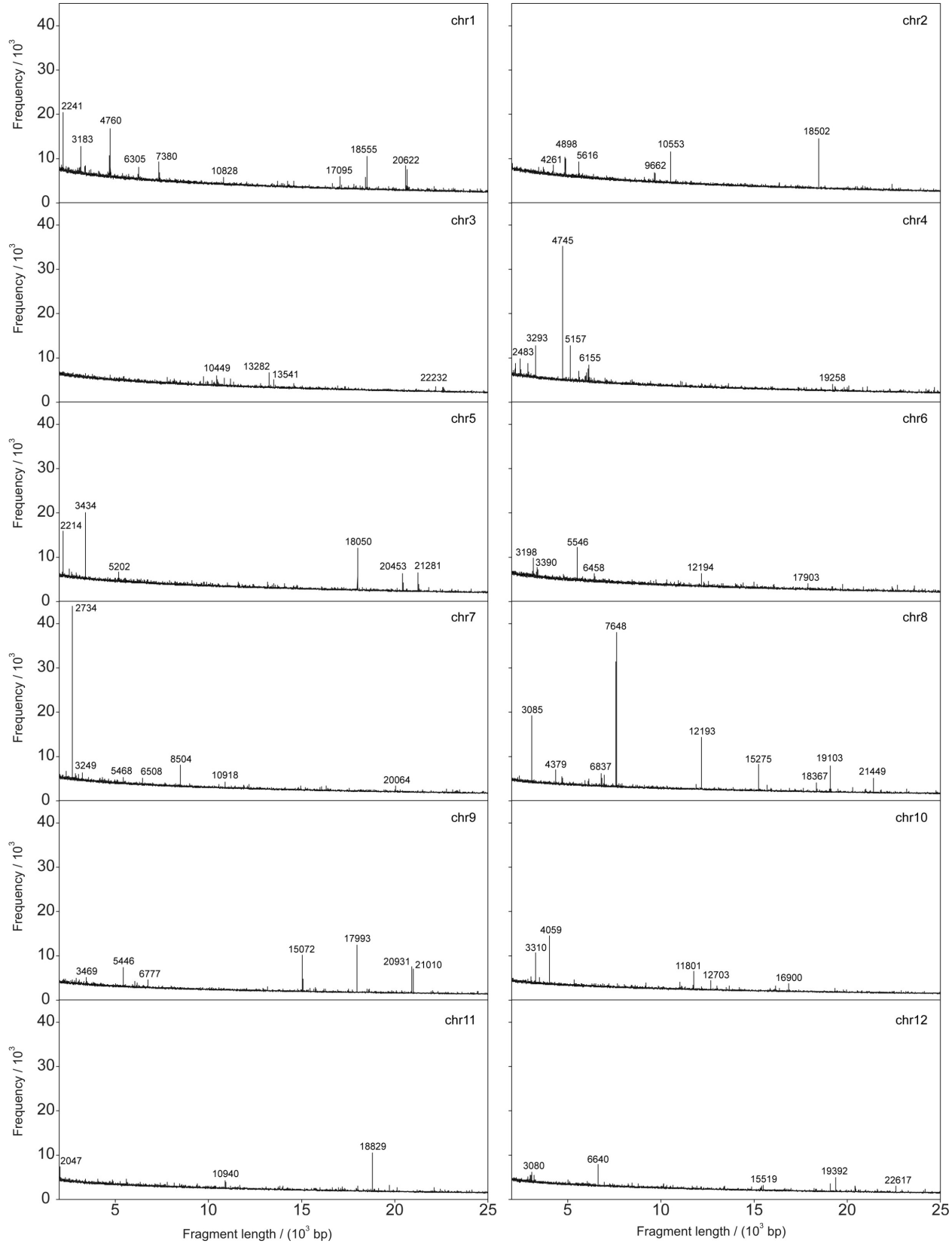

Figure 7. GRM diagrams for Build 37.2 assembly of human chromosomes 1 to 12 in the interval of fragment lengths 2 kb – 25 kb bp.

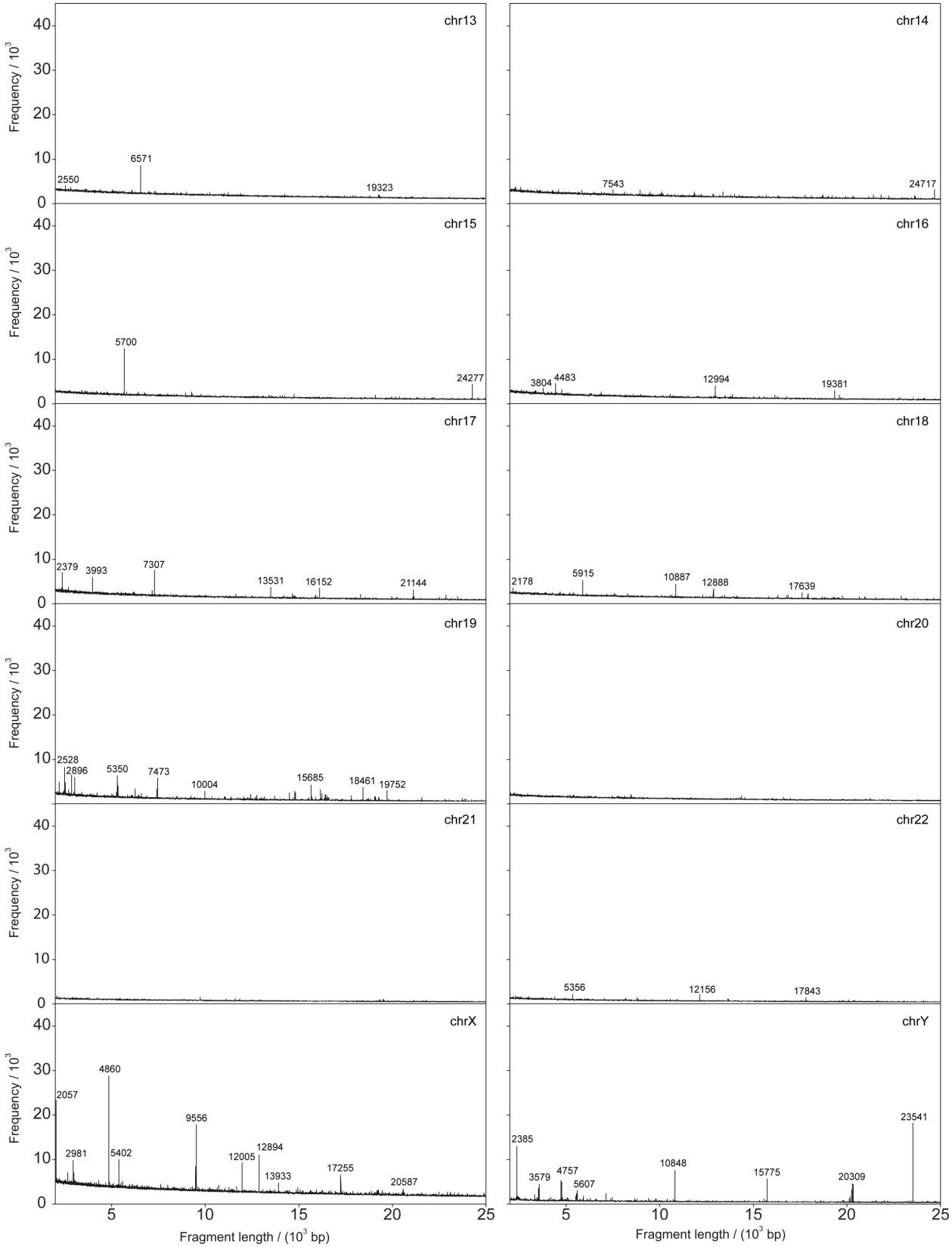

Figure 8. GRM diagrams for Build 37.2 assembly of human chromosomes 13 to Y in the interval of fragment lengths 2 kb – 25 kb bp.

