

Efficient genome monomer higher order structure annotation and identification using the GRMhor algorithm

Matko Glunčić^{1*}, Domjan Barić¹, Vladimir Paar^{1,2}

¹Faculty of Science, University of Zagreb, 10000 Zagreb, Croatia.

²Croatian Academy of Sciences and Arts, 10000 Zagreb, Croatia.

*corresponding author: matko@phy.hr

ABSTRACT

Tandem monomeric units, integral to eukaryotic genomes, form higher-order repeat (HOR) structures with dual structural and functional significance. The recent complete assembly of the human genome (T2T-CHM13) provides an unparalleled opportunity to study these repeats, which, due to their complex structure, were previously under-sequenced. Here, we introduce the GRMhor algorithm, capable of identifying canonical and variant HORs within tandem sequences. Utilizing a concept akin to Southern blotting, extended to monomeric space, the algorithm visually represents HORs through diagrams and aligned schemes. To elucidate the newly discovered types of HORs derived from our analysis, we introduce two fundamental categories: Willard's HORs, distinguished by the presence of various monomer types within each HOR copy, and cascading HORs, characterized by the repetition of specific monomer types within canonical HOR units. We apply GRMhor to all monomeric alpha satellite arrays in the T2T-CHM13 human chromosome 20 assembly, revealing six distinct HOR arrays, including cascading 16mer, cascading 11mer, and conventional Willard's type 8mer HORs. Additionally, we identify the cascading 8mer HOR, cascading 26mer HOR, and highly variant 18mer HOR. The analysis unveils the intricate architecture of centromeric HORs, elucidating their organization and evolution, with potential implications for chromosome segregation and stability.

INTRODUCTION

Monomer arrays, typically located in heterochromatin, play crucial roles in forming essential chromosome structures such as centromeres and telomeres (1). Despite their significance in these pivotal structures, monomers exhibit remarkable variation in both sequence and copy number across species, even among close relatives (2), indicating rapid evolutionary changes. Several models of monomer evolution have been proposed to account for this variation, yet genome-wide testing of these models has been hampered by technological and computational limitations in assessing the repetitive genome portion. Understanding the mechanisms driving monomer DNA variation among individuals and species is crucial, given the established associations between monomers and phenotypes in diverse organisms, including humans (3). For instance, monomer derepression is linked to cancer outcomes (4), chromosome mis-segregation, aneuploidy (5), and aging (6). Furthermore, variation in monomer copy number has been associated with genetic incompatibilities between species (7), differences in gene expression (8-10) and evolutionary development (11-13). Due to the aforementioned facts, the identification and analysis of various types of monomers have arisen as subjects of considerable interest. Nevertheless, the examination of human monomer DNA and RNA poses diverse challenges, emphasizing the necessity for technological advancements to enhance our comprehension of this predominantly unexplored portion of the genome (13,14).

Monomer arrays are composed of primary repeat units, which consist of divergent monomers arranged in a head-to-tail configuration. Individual monomers exhibit a sequence divergence of 20-40%. However, the majority of monomers are organized hierarchically into higher-order repeats, secondary repeat units, in which the monomers repeat as structures with high sequence identity (>95%) (1,15-21). As depicted in Figure 1, within a single HOR, all monomers exhibit a variation of 20-40%, whereas corresponding pairs of monomers across different HORs display less than 5% variation.

The most prevalent HOR copy with n constituting monomers is termed canonical n mer HOR (3mer HOR in Fig 1). HOR units within the same HOR array that contain inserts or deletions compared to the canonical HOR unit are known as variants HOR units (for instance, HOR2 with $t4$ insertion and HOR4 with $t2$ deletion in Fig1).

In this paper, we introduce two basic types of HORs: (i) Willard's HORs, where monomers of different types are found within each HOR copy (HOR1, HOR2, HOR3, and HOR4 in Fig 1). (ii) Cascading HORs, where specific monomer types are reiterated within a canonical HOR copy (HOR5 in Fig1).

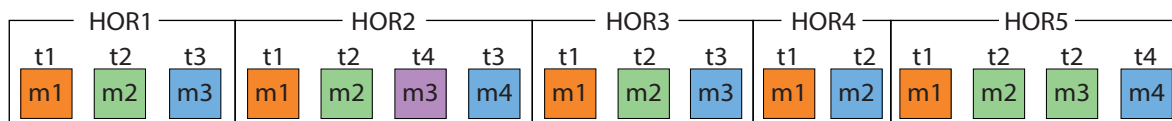


Figure 1. Schematic representation of a monomer array and HORs. Each monomer is represented by a single square. Monomers within HOR unit are labelled as $m1, m2, \dots$, in order of their appearance (from left to right within each HOR). Monomers exhibiting <5% sequence divergence are depicted in the same color and labeled with the same identifier ($t1, t2, \dots$). A group of three monomers is sequentially repeated to form a higher-order structure known as a 3mer canonical HOR. HOR2, HOR4 and HOR5 are variant HORs due to the insertion (monomer $t4$ in HOR2, monomer $t2$ in HOR5) and deletion (monomer $t2$ in HOR4) of one monomer.

Monomer HORs in human and nonhuman primates were initially identified through hybridization techniques (15-17,22-25), and subsequently by bioinformatics tools. While various existing software applications effectively identify regions with tandem repeats (26-34), they fall short of providing precise annotations for individual repeat locations or HORs. Similarly, more recent tools designed for annotating human HORs within genomic sequences (34-39) have limited broader applicability (21). On the other hand, a specific set of software has been developed for the accurate identification of Willard's type HORs (21,40-42). In the context of the complete assembly of human chromosomes, alpha satellite HORs were initially computed using the NTRprism algorithm (43), which bears resemblance to the 2007 version of GRM (40).

Here we introduce our novel GRMhor algorithm and its accompanying application, designed to identify all HORs, including both canonical and variant types, as well as Willard's and Cascading HORs, within monomeric tandem sequences, and graphically display them in the form of diagrams (see Figure 2) and aligned schemes. The algorithm consists of three complementary components: the GRM diagram, which is based on the concept of the traditional Southern blotting molecular biology technique extended to the monomeric space; the Monomer Distance diagram (MD diagram), which precisely depicts the spatial distribution of periods of monomeric repetitions within the monomeric array; and a aligned schematic representation of HORs array, providing an in-depth visualization of the organization and arrangement of monomers within sequences into HOR structures.

In this study, we report the outcomes of utilizing the GRMhor algorithm for analyzing alpha satellite monomers. Furthermore, in the Discussion section, we provide insights into its application on the Neuroblastoma Break Family monomers as reported in our referenced articles (44,45). Notably, the GRMhor algorithm demonstrates equal efficacy in identifying and analyzing HOR structures of any type of monomer across various genomic sequences.

MATERIALS AND METHODS

In parallel, we will describe the working principles of all three parts of algorithm as they are integrated into a single GRMhor application that utilizes the same input data. Throughout the text, our focus will be on alpha satellite monomers, although the algorithm and application perform equally well for any monomeric repetitions.

Algorithm outline

In the first step, we construct an N -dimensional array, $\mathbf{M} = \{\mathbf{m}^1, \dots, \mathbf{m}^N\}$, consisting of two-dimensional vectors

$$\mathbf{m}^i = (m_1^i, m_2^i), i \in [1, N] \quad (1)$$

where N is the length of the input monomeric array. The first component of each vector in the array represents the monomer's position in the sequence ($m_1^i = i$), while the second component represents the distance of the monomer at position i to the first adjacent monomer in the sequence that differs from it by less than 5% ($m_2^i = \text{position of first similar monomer} - i$) (Fig. 2). Similarities (differences) between monomers are calculated using the Needleman-Wunsch algorithm (46), or alternatively, using the Edlib (47) algorithm. For example, if we consider the i -th monomer and find that the $i + 1$, $i + 2$, and $i + 3$ monomers differ from it by more than 5%, but the $i + 4$ monomer differs by less than 5%, then $m_2^i = i + 4 - i = 4$.

