

# Regular higher order repeat structures in beetle *Tribolium castaneum* genome

---

Vlahović, Ines; Glunčić, Matko; Rosandić, Marija; Ugarković, Đurđica; Paar, Vladimir

Source / Izvornik: **Genome Biology and Evolution**, 2016, 9, 2668 - 2680

Journal article, Published version

Rad u časopisu, Objavljena verzija rada (izdavačev PDF)

<https://doi.org/10.1093/gbe/evw174>

Permanent link / Trajna poveznica: <https://urn.nsk.hr/urn:nbn:hr:217:107914>

Rights / Prava: [Attribution-NonCommercial 4.0 International/Imenovanje-Nekomercijalno 4.0 međunarodna](#)

Download date / Datum preuzimanja: **2025-03-12**



Repository / Repozitorij:

[Repository of the Faculty of Science - University of Zagreb](#)



# Regular Higher Order Repeat Structures in Beetle *Tribolium castaneum* Genome

Ines Vlahović<sup>1,†</sup>, Matko Glunčić<sup>1,†,\*</sup>, Marija Rosandić<sup>2</sup>, Đurđica Ugarković<sup>3</sup>, and Vladimir Paar<sup>1,2</sup>

<sup>1</sup>Faculty of Science, University of Zagreb, Zagreb, Croatia

<sup>2</sup>Croatian Academy of Sciences and Arts, Zagreb, Croatia

<sup>3</sup>Division of Molecular Biology, Institute Ruđer Bošković, Zagreb, Croatia

†These authors contributed equally to this work.

\*Corresponding author: E-mail: matko@phy.hr.

Accepted: July 27, 2016

## Abstract

Higher order repeats (HORs) containing tandems of primary and secondary repeat units (head-to-tail “tandem within tandem pattern”), referred to as regular HORs, are typical for primate alpha satellite DNAs and most pronounced in human genome. Regular HORs are known to be a result of recent evolutionary processes. In non-primate genomes mostly so called complex HORs have been found, without head to tail tandem of primary repeat units. In beetle *Tribolium castaneum*, considered as a model case for genome studies, large tandem repeats have been identified, but no HORs have been reported. Here, using our novel robust repeat finding algorithm Global Repeat Map, we discover two regular and six complex HORs in *T. castaneum*. In organizational pattern, the integrity and homogeneity of regular HORs in *T. castaneum* resemble human regular HORs (with *T. castaneum* monomers different from human alpha satellite monomers), involving a wider range of monomer lengths than in human HORs. Similar regular higher order repeat structures have previously not been found in insects. Some of these novel HORs in *T. castaneum* appear as most regular among known HORs in non-primate genomes, although with substantial riddling. This is intriguing, in particular from the point of view of role of non-coding repeats in modulation of gene expression.

**Key words:** insect genome, *Tribolium castaneum*, higher order repeats, Global Repeat Map, regulatory elements.

## Introduction

*Tribolium castaneum* was noted as sophisticated genetic model organism for studies of insect development (Sokoloff 1972; Denell 2008; Roth and Hartenstein 2008; Feliciello et al. 2015a). It is known as a member of the most species-rich eukaryotic order and important pest of stored products. During the last two decades, the *T. castaneum* genome has been sequenced and studied (Ugarković et al. 1996a; Richards et al. 2008; Wang et al. 2008; Feliciello et al. 2011, 2015a). Contigs have been assembled into ten linkage groups (LG), and the remaining sequence represented as unplaced scaffolds (DS) and unplaced singletons (GG) (Tcas\_3.0 assembly) ([www.ncbi.nlm.nih.gov/bioproject/12540](http://www.ncbi.nlm.nih.gov/bioproject/12540); <http://beetlebase.org>). This genome can provide useful information on genome evolution and functions linking genomes of insects and vertebrates. Tandemly repeated DNA sequences, known as satellite DNAs, are among most rapidly evolving sequences

in eukaryotic genomes, usually differing significantly among evolutionary closely repeated species and expected to drive population and species divergence (Feliciello et al. 2015a). It was hypothesized that satellite DNAs could influence nearby gene expression and thereby acting as regulatory elements (Ugarković 2005; Palomeque and Lorite 2008; Pezer et al. 2012). *Tribolium castaneum* genome revealed the presence of highly repetitive families (Wang et al. 2008). The pronounced satellite-like 0.3–0.4 kb repeat units, called TCAST satellites, have been found abundant in *T. castaneum* genome (Ugarković et al. 1996a; Feliciello et al. 2011; Brajković et al. 2012; Feliciello et al. 2015a). They make up 35% of sequenced genome, encompassing pericentromeric and centromeric regions, and a small portion is dispersed within euchromatin (Brajković et al. 2012) where it affects expression of genes under specific environmental conditions of heat stress (Feliciello et al. 2015b). The genes are suppressed due to transient formation of heterochromatin at

dispersed satellite repeats and this represents the first experimental proof for the gene modulatory role of satellite DNA repeats. Besides, this novel mode of gene regulation does not seem to be unique to the specific satellite DNA/gene, and it could be hypothesised that other non-coding repeats dispersed within euchromatin, could influence expression of associated genes by the similar mechanism.

### Regular HORs

Human and non-human primate centromeres are characterized by well-known prototype of higher order repeats (HORs) in non-coding sequences, based on diverged ~171 bp alpha satellite monomers (Manuelidis 1978; Tyler-Smith and Brown 1987; Willard and Wayne 1987; Warburton and Willard 1996; Lee et al. 1997; Rudd and Willard 2004; Rosandić et al. 2006, 2013; Rudd et al. 2006; Schueler and Sullivan 2006; Warburton et al. 2008; Alkan et al. 2011; Paar et al. 2011a, 2011b). Head-to-tail tandems of alpha satellite monomers form HOR units. Within each HOR unit, the constituent alpha satellite monomers exhibit substantial intermonomeric sequence divergence (20–35%). HOR units also organize themselves in head-to-tail tandem repeats as shown in figure 1A. The range of their length variations is small (~2% of the average HOR length). Divergence between HOR units is less than ~5%, which is sizably smaller than divergence between monomers within each HOR unit (20–35%) (Warburton and Willard 1996).

Such “tandem within tandem” pattern will be referred to as regular HOR. A constituent monomer in a primary tandem (alpha satellite monomer in primates) is called primary repeat unit, and constituent HOR unit in the secondary tandem (alpha satellite HOR unit in primates) is called secondary repeat unit.

Human alpha satellite HORs represent a model case of regular HORs. They characterize primate genomes, and are species specific. For example, this was shown for chromosome 5 with species-specific alpha satellite HORs: 13mer in human (Baldini et al. 1989), 13mer in Neanderthal (Vlahović et al. 2015, unpublished data), 5mer in chimpanzee, 14mer in orangutan and 3mer in macaque (Rosandić et al. 2013). In assemblies of chromosomes of lower primates several pronounced regular alpha satellite HORs were recently found: 6mer or 4mer in siamang (Terada et al. 2013), 31mer in Hoolock, 9mer in *Nomascus*, 8mer and 13mer in *Hylobates* (Koga et al. 2014), 9mer in owl monkey and 12mer in marmoset (Sujiwattanarat et al. 2015).

In non-primate genomes, the regular HORs with more than two monomers in secondary repeat unit are rare (Horz and Altenburger 1981; Lee et al. 1997; Kato 1999; Komissarov et al. 2011).

In cave beetle *Pholeuon proserpinae* a 532-bp dimer HOR unit composed of two types of 266 bp monomers was found (Pons et al. 2003). In *Pimelia radula ascendes* a 712 bp dimer

HOR unit is composed of two types of 356 bp monomers (Pons et al. 2002). In satellite DNA from the phytophagous beetle *Chrysolina carnifex* the 211-bp monomers are organized in the form of dimers or 3mers (Palomeque et al. 2005; Palomeque and Lorite 2008). Some cases of regular HORs have been also found in insects *Tribolium brevicornis* and *Tribolium madens* (Ugarković et al. 1996a; Mravinac et al. 2005).

Prior to this study, HORs have not been identified in *T. castaneum*.