Table 4. Alpha Satellite HOR annotation of the Build 37.2 data for human genome using Global Repeat Map algorithm

Chr.	Structure	Position NCBI Build 37.2	HOR consensus length (bp)	Orientation	CENP-B Box (monomer No)	pJD (monomer No)
1	11mer	ch1:121351617-121484037	1866	RC	2,4,11	6,7,8
2[(a)]	10mer	ch2:92275939-92305275	1708	D	4,6,8,10	1,2,5,7,9
4	13mer	ch4:52660162-52683268	2211	RC	7,8,9,11	1,4,6
5	13mer	ch5:49405714-49441477	2214	RC	-	-
7	16mer	ch7:61097241-61245558	2734	D	12	1
8	11mer	ch8:43820962-43838799 ch8:46838951-46857798	1867	D D	1,5,7,9,11	2,4,6,8,10
9[(a)]	7mer	ch9:not placed	1192	RC	-	1,2,5,7,8,9,10
9[(a)]	11mer	ch9:not placed	1872	D	3,5	2,6
10	18mer	ch10:42533098-42546683	3058	RC	1,3,5,7,9,11,15,17	13
11	12mer	ch11:51578629-51594143	2047	RC	3,7	1,11
17	14mer	ch17:22244535-22262545	2379	D	2,3,6,9,11,12	7,14
19	13mer	ch19:not placed	2215	RC	-	-
19	17mer	ch19:24603482-24631755	2895	D	1	17
X	12mer	chX:61682154-61725861	2042	RC	2,5,8,9	4,6,11
Y	45mer	chY:10083803-10104550 chY:13104583-13131941	7661	RC	-	6,7,12,13,15,16,17,18,20, 22,23,26,27,29,31,33,34, 36,37,38,39,40,41,42,43

[(a)] Human alpha satellite HORs reported here for the first time. Precise values of HOR consensus lengths are the same as obtained in[48,49] using earlier Build assemblies.

Table 5. Suprachromosomal family (SF) classification of HOR sequences from Table 3 and Ref. 55. To each monomer from HOR we assign the SF classification of closest SF consensus monomer defined in Ref. 8 and Ref. 61

Chr.	SF	Monomer family / Div (%)[55]
1	3	w5 17.9 w1 19.9 w1 15.1 w1 18.0 w5 12.9 w4 12.3 w1 17.0 w4 11.7 w3 14.0 w1 17.3 w1 14.0
2	2	d2 15.8 d1 9.4 d2 15.2 d1 12.9 w4 19.2 d1 12.9 d1 10.5 d1 10.5 d2 12.9 d1 12.9
4	2, 3, 4	m1 11.1 m1 15.1 m1 15.1 m1 12.9 d2 17.4 m1 14.5 w1 16.4 d2 17.5 d1 21.3 d1 21.1 w1 17.4 w3 21.4 w4 16.2
5	2,4	d1 13.5 m1 14.0 d1 12.3 w4 12.2 m1 15.2 m1 18.0 m1 16.4 m1 13.5 m1 16.4 d1 14.0 m1 13.5 m1 13.5 m1 9.9
7	2, 3, 4	w4 15.2 d1 14.5 m1 13.4 w4 18.1 m1 9.9 w4 14.0 d1 14.6 m1 12.2 d1 14.0 m1 11.7 w4 15.8 d2 14.0 m1 15.8 m1 14.6 m1 14.5 w4 9.4
8	2	d1 13.5 d2 16.4 d1 15.2 d2 12.3 d1 20.5 d2 16.4 d1 18.1 w4 18.6 d1 18.7 d2 14.6 d1 14.0
9	4	m1 5.9 m1 7.6 m1 11.1 m1 15.2 m1 15.8 m1 12.3 m1 9.9 m1 10.5 m1 6.4 m1 10.5 m1 8.8
9	2	d1 9.9 d2 11.1 d1 11.7 w4 13.4 d1 11.1 d2 11.1 d1 16.4
10	1	j1 19.9 j1 15.8 j2 16.4 j1 15.2 j2 13.5 j1 16.4 j2 15.1 j1 16.4 j2 15.8 j1 15.0 j1 21.6 j1 14.6 j2 18.7 j1 15.8 j2 15.2 j1 18.1 j2 17.0 j1 16.9
11	3	w4 17.0 w1 14.6 w1 11.6 w1 17.0 w4 12.9 w4 11.7 w2 10.5 w1 12.8 w5 18.0 w4 13.7 w4 12.9 w1 12.9
17	3	w4 11.1 w1 14.0 w1 12.2 w1 13.5 w4 16.4 w1 13.5 w4 8.8 w4 12.9 w1 17.5 w1 15.7 w5 18.7 w1 12.8 w2 14.0 w4 13.5
19	2, 4	m1 8.8 d1 12.3 w4 12.8 d1 10.5 w4 11.1 m1 9.9 m1 13.5 m1 12.9 m1 10.5 m1 14.6 d1 11.7 m1 11.7 d2 11.0
19	2, 3, 4	w1 15.5 m1 12.9 m1 9.4 m1 10.5 m1 14.0 w4 12.3 m1 16.4 m1 13.5 m1 12.9 m1 15.2 m1 11.7 w4 12.2 d1 10.5 w4 12.8 d1 11.1 m1 11.7 w4 14.6
X	3	w1 14.5 w1 17.5 w4 9.4 w4 14.6 w1 11.7 w4 22.4 w1 14.0 w1 15.7 w5 16.9 w4 16.4 w4 7.0 w2 10.5
Y	4	m1 14.9 m1 12.9 m1 14.0 m1 14.6 m1 12.3 m1 12.1 m1 15.2 m1 14.0 m1 12.9 m1 16.3 m1 10.5 m1 7.0 m1 9.3 m1 14.0 m1 9.9 m1 11.1 m1 8.7 m1 8.1 m1 10.5 m1 14.0 m1 10.5 m1 11.1 m1 14.0 m1 4.7 m1 9.9 m1 14.0 m1 15.1 m1 5.3 m1 12.2 m1 17.0 m1 7.0 m1 11.1 m1 11.1 m1 5.9 m1 10.5 m1 7.6 m1 6.4 m1 12.9 m1 7.6 m1 12.3 m1 9.9 m1 14.0 m1 13.5 m1 8.8 m1 15.8

The SF classification of alpha monomers or HORs is used as a basis for discussion of CENP B box and pJα motif distributions in alpha monomers.

CENP-B box and pJα Motif Distribution

The consensus alpha satellite monomers for basic types A and B have only seven differences, five of which are concentrated in a 16 bp region of alpha satellite monomer. Such clustering indicates that these mutations are not random, but are affected by a selection.[8,61] Indeed, the alternative A and B configurations match the binding sites of two alpha satellites-binding proteins, pJα (5'-TTCCTTTTPyCACCPuTAG-3') and CENP-B (5'-PyTTCGTTGGAAPuCGGGA-3').[61,62] Ohzeki *et al.*[63] and Warburton[64] have shown that only a combination of both the CENP-B box and HOR pattern provided successful centromere binding to kinetochore complex during mitotic processes. CENP-B box appears only in alpha satellite HOR[8,65,66] while no CENP-B boxes were detected in monomeric alpha satellites.[67,68] The pJα motif reflects some of nucleotides derived from alpha satellite monomer which were shown to be effective in binding experiments. A shorter pJα core sequence CCTTTTPyC,[61] 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.

After determining the SF classification of monomers in consensus HORs, we investigate the appearance of CENP-B box and pJα motif in these monomers. We find that only the monomers in 13mer HOR in chromosome 5 and monomers in 13mer HOR in chromosome 19 are without any CENP-B box and pJα motif (Table 4). This is an exception to the general pattern found for human chromosomes.[69] In the next chapter we will see that these two HORs are highly homologous, what could be a consequence of interchromosomal transition or orchestrated interchromosomal homogenization. Another consensus HOR from chromosome 19, a 17mer, has one CENP-B box and one pJα motif. The consensus 18mer HOR in chromosome 10 has eight CENP-B boxes, located in every other monomer except one. In chromosome 2 a new 10mer consensus HOR has four CENP-B boxes in every other monomer except one. In chromosome 4 a 13mer consensus HOR has CENP-B box in three consecutive monomers. In chromosome 9 a new 7mer consensus HOR has the pJα motif in four consecutive monomers. Moreover, we find in chromosome Y a first reported case of HOR with only pJα motif and no CENP-B box.

Since the CENP-B box and pJα motif are essential for protein binding, an interesting question is whether the monomers with and without CENP-B box and pJa motif have different sequence divergences. In this respect, we find that the pairwise divergence among monomers shows no dependence on the presence or absence of the CENP-B box or pJα motif.

Homogenization within Consensus Higher-Order Alpha Satellite Monomers

To explore the evolutionary relationships of higher-order alpha satellite monomers in human genome, we compared all consensus higher-order alpha satellite monomers from Table 3 and Supplementary Table 1 to each other. We performed Needleeman-Wunsch alignments[58] between all possible pairwase combinations of monomers (223 monomers, 49729 alignments). The relationship between monomers in consensus HORs is presented graphically in a heatmap (Figure 9), where each divergence is depicted according to a given color scale.