In the second step, we construct new L -dimensional array, $\mathbf{P} = \{\mathbf{p}^j, \dots, \mathbf{p}^L\}$, consisting of two-dimensional vectors

$$\mathbf{p}^j = (p_1^j, p_2^j), j \in [1, L] \quad (1)$$

where L is the maximum distance between any two similar monomers (differs < 5%). The first component of the new vector represents the distance between two similar monomers ($p_1^j = j, j = 1, \dots, L$), while the second component represents the frequency of occurrence of this distance in the N -dimensional array \mathbf{M} , $p_2^j = \sum_{i=1}^N \delta(p_1^j, j)$, where $\delta(p_1^j, j)$ represents the delta function.

In the third step, using the array \mathbf{M} , we form groups of monomers such that each group contains monomers differing from each other by less than 5%, and assign each group a name starting from the first, $m1$, to the last, mk . This way, each monomer in the group is assigned a name, thereby determining its position in the scheme of structural monomer distribution in the third algorithm.

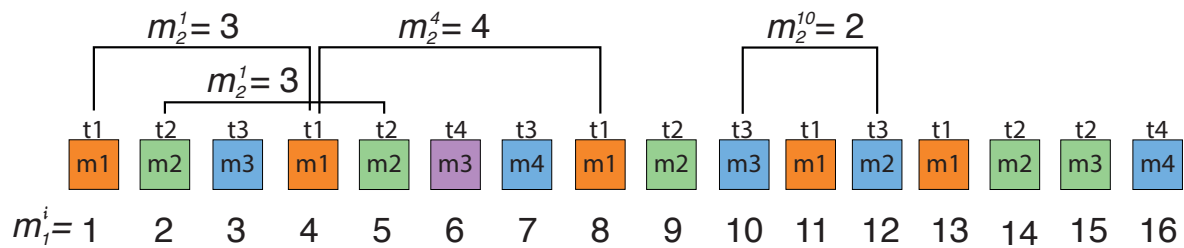


Figure 2. Scheme of an example monomer sequence with 16 monomers, illustrating the first step of the algorithm. The array \mathbf{M} consists of 16 two-dimensional vectors, $\mathbf{M} = \{(1,3), (2,3), (3,4), (4,4), (5,4), (6,0), (7,3), (8,3), (9,5), (10,2), (11,2), (12,4), (13,0), (14,1), (15,0), (16,0)\}$. Each monomer is represented by a single square. Monomers within HOR unit are labelled as $m1, m2, \dots$, in order of their appearance (from left to right within each HOR). Monomers exhibiting <5% sequence divergence are depicted in the same color and labeled with the same identifier ($t1, t2, \dots$).

Finally, utilizing the data obtained from the previous steps, we generate two graphs and a schematic representation: (i) the GRM diagram, where we plot the repeat period of monomers, p_1^j , on the x-axis, and the frequency of occurrence of each repeat period in the monomeric sequence, p_2^j , on the y-axis; (ii) the MD diagram, where we plot the ordinal number of monomers in the sequence, m_1^i , on the x-axis, and the distance to the first similar monomer in the sequence, m_2^i , on the y-axis; (iii) a aligned schematic representation of the organization of monomers in the sequence, where all monomers from the same group in step three are placed in the same column, sharing the same x-coordinate. In the graphical representation, these monomers are depicted by squares of the same color. The squares (monomers) are arranged from left to right and top to bottom according to their appearance

index in the sequence, m_1^i , with the condition that when a monomer from the same group appears in the same row, its y-coordinate is increased by one, causing it to move to a new row to ensure placement in a column with monomers from its group.

Application usage and output

The input data for our algorithm consists of a series of tandem monomers, which can be obtained in various ways. For the case study presented in the following text, we employed our MonFinder tool (<https://github.com/domjanbaric/GRMhor/tree/main>), which takes genomic sequences (subject) and consensus sequence (query) as input and delivers a list of detected monomers. This algorithm utilizes the Edlib open-source C/C++ library for precise pairwise sequence alignment (47). Within the MonFinder algorithm, the subject sequence is searched in both the direct and reverse complement directions to identify all monomers. In this study, a unique consensus sequence of 171 base pairs (bp) in length (the consensus sequence is located within the MonFinder code on GitHub), derived from over 1,000,000 different alpha satellites across all higher primates, including humans, was utilized as a query for detecting all alpha satellites in the genomic sequence under investigation. In a similar manner, a variety of different tools can be utilized, for instance BLASTN algorithm (48).

The Python program GRMhor (<https://github.com/domjanbaric/GRMhor/tree/main>) is executed with a file containing a sequence of monomers as the input parameter and optional additional parameters such as the starting monomer in the sequence (default = 0), the maximum value of the displayed period (default = 60), and printing the genomic position of the first monomer in the HOR. After loading the monomer array, the application autonomously proceeds through the steps described in the Algorithm outline, ultimately generating a GRM diagram, MD diagram, and aligned schematic representation of the monomer organization in the array of monomers (Figure 3). Each generated visualization is automatically saved in three distinct .ps files in the initial directory. In the following chapter, through several case studies, first with artificial arrays of monomers, and then with monomers from real sequences of the human genome (T2T CHM13), we will elucidate how to interpret each of these visualizations and easily identify and analyze HORs types, organization and structure in detail.

RESULTS

In the following four artificial case studies, we utilized actual monomers from of the T2T-CHM13 assembly of human chromosome 3, selecting 10 distinct alpha satellites (with a mutual difference > 20%), to construct various artificial monomer arrays. All artificial monomer arrays are available for testing on <https://github.com/domjanbaric/GRMhor/tree/main>. Each of these artificial arrays was then subjected to our algorithm, with a detailed discussion of the results provided. Subsequently,

we conducted an analysis of the entire real sequence of the T2T-CHM13 assembly of human chromosome 20. In the following text, when we use the term "similar monomers" we are referring to a difference between monomers that is less than 5%.

Case study: artificial sequence of Willard's canonical alpha satellite HORs

We replicated a set of ten distinct monomers ten times, resulting in a sequence of 100 monomers, where each monomer possesses ten identical copies. The analysis result is depicted in Fig 3. In the GRM diagram (Fig. 3a), a distinct peak corresponding to a period of ten is observed, providing clear evidence that all similar monomers are spaced at a distance of ten monomers from each other. The same conclusion can be drawn from the MD diagram (Fig. 3b), where each point represents a vector (m_1^i, m_2^i) , namely a function of the monomer's position in the sequence and the distance to the first similar monomer. All points lie on the ordinate $y = 10$, indicating that any two similar monomers are situated at a distance of ten monomers in the monomeric array. Together, we can conclude that our monomers form a 10-order HOR, i.e., a 10mer HOR, as also evident from the schematic representation of the organization of monomers in Fig. 3c. To facilitate the description of more complex structures in the following case studies, we will introduce two distinct labels for monomers within the HOR unit (Fig. 3c). With the label $t\tau$, we will denote all similar monomers at position τ within the HOR unit, while with the label mn , we will denote the ordinal number of the monomer within the HOR unit. In Willard-type HORs, the labels of these two designations for each monomer within the HOR unit are identical ($\tau = n$) (Fig. 3c). In the MD diagram the last ten monomers (Index 91-100) exhibit a period of zero as none of them finds a similar monomer to the end of the sequence.