### Complex HORs

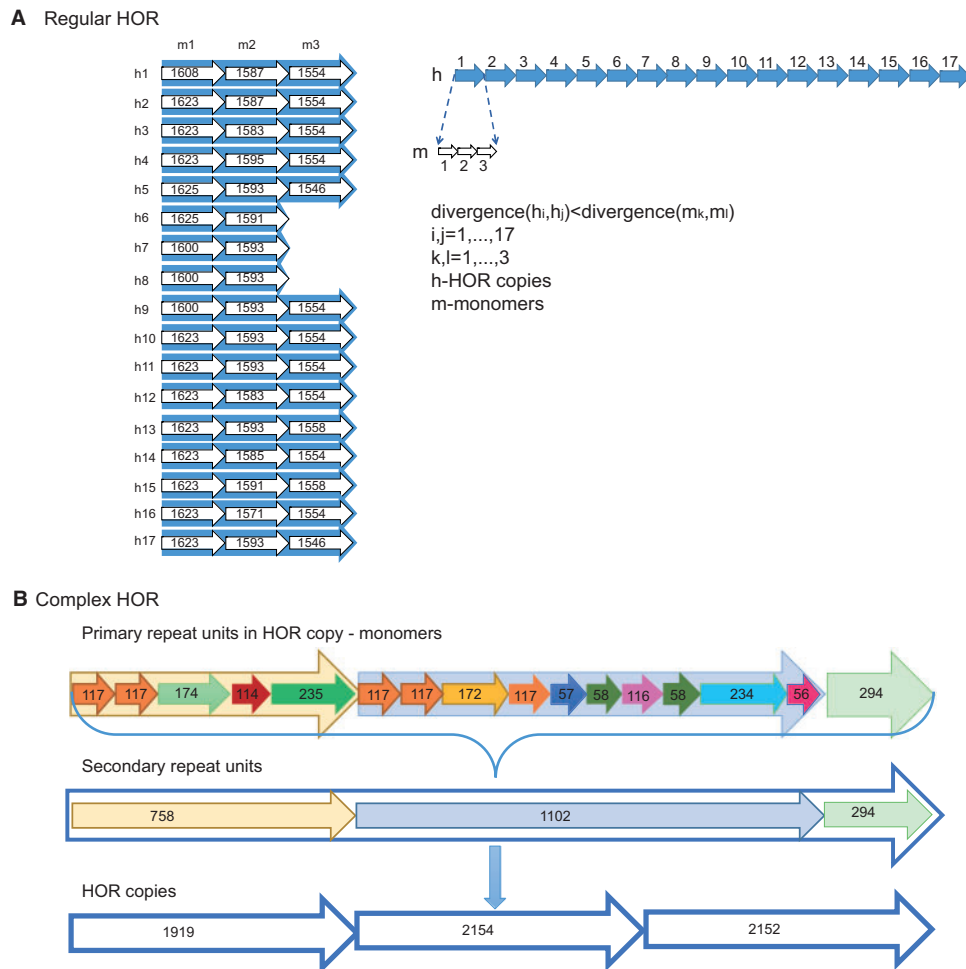
In non-primates the internal structure of some tandem repeats has internal subunits, but subunits are not organized into a head-to-tail tandem. Usually, they are formed from interspersed and/or inversely oriented monomers and frequently with extraneous sequence elements. In the literature such repeats are called complex HORs or simply HORs, what can cause misunderstandings with respect to “tandem-within-tandem” definition of regular HORs in human genome. While regular HORs have both primary and secondary tandem repeat pattern (associated with monomer unit and the corresponding HOR unit, respectively), complex HORs contain no tandem of primary monomers.

Complex HORs have been frequently found in non-human mammals, for example in mouse (Komissarov et al. 2011) (fig. 1B), swine (Janzen et al. 1999), bovids (Modi et al. 2004), horse, dog, and elephant (Alkan et al. 2011).

Satellite DNAs are abundant in insects (Palomeque and Lorite 2008), and various complex HORs were found, for example, in *T. brevicornis*, *T. madens* and *T. audax* (Ugarković et al. 1996b; Mravinac et al. 2005; Mravinac and Plohl 2007; Mravinac and Plohl 2010).

### Non-coding Regulatory Elements

Non-coding sequences are gaining significance in light of revealing large-scale regulatory architecture of genomes. Eukaryotic gene expression is often controlled by distant regulatory elements (King and Wilson 1975; Pennachio and Rubin 2001; Haygood et al. 2010; Noonan and McCallion 2010; Anderson and Hill 2013; Andersson et al. 2015; Braikia et al. 2015; Erokhin et al. 2015; Thakurela et al. 2015) including those present within non-coding repetitive DNAs such as transposons or satellite DNAs (Ugarković 2005; Tomilin 2008; Eichten et al. 2012; Pezer et al. 2012). The gene modulatory effect of satellite DNA is positively correlated with the copy number of repeats within gene-associated satellite DNA elements (Felicciello et al. 2015b) and in this context, accelerated HOR patterns may be of significant interest regarding evolution of gene regulatory networks as well as species evolution (Paar et al. 2011b). It was suggested that HORs might also have a role as possible encryption key for microtubule-centromere interaction (Rosandić et al. 2008).



**Fig. 1.**—Regular and complex HORs. (A) An example of regular HOR is 17 copy 3mer HOR~4770 bp based on monomer length of ~1600 bp identified in human chromosome 1 in NBPf family gene. This regular HOR satisfy divergence between monomers of ~20% and divergence between HOR copies of ~0.5%. Numbers in figure of monomer and HOR lengths are represented in base pairs. Structure of this HOR is explained in Paar et al. (2011b). (B) An example of complex HOR structure is MaSat ~2154 bp found in mice. The complex HOR is composed from smaller subunits—primary repeats of monomers of different lengths that make secondary repeat units with three types of monomers that ultimately make HOR. Different colors represents different types of monomers. Numbers in figure of monomer and HOR lengths are represented in base pairs This HOR structure was published in Komissarov et al. (2011).

## Materials and Methods

Our studies of detecting and analyzing HORs in genome sequences of *T. castaneum* (Assembly Name: Tcas\_3.0; Submitter: Baylor College of Medicine; GenBank Assembly ID: GCA\_000002335.1; and *T. castaneum* strain Georgia GA2, whole genome shotgun sequencing project; GenBank ID: AAJJ00000000.1) (Richards et al. 2008; Kim et al. 2010) are performed here using new robust repeat finding algorithm Global Repeat Map (GRM) (Rosandić et al. 2003; Paar et al. 2005, 2011a, 2011b; Glunčić and Paar 2012). The assembled genome can be downloaded in fasta format from: <ftp://ftp.bioinformatics.ksu.edu/pub/BeetleBase/3.0>. The novelty of GRM approach is a direct mapping of symbolic DNA sequence into frequency domain using complete *K*-string ensemble instead of statistically adjusted individual *K*-strings optimized

locally. In this way, GRM provides a straightforward identification of DNA repeats using frequency domain, but avoiding mapping of symbolic DNA sequence into numerical sequence, and uses *K*-string matching, but avoiding statistical methods and locally optimizing individual *K*-strings (Glunčić and Paar 2012). The GRM algorithm is an extension of KSA (Key String Algorithm) framework (Glunčić and Parr 2012) which 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 where the key string appears within a given genomic sequence. The ensuing KSA fragments form the KSA length array that could be compared with an array of lengths of restriction fragments resulting from hypothetical complete digestion cutting genomic sequence at recognition sites corresponding to the KSA key string. The GRM algorithm consists of the following steps: Step 1, computes the frequency versus

fragment length distribution for a given genomic sequence by superposing results of consecutive KSA segmentations computed for an ensemble of all 8-bp key strings (48 = 65,536 key strings) (Glunčić and Parr 2012). In the GRM diagram, each pronounced peak corresponds to one or more repeats at that length, tandem or dispersed. Step 2, determines the dominant key string corresponding to the 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, 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, aligns all sequences of repeat copies from Step 3 and constructs consensus sequence. Step 5, computes divergence between each repeat copy from step 3 and consensus sequence from step 4 using the Needleman–Wunsch (Needleman and Wunsch 1970) algorithm. Step 6, the pattern in divergences between repeat copies and with respect to consensus reveals tandem monomer sequence or HOR structure. Diagram outlining the main steps in the GRM process is shown in figure 2. The GRM program is publicly available at <http://www.hazu.hr/grm/tools.html>.

## Results

### GRM Diagrams for *Tribolium castaneum* Genome

Using GRM algorithm, in the first step we compute GRM diagrams for the whole Tcas\_3.0 genomic assembly of *T. castaneum*. The computed results for repeat unit lengths up to 3000 bp are displayed (fig. 3A–D), and repeat unit lengths assigned to pronounced GRM peaks. The most frequent

repeat unit lengths in the whole Tcas\_3.0 assembly are 180 bp, 166 bp (fig. 3B), 3 bp (fig. 3A), and a family of repeat units in the length interval from ~300 to 400 bp (more precisely, the interval from 309 to 381 bp) (fig. 3C). Ensemble of monomers from 309 to 381 bp is denoted Tcast-360. Moreover, in addition to known monomeric tandem repeats, we discover novel HORs.