Within each of fifteen HORs intrachromosomal monomer divergence varies from ≈17% to ≈25% (Table 6 and Figure 10). Monomer divergence is lowest in 13mer in chromosome 19 (16.9 % ± 2.4 %), 11mer in chromosome 9 (17.9 % ± 3.5 %) and in 17mer in chromosome 19 (18.8 % ± 2.9 %) reflecting their higher homogeneity in comparison with, for example, 11mer in chromosome 1 (25.3 % ± 4.0 %) or 11mer in chromosome 8 (25.3 % ± 6.4 %).

There is difference between weaker homogeneity within HOR alpha satellite consensus monomers and before mentioned stronger homogeneity within array of alpha satellite HOR copies which have typical divergence between copies of 1−5 % (for example of chromosome 4 see Table 1), which is a consequence of evolutionary concerted processes. Similarity between various consensus monomers within one higher-order repeat unit is derived from common ancestral alpha satellite monomer. Creation of a large tandemly repeated alpha satellite array may occur through abruptly amplification of ancestral alpha satellite monomer or through a step by step series of unequal crossovers and/or gene conversion events that initially create duplication and then expand.[8] Such hypothesis of one pra-ancestral alpha satellite monomer is additionally supported by calculations of consensus sequences of all monomers within any consensus HOR and their mutual alignment (Table 8). Very low mutual divergences between consensus monomers from consensus HORs in Table 8 reveal that calculation of consensuses is like traveling back in time: all higher-order monomers in every chromosome are evidently descendants of one - ancestral alpha satellite. After a specific alpha satellite array has been created by amplification, there are no mechanisms, like “homogenization” processes in a case of HOR copies, to maintain its homogeneity and copies just passively accumulate mutations. In a more realistic situation, there are homogenization mechanisms in both, monomeric and HOR sequences, but homogenization processes within HOR copies will occur more

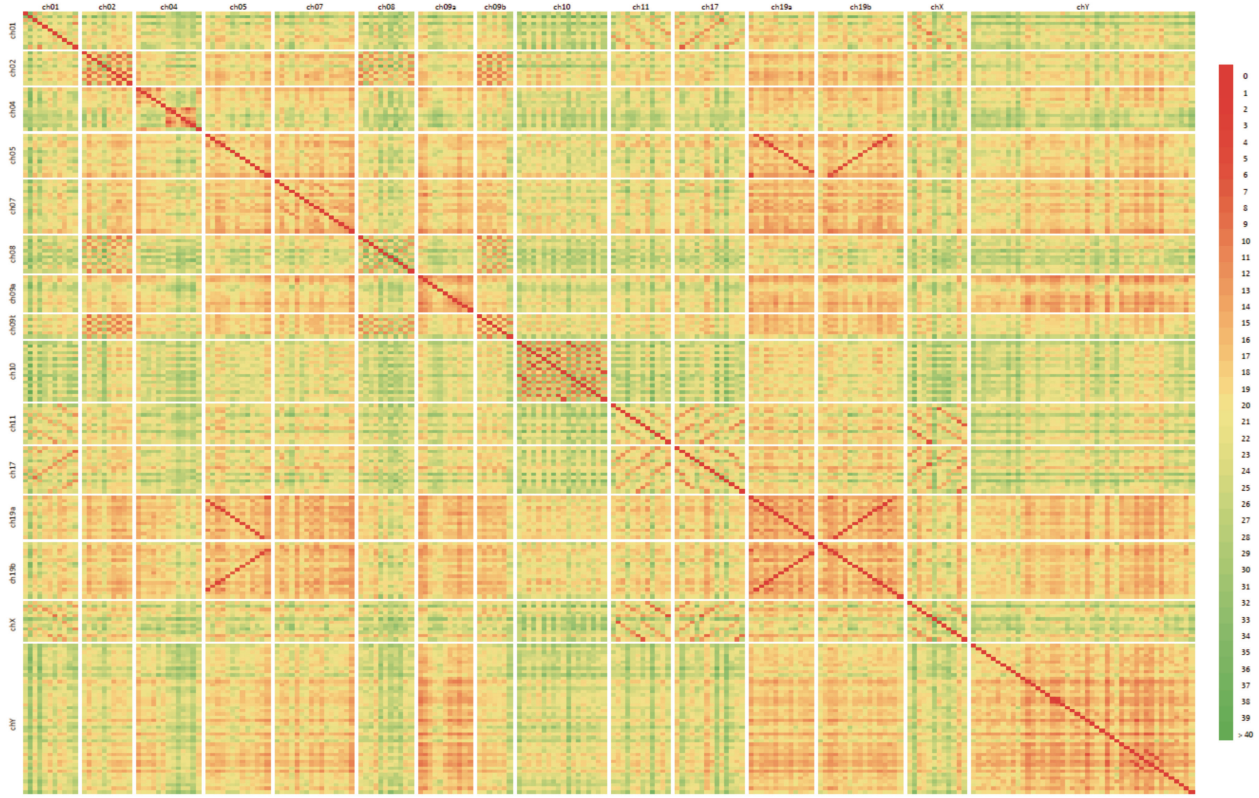

Figure 9. Graphical presentation („heatmap“) of divergence between monomers in consensus HORs. Monomers in consensus HORs are displayed both horizontally and vertically. The color of the intersection of the horizontal band corresponding to the n-th monomer in the i-th chromosome and the vertical band corresponding to the m-th monomer in j-th chromosome represents the divergence (in percents) between these two monomers (vertical scale on the right side).

frequently than between monomeric units.[25] In both cases, the mean divergences indicate the time of creation of initial alpha satellite arrays, and in the case of isotropic random mutations, standard deviations indicate rate of creation processes alone. On the other hand, the mean divergence could also be a good recipe for estimation of the age of HOR creation, because after a specific HOR unit has been created, the processes of effective “homogenization” have started and present mutations have been fixated.

Analyses of interchromosomal (or inter-HOR, if there are two different higher-order alpha satellite units in the same chromosome) mean divergences reveal (Table 6) a few possible similar higher-order repeat units, *e.g.*, 13 mer in chromosome 5 and 13mer in chromosome 19, 13mer in chromosome 5 and 17mer inchromosome 19, 16 mer in chromosome 7 and 13mer in chromosome 19, 16mer in chromosome 7 and 17mer in chromosome 19, 11mer in chromosome 9 and 45mer in chorosome Y, and so on. It is important to notice that real information of mean divergences between two different higher-order repeat units are masked in Table 6 because of monomeric heterogeneity within one higher-order repeat unit.

To overcome above-mentioned problem we performed modified estimation of mean divergence; to each consensus monomer in one higher-order repeat

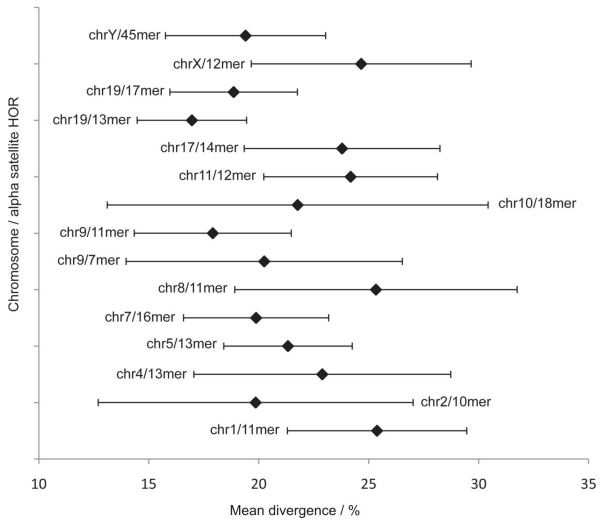

Figure 10. Mean divergence and standard deviation among human alpha satellite monomers in consensus HOR.