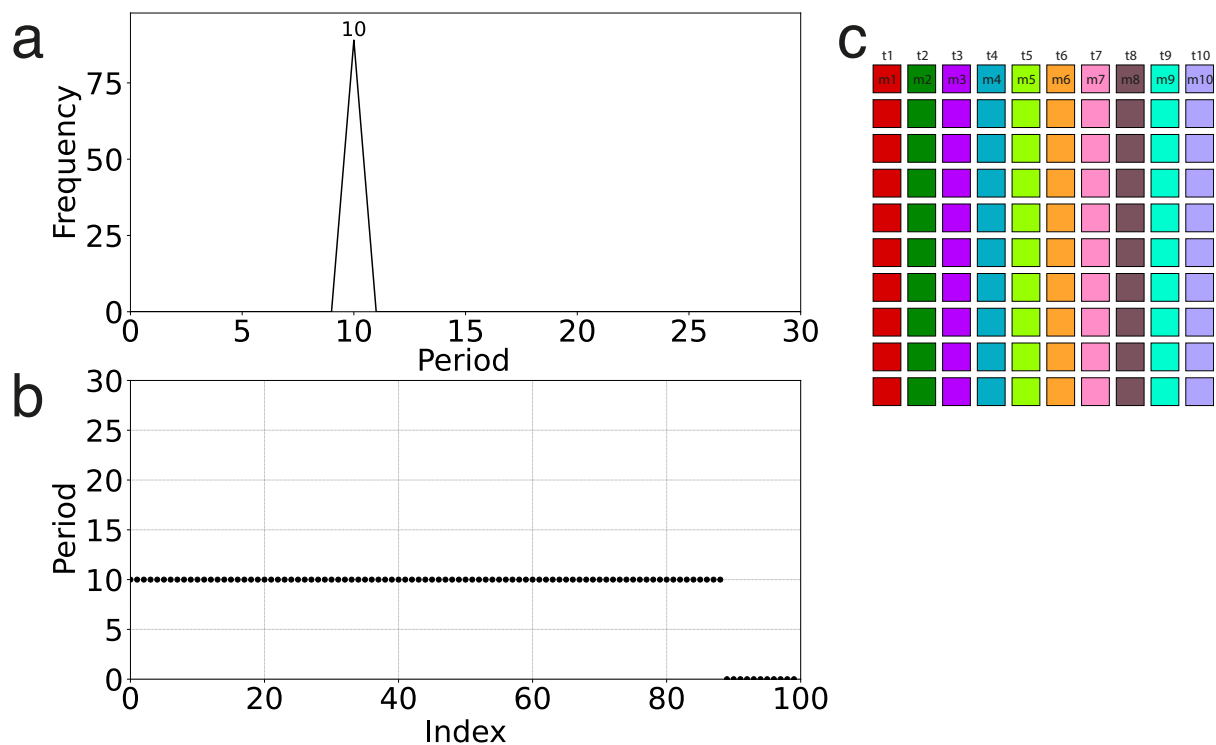


Figure 3. Resulting diagrams and higher-order repeat (HOR) scheme for 10 perfect Willard HORs of length 10 monomers. (a) GRM diagram. (b) MD diagram. Period denotes the distance between two similar monomers in monomer units. Index denotes the ordinal number of the monomer in the monomeric array. **(c) Aligned scheme for Willard's HOR alignment ($n = 10, \tau = 10$) (10 monomers of 10 different types).** Monomers within HOR unit are labelled as m1, m2, ... m10, in order of their appearance from left to right within each row and from top to bottom). Each monomer is depicted by a coloured box, with distinct colours corresponding to different monomer types. Monomers are organized into columns based on their monomer types: monomer type t1 in the first column, monomer type t2 in the second column, and so forth. The number of columns, i.e., the number of different monomer types in the canonical HOR unit, is denoted by τ .

Case study: artificial sequence of Willard's canonical and variant alpha satellite HORs

In the sequence of 100 monomers from the previous case study (subsection 3.1), we made modifications by deleting the 18th, 19th, 38th, and 39th monomers and inserting two new monomers (distinct from the initial ten) after the 66th and 86th monomers (see Fig. 4c). This was done to simulate variant Willard's HORs with deletions and insertions. The dominant peak on the GRM diagram (Fig 4a) remains at a period of 10, albeit with a slightly lower frequency. New peaks emerge at periods, in order of frequency, 12, 8, 18, and 22. In the MD diagram (Fig 4b), alongside the highest concentration of points distributed at $y = 10$, new sequences of points also appear at the corresponding new periods.

The peaks at periods 8 and 18 correlate with an additional set of points on the left side of the MD diagram, indicating that these periods result from the emergence of new HOR variants through the deletion of monomers. Fig. 4c reveals that the first seven monomers in the second, variant HOR (second row in Fig 4c) now repeat not after 10, but after 8 monomers, due to the absence of the deleted monomers t_7 and t_8 in this HOR. The same pattern is observed with the fourth, variant HOR unit. Consequently, two sets of eight points at period $y = 8$ appear on the MD diagram. Furthermore, monomers t_7 and t_8 in the first, canonical HOR unit lack similar copies in the second, variant HOR unit, and their similar copies are only found in the third, canonical HOR unit, repeating after $8 + 10 = 18$ monomers. A similar scenario applies to monomers t_7 and t_8 in the third, canonical HOR unit. Consequently, two sets of two points at period $y = 18$ appear on the MD diagram.

The peaks at periods 18 and 22 in the GRM diagram and the series of points at the same ordinates in the MD diagram are the result of the insertion of two new monomers in variant HOR units. It is evident that in these HOR units, due to the two additional monomers, the first six monomers are repeated only after 12 monomers. Additionally, the first pair of two additional monomers finds its similar monomers only after a sequence of $5 + 10 + 7 = 22$ monomers. The second pair of two monomers does not find similar monomers until the end

of the sequence; therefore, in the MD diagram, we only have one set of two points at a $y = 22$.

In conclusion, it is quite straightforward from the GRM diagram and the MD diagram to conclude that, in general, we are dealing with a n mer canonical HOR of Willard's type, along with some variant HORs obtained through deletions or insertions of new monomers.