### Regular 5mer HOR in Unplaced Singleton GG695826.1

We discover that the singleton GG695826.1 from Tcas\_3.0 is segmented by GRM algorithm into satellite-like monomers of five types, denoted *m1* to *m5* (consensus sequences: [supplementary table S1a, Supplementary Material](#) online; monomer composition of HORs: [supplementary table S2, Supplementary Material](#) online). Consensus lengths of these monomers (331, 362, 369, 361, and 369 bp) differ by up to 11% (fig. 4A), corresponding to pronounced peaks in GRM diagram of GG695826.1 (fig. 3E). These five monomer types in the length interval 300–400 bp belong to the so called Tcast-360 monomers (fig. 3G).

Average divergence between monomer copies of each monomer type is low (below 2%) as shown by divergence matrix (fig. 4B, diagonal); in contrast, average divergence between monomer copies of different monomer types (*m1* vs. *m2*, *m2* vs. *m3*, etc.) is substantially higher, in the range 18–26% (fig. 4B, off diagonal). Consequently, divergence between two different HOR copies is substantially lower than divergence between monomer copies within each HOR copy, which is a characteristic of HOR pattern. More specifically, divergence values for monomer copies *m2* versus *m4* and *m3* versus *m5* are intermediate (8% and 5%, respectively), but still substantially higher than within each monomer type. Thus, diagonal values in divergence matrix are sizably smaller than off diagonal values. In this way, the GG695826.1 sequence is segmented into an array of *m1m2m3m4m5* copies, representing a pronounced 5mer HOR pattern (9 HOR copies), displayed schematically by aligning five columns of constituent monomer copies (each represented by a heavy horizontal line) (fig. 4C). To each HOR copy we attribute a HOR copy number (from 1 to 9). Six of these HOR copies (Nos. 3, 5–9) are complete, each containing all five constituent monomer types (*m1*–*m5*). In each of the remaining three HOR copies (Nos. 1, 2, 4) three monomers are missing, is referred to as riddled, in analogy to situation in some human HORs. A group of peaks in GRM diagram of GG695826.1 characterizes this riddling in the length interval from 500 to 3000 bp (fig. 3F), which is called the GRM signature of riddled HOR.

The concept of HOR signature was previously introduced for riddled HORs in alpha satellites of human Y chromosome (Paar et al. 2011a).

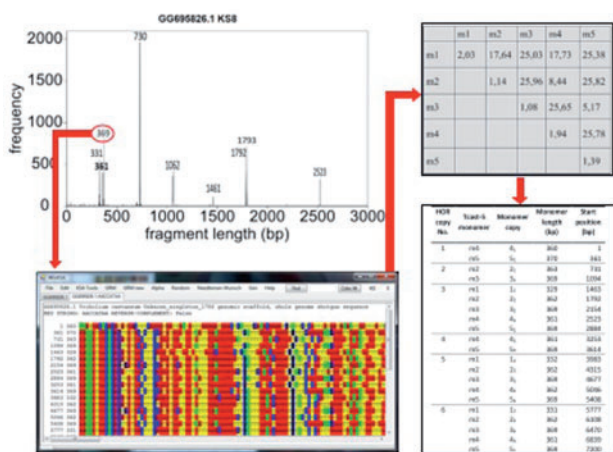
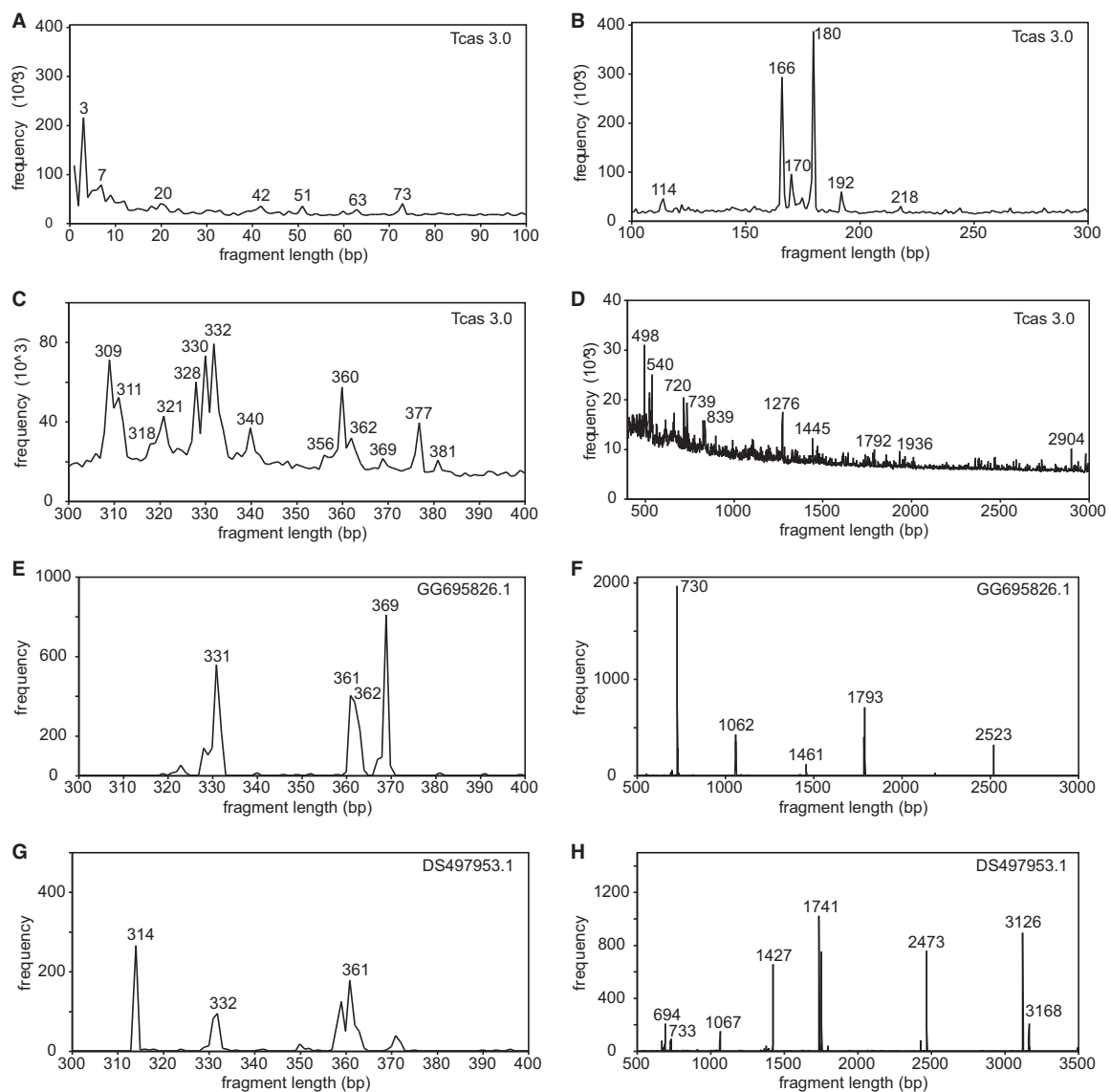


Fig. 2.—Diagrammatic outline of six steps in the process of applying the GRM algorithm.