Table 6. Mean divergence among human consensus alpha satellite HOR monomers

	chr1/11mer	chr2/10mer	chr4/13mer	chr5/13mer	chr7/16mer	chr8/ 11mer	chr9/7mer	chr9/11mer	chr10/18mer	chr11/12mer	chr17/14mer	chr19/13mer	chr19/17mer	chrX/12mer	chrY/45mer
chr1/11mer	25.3 ± 4.0	24.8 ± 3.4	26.9 ± 3.4	24.9 ± 3.7	25.0 ± 3.8	27.5 ± 3.5	24.7 ± 3.5	25.5 ± 3.7	28.1 ± 3.7	4.2 ± 4.4	23.8 ± 4.9	23.1 ± 3.6	24.0 ± 3.7	24.2 ± 4.9	26.4 ± 3.7
chr2/10mer		19.8 ± 7.1	24.2 ± 3.7	22.0 ± 2.8	21.9 ± 3.3	22.4 ± 6.3	19.6 ± 6.2	22.9 ± 2.8	25.2 ± 3.7	24.0 ± 3.2	23.9 ± 3.3	20.0 ± 2.8	20.9 ± 3.0	24.4 ± 3.7	23.8 ± 3.0
chr4/13mer			22.8 ± 5.8	23.7 ± 3.2	23.0 ± 3.2	27.1 ± 3.3	24.0 ± 3.3	24.3 ± 3.8	26.7 ± 2.9	26.6 ± 2.8	26.1 ± 2.9	21.6 ± 3.0	22.5 ± 3.1	26.3 ± 3.6	24.9 ± 3.8
chr5/13mer				21.3 ± 2.9	20.8 ± 2.8	24.9 ± 3.1	22.1 ± 3.0	21.1 ± 2.9	24.8 ± 2.5	24.2 ± 3.4	24.0 ± 3.6	18.1 ± 4.7	19.3 ± 4.3	24.4 ± 3.8	22.2 ± 2.9
chr7/16mer					19.8 ± 3.3	24.8 ± 2.9	21.5 ± 3.0	21.0 ± 3.2	24.5 ± 2.8	24.2 ± 3.4	24.0 ± 3.8	18.6 ± 2.7	19.6 ± 3.0	24.2 ± 4.0	21.9 ± 3.3
chr8/11mer						25.3 ± 6.4	22.3 ± 6.2	25.6 ± 3.1	27.8 ± 2.6	26.9 ± 2.9	26.5 ± 3.3	23.2 ± 3.0	24.1 ± 3.3	27.1 ± 3.4	26.2 ± 3.2
chr9/7mer							20.2 ± 6.2	22.9 ± 2.9	24.6 ± 2.9	24.1 ± 3.0	23.7 ± 3.3	20.0 ± 2.6	20.9 ± 3.0	24.4 ± 3.8	23.7 ± 3.0
chr9/11mer								17.9 ± 3.5	25.5 ± 2.8	25.1 ± 3.2	25.0 ± 3.7	19.1 ± 2.7	20.1 ± 3.0	25.1 ± 3.7	19.3 ± 3.7
chr10/18mer									21.7 ± 8.6	27.7 ± 3.2	27.0 ± 3.4	22.8 ± 2.4	23.9 ± 2.7	27.7 ± 3.6	26.1 ± 3.0
chr11/12mer										24.1 ± 3.9	23.3 ± 4.7	22.4 ± 3.3	23.2 ± 3.4	23.4 ± 5.7	25.7 ± 3.2
chr17/14mer											23.7 ± 4.4	22.2 ± 3.5	23.0 ± 3.7	23.6 ± 5.4	25.6 ± 3.7
chr19/13mer												16.9 ± 2.4	17.0 ± 4.3	22.5 ± 3.8	20.3 ± 2.8
chr19/17mer													18.8 ± 2.9	23.4 ± 3.9	21.1 ± 2.9
chrX/12mer														24.6 ± 5.0	25.6 ± 3.6
chrY/45mer															19.4 ± 3.6

Table 7. Modified mean divergence among human alpha satellite monomers from consensus HORs. To each monomer from each HOR a monomer with lowest divergence (%) from other consensus HORs is assigned and then the mean divergence calculated

	chr1/11mer	chr2/10mer	chr4/13mer	chr5/13mer	chr7/16mer	chr8/11mer	chr9/7mer	chr9/11mer	chr10/18mer	chr11/12mer	chr17/14mer	chr19/13mer	chr19/17mer	chrX/12mer	chrY/45mer
chr1/11mer		20.8 ± 1.7	22.8 ± 2.8	20.8 ± 2.6	20.8 ± 3.2	23.0 ± 3.2	20.8 ± 2.1	22.3 ± 3.5	23.6 ± 1.8	16.4 ± 2.1	13.9 ± 2.9	19.7 ± 2.6	20.0 ± 2.6	15.0 ± 2.2	21.5 ± 3.9
chr2/10mer	20.8 ± 1.7		19.2 ± 2.7	18.2 ± 2.4	18.2 ± 2.7	13.7 ± 2.2	11.9 ± 2.0	19.6 ± 2.4	21.3 ± 4.2	20.4 ± 2.1	19.7 ± 2.2	16.8 ± 2.6	17.2 ± 2.2	19.1 ± 2.1	18.6 ± 2.2
chr4/13mer	22.8 ± 2.8	19.2 ± 2.7		19.7 ± 2.2	19.0 ± 2.6	22.3 ± 2.1	19.0 ± 1.6	19.0 ± 2.8	23.2 ± 2.1	22.5 ± 2.2	22.0 ± 3.0	17.0 ± 1.6	18.4 ± 2.5	21.8 ± 2.9	19.9 ± 3.8
chr5/13mer	20.8 ± 2.6	18.2 ± 2.4	19.7 ± 2.2		16.7 ± 2.2	20.3 ± 2.8	17.7 ± 2.6	17.4 ± 2.8	16.7 ± 2.2	19.9 ± 2.5	18.4 ± 2.5	4.7 ± 1.1	6.1 ± 1.8	20.2 ± 3.4	16.8 ± 2.1
chr7/16mer	20.8 ± 3.2	18.2 ± 2.7	19.0 ± 2.6	16.7 ± 2.2		21.2 ± 2.6	17.5 ± 1.6	16.1 ± 2.9	21.2 ± 2.2	20.0 ± 3.2	19.1 ± 3.4	14.2 ± 1.4	15.4 ± 2.9	20.2 ± 4.1	16.5 ± 2.5
chr8/11mer	23.0 ± 3.2	13.7 ± 2.2	22.3 ± 2.1	20.3 ± 2.8	21.2 ± 2.6		13.6 ± 2.3	22.1 ± 2.9	24.1 ± 1.8	23.2 ± 1.9	22.0 ± 2.9	19.5 ± 1.9	19.9 ± 2.4	22.1 ± 2.8	21.4 ± 2.7
chr9/7mer	20.8 ± 2.1	11.9 ± 2.0	19.0 ± 1.6	17.7 ± 2.6	17.5 ± 1.6	13.6 ± 2.3		19.6 ± 3.1	20.8 ± 2.4	20.3 ± 1.6	18.8 ± 2.5	16.7 ± 2.0	16.6 ± 2.0	18.5 ± 2.1	19.0 ± 2.2
chr9/11mer	22.3 ± 3.5	19.6 ± 2.4	19.0 ± 2.8	17.4 ± 2.8	16.1 ± 2.9	22.1 ± 2.9	19.6 ± 3.1		22.4 ± 2.3	21.7 ± 2.3	189.0 ± 2.9	15.7 ± 2.5	16.7 ± 2.2	19.2 ± 2.6	14.0 ± 3.3
chr10/18mer	23.6 ± 1.8	21.3 ± 4.2	23.2 ± 2.1	16.7 ± 2.2	21.2 ± 2.2	24.1 ± 1.8	20.8 ± 2.4	22.4 ± 2.2		23.7 ± 1.8	23.2 ± 1.6	19.5 ± 1.8	20.3 ± 2.3	23.4 ± 2.0	21.0 ± 1.8
chr11/12mer	16.4 ± 2.1	20.4 ± 2.1	22.5 ± 2.2	19.9 ± 2.5	20.0 ± 3.2	23.2 ± 1.9	20.3 ± 1.6	21.7 ± 2.3	23.7 ± 1.8		13.7 ± 2.2	19.0 ± 2.7	19.1 ± 2.9	11.2 ± 3.8	20.6 ± 3.1
chr17/14mer	13.9 ± 2.9	19.7 ± 2.2	22.0 ± 3.0	18.4 ± 2.5	19.1 ± 3.4	22.0 ± 2.9	18.8 ± 2.5	19.0 ± 2.9	23.2 ± 1.6	13.7 ± 2.2		16.8 ± 1.6	19.0 ± 2.8	13.9 ± 4.4	20.4 ± 3.6
chr19/13mer	19.7 ± 2.6	16.8 ± 2.6	17.0 ± 1.6	4.7 ± 1.1	14.2 ± 1.4	19.5 ± 1.9	16.7 ± 2.0	15.7 ± 2.5	19.5 ± 1.8	19.0 ± 2.7	16.8 ± 1.6		3.2 ± 1.8	19.2 ± 3.6	14.6 ± 1.4
chr19/17mer	20.0 ± 2.6	17.2 ± 2.2	18.4 ± 2.5	6.1 ± 1.8	15.4 ± 2.9	19.9 ± 2.4	16.6 ± 2.0	16.7 ± 2.2	20.3 ± 2.3	19.1 ± 2.9	19.0 ± 2.8	3.2 ± 1.8		19.6 ± 3.2	15.8 ± 2.0
chrX/12mer	15.0 ± 2.2	19.1 ± 2.1	21.9 ± 2.9	20.2 ± 3.4	20.2 ± 4.1	22.1 ± 2.8	18.5 ± 2.1	19.2 ± 2.6	23.4 ± 2.0	11.2 ± 3.8	13.9 ± 4.4	19.2 ± 3.6	19.6 ± 3.2		20.9 ± 3.5
chrY/45mer	21.5 ± 3.9	18.6 ± 2.2	19.9 ± 3.8	16.8 ± 2.1	16.5 ± 2.5	21.4 ± 2.7	19.0 ± 2.2	14.0 ± 3.3	21.0 ± 1.8	20.6 ± 3.1	20.4 ± 3.6	14.6 ± 1.4	15.8 ± 2.0	20.9 ± 3.5	