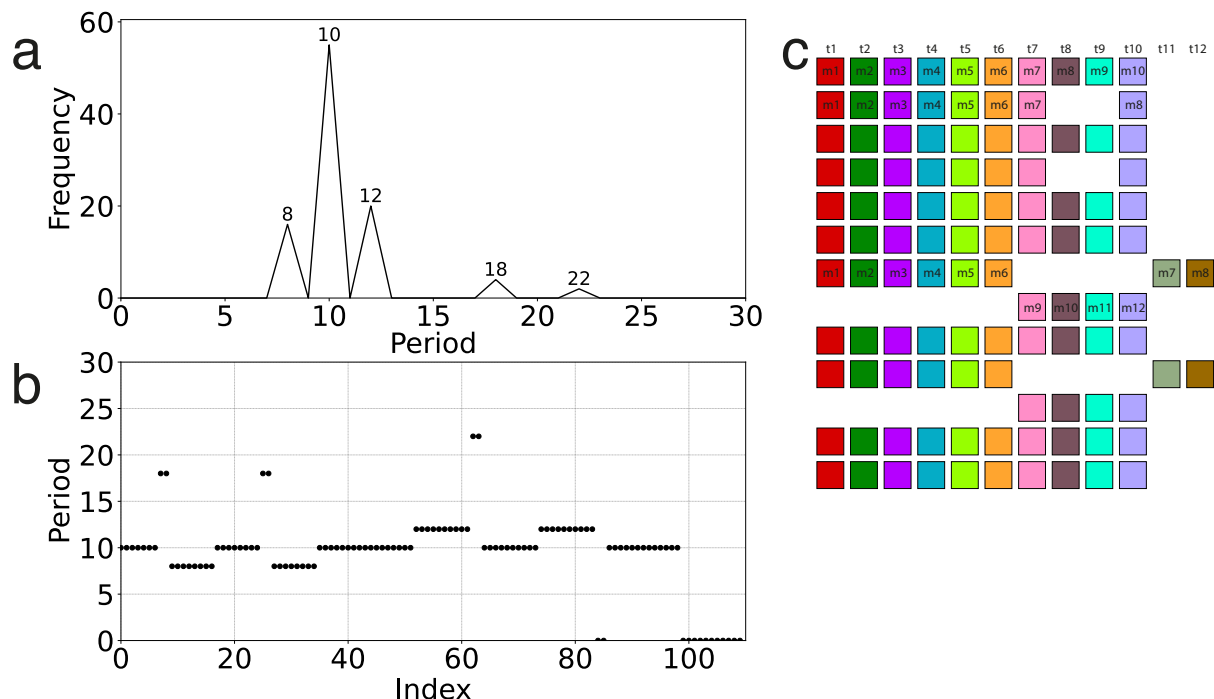


Figure 4. Resulting diagrams and higher-order repeat (HOR) scheme for 6 Willard's canonical and 4 variant HORs. (a) GRM diagram. (b) MD diagram scheme. Period denotes the distance between two similar monomers in monomer units. Index denotes the ordinal number of the monomer in the monomeric array. (c) Aligned scheme for Willard's canonical and variant HOR alignment ($n = 12, \tau = 12$) (12 monomers of 12 different types). Monomers within HOR unit are labelled as $m1, m2, \dots, m12$, in order of their appearance (from left to right within each row and from top to bottom). Each monomer is depicted by a coloured box, with distinct colours corresponding to different monomer types. Monomers are organized into columns based on their monomer types: monomer type $t1$ in the first column, monomer type $t2$ in the second column, and so forth. The number of columns, i.e., the number of different monomer types in the canonical HOR unit, is denoted by τ .

Case study: artificial sequence of cascading alpha satellite canonical HORs

In the sequence of 100 monomers from the previous case study (subsection 3.1), we made modifications by inserting monomer $t2$ into each HOR after monomer $t6$, so that the monomeric sequence in each HOR resembles consensus HOR shown in Fig. 5c. Now, the

dominant peak in the GRM diagram is at period 11, with significantly lower peaks at periods 5 and 6 (Fig 5a). Accordingly, the majority of points in the MD diagram are also found at period 11, with fewer at periods 5 and 6 (Fig 5b). From the scheme in Fig. 5d, it is clear that all monomers, except m_2 , in this situation encounter a similar monomer after 11 monomers. Furthermore, the first copy of m_2 in each HOR encounters a similar monomer after 5 other monomers, and the second copy of m_2 in each HOR encounters a similar monomer in the next HOR after 6 monomers. This accounts for the peaks at 5 and 6 in the GRM diagram, or the points at $y = 5$ and $y = 6$ in the MD diagram. Altogether, both diagrams clearly indicate an 11mer Cascading HOR with a duplicated single similar monomer.

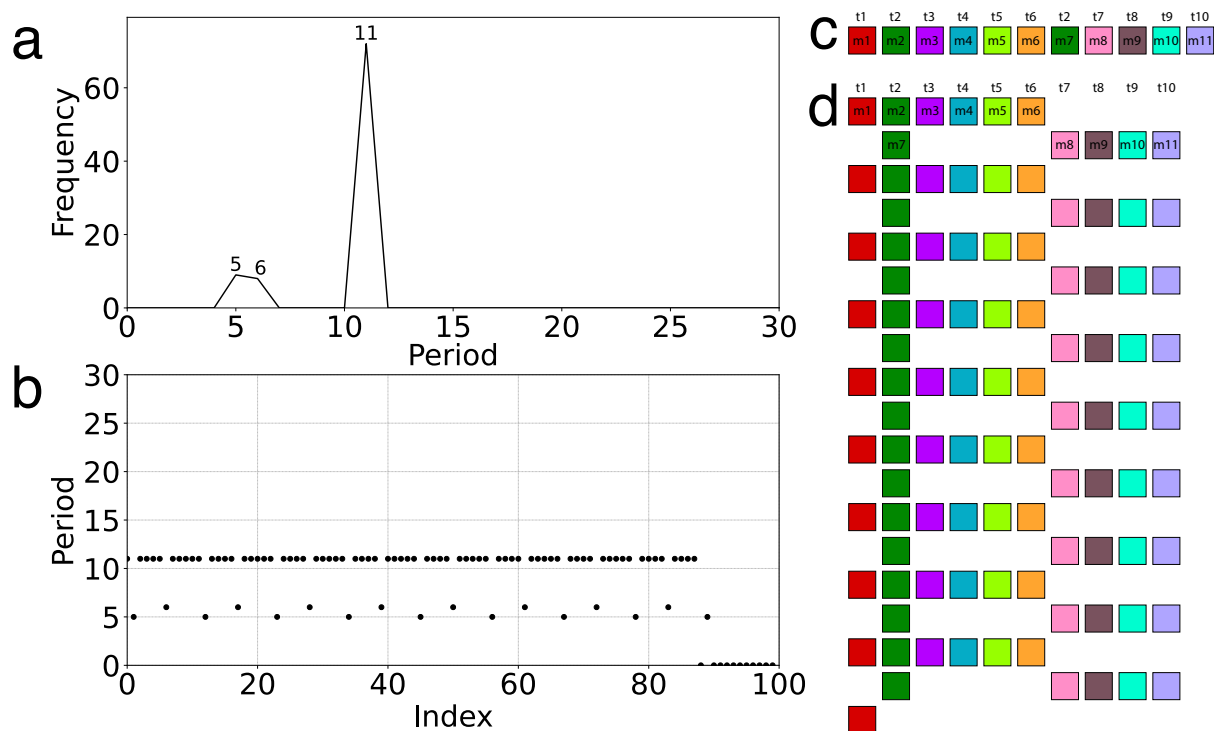


Figure 5. Resulting diagrams and higher-order repeat (HOR) scheme for 10 Cascading HORs. (a) GRM diagram. (b) MD diagram. Period denotes the distance between two similar monomers in monomer units. Index denotes the ordinal number of the monomer in the monomeric array. **(c) Aligned scheme for Cascading HOR alignment ($n = 10, \tau = 11$) (10 monomers of 11 different types).** Monomers within HOR unit are labelled as m_1, m_2, \dots, m_{10} , in order of their appearance (from left to right within each row and from top to bottom). Each monomer is depicted by a coloured box, with distinct colours corresponding to different monomer types. Monomers are organized into columns based on their monomer types: monomer type t_1 in the first column, monomer type t_2 in the second column, and so forth. The number of columns, i.e., the number of different monomer types in the canonical HOR copy, is denoted by τ .

Case study: artificial sequence of randomly distributed alpha satellite monomers

As our final artificial case study, from an initial sample of 10 distinct monomers, we constructed a series of 100 tandem monomers by duplicating them randomly using Python's default random number generator based on the Mersenne Twister algorithm. Considering that the examined sequence resulted from the random duplication of 10 initial monomers, we anticipate that the distribution of peaks in the GRM diagram will be highest at small periods. Both the GRM diagram and the MD diagram in this instance indicate a lack of higher-order organization, which is further illustrated in the schematic in Figure 6c.