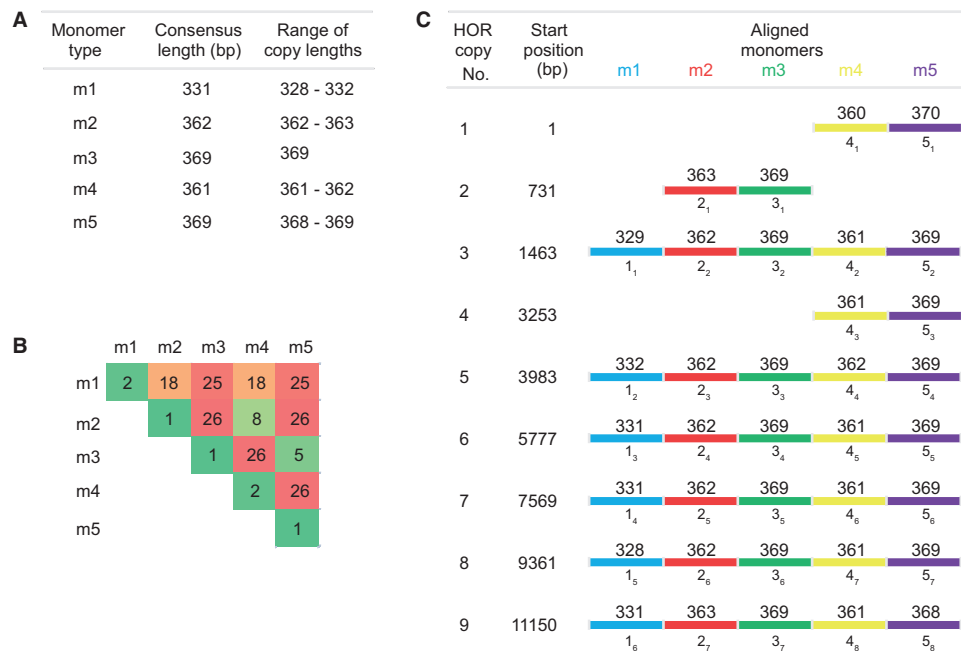


**Fig. 3.**—GRM diagram of *Tribolium castaneum* genome (Tcas\_3.0 assembly). (A–D) GRM diagrams of Tcas\_3.0 assembly (repeat unit lengths in the range 1–3000 bp). For pronounced frequency peaks, the lengths of the corresponding repeat units are shown. (E–F) GRM diagrams of GG695826.1. Peaks in the range 300–400 bp at 331, 361, 362, and 369 bp (E) represent approximate repeat unit lengths of constituent Tcast-360 monomers. Peaks in the range 500–3000 bp at 730, 1062, 1793, and 2523 bp (F) represent GRM signature of riddled 5mer HOR (table 1). (G–H) GRM diagrams of DS497953.1. The pronounced peaks at 314, 332, and 361 bp (G) correspond to approximate repeat unit lengths of constituent Tcast-360 monomers and peaks at 694, 733, 1067, 1427, 1741, 2473, 3126, and 3168 bp (H) represent GRM signature of riddled 4mer HOR (table 2).

Lengths at which these signature peaks appear in figure 3F are approximately equal to distances between start positions of neighboring homologous monomers (table 1), which give rise to HOR signature peaks. For example, the monomer copy 5<sub>2</sub> (369 bp), (starting at position 2884), is highly homologous to monomer copy 5<sub>3</sub> (starting at position 3614) and between them is the monomer copy 4<sub>3</sub> (361 bp). Thus, the distance between the first nucleotides in the 5<sub>2</sub> and 5<sub>3</sub> monomers is 369 + 361 = 730 bp. This gives rise to a peak at fragment length 730 bp in GRM diagram (fig. 3F).

### Complex 4mer HOR in Unplaced Scaffold DS497953.1

Another HOR which contains monomers in the range 300–400 bp (which also belong to Tcast-360 monomers), we find in the scaffold DS497953.1. The GRM algorithm segments this sequence into four monomers (consensus sequences: [supplementary table S1b](#), [Supplementary Material](#) online). Three of them belong to Tcast-360 monomers (300–400 bp). They correspond to three GRM peaks, at ~361, ~314, and ~332 bp (fig. 3G), denoted by m1, m3, and m4, respectively.



**Fig. 4.**—5mer HOR in *Tribolium castaneum* sequence GG695826.1. (A) Five Tcast-360 monomer types constituting 5mer HOR. (B) Average divergence matrix for monomer copies of five Tcast-360 monomer types in 5mer HOR (%). (C) Aligned monomer structure of 5mer HOR. Similar method of HOR schematic presentation was previously used for human chromosomes (Paar et al. 2011a). Monomer types constituting HOR are denoted m1, m2, m3, m4, and m5 (top enumeration of columns). Horizontal bars in each column: monomer copies of the same type. Length (in bp) is assigned to each monomer copy. In each column enumeration of monomers 1, 2, 3, ... (subscript) goes from top to bottom. Example: monomer of type m1 in HOR copy No. 5 is denoted by 1<sub>2</sub>; its length is 332 bp. Only genomic sequences belonging to Tcast-360 or their segments of at least 180bp and having >70% identity to reference sequence are displayed. Consensus monomers: [supplementary table S1a, Supplementary Material online](#). Monomer structure of 5mer HOR: [supplementary table S2, Supplementary Material online](#).

**Table 1**

Illustration of Origin of Pronounced HOR-signature GRM Peaks in Figure 3F as Distances between Start Positions of Approximately Homologous Monomers in 5mer HOR Array from GG695826.1 (fig. 4C)

Distance between start of monomers of the same type (bp)	HOR-signature peak (bp)
$5_2/5_3 = 369 + 361 = 730$	730
$5_3/5_4 = 369 + 332 + 362 + 369 + 362 = 1794$	1793
$2_1/2_2 = 363 + 369 + 329 = 1061$	1062
$5_1/5_2 = 370 + 363 + 369 + 329 + 362 + 369 + 361 = 2523$	2523

Labeling of monomer copies: see figure 4C. Example: 5<sub>2</sub> denotes the second m5-type monomer (counting from top to bottom) in m5 column in figure 4C (HOR copy No. 3); 5<sub>3</sub> denotes the third m5-type monomer in m5 column (HOR copy No. 4).

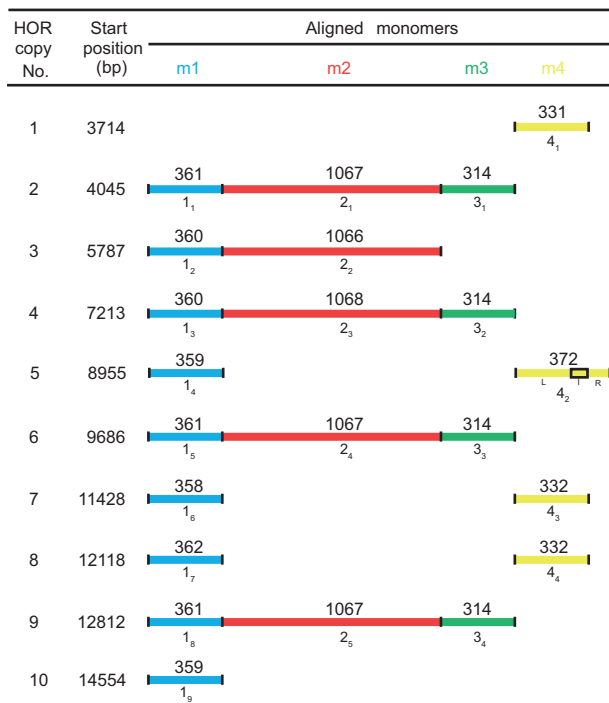
Divergence among homologous monomers of types m1, m3, and m4 are  $4.66 \pm 2.02\%$ ,  $3.92 \pm 3.87\%$ , and  $5.61 \pm 1.74\%$ , while divergence of copies of this three types of monomers is  $18.51 \pm 10.99\%$ .

The 372 bp monomer in HOR copy No. 5 belongs to the m4 type of monomers. It can be approximately obtained from ~332 bp m4 monomer by inserting after position 255 a 48 bp extraneous insertion, denoted I. Thus, in the structure

of 372 bp monomer, we find three segments: left segment denoted L (from position 1 to 255), inserted extraneous middle segment denoted I (from position 256 to 305), and right segment denoted R (from position 306 to 372) (fig. 5). Excluding the inserted segment I, the sequence consisting of only R and L segments is approximately homologous to ~332 bp m4 monomers (at the level of more than 90% identity).

The three monomers m1, m3, and m4 corresponding to DS497953.1, differ from monomers in GG695826.1 from the preceding section. Together, they belong to a family of so called Tcast-360 monomers which lie in the interval from 300 to 400 bp ([supplementary table S4, Supplementary Material online](#)).

The fourth constituting monomer of 4mer HOR in DS497953.1, denoted by m2, is much longer (~1067 bp). However, this long monomer is related to Tcast-360 monomers. The first 150 bp segment of m2 is approximately homologous to the first 150 bp segment of m1 monomer, while the largest part of m2 (after position ~150) has no similarity to m1, m3, and m4 monomers. Divergence among homologous monomers of this type is  $1.24 \pm 0.33\%$ . Furthermore, monomers m3 (~314 bp) approximately correspond to a segment of



**FIG. 5.**—Aligned monomer structure of 4mer HOR in DS497953.1 constituted from three Tcast-360 monomer types (m1, m3, and m4) and one ~1067 bp monomer type (m2). Monomer types constituting HOR are denoted by m1, m2, m3, and m4 (top enumeration of columns). The monomer 4<sub>2</sub> (of length 372 bp) from HOR copy No. 5 has a 48 bp extraneous insertion (indicated by a short rectangle along a part of horizontal bar representing this monomer and labeled I). Consensus monomers: [supplementary table S1b, Supplementary Material online](#). Monomer structure of 4mer HOR: [supplementary table S3, Supplementary Material online](#). The first 150 bp segment of long monomer m2 is approximately homologous to the first 150 bp segment of m1; the remaining part of m2 has no similarity to monomers m1, m3, and m4.