Table 8. Mean divergences among alpha satellite monomers from consensus HOR sequences in human chromosomes. For example, to a small square at the intersection of the first horizontal band corresponding to chromosome 1 and the fifth vertical band corresponding to the chromosome 7, divergence between consensus sequence of all 11 monomers from consensus 11mer HOR in chromosome 1 and from consensus sequence of all 16 monomers from consensus 16mer HOR in chromosome 7 is assigned (5%)

	chr1/11mer	chr2/10mer	chr4/13mer	chr5/13mer	chr7/16mer	chr8/11mer	chr9/7mer	chr9/11mer	chr10/18mer	chr11/12mer	chr17/14mer	chr19/13mer	chr19/17mer	chrX/12mer	chrY/45mer
chr1/11mer	0	9	8	5	5	10	7	11	14	4	5	5	5	4	7
chr2/10mer		0	8	7	6	5	9	6	14	8	9	6	6	9	10
chr4/13mer			0	4	5	8	5	8	10	8	9	5	5	9	6
chr5/13mer				0	2	7	2	8	10	6	7	1	1	7	3
chr7/16mer					0	7	5	7	11	5	6	2	2	6	5
chr8/11mer						0	9	5	11	8	8	6	6	8	10
chr9/7mer							0	9	12	9	9	3	3	9	1
chr9/11mer								0	12	7	8	8	8	8	10
chr10/18mer									0	13	12	11	11	13	12
chr11/12mer										0	2	6	6	1	9
chr17/14mer											0	6	6	2	10
chr19/13mer												0	0	6	4
chr19/17mer													0	6	4
chrX/12mer														0	9
chrY/45mer															0

unit a monomer with lowest alignment divergence from other higher-order repeat unit has been assigned and then the mean divergence of, so obtained monomer pairs, has been calculated (Table 7). This modified mean percent divergence we called minimal divergence.

The 13mer and 17mer HOR in chromosomes 19 have the lowest mutual minimal divergence; we proposed that one higher-order unit is derived from the other, although more complex explanations, with both higher-order units derived from a third unknown higher-order unit is also possible.[50] It is very unlikely that the 17mer unit arose from 13mer unit by addition of four monomers, because monomers alignment excluded possibility that the four additional monomers in 17mer unit are duplications of any monomers from 13mer unit (see chromosome 19 13mer and 17mer alignment in a heatmap from Figure 9). Therefore, we hypothesized that the shorter, 13mer higher-order repeat unit arose from the longer 17mer higher-order unit by deletion of four alpha satellite monomers which are all distinct from the monomers in 13mer. This is consistent with a general view[7] that a type of polymorphism found in alphoid arrays can be related to HOR units that differ by an integral number of alphoid monomers. It should be noticed that, in addition to the chromosome 19 case of two similar higher-order units on the same centromere, there is a sample of two completely different higher-order alpha satellite units on the centromere of chromosome 9. It is obvious, from Figure 9 and Table 7, that these higher-order alpha satellite structures have completely different building units (monomers) and, from Table 6, that they are created in different moments and with different rates.

Moreover, 13mer in chromosome 5 is similar to the both 13mer and 17mer in chromosome 19 what could be a consequence of two possible processes: (1) these sequences were subject to intrachromosomal homogenization mechanisms or (2) blocks of higher-order 13mer alpha satellite may have undergone exchanges via transposition mechanisms.[25]

In addition to this group of three very similar alpha satellite HORs there are a few groups of HORs with somewhat greater mutual diversity: the group of 11mer in chromosome 1, 12mer in chromosome 11, 14mer in chromosome 17, and 12mer in chromosome X, or the group of 10mer in chromosome 2, 11mer in chromosome 8, and 7mer in chromosome 9, and so on, with mutual divergence of about 13 %. If we assume that the transposition mechanisms are more probable to be responsible for these similarities it is very easy (from

Table 9. Alpha satellite monomer structure of 21mer HORs in contig NW_003456961.1 in chimpanzee chromosome 4

HOR Copy No.	Position / bp	Divergence / %	Composition
1	11758218	1.4	mc01, ..., mc21
2	11761827	1.4	mc01, ..., mc21
3	11765430	1.3	mc01, ..., mc21
4	11769037	1.0	mc01, ..., mc17, 640 bp N, mc21
5	11772772	0.9	mc01, ..., mc21
6	11776379	0.8	mc01, ..., mc21
7	11779985	1.9	mc01, ..., mc08, 171 bp*, 168 bp N, mc06, ..., mc21
8	11784611	0.9	mc01, ..., mc21
9	11788220	1.3	736 bp N, m06, ..., m21
10	11791705	1.1	mc01, ..., mc21
11	11795310	1.7	mc01, ..., mc05, 712 bp N, mc14, ...,mc21

N – unsequenced segment inside contig NW_003456961.1
* 171 bp segment with divergence of 53.5 % with respect to mc18 consensus sequence

Table 7) to follow the time and space pattern of human alpha satellite HORs creation.

HORs in Chimpanzee Chromosome 4

Applying GRM to the chimpanzee chromosome 4 (NCBI Build 2.1 assembly), we find 21mer HORs in chimpanzee contig NW_003456961.1 (Table 9). The corresponding basic consensus monomers are denoted as mc01, …, mc21. The corresponding consensus HOR array are shown in Table 10. The consensus length of 21mer HOR (secondary periodicity) repeat unit is 3606 bp. Previously, an alpha satellite subset organized as a series of pentameric (higher-order) repeats was reported at the centromere of the chimpanzee chromosome 4.[41]

Comparison of Monomers from Consensus HORs in Human and Chimpanzee Chromosome 4

In the first step, we compute divergences between 13 human monomers from consensus 13mer HOR and 21 chimpanzee monomers from consensus 21mer HOR (Table 11). It is obvious that divergences are scattered and there is no small divergence between any human and chimpanzee alpha monomers. Thereafter, we conclude that none of chimpanzee monomers can be assigned to a particular human monomer.

The mean values of divergence between monomer sequences from consensus HORs are:
div (13 human vs. 13 human) = 22.9 %,
div (21 chimp vs. 21 chimp) = 25.7 %,
div (13 human vs. 21 chimp) = 32.9 %,
where (13 human) denotes the set of alpha satellite monomers from human consensus 13mer HOR, and (21 chimp) from chimpanzee consensus 21mer HOR.

The number of different monomers constituting consensus HOR in human chromosome 4 (13 monomers) is different than in the chimpanzee chromosome 4 genome (21 monomers). All monomers constituting human 13mer HOR are different from monomers constituting chimpanzee 21mer HOR by ≈33 %, which is considerably larger than divergence between monomers within a single HOR copy (Table 6). This results show that alpha satellite HORs in human and chimpanzee chromosome 4 have been created after the human-chimpanzee separation. Nevertheless, because the mean difference is still sizably lower than difference between two random sequences, it is reasonable to assume that both human and chimpanzee alpha satellites originate from a common ancestor that predated the human-chimpanzee separation.