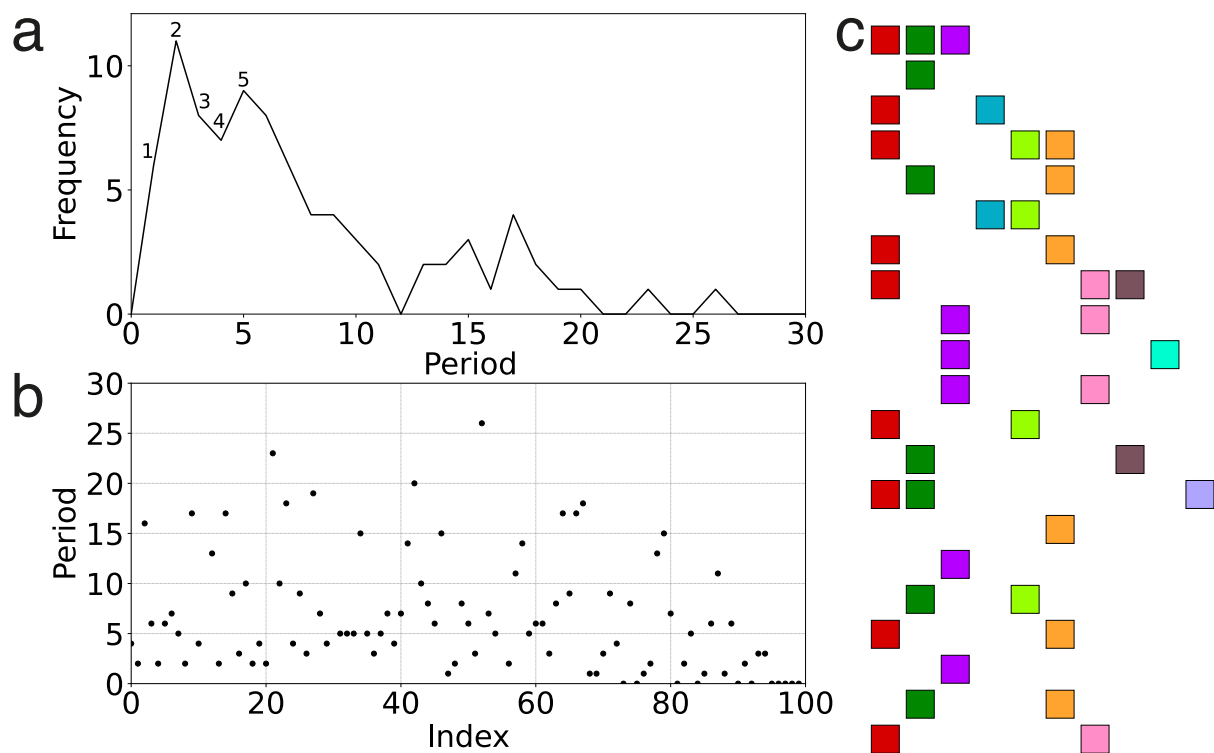


Figure 6. Resulting diagrams and higher-order repeat (HOR) scheme for artificial sequence of randomly distributed monomers. (a) GRM diagram. (b) MD diagram. Period denotes the distance between two similar monomers in monomer units. Index denotes the ordinal number of the monomer in the monomeric array. (c) **Aligned scheme for randomly distributed monomers.**

Case study: Alpha satellite monomers HORs in the T2T-CHM13 assembly of human chromosome 20

Using our MonFinder algorithm, we have isolated all alpha satellites in the T2T-CHM13 assembly of human chromosome 20. As a result, we identified 24,128 alpha satellites, with the majority located in several blocks of tandem repeats.

From the MD diagram (Fig. 7a), we can straightforwardly identify six distinct HOR regions that generate various prominent peaks on the GRM diagram (Fig. 7b). Region A comprises 8mer Willard HORs with a minor proportion of variant HORs, region B consists of 16mer Cascading HORs, region C contains 11mer Cascading HORs, region D encompasses 8mer Cascading HORs, region E comprises highly variant 18mer Cascading HORs, and region F contains 26mer Cascading HORs. In regions containing multiple variant HORs, we determine the dominant *n*meric HOR based on the highest number of dots at a specific period in the MD diagram and the most frequent pattern in the schematic representation (Supplementary Figures). A comprehensive schematic representation of the HORs, along with the first monomer positions of HORs within the genomic sequence, is provided in the Supplementary Materials due to the extensive lengths of the sequences (Supplementary Fig. S1-S6, for regions A-F, respectively). For a clearer presentation of each aligned scheme in supplementary Figs. S1-S6, individual blocks of monomers were extracted from each region according to the indices in the MD diagram and reprocessed through the GRMhor algorithm.

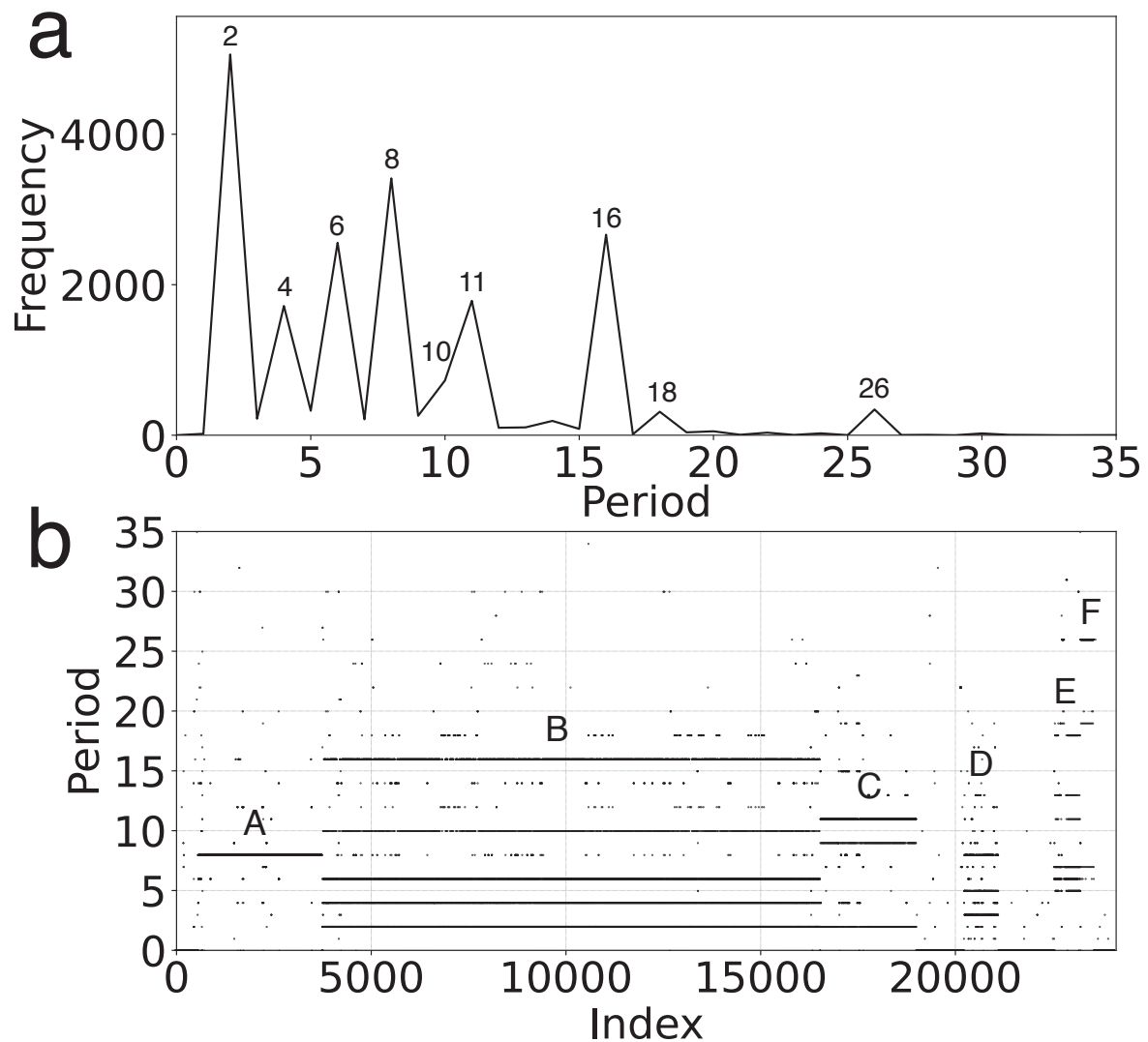


Figure 7. (a) Global Repeat Map (GRM) diagram for tandemly arranged alpha satellite monomers in the complete T2T-CHM13 assembly of human chromosome 20. Horizontal

axis: GRM periods (in monomer units). Vertical axis: frequency of monomer repeats period. Identified GRM peaks exhibit periods 2, 4, 6, 8, 10, 11, 16, 18 and 26. The significance of these GRM peaks (HORs or associated subfragment repeats) can be inferred from the Monomer Distance (MD) diagram. **(b) MD diagram.** Horizontal axis: enumeration of tandemly organized alpha satellite monomers, in sequential order as revealed by GRM analysis of the T2T assembly. Vertical axis: period (the distance between start of a monomer and of the next monomer of the same type (see Fig. 2)). Four distinct regions of monomer tandems are denoted A, B, C, D, E and F. Additionally, there are sporadic MD points that do not correspond to HORs or their subfragments.