314 nucleotides from m1 (~361 bp) monomers after deletion of ~45 nucleotides. Using dominant key string ACTCCTAT, we segment DS497953.1 into complex array of monomers m1–m4 with substantial riddling, displayed schematically by aligning the constituent monomers (fig. 5).

Average divergence between copies within each monomer type is low (~1–3%), while divergence between monomer copies of different types is substantially larger (~20%, or more). Pronounced GRM peaks in the length interval 500–3500 bp (fig. 3H) represent the corresponding GRM signature (table 2).

Consensus sequence of m1 monomer copies in DS497953.1 is referred to as Tcast-360/1 ([supplementary table S4a, Supplementary Material online](#)). It serves as a reference sequence for Tcast-360 monomers. Slightly different reference sequence is obtained by GRM segmentation of DS497953.1 using the key string AACCATAA. Of monomer sequences obtained in this way, the one, which is closest to

**Table 2**

Illustration of Origin of Pronounced HOR-Signature GRM Peaks in Figure 3H as Approximately Homologous Distances between Start of Monomers from 4mer HOR Array in DS497953.1 (fig. 5)

Distance between start of monomers of the same type (bp)	HOR-signature peak (bp)
$1_1/1_2 = 361 + 1067 + 314 = 1742$	1741
$3_3/3_4 = 314 + 358 + 332 + 362 + 332 + 361 + 1067 = 3126$	3126
$3_2/3_3 = 314 + 359 + 372 + 361 + 1067 = 2473$	2473
$1_2/1_3 = 360 + 1066 = 1426$	1427
$3_1/3_2 = 314 + 360 + 1066 + 360 + 1068 = 3168$	3168
$4_3/4_4 = 332 + 362 = 694$	694
$4_1/4_2 = 331 + 361 + 1067 + 314 + 360 + 1066 + 360 + 1068 + 314 + 359 = 5600$	5600

Illustration of origin of pronounced HOR-signature GRM peaks in figure 3H (displayed for frequency interval from 500 to 3500 bp) as approximate distances between start of monomers of the same type in 4mer HOR array from DS497953.1 (fig. 5). The last peak at 5600 bp is also present in GRM diagram (lying outside of frequency interval displayed in fig. 3H).

the experimental Tcast1b from Ugarković et al. (1996a), is referred to as Tcast-360/2 ([supplementary table S4b, Supplementary Material online](#)). Divergence between Tcast-360/1 and Tcast-360/2 is 6% and the minimal divergence between Tcast-360/2 and experimental Tcast1b is 5%.

### Complex 4mer HOR in Linkage Group CM000279.1

In the linkage group CM000279.1, we discover 4mer HOR constituted of three ~170 bp monomer types and one ~940 bp monomer (fig. 6A) (consensus sequences: [supplementary table S1c, Supplementary Material online](#)). The ~940 bp monomer has no significant overlap with ~170 bp monomers. In addition, within the ~14 kb interval in front of this HOR array, we find dispersed riddled HOR copies (mostly containing m1 and m3 monomers, while m2 and m4 monomers are deleted) (fig. 6B). Divergence among monomers from 4mer HOR is in accordance with regular HOR pattern (fig. 6C). In this sense, this structure resembles a pattern evolving from regular HOR, with an extraneous extension.

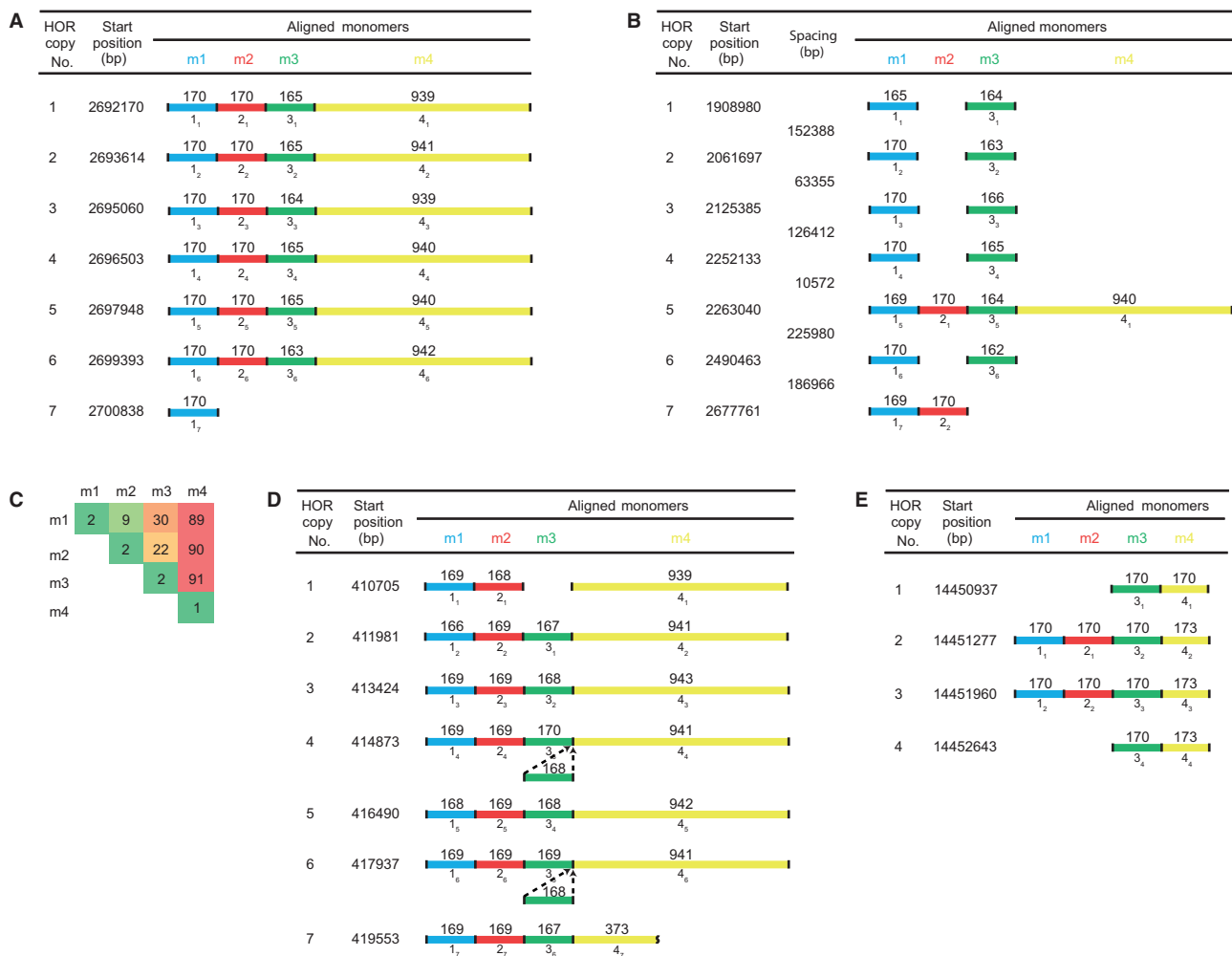
### Complex 4mer HOR in Linkage Group CM000277.2

In CM000277.2, we identify another complex 4mer HOR (fig. 6D). The constituent monomers, three ~170 bp and one ~940 bp monomers (consensus sequence: [supplementary table S1d, Supplementary Material online](#)), approximately correspond to reverse complements of monomer copies from 4mer HOR in CM000279.1 (fig. 6A).