Comparison of Codon-Like Trinucleotides Extension in Human and Chimpanzee HORs in Chromosomes 4

In the next step, a new method of codon-like trinucleotides (CLTs) extensions[55] is applied to analyze the differences in structure of human and chimpanzee alpha satellite monomers in HORs in chromosome 4. Inspired by much interest in gene regulators, we analyzed trinucleotides corresponding to the start (ATG) and stop (TGA, TAA, TAG) CLTs in monomer sequences. It was shown[55] that the specificity and the level of extension of start/stop-CLTs distinguish human alpha satellite sequences from non-alpha satellite HORs and non-repeat sequences. As a measure of CLT-clustering of genomic sequences the corresponding extension factor r was introduced.[55]

Recently, we have identified HORs in chromosome 5 of chimpanzee (5mer), orangutan (14mer), and macaque (two 3mers).[70]

Table 10. Consensus sequence of chimpanzee alpha satellite 21mer HOR in chromosome 4 determined by the GRM analysis of chimpanzee chromosome 4 (contig NW_003456961.1). Monomers are denoted by mc01, ..., mc21. Consensus length of 21mer HOR is 3606 bp

```
mc01 170 TGTGAAATGGATTTGTGATGTGTGGATTTATCTCACAGAGGTAAACCTTTCTTTTGATTCAGCAGGTTGGAAACACTTTTTCTAGGATCTGCAAAGGGAAATTTGGAAGCCCATTGAGGCCTATAGTGAAAAACCAAGTATCCCCAGATAAAAACTAGAAAGATGCGATC
mc02 169 TGTGAAACTGCCTTGTGATGTGTGGATTCACCTCAGAGAGTTAAACCTTCTTTTATTCAGCAGGGTGGAAACACTGTCTCTGGTGAATCTGTGTAGAAAACATTTGGGAGCCCATAGAGTCCTATGGTGAAAAACAGAATATCCTTCTTTAAAAACTAGAAGAAGGTATA
mc03 171 TCTGAAATGGCTTTGCGATGTATGGATTCATCTCACAGAGTTAACACCTACATTTGATTCAAACTTTTGGCAACACTTTTTTTGTAAAATTTGTGAAGGGATATTTGAAAGCCACGTGAGGCCTAAGGGTTAAGCCGTATATACTGACATAAAAACTAGAAAGAAACTATC
mc04 172 TGTTAAATAGCTTCATAATCTGTGTATTCATCTCACAGGGTCAAACCTTTCTTTCGGTTCAGCAGGTTGTAAACACTCTTTTTGTAGAATCTGCAAATGGACATTTGGAATCCCTTTGAGGCTTATGGTGAAGAAATGAATATCCCCTGATAAAAACTAGAAATAAACTATC
mc05 172 TATGAAACTGCTTTGTGATGTGAAGATTCATCTAACAGAGGTAAACCTTTCTTTTGATTCAGTACATTGGAAACAATCTTTTTGTAGAATCTGTGTAGGGACATTTGGGAGCTCATTGTGGCCTATGGTAAGAAACTGAATATTCCCAGATAAAAACCAGAAAGAAGCTATC
mc06 172 TGTGAAACTCCTTTGTGATGTGTGAATTCATCTCACAGAACTAAATCTTTCTTTTGGTTCAGCAGGTTGGAAACACTTCTTTTGTGGAATATTTGAAGTCACATTTTAAAGCCCATTGTAGCACATGGTGAAAAACAGAATATCCCCAGATAAAAGCTAGAAGGAAGCTTTC
mc07 172 TGTAAAACTTCTTTGTGATGTATGTTTTCATCTCAAATAGTTAAACCTTGCTGTTGAATCCGTCGGTTGGAAATGCTCTTTGTGTAGGATCTGCAAAGAAACATTTGGGAGCTCATTGAGGGGTATGGTGAAAAACAGAATATCCTCAGATAAAAACCAGAAAGAAGGTATC
mc08 170 TGTGAAACAGCTTTGCAGTCTGTGGATTCACATCACAGAGATAATATTCATTTTGATACAGTAGTTTGGAAACTCTCTTTTTGTAAATCCTGTGAAGGGACATTGGGAGTCCTTGTAGGCTCTTGGAGGAAGACTGAATATCCCCAGATAAAAACTAGAAAGAAGCTACC
mc09 173 TGTGAAATAGATTTGTGCTGTTTGGATTCCTCTCCCAGAGTTAAACATTTCTGTCAATTCCACAGGTTGGAAACACTCTTTTAGTAGAATCTGCAAAGGGATATTTGGGAGCCCTTTGAGAGCAATGGTGAAAAAGTGAATATCCTCAGATAAAAACTAGAAAGATGTTATCT
mc10 171 GTGAAACAGCTTTGTTTAGGTGTGGATTCATCTCACAGAGTTAAACATTTCTTTAGAACCAGCAGCTTGGAACACTCTTCTGGTAGAAGTTGTGAAGGGCCATTCAGGAGTCCATTGAGGCTTATGGTGAAAAACCAAATATCACCAGATAAAAACAAGAAACAACTTATG
mc11 173 TATGAAACTGCTTTGTAATGTGTAGATTAATCTCACAGAGTTAAACCTTTTTTTTTGATATAGCATGTTGGAAACACTCTTTAGTAGAAGCAGCAAAGGGACATTAGGGAGCACATTGACGCCTACTATGAAAAACAAAATACATGCAGATAAAACCTAGAAAGAAGCTATC
mc12 176 TGTGAAACTGCTTTGTGATATGTGGATTCATCTCATAGAGTTAAACCTTTTTTTTTTTTTGATTCAGCATGTTGGAAACACTGTTTTTGTAGAGTCTGTGAAAAGACATATGGAAGCGTATTGAGGCCTACAGTGGAAAACCAAATATCCTTAGATTAAACTAGAAAGAAGTTATC
mc13 170 CCTGAAACTGCTCTGTTATGTGTGCCTTCACCTCCCACATTTAAAAAATCTTTTGATTCAGCAGGTTGGAAACTCTCTTTTGATAGAATCTGTGAAGGGACATTCCAGACCCCAATGAGGCTCTGGTGAAAAACAGACTACTCATGGACAACAACTTGAAAGAAGCCATT
mc14 173 TGTGAAACTGTTTTTTGATGTGTGGATATGTCTAACAGAGTTAAACTTTTCTTTTTTTTTCAGCCATTTGAAAACACTCTTTTAGTAGTATCTGTGAAAAAGCATTTGGGGGTCCACTGAAGCCTGAGGTTAACAACTGAATATACCCAGATAAAAATTTAAAAACATATACT
mc15 172 TGTGAAGCTGCTTTGTGATGTGTGGATTCATCTCACAGAGTTAAACTTTTCTATTGATTCAGCAAGCTTGAAAAACTATTTTTGGATAATCTGTGAAAAGACATTTAGGGGCCCATTTATGCCTATGTTGAAAAACGGAATATTCCCATAGGAAAACTAGAAGGACAATATC
mc16 174 GGTAAAATGGCTCTGTGATGTAGGGATTCATCGTACAGAGTTAAACCTTTGTTTTGATTCCACAGGTAGAAAACACACTTTTTGTAGAATCAGGGAAGTGACATTCTGGAGCCCATTGAGGCCTGTGGTGAAAATCAGAATAGCCCCAGATAAAAACTAGAAAAAAGTAATCTG
mc17 171 TTAAAACTGCTTTGTGATGTGTGGATTAACCACACAGCATTAAACCTTTCTTTTGATTTATCAAGTTGGAAACACTGGTTTGGTAGGATCTGTGAAGAGACATTTAGTTGCCCTATGAGGCCTATGGTGAGAAACTGAATGTCCCCAGAAAAACATTAGAAAGAAGCTATC
mc18 171 TGTTAAAGTGCATTTTGATGTGTGGATTCATCTTACAGAATTAAATTCTTGTTTTGATTCAGCAGGTTAGAAACACTCTTTTTGTAGAATCTGCAAAGGGACATTTAAGAGCCCATTGAGGCCTATGGTAAAACACAGAATATCCAAAGATAAAACTACACAGAATCTATG
mc19 172 AGTGAGACTGATTTGTGATGTGTGGATTCATCTCACAGAGTTAAACGTTTCTTTTGTTTAAGCAGTTTGAAAACAATCTTTTTGTAGAATATGTGAAGGGACATTCAAGAGCCCATTGAGGCCTATGGTAAAAAACGGAATATCCCCAGAGGAAAACTAGAAAGAAGCTACC
mc20 173 TGTGAAACAGCTTTGTGATTTGTGGATTCATCACACAGGTTTAAGCCTTTCTTTTTATTCCACAAAATGGAGACTCTCTTTTTGTAGAATCTGTGAAAGGACATGTGGGGGCCCACTGAGGCATATGGTGAAAATCCAAATATTCACAGATTAAAAACTAGAAAGAAGCTATT
mc21 169 TGTGAAAATGCTTTGCGATGTGTTAATTCATCTCACATAGTTAAATCATTCTTTTGATTCAGCAGGTTGGAAACACTTTTTGTAGAATCTTTGAAGGAACATATGAGAACCCATTGAGGTATGTGGTGAAAGACTGAATATTTCCAAATAAAAACTAGAAAGAAGCTAT
```