We will provide concise remarks on each of the HOR units, the peaks they generate in the GRM diagram, and the distribution of points on the MD diagram, utilizing representative samples from each region (see Fig. 8). In region A (Fig. 8a, Fig. S1), the HOR units are predominantly Willard's consensus HORs, and it is evident from the MD diagram that they generate peak 8 in the GRM diagram.

The HOR units in region B are Cascading 16mer HORs (Fig. 8b, Fig. S2), consisting of a large number of duplications of monomers t_1 and t_2 (red and green squares in Fig. 8b and Fig. S2). These duplications of two monomers within the same HOR result in peaks at periods 2, 4, and 6. The peak and series of points at period 10 arise from variant HOR units in this region, occasionally involving the deletion of 6 monomers from canonical HOR unit ($m_9, m_{10}, m_{11}, m_{12}, m_{13}$, and m_{14}).

The HOR units in region C are Cascading 11mer HORs with a smaller number of variant copies. Consequently, these HOR units generate two additional peaks, at periods 9 and 2. Specifically, due to the duplication of monomer t_9 , the first copy of t_9 (m_9) repeats after two monomers, while the second copy of t_9 (m_{11}) repeats after 9 monomers (see Fig. 8c). All other monomers repeat after 11 copies, making the peak at period 11 the most prominent in this region.

The HOR units in region D are Cascading 8mer HORs with duplicated monomers t_1 and t_2 . Consequently, these HORs, with a dominant peak in the GRM diagram and the densest distribution of points in the MD diagram at period 8, also generate a peak at period 5 because both monomers (t_1 and t_2) repeat for the first time after 5 monomers (m_1 and m_2), and a peak at period 3 because both monomers repeat again after 3 monomers (m_6 and m_7). These two peaks (3 and 5) are not prominent in the GRM diagram due to the short length of the region occupied by this HOR compared to other regions. However, increased distributions of points at periods 3 and 5 are clearly visible in region D in the MD diagram.

The highly Cascading 18mer HOR units in the E region exhibit significant complexity, featuring five duplicated monomers, namely $t_1(\times 2), t_4(\times 2), t_5(\times 3), t_6(\times 3)$, and $t_8(\times 2)$. In addition to the dominant peak at a period of 18, different combinations of these duplicated monomers within the same HOR and across neighboring HORs result in an increased

distribution of points at periods 13 (e.g., $t8: m14$ in $m9$ of the adjacent HOR), 11 (e.g., $t1: m8$ in $m1$ of the subsequent HOR), 7 (e.g., $t1: m1$ in $m8$), 6 (e.g., $t5: m5$ in $m11$ and $t6: m6$ in $m12$), and 5 (e.g., $t8: m9$ in $m14$) in the MD diagram.

In the region F, a complete set of canonical 26mer Cascading HOR units with only one recurring monomer ($t1$) is found. In addition to the dominant peak and increased point distribution at the 26 period, the first repetition of the similar monomer $t1$ ($m1$ in $m20$) generates additional points at the 19 period (Fig. 7b), while the second repetition of the similar monomer $t1$ ($m20$ in $m1$ in the subsequent HOR unit) generates additional points at the 7 period (Fig. 7b).

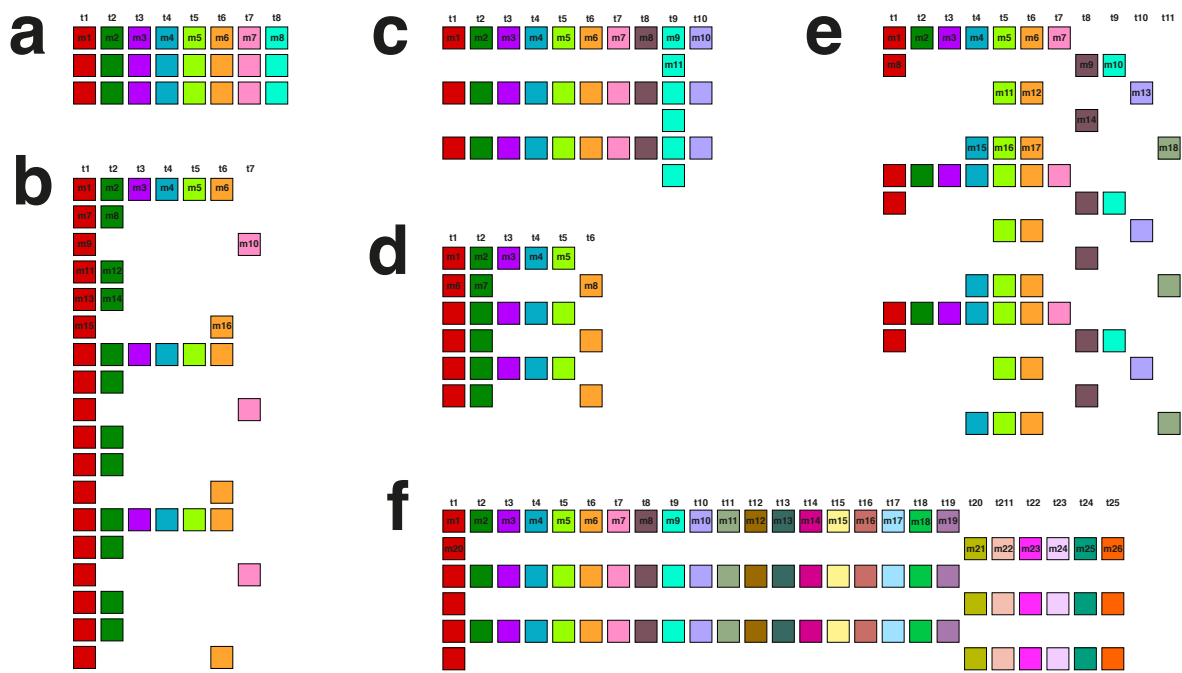


Figure 8. Aligned schemes of the three selected tandem canonical HOR units from all HOR regions in the T2T-CHM13 assembly of human chromosome 20. (a) 8mer Willard-type HOR in region A. (b) 16mer Cascading HOR in region B. (c) 11mer Cascading HOR in region C. (d) 8mer Cascading HOR in region D. (e) 18mer Cascading HOR in region E. (f) 26mer Cascading HOR in region F. Monomers within HOR copy are labelled as $m1, m2, \dots, mn$, in order of their appearance (from left to right within each row and from top to bottom). Each monomer is depicted by a coloured box, with distinct colours corresponding to different monomer types. Monomers are organized into columns based on their monomer types: monomer type $t1$ in the first column, monomer type $t2$ in the second column, and so forth. The number of columns, i.e., the number of different monomer types in the canonical HOR copy, is denoted by τ .

Table 1. Alpha satellite HOR arrays in T2T-chm13 assembly of human chromosome 20 determined using GRMhor algorithm.