### Regular/Complex 4mer HOR in Linkage Group CM000277.2

In CM000277.2, we discover a short segment of another 4mer HOR, which consists of four ~170 bp monomers (consensus sequence: [supplementary table S1e, Supplementary Material online](#)). Its pattern of HOR scheme corresponds to





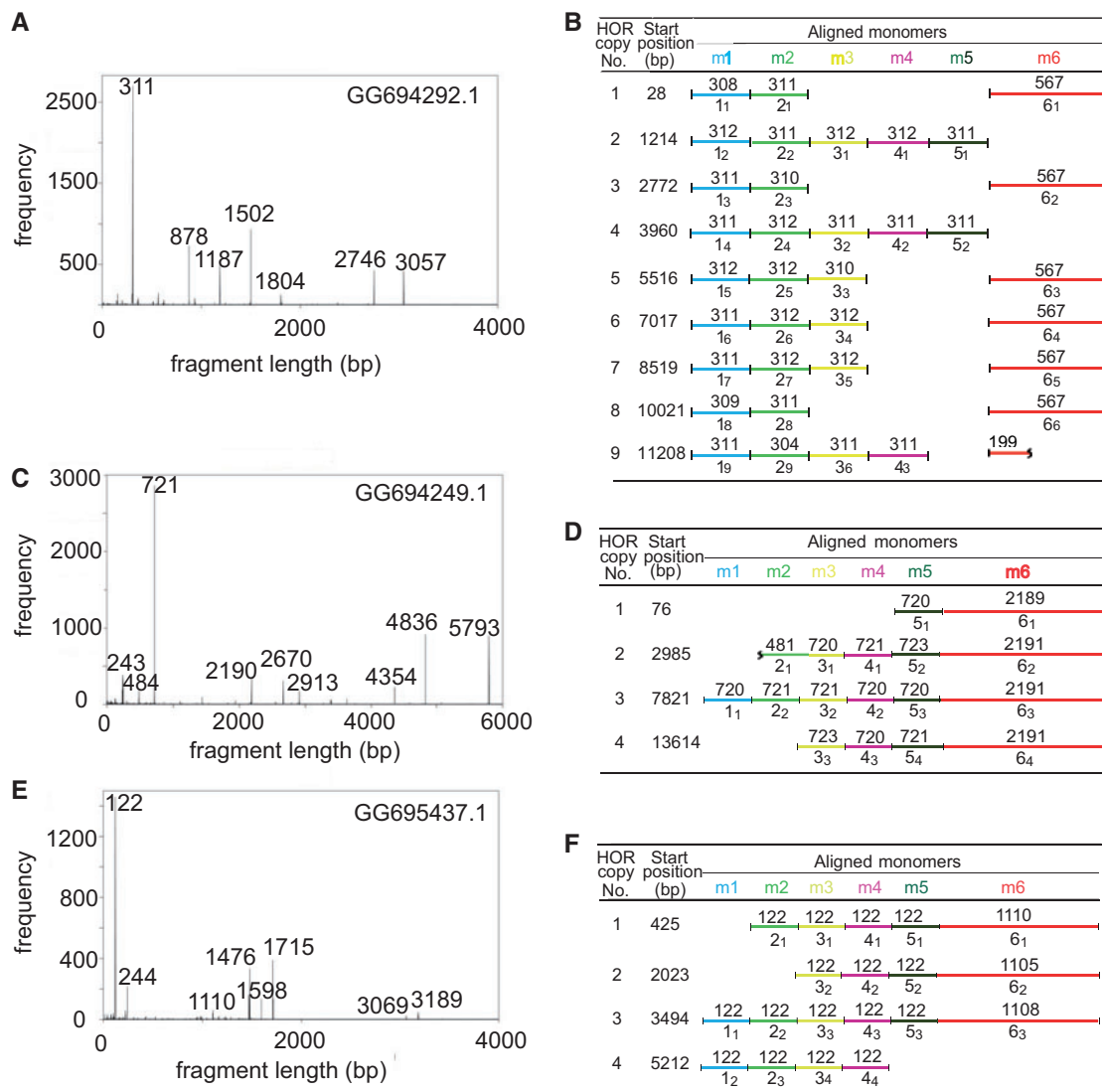
**Fig. 6.**—HORs in CM000279.1 and CM000277.2. (A) 4mer HOR in CM000279.1 consisting of three ~170 bp and one ~940 bp monomers. Consensus monomers: [supplementary table S1c, Supplementary Material](#) online. (B) Dispersed HOR copies in CM000279.1. Third column: spacing between neighboring dispersed copies. (C) Divergence matrix: average divergence between monomers from figure 6A (%). (D) 4mer HOR in CM000277.2 constituted of three ~170 bp and one ~940 bp monomers. In the 373 bp sequence (column m4, HOR copy No. 7) the first 118 bp and the last 204 bp nucleotides correspond to the first 118 bp and the last 204 bp segments from consensus m4, respectively. Consensus monomers: [supplementary table S1d, Supplementary Material](#) online. Monomers in this HOR approximately correspond to reverse complements of monomers from figure 6A. (E) 4mer HOR in CM000277.2 constituted of four ~170 bp monomers. Consensus monomers: [supplementary table S1e, Supplementary Material](#) online.

regular HOR (fig. 6E). On the other hand, these HORs are “weak” in the sense that divergence between some of neighboring monomers is only slightly larger than divergence between monomers of the same type in different HOR copies. For example, divergence m1 versus m2 is only ~50% higher than divergence between m1 monomers in different HOR copies.

### Complex 6mer HOR in Unplaced Singleton GG694292.1

In GG694292.1, we discover a sequence of nine copies of complex 6mer HOR, each constituted of five almost homologous ~311 bp monomers (referred to as type 1, consensus sequence: [supplementary table S1f, Supplementary Material](#)

online) and one 567 bp monomer (referred to as type 2, consensus sequence: [supplementary table S1f, Supplementary Material](#) online) (HOR scheme in fig. 5B). In analogy to tables 1 and 2, the peaks in GRM diagram (fig. 7A) are approximately equal to distances between start positions of neighboring homologous monomers, giving rise to HOR signature peaks at 1502, 878, 1187, 2746, 3057, and 1804 bp (Table 3). Due to high level of homology among type 1 monomers, the highest frequency peak in GRM diagram (fig. 7A) is at ~311 bp due to pairs of neighboring ~311 bp monomers. It is interesting that consensus of type 2 monomers contains two shorter segments from consensus type 1 monomers. A 69 bp segment at the start of type 1 consensus is highly homologous to 69 bp segment starting at position 207 within type 2 consensus, and



**Fig. 7.**—HORs in GG694292.1, GG694249.1, and GG695437.1. (A) GRM diagram for GG694292.1. (B) 6mer HOR scheme in GG694292.1. Five columns m1–m5 display ~311 bp monomers that are mutually similar, referred to as copies of type 1. The sixth column m6 displays the ~567 bp monomers, referred to as monomer type 2. Mutual divergence among type 1 monomers is ~3% (in the range from 0.3% to 6%) and between type 2 monomers divergence is below 0.2%. (C) GRM diagram for GG694249.1. (D) 6mer HOR scheme in GG694249.1: five columns correspond to ~720 bp monomers (type 1) and the sixth column to ~2191 bp (type 2). This HOR can be also characterized as weak. (E) GRM diagram for GG695437.1. (F) 6mer HOR scheme in GG695437.1: five columns correspond to 122 bp monomers (type 1) and the sixth column to ~1108 bp monomers (type 2). This HOR can be also characterized as weak.

38 bp segment at the end of type 1 consensus is highly homologous to 38 bp segment near the end of type 2 consensus. For this HOR pattern, the intra-HOR monomer divergence is only slightly larger than inter-HOR. In this sense this HORs can be characterized as weak.

#### Complex 6mer HOR in Unplaced Singleton GG694249.1

In GG694249.1, we discover a segment of complex 6mer HOR, constituted of five ~720 bp monomers (type 1, consensus sequence: [supplementary table S1g](#), [Supplementary](#)

[Material online](#)) and one ~2191 bp monomer (type 2, consensus sequence: [supplementary table S1g](#), [Supplementary Material online](#)) (GRM diagram in [fig. 7C](#), and HOR scheme in [fig. 7D](#)). Similarly as in the previous case, this HOR can be characterized as weak.

#### Complex 6mer HOR in Unplaced Singleton GG695437.1

In GG695437.1, we discover a segment of another complex 6mer HOR, constituted of five 122 bp monomers (type 1, consensus sequence: [supplementary table S1h](#), [Supplementary](#)

Material online) and one ~1108 bp monomer (type 2, consensus sequence: [supplementary table S1h](#), [Supplementary Material](#) online) (GRM diagram in [fig. 7E](#), and HOR scheme in [fig. 7F](#)). The first 110 bp segment and the last 84 bp segment in the ~1108 bp monomer are approximately homologous to the corresponding segments in 122 bp monomers. Thus, the ~1108 bp monomer could be related to a large extraneous insertion into 122 bp monomer. Similarly as in the previous two cases, this HOR can be characterized as weak.

### Positions of HOR Arrays and Dispersed HOR Copies in Relation to Gene Positions

We located different HOR arrays in CM000277.2 and CM000279.1 from [figure 6](#) in relation to nearby genes in these two chromosomes. In CM000277.2 4mer ~1447 bp HOR array from [figure 6D](#) is positioned between two genes, Kv channel interacting protein 1 gene and ATP-dependent RNA helicase p62-like protein gene, while 4mer ~ 680 bp HOR array from [figure 6E](#) is located in intron between exons 5 and 6 in decaprenyl diphosphate synthase subunit 1 gene (for more details see [fig. 8](#)). Three dispersed HOR copies of ~1447 bp HOR in CM000279.1 from [figure 6B – b, c and f](#), are located in introns of VSX gene, mRNA transcription factor

AP-4 and Dachshund homolog 1 genes, respectively ([fig. 9](#)). 4mer ~1447 bp HOR array from [figure 6A](#) is located in-between two genes, leucine rich repeat protein soc2 homolog gene and fascilin-2 gene (for more details see [fig. 9](#)).