Table 11. Divergence between monomers from consensus alpha satellite 13mer HOR in human chromosome 4 ({*m*}) (from Table 3) and monomers from 21mer HOR in chimpanzee chromosome 4 ({*mc*}) (from Table 10)

human/ chimpanzee	cm01 170	cm02 169	cm03 171	cm04 172	cm05 172	cm06 172	cm07 172	cm08 170	cm09 173	cm10 171	cm11 173	cm12 176	cm13 170	cm14 173	cm15 172	cm16 174	cm17 171	cm18 171	cm19 172	cm20 173	cm21 169
m01 171	26	28	30	27	26	27	30	30	27	29	25	26	32	36	27	29	29	27	26	30	27
m02 172	29	30	35	30	29	29	33	33	33	31	29	31	35	36	31	34	31	32	31	31	30
m03 171	31	33	34	32	30	33	34	34	32	34	29	33	35	37	32	33	32	31	31	34	32
m04 171	27	31	35	32	31	33	33	34	33	33	32	31	36	39	33	35	32	31	30	36	34
m05 171	30	32	34	31	30	29	34	33	31	33	31	31	38	38	31	34	32	33	31	36	32
m06 171	31	33	34	33	33	33	34	36	34	35	31	31	36	38	31	35	34	33	30	36	34
m07 169	30	32	32	35	31	30	32	33	33	35	31	32	34	37	30	36	32	34	31	34	30
m08 170	30	32	32	32	29	30	32	32	32	32	31	32	35	36	31	34	30	32	28	32	29
m09 170	35	39	36	38	36	36	38	36	36	39	36	38	38	40	38	38	38	36	34	39	34
m10 168	31	31	35	35	31	31	33	33	35	34	32	33	34	38	34	37	30	34	32	34	31
m11 171	31	33	32	34	31	33	32	33	32	35	31	33	33	35	31	35	30	33	31	33	30
m12 166	36	37	37	38	36	35	36	37	37	39	35	37	37	38	34	41	36	38	33	39	33
m13 170	30	32	34	30	29	30	34	30	31	33	32	31	35	39	31	35	32	31	31	32	29

Table 12. Extended and non-extended start/stop-CLTs in alpha satellite HORs in human and chimpanzee chromosome 4

			Start/Stop CLTs		TGA				TAG				TAA				ATG			
	Chr.	Structure	nt all E&NE[(a)]	all E[(b)]	nt E&NE[(c)]	nt E[(d)]	nt NE[(e)]	r[(f)]	nt E&NE	nt E	nt NE	r	nt E&NE	nt E	nt NE	r	nt E&NE	nt E	nt NE	r
human	4	13mer	37	71	22	20	1	17	5	2	3	0	2	2	0	7	8	6	2	3
chimp	4	21mer	49	71	21	19	2	10	7	4	3	1	9	8	1	11	11	8	3	3

[(a)] nt all E&NE - percentage of nucleotides in all extended and non-extended start/stop-CLTs in genomic sequence; [(b)] all E - percentage of extended start/stop-CLTs with respect to all extended and non-extended start/stop-CLTs in genomic sequence; [(c)] nt E&NE - percentage of nucleotides in extended and non-extended stop-TGA CLTs in genomic sequence; [(d)] nt E - percentage of nucleotides in extended stop-TGA CLTs in genomic sequence; [(e)] nt NE - percentage of nucleotides in non-extended stop-TGA CLTs in genomic sequence; [(f)] r – quotient of nt E and nt NE. In the remaining 3×4 columns the analog results are given for the other three start/stop CLTs. If smaller than 1, the value of r-factor was rounded off to 0 according to Ref. 55.

In alpha satellite HORs in human chromosome 4 the extensions of start-ATG CLT are significantly smaller than extensions of stop-TGA CLT, and the extensions of stop-TAG CLT are absent (Table 12). The extensions of stop-TAA CLT are slightly larger, but still ≈2.4 times smaller then extensions of stop-TGA CLT. This is in accordance with dominant contribution from stop-TGA CLT found in Ref 55.

On the other hand, the extensions of stop-TGA CLT are sizably smaller and of stop-TAA CLT sizably larger in chimpanzee chromosome 4 alpha satellites: chimpanzee alpha satellites have by a factor ≈0.7 reduced extensions of stop-TGA CLT and by a factor of ≈1.6 increased extensions of stop-TAA CLT with respect to human alpha satellites (Table 12). If we compare these results with base to base divergences, we can conclude that the main difference between human and chimpanzee alpha satellites lies in extensions of codon-like trinucleotides. Having in mind that small and seemingly insignificant differences in a nonlinear network of genes and regulators could produce significant functional differences,[55] we hypothesize that these start/stop-CLTs differences could have been important factor in human evolution and separation of human from other primates.

CONCLUSION

We apply Global Repeat Map algorithm for identification and analysis of tandem and dispersed repeats in genomic sequences of all human chromosomes from Build 37.2 assembly. GRM analysis of repeats is done without use of any prior knowledge of a period or pattern and without involving any numerical parameter. The GRM analysis identifies HORs in the presence of insertions and/or deletions and gives a full list and structure of insertions, deletions and point mutations within HORs.

Applying GRM, in this paper, the whole available human genome sequence (Build 37.2 assembly) is investigated, and major alpha satellite higher-order repeats are identified. Fifteen different alpha satellite HORs in thirteen human chromosomes are identified, three of them novel, not reported previously. Detailed monomer scheme and consensus sequences for all alpha satellite HORs in human and chimpanzee chromosome 4 are identified and extensive study and research of their structure, development and evolutionary relationships are performed.

Supplementary Materials. – Supporting informations to the paper are enclosed to the electronic version of the article. These data can be found on the website of *Croatica Chemica Acta* (http://public.carnet.hr/ccacaa).