HOR	n	τ	No. of HOR copies	No. of canonical HOR copies	No. of variant HOR copies	Type of HOR
8mer	8	8	388	361	27	Willard's type
16mer	16	7	805	674	131	Cascading
11mer	11	10	235	203	32	Cascading
8mer	8	6	107	69	38	Cascading
18mer	18	11	37	21	16	Cascading
26mer	26	25	14	13	1	Cascading

n denotes number of monomers in canonical n mer HOR and τ denotes the number of different monomer types in canonical Cascading n mer HOR. The scheme of all HOR copies identified by GRMhor algorithm are presented in Supplementary Figures S1-6.

DISCUSSION

In artificial case studies, we demonstrated that the GRMhor algorithm can effectively detect all types of HOR arrays, whether they are fully canonical or exhibit various variant modifications. To describe the full spectrum of HOR arrays, we introduce the innovative concept of Cascading HORs, differing from Willard's HORs in that within the HOR, at least one constituent monomer appears in two or more copies. Due to the duplications of individual monomers in the schematic representation, such HORs are depicted in multiple rows, hence the intuitive name, Cascading HOR.

Subsequently, we showed that even in the complex structure of the T2T-CHM13 assembly of human chromosome 20, the GRMhor algorithm successfully identifies alpha satellite HOR arrays and reveals their internal structure. Six distinct HOR arrays were delineated: the Cascading 16mer HOR, comprising 805 copies (83.7% canonical) ; the Cascading 11mer HOR, containing 235 copies (86.4% canonical); the conventional Willard's type 8mer HOR, consisting of 388 HOR copies (93.0% canonical); Cascading 8mer HOR, with 107 copies (64.5% canonical); Cascading and almost complete canonical 26mer HOR, with 14 copies (92.9% canonical); and highly Cascading and highly variant 18mer HOR with 37 copies (56.8%

canonical) (Table 1). Supplementary Figures S1-S6 provide a comprehensive visual representation of all identified HOR copies using the GRMhor algorithm.

Let us provide commentary on the comparison with other established computer tools utilized for the identification and analysis of higher order structures. The two most recent tools for the automatic annotation of centromere structure are NTRPrism (43) and HiCAT (37). The study by Altemose et al. (2022) corroborates the identification of the same HOR structures as the GRMhor algorithm, particularly 16mer, 8mer, 11mer, 8mer, 18mer, 26mer, and 6mer. The only difference lies in the 6mer HOR, which in GRMhor algorithm, unlike in Ref (43), does not possess the status of a distinct HOR, as it represents a variant of the 16mer and 18mer HORs, as evident in Fig. 7a and Fig. 7b. In Ref (43), it is stated that this 6mer HOR is divergent, and the region occupied by its repetitions is very short (19996 bp), corresponding to 117 monomers, or 20 variant HORs.

It is noteworthy that the HOR annotation methodology implemented by the NTRprism algorithm, as delineated in (43), bears a striking resemblance to the 2007 iteration of GRM (40), which was specifically tailored for identifying Willard-type HORs. Consequently, this approach demonstrates limited efficacy in discerning more intricate HOR arrangements, such as variant HORs and combinations of distinct HORs within the same genomic region. Let's, for example, consider the HOR in region E. Utilizing the GRMhor algorithm, it is straightforward from the MD diagram that after 5 canonical and 5 variant copies of the 18mer HOR within region E, three copies of the 26mer HOR appear, followed by variant and canonical copies of the 18mer HORs. This internal substructure is readily discernible and depicted schematically in Supplementary Fig. S5.

In the computation using HiCAT algorithm, five HORs were reported in chromosome 20: R1L16 - 16mer, R2L14 - 14mer, R3L14 - 14mer, R4L2-2mer and R5L8 - 8mer (37). In comparison to HORs identified using the GRMhor algorithm and the NTRPrism algorithm, the 26mer, 11mer, and 18mer HORs, as well as another version of the 8mer HOR, are missing. Additionally, 14mer and 2mer HORs appear. By comparing HOR regions and copy numbers, we can conclude that two versions of the 14mer HOR identified by HiCAT correspond to the 18mer and 26mer HORs, respectively, while the 2mer HOR corresponds to variant substructures of the Cascading 8mer HOR (see Supplementary Fig. S4).

These two comparisons with the latest tools clearly highlight the precision and thoroughness of the GRMhor algorithm, enabling it to effortlessly detect all types of HORs, irrespective of their divergence (variant HORs) or the number of monomer repetitions within a single HOR unit (Cascading HORs). Furthermore, as demonstrated in Refs. (44,45), the algorithm is applicable to any type of repetitive units, not only alpha satellite monomers. In these two articles, we adopted the Neuroblastoma Break Point Family (NBPF) consensus sequence, a monomer of approximately ~1700 bp length. Employing the GRMhor algorithm, we identified 3mer HOR structures within several NBPF genes.

In addition, as evident from Fig. 8 and Supplementary Figures, the algorithm, coupled with precise identification of higher-order structures, also ensures a comprehensive schematic representation of HORs. Such visualization of higher-order structures enables accurate analysis of HOR length, number of HORs, variant copy statistics, and all other parameters necessary for characterizing HOR regions (see Table 1).

Our findings underscore a noteworthy concordance between bioinformatic analyses and traditional molecular methodologies, with significant implications for the field of bioinformatics. Specifically, the identification of various bands using Southern blotting, employing satellite monomers as probes, closely parallels our bioinformatic discovery of major HOR structures featuring varying numbers of monomers. This alignment not only validates the robustness of bioinformatic approaches but also underscores their compatibility and complementarity with conventional molecular techniques. By bridging these methodologies, our study not only enhances our comprehension of genomic structures but also underscores the importance of integrating diverse scientific approaches to unravel complex biological phenomena.

While some identified HOR sequences have been previously documented, this article represents a significant leap forward as it unveils their precise internal architecture for the first time. The findings presented herein lead to the following key conclusions: (i) human chromosome centromeres harbor a remarkably diverse spectrum of higher-order structures; (ii) HOR configurations consist of tandem repeats occurring in numerous copies; (iii) within canonical HOR arrangements, individual monomers, originating from identical monomer sequences, assemble into cascading formations. Although the functional significance of these cascade patterns remains elusive, the results provide novel insights into the intricate makeup of human centromeres. Our study underscores the vital distinction between these HORs and the conventional Willard-type HORs. This revelation elucidates the intricate architecture of these HORs within the centromere, shedding light on their potential role in conveying essential genetic information, with potential implications for chromosome segregation and genetic stability.

Data Availability The MonFinder and GRMhor (python applications) is freely available at <https://github.com/domjanbaric/GRMhor/tree/main>. All artificial monomer arrays are available for testing on <https://github.com/domjanbaric/GRMhor/tree/main>. Reference genome sequences T2T CHM13v2 used to test the application are freely available at the National Center for Biotechnology Information official website https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1/.

Acknowledgements This research was funded by the project “Implementation of cutting-edge research and its application as part of the Scientific Center of Excellence for Quantum and Complex Systems, and Representations of Lie Algebras”, grant number PK.1.1.02,

European Union, European Regional Development Fund and by the Croatian Science Foundation, grant number IP-2019-04- 2757.

Conflict of Interest Disclosure All authors of this article declare that they have no conflicts of interest.

Author Contributions M.G. performed the computations. M.G. and D.B. wrote GRMhor algorithm. V.P. supervised the study. All authors analyzed computational results. M.G. and V.P. wrote the manuscript. All authors read and approved the final version of the manuscript.