For HORs identified in unlinked multi-component scaffold DS497953.1 and unknown singleton scaffolds GG695826.1, GG694292.1, GG694249.1, and GG695437.1 in Tcas\_3.0 build, we could not identify nearby genes locations because this components in this version of *T. castaneum* genomes was not placed on chromosomes during assembling of the genome sequence.

### Discussion

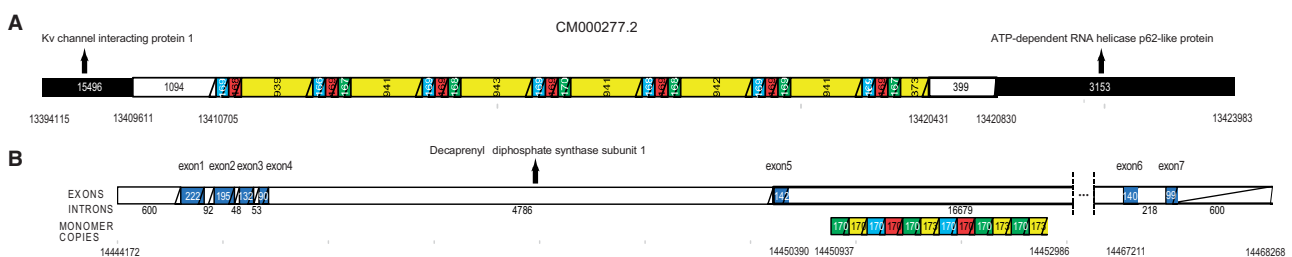
In Tcast-360 5mer HOR ([fig. 4C](#)), the constituent monomers in each HOR copy exhibit substantial intermonomeric sequence divergence, while divergence between monomers of the same type in different HOR copies is an order of magnitude lower. That is a pronounced HOR feature, analogous as in human HORs. Also, deletion of integer number of monomers is present in human HORs too ([Warburton and Willard 1996](#)). Furthermore, extraneous insertions in HOR copies are even scarcer in *T. castaneum* Tcast-360 HOR copies than in some human alpha satellite HORs. In this respect, these *T. castaneum* HORs could be considered as textbook examples analogous to human chromosomes. However, the monomer lengths within HOR copies in *T. castaneum* have significantly larger relative spreading than within human HORs.

So far, the regular HORs have been identified only in primates, most often having alpha satellites as primary repeat units ([Willard and Waye 1987](#)). In other eukaryotes only complex HORs have been identified ([Janzen et al. 1999](#); [Pertile et al. 2009](#); [Alkan et al. 2011](#); [Komissarov et al. 2011](#)). In this sense, the appearance of regular HORs in so distant species as insect *T. castaneum* is interesting, since regular HOR pattern of more recent origin in primates appears also in evolutionary so distant species as *T. castaneum*, while so far it was not found between primates and *T. castaneum* along the evolutionary tree.

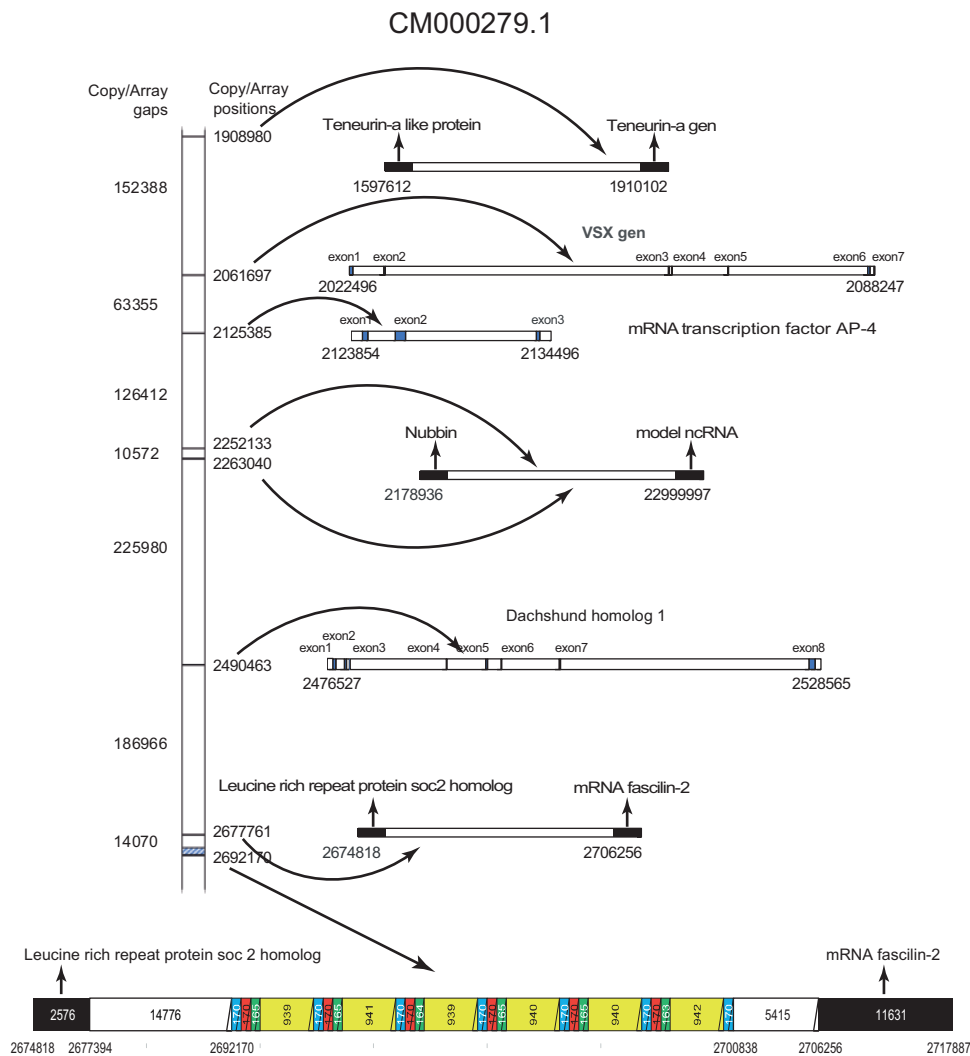
**Table 3**

Illustration of Origin of Pronounced HOR-Signature GRM Peaks in [Figure 7A](#) as Distances between Start Positions of Approximately Homologous Monomers from GG694292.1 ([fig. 7B](#))

Distance between start of monomers of the same type (bp)	HOR-signature peak (bp)
$6_3/6_4 = 567 + 311 + 312 + 312 = 1502$	1502
$2_1/1_2 = 311 + 567 = 878$	878
$1_1/1_2 = 308 + 311 + 567 = 1186$	1187
$6_1/6_2 = 567 + 312 + 311 + 312 + 311 + 312 + 311 + 310 = 2746$	2746
$6_2/6_3 = 567 + 311 + 312 + 311 + 311 + 311 + 312 + 312 + 310 = 3057$	3057
$6_6/6_7 = 567 + 311 + 304 + 311 + 311 = 1804$	1804



**Fig. 8.**—Positions of (A) 4mer ~1447 bp HOR from [figure 6D](#) and (B) 4mer ~ 680 bp HOR from [figure 6E](#) in CM000277.2 relatively to positions of protein coding genes (Tcas\_3.0 NCBI database). (A) 4mer ~1447 bp HOR is located at an intergenic region: 1094 bp from 3' end of KV channel-interacting protein 1 gene (LOC654909) and 399 bp from 5' end of a probable ATP dependent RNA helicase DDX56 protein gene (LOC655034). (B) 4mer ~680 bp HOR, based on 170 bp monomers, is located in the sixth intron of decaprenyl-diphosphate synthetase subunit 1 gene (LOC662107).