REFERENCES

1. H. F. Willard, *Curr. Opin. Genet. Dev.* **8** (1998) 219−225.
2. H. F. Willard and J. S. Waye, *Trends Genet.* **3** (1987) 192−198.
3. D. T. Murphy and G. H. Karpen, *Cell* **93** (1998) 317−320.
4. J. C. Lamb and J. A. Birchler, *Genome. Biol.* **4** (2003) 214.
5. P. E. Warburton, C. A. Cooke, S. Bourassa, O. Vafa, B.A. Sullivan, G. Stetten, G. Gimelli, D. Warburton, C. Tyler-Smith, K. F. Sullivan, G. G. Poirier, and W. C. Earnshaw, *Curr. Biol.* **7** (1997) 901−904.
6. H. F. Willard, *Curr. Opin. Genet. Dev.* **1** (1991) 509−514.
7. P. E. Warburton and H. F. Willard, *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, (1996) pp. 121−145.
8. I. A. Alexandrov, A. Kazakov, I. Tumeneva, V. Shepelev, and Y. Yurov, *Chromosoma* **110** (2001) 253−266.
9. R. Wevrick, V. P. Willard, and H.F. Willard, *Genomics* **14** (1991) 912–923.
10. H. F. Willard, *Amer. J. Hum. Genet.* **37** (1985) 524−532.
11. C. Tyler-Smith, *Development* **101** (1985) 93−100.
12. C. Tyler-Smith and W. R. A. Brown, *J. Mol. Biol.* **195** (1987) 457−470.
13. A. L. Jorgensen, C. J. Bostock, and A. L. Bak, *J. Mol. Biol.* **187** (1986) 185−196.
14. T. Haaf and H. F. Willard, *Genomics* **13** (1992) 122−128.
15. K. H. A. Choo, *The Centromere*, Oxford University Press, Oxford (1997).
16. H. Rosenberg, M. Singer, and M. Rosenberg, *Science* **200** (1978) 394–402.
17. G. Alves, H. N. Seuanez, and T. Fanning, *Chromosoma* **103** (1994) 262–267.
18. G. Alves, H. N. Seuanez, and T. Fanning, *Mol. Phylogenet. Evol.* **9** (1998) 220–224.
19. J. E. Horvath, L. Viggiano, B. J. Loftus, M. D. Adams, N. Archidiacono, M. Rocchi, and E. E. Eichler, *Hum. Mol. Genet.* **9** (2000) 113–123.
20. M. G. Schueler, A. W. Higgins, M. K. Rudd, K. Gustashaw, and H. F. Willard, *Science* **294** (2001) 109−115.
21. J. Guy, T. Hearn, M. Crosier, J. Mudge, L. Viggiano, D. Koczan, H. J. Thiesen, J. A. Bailey, J. E. Horvath, E. E. Eichler, M. E. Earthrowl, P. Deloukas, L. French, J. Rogers, D. Bentley, and M. S. Jackson, Genome Res. **13** (2003) 159−172.
22. M. K. Rudd and H. F. Willard, *Trends Genet.* **20** (2004) 529−533.
23. M. G. Schueler, J. M. Dunn, C. P. Bird, M. T. Ross, L. Viggiano, M. Rocchi, H. F. Willard, and E. D. Green, Proc. Natl. Acad. Sci. **102** (2005) 10563−10568.
24. A. E. Kazakov, V. A. Shepelev, I. G. Tumeneva, A. A. Alexandrov, Y. B. Yurov, and I. A. Alexandrov, Genomics **82** (2003) 619−627.
25. M. K. Rudd, G. A. Wray, and H. F. Willard, *Genome Res.* **16** (2006) 88−96.
26. J. M. Spence, R. Critcher, T. A. Ebersole, M. M. Valdivia, W. C. Earnshaw, T. Fukagawa, and C. J. Farr, *EMBO J.* **21** (2002) 5269−5280.
27. O. Vafa and K. F. Sullivan, *Curr. Biol.* **7** (1997) 897–900.
28. S. Ando, H. Yang, N. Nozaki, T. Okazaki, and K. Yoda, *Mol. Cell. Biol.* **22** (2002) 2229–2241.
29. J. J. Harrington, G. Van Bokkelen, R. W. Mays, K. Gustashaw, and H. F. Willard, *Nature Genet.* **4** (1997) 345−355.
30. M. Ikeno, B. Grimes, T. Okazaki, M. Nakano, K. Saitoh, H. Hoshino, N. I. McGill, H. Cooke, and H. Masumoto, *Nature Biotechnol.* **16** (1998) 431−439.

31. E. M. Southern, *J. Mol. Biol.* **94** (1975) 51−69.
32. G. P. Smith, *Science* **191** (1976) 528−535.
33. C. Alkan, E. E. Eichler, J. A. Bailey, S. C. Sahinalp, and E. Tuzun, *J. Comp. Biol.* **11** (2004) 933−944.
34. E. E. Eichler, R. A. Clark, and X. She, *Nat. Rev. Genet.* **5** (2004) 345−354.
35. R. J. Britten and D. E. Kohne, *Science* **161** (1968) 529−540.
36. R. J. Britten and E. H. Davidson, *Science* **165** (1969) 349−357.
37. R. J. Britten and E. H. Davidson, *Quart. Rev. Biol.* **46** (1971) 111−138.
38. E. H. Davidson and R. J. Britten, *Science* **204** (1979) 1052−1059.
39. L. A. Lettice, T. Horikoshi, S. J. H. Heaney, M. J. van Baren, H. C. van der Linde, G. J. Breedveld, M. Joosse, N. Akarsu, B. A. Oostra, N. Endo, M. Shibata, M. Suzuki, E. Takahashi, T. Shinka, Y. Nakahori, D. Ayusawa, K. Nakabayashi, S. W. Scherer, P. Heutink, R. E. Hill, and S. Noji, *Proc. Natl. Acad. Sci. USA* **99** (2002) 7548−7553.
40. J. A. Shapiro and R. von Sternberg, *Biol. Rev.* **80** (2005) 227−250.
41. T. Haaf and H. F. Willard, *Chromosoma* **106** (1997) 226−232.
42. T. Haaf and H. F. Willard, *Mammalian Genome* **9** (1998) 440−447.
43. R. Toder, Y. Xia, and E. Bausch, *Chromosomal Res.* **6** (1998) 487−494.
44. R. Toder, F. Grutzner, T. Haaf, and E. Bausch, *Chromosome Res.* **9** (2001) 431−435.
45. C. Alkan, M. Ventura, N. Archidiacono, M. Rocchi, S. C. Sahinalp, and E. E. Eichler, *PLOS Comput. Biol.* **3** (2007) e181.
46. C. Alkan, M. F. Cardone, C. R. Catcchio, F. Antonaccio, S. J. O'Brien, O. A. Ryder, S. Purgato, M. Zoli, G. Della Valle, E. E. Eichler, and M. Ventura, *Genome Res.* **21** (2011) 137−145.
47. F. S. Collins, A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, and L. Walters, *Science* **282** (1998) 682–689.
48. V. Paar, N. Pavin, M. Rosandić, M. Glunčić, I. Basar, R. Pezer, and S. Durajlija Žinić, *Bioinformatics* **21** (2005) 846−852.
49. V. Paar, I. Basar, M. Rosandić, and M. Glunčić, *Curr. Genomics* **8** (2007) 93−111.
50. V. Paar, M. Glunčić, I. Basar, M. Rosandić, P. Paar, and M. Cvitković, *J. Mol. Evol.* **72** (2011) 34−55.
51. V. Paar, M. Glunčić, M. Rosandić, I. Basar, and I. Vlahović, *Mol. Biol. Evol.* **28** (2011) 1877−1892.
52. M. Glunčić and V. Paar, *Nucleic Acids Res.* **2012** 1−17.
53. M. Rosandić, V. Paar, M. Glunčić, I. Basar, and N. Pavin, *Croat. Med. J.* **44** (2003b) 386−406.
54. M. Rosandić, V. Paar, I. Basar, M. Glunčić, N. Pavin, and I. Pilaš, *Chromosome Res.* **14** (2006) 735−753.
55. M. Rosandić, V. Paar, and M. Glunčić, *Croat. Chem. Acta* **84** (2011) 331−341.
56. G. Benson, *Nucleic Acids Res.* **27** (1999) 573−580.
57. P. E. Warburton, D. Hasson, F. Guillem, C. Lescale, X. Jin, and G. Abrusan, *BMC Genomics* **9** (2008) 533.
58. S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* **48** (1970) 443–453.
59. P. E. Warburton, J. S. Waye, and H. F. Willard, *Mol. Cell. Biol.* **13** (1993) 6520–6529.
60. T. Mashkova, N. Oparina, I. Alexandrov, O. Zinovieva, A. Marusina, Y. Yurov, M. H. Lacroix, and L. Kisselev, *FEBS Lett.* **441** (1998) 451–457.
61. L. Y. Romanova, G. V. Deriagin, T. D. Mashkova, I. G. Tumeneva, A. R. Mushegian, L. L. Kisselev, and I. A. Alexandrov, *J. Mol. Biol.* **261** (1996) 334−340.
62. C. Gaff, D. du Sart, P. Kalitsis, R. Iannello, A. Nagy, and K. H. Choo, *Hum. Mol. Genet.* **3** (1994) 711–716.
63. J. Ohzeki, M. Nakano, T. Okada and H. Masumoto, *J. Cell. Biol.* **159** (2002) 765–775.
64. P. E. Warburton, *Chromosome Res.* **12** (2004) 617–626.
65. H. Masumoto, H. Masukata, Y. Muro, N. Nozaki, and T. Okazaki, *J. Cell Biol.* **109** (1989) 1963−1973.
66. H. Masumoto, M. Nakano, and J. Ohzeki, *Chromosome Res.* **12** (2004) 543−556.
67. H. E. Trowell, A. Nagy, B. Vissel, and K. H. Choo, *Hum. Mol. Genet.* **2** (1993) 1639–1649.
68. M. Ikeno, H. Masumoto, and T. Okazaki, *Hum. Mol. Genet.* **3** (1994) 1245–1257.
69. T. Haaf, A.G. Mater, J. Weinberg, and D. C. Ward, *J. Mol. Evol.* **41** (1995) 487−491.
70. M. Rosandić, M. Glunčić, and V. Paar, *J. Theor. Biol.* 2012 http://dx.doi.org/10.1016/j.jtbi.2012.09.022