REFERENCES

1. Garrido-Ramos, M.A. (2017) Satellite DNA: An Evolving Topic. *Genes (Basel)*, **8**.
2. Jagannathan, M., Warsinger-Pepe, N., Watase, G.J. and Yamashita, Y.M. (2017) Comparative Analysis of Satellite DNA in the *Drosophila melanogaster* Species Complex. *G3 (Bethesda)*, **7**, 693-704.
3. Lower, S.S., McGurk, M.P., Clark, A.G. and Barbash, D.A. (2018) Satellite DNA evolution: old ideas, new approaches. *Curr Opin Genet Dev*, **49**, 70-78.
4. Bersani, F., Lee, E., Kharchenko, P.V., Xu, A.W., Liu, M., Xega, K., MacKenzie, O.C., Brannigan, B.W., Wittner, B.S., Jung, H. *et al.* (2015) Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. *Proc Natl Acad Sci U S A*, **112**, 15148-15153.
5. Aldrup-MacDonald, M.E., Kuo, M.E., Sullivan, L.L., Chew, K. and Sullivan, B.A. (2016) Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Res*, **26**, 1301-1311.
6. Zhang, W., Li, J., Suzuki, K., Qu, J., Wang, P., Zhou, J., Liu, X., Ren, R., Xu, X., Ocampo, A. *et al.* (2015) Aging stem cells. A Werner syndrome stem cell model unveils heterochromatin alterations as a driver of human aging. *Science*, **348**, 1160-1163.
7. Ferree, P.M. and Barbash, D.A. (2009) Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. *PLoS Biol*, **7**, e1000234.
8. Lemos, B., Branco, A.T. and Hartl, D.L. (2010) Epigenetic effects of polymorphic Y chromosomes modulate chromatin components, immune response, and sexual conflict. *Proc Natl Acad Sci U S A*, **107**, 15826-15831.
9. Feliciello, I., Akrap, I. and Ugarkovic, D. (2015) Satellite DNA Modulates Gene Expression in the Beetle *Tribolium castaneum* after Heat Stress. *PLoS Genet*, **11**, e1005466.
10. Joshi, S.S. and Meller, V.H. (2017) Satellite Repeats Identify X Chromatin for Dosage Compensation in *Drosophila melanogaster* Males. *Curr Biol*, **27**, 1393-1402 e1392.
11. Pennacchio, L.A. and Rubin, E.M. (2001) Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet*, **2**, 100-109.
12. Visel, A., Akiyama, J.A., Shoukry, M., Afzal, V., Rubin, E.M. and Pennacchio, L.A. (2009) Functional autonomy of distant-acting human enhancers. *Genomics*, **93**, 509-513.

13. Noonan, J.P. and McCallion, A.S. (2010) Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet*, **11**, 1-23.
14. Miga, K.H., Newton, Y., Jain, M., Altemose, N., Willard, H.F. and Kent, W.J. (2014) Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res*, **24**, 697-707.
15. Willard, H.F. and Waye, J.S. (1987) Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. *J Mol Evol*, **25**, 207-214.
16. Willard, H.F. (1991) Evolution of alpha satellite. *Curr Opin Genet Dev*, **1**, 509-514.
17. Warburton, P.E. and Willard, H.F. (1996), *Human Genome Evolution*. BIOS Scientific Publisher, pp. 121-145.
18. Choo, K.H., Vissel, B., Nagy, A., Earle, E. and Kalitsis, P. (1991) A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. *Nucleic Acids Res*, **19**, 1179-1182.
19. Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V. and Yurov, Y. (2001) Alpha-satellite DNA of primates: old and new families. *Chromosoma*, **110**, 253-266.
20. Sullivan, L.L., Chew, K. and Sullivan, B.A. (2017) alpha satellite DNA variation and function of the human centromere. *Nucleus*, **8**, 331-339.
21. Wlodzimierz, P., Hong, M. and Henderson, I.R. (2023) TRASH: Tandem Repeat Annotation and Structural Hierarchy. *Bioinformatics*, **39**.
22. Willard, H.F. (1985) Chromosome-specific organization of human alpha satellite DNA. *Am J Hum Genet*, **37**, 524-532.
23. Jorgensen, A.L., Bostock, C.J. and Bak, A.L. (1986) Chromosome-specific subfamilies within human alphoid repetitive DNA. *J Mol Biol*, **187**, 185-196.
24. Tyler-Smith, C. and Brown, W.R. (1987) Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J Mol Biol*, **195**, 457-470.
25. Alexandrov, I.A., Mashkova, T.D., Akopian, T.A., Medvedev, L.I., Kisselev, L.L., Mitkevich, S.P. and Yurov, Y.B. (1991) Chromosome-specific alpha satellites: two distinct families on human chromosome 18. *Genomics*, **11**, 15-23.
26. Smit, A.F.A., Hubley, R. and Green, P. RepeatMasker Open-3.0. 1996–2010.
27. Novak, P., Neumann, P. and Macas, J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.
28. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, **27**, 573-580.
29. Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C. and Smit, A.F. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*, **117**, 9451-9457.
30. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, **30**, 772-780.
31. Novak, P., Avila Robledillo, L., Koblizkova, A., Vrbova, I., Neumann, P. and Macas, J. (2017) TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res*, **45**, e111.
32. Novak, P., Neumann, P., Pech, J., Steinhaisl, J. and Macas, J. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792-793.

33. Schaper, E., Korsunsky, A., Pecerska, J., Messina, A., Murri, R., Stockinger, H., Zoller, S., Xenarios, I. and Anisimova, M. (2015) TRAL: tandem repeat annotation library. *Bioinformatics*, **31**, 3051-3053.
34. Sevim, V., Bashir, A., Chin, C.S. and Miga, K.H. (2016) Alpha-CENTAURI: assessing novel centromeric repeat sequence variation with long read sequencing. *Bioinformatics*, **32**, 1921-1924.
35. Kunyavskaya, O., Dvorkina, T., Bzikadze, A.V., Alexandrov, I.A. and Pevzner, P.A. (2022) Automated annotation of human centromeres with HORmon. *Genome Res*, **32**, 1137-1151.
36. Bzikadze, A.V. and Pevzner, P.A. (2020) Automated assembly of centromeres from ultra-long error-prone reads. *Nat Biotechnol*, **38**, 1309-1316.
37. Gao, S., Yang, X., Guo, H., Zhao, X., Wang, B. and Ye, K. (2023) HiCAT: a tool for automatic annotation of centromere structure. *Genome Biol*, **24**, 58.
38. Dvorkina, T., Kunyavskaya, O., Bzikadze, A.V., Alexandrov, I. and Pevzner, P.A. (2021) CentromereArchitect: inference and analysis of the architecture of centromeres. *Bioinformatics*, **37**, i196-i204.
39. Dvorkina, T., Bzikadze, A.V. and Pevzner, P.A. (2020) The string decomposition problem and its applications to centromere analysis and assembly. *Bioinformatics*, **36**, i93-i101.
40. Paar, V., Basar, I., Rosandic, M. and Gluncic, M. (2007) Consensus higher order repeats and frequency of string distributions in human genome. *Curr Genomics*, **8**, 93-111.
41. Paar, V., Gluncic, M., Rosandic, M., Basar, I. and Vlahovic, I. (2011) Intragenic higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees. *Mol Biol Evol*, **28**, 1877-1892.
42. Gluncic, M. and Paar, V. (2013) Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. *Nucleic Acids Res*, **41**, e17.
43. Altemose, N., Logsdon, G.A., Bzikadze, A.V., Sidhwani, P., Langley, S.A., Caldas, G.V., Hoyt, S.J., Uralsky, L., Ryabov, F.D., Shew, C.J. *et al.* (2022) Complete genomic and epigenetic maps of human centromeres. *Science*, **376**, eabl4178.
44. Gluncic, M., Vlahovic, I., Rosandic, M. and Paar, V. (2023) Tandemly repeated NBPF HOR copies (Olduvai triplets): Possible impact on human brain evolution. *Life Sci Alliance*, **6**.
45. Gluncic, M., Vlahovic, I., Rosandic, M. and Paar, V. (2023) Tandem NBPF 3mer HORs (Olduvai triplets) in Neanderthal and two novel HOR tandem arrays in human chromosome 1 T2T-CHM13 assembly. *Sci Rep*, **13**, 14420.
46. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**, 443-453.
47. Sobic, M. and Sikic, M. (2017) Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, **33**, 1394-1395.
48. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.