**Fig. 9.**—Positions of HOR copies from figure 6A and B in CM000279.1 relatively to positions of protein coding genes (Tcas\_3.0 NCBI database). The first HOR copy in this figure (fig. 6B HOR copy No 1) is located 299210 bp from 3' end of teneurin-a like protein gene and 3122 bp from 5' end of teneurin-a gene (LOC654865). The second HOR copy (fig. 6B HOR copy No 2) is located in the second intron of visual system homeobox protein gene (VSX gen, LOC655933). The third HOR copy (fig. 6B HOR copy No 3) is located within second intron of mRNA transcription factor AP-4 gene (LOC656093). The fourth and the fifth HOR copies (fig. 6B HOR copy No 4 and HOR copy No 5) are positioned 26,156 and 37,063 bp, respectively, from 3' end of nubbin gene (LOC656845) as well as 48,199 and 38,400 bp, respectively, from 5' end of ncRNA gene (LOC103312583). The sixth HOR copy is completely located in fifth intron of dachshund homolog 1 gene (LOC652934). The seventh HOR is located 367 bp from 3' end of leucine rich repeat protein soc2 homolog gene (LOC103312583) and 39,614 bp from 5' end of fascilin-2 gene (LOC664545). Seven copies of 4mer HOR ~1447 bp (fig. 6A) are located 14,776 bp from 3' end of leucine rich repeat protein soc2 homolog gene, described above in the seventh HOR copy, and 5,415 bp from 5' end of fascilin-2 gene which is also described above. We calculated positions of genes after downloading the sequence of nearby genes of HORs (according to NCBI positions for HORs in Tcas\_3.0) from <http://metazoa.ensembl.org/index.html> and aligning them, as query in blastn, with CM000277.2 and CM000279.1 from Tcas\_3.0, as subject sequences.

Appearance of regular HORs in insects is surprising, since the human alpha satellite HORs were generally considered as result of recent evolutionary processes (Rudd et al. 2006). It was proposed that human alpha satellite HORs appeared after the split of human from chimpanzee lineage, estimated to occur 2–7 Myrs ago (Schueler and Sullivan 2006). This is in accordance with observed large differences between HORs in human and chimpanzee (Paar et al. 2011a, 2011b; Rosandić

et al. 2013). On the other hand, Tcast-360 bp satellite is specific for *T. castaneum* and has been amplified within its genome since the split from closest species *T. freemani*, which occurred 12–47 Myrs ago (Angelini and Jockusch 2008). However, since tandemly repeated satellite DNA represents very unstable part of genome, prone to constant turnover and evolution, it is possible that HORs in *T. castaneum* genome occurred later during the course of evolution, not

**Table 4**List of HORs in *Tribolium castaneum* Genome Identified in this Work

HOR	Constituting monomers	Location
5mer	Five Tcast-360	GG695826.1
4mer	Three Tcast-360 + one ~1067 bp	DS497953.1
4mer	Three ~170 bp + one ~941 bp	CM000279.1
4mer	Three ~170 bp + one ~941 bp	CM000277.2
4mer	Four ~170 bp	CM000277.2
6mer	Five ~ 311 bp + one 567 bp	GG694292.1
6mer	Five ~720 bp + one ~ 2191 bp	GG694249.1
6mer	Five 122 bp + one ~ 1108 bp	GG695437.1

exceeding in age human alpha satellite HORs. The regularity, integrity and low divergence among copies speak in favor of recent evolutionary origin of Tcast satellite HORs.

Although in general the formation of HORs can be explained by stochastic recombination processes (Charlesworth et al. 1994), preferential location of complex HORs within human centromeres and their specific structural features such as binding sites for kinetochore protein CENP-B suggest possible functional significance for centromere establishment (Ugarković 2009). In addition, HORs could substantially increase the rate of genome divergence among species and contribute to speciation processes. For example, HORs contribute substantially to divergence between human and chimpanzee Y chromosomes (Paar et al. 2011a). As presented in this paper, HORs in *T. castaneum* are associated with genes and considering previously shown role of satellite repeats in modulation of gene expression (Felicciello et al. 2015b), we propose that HORs could act as gene regulatory elements. In this case, variation in HOR composition among individuals or populations can generate gene expression diversity and contribute to the evolution of gene regulatory network. HORs in *T. castaneum* genome have not been identified in previous analyses using standard bioinformatics tools. Difficulties were probably due to significant spreading of relative lengths of constituting monomers. In HORs based on monomers of 300–400 bp, the spreading of monomer lengths is sizeable, up to ~20% of monomer length, mostly due to deletion of sizable segments from monomers. This spreading in *T. castaneum* is much larger than ~2% spreading of monomer lengths in human alpha satellites.

In summary, we discover eight novel HORs in *T. castaneum* (table 4), while prior to this work none was known.

The presence of HORs and in particular regular HORs within insect genome demonstrates a wide evolutionary spreading of HOR pattern hierarchical organization and their appearance in evolutionary distant eukaryotes. It is possible that using robust repeat finding GRM algorithm regular HORs could be discerned in some other invertebrate genomes too. Further studies are necessary to show whether regular HORs are to an extent restricted to some “islands” along the evolutionary

chain, like higher primates and insects as discussed here, or are more broadly dispersed and to study their possible role as regulatory elements.

## Acknowledgments

The authors are grateful for thoughtful comments of reviewers helping to improve the manuscript. This work is supported by Grant IP-09-2014 from Croatian Science Foundation (HRZZ).

## Supplementary Material

Supplementary tables S1–S4 are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org>).

## Literature Cited

- Alkan C, et al. 2011. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. *Genome Res.* 21:137–145.
- Anderson E, Hill RE. 2013. Long range regulation of the sonic hedgehog gene. *Curr Opin Genet Dev.* 27:54–59.
- Andersson R, Sandelin A, Danko CG. 2015. A unified architecture of transcriptional regulatory elements. *Trends Genet.* 31:426–433.
- Angelini DR, Jockusch EL. 2008. Relationships among pest flour beetles of the genus *Tribolium* (Tenebrionidae) inferred from multiple molecular markers. *Mol Phyl Evol.* 46:127–141.
- Baldini A, Smith DI, Rocchi M, Miller OJ, Miller DA. 1989. A human alphoid DNA clone from the EcoRI dimeric family: genomic and internal organization and chromosomal assignment. *Genomics* 5:822–828.
- Braikia FZ, et al. 2015. Developmental switch in the transcriptional activity of a long-range regulatory element. *Mol Cell Biol.* 35:3370–3380.
- Brajković J, Felicciello I, Bruvo-Madžarić B, Ugarković Đ. 2012. Satellite DNA-like elements associated with genes within euchromatin of the beetle *Tribolium castaneum*. *G3 Genes Genomes Genet.* 2:931–941.
- Charlesworth B, Sniegowski P, Stephan W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371:215–220.
- Denell R. 2008. Establishment of *Tribolium* as a genetic model system and its early contributions to evo-devo. *Genetics* 180:1779–1786.
- Eichten SR, et al. 2012. Spreading of heterochromatin is limited to specific families of maize retrotransposons. *PLoS Genet.* 8:e1003127.
- Erokhin M, Vassetzky Y, Georgiev P, Chetverina D. 2015. Eukaryotic enhancers: common features regulation, and participation in diseases. *Cell Mol Life Sci.* 72:2361–2375.
- Felicciello I, Chinali G, Ugarković Đ. 2011. Structure and evolutionary dynamics of the major satellite in the red flour beetle *Tribolium castaneum*. *Genetica* 139:999–1008.
- Felicciello I, Akrap I, Ugarković Đ. 2015b. Satellite DNA modulates gene expression in the beetle *Tribolium castaneum* after heat stress. *PLoS Genet.* 11:e1005466.
- Felicciello I, Akrap I, Brajković J, Zlatař I, Ugarković Đ. 2015a. Satellite DNA as a driver of population divergence in the red flour beetle *Tribolium castaneum*. *Genome Biol Evol.* 7:228–229.
- Glunčić M, Paar V. 2012. Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. *Nucleic Acids Res.* 41:e17.
- Haygood R, Babbitt CC, Fedrigo O, Wray GA. 2010. Contrasts between adaptive coding and noncoding changes during human evolution. *Proc Natl Acad Sci U S A.* 107:7853–7857.
- Horz W, Altenburger W. 1981. Nucleotide sequence of mouse satellite DNA. *Nucleic Acids Res.* 9:683–696.
- Janzen MA, Buoen LB, Zhao F, Louis CF. 1999. Characterization of a swine chromosome-specific centromeric higher-order repeat. *Mamm Genome* 10:579–584.

- Kato M. 1999. Structural bistability of repetitive DNA elements featuring CAT/G dinucleotide steps and mode of evolution of satellite DNA. *Eur J Biochem.* 265:204–209.
- Kim HS, et al. 2010. BeetleBase in 2010: revisions to provide comprehensive genomic information for *Tribolium castaneum*. *Nucleic Acids Res.* 38:D437–D442.
- King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188:107–116.
- Koga A, et al. 2014. Evolutionary origin of higher-order repeat structure in alpha-satellite DNA of primate centromeres. *DNA Res.* 21:407–415.
- Komissarov AS, Gavrilova EV, Demin SJ, Ishov AM, Podgornaya OI. 2011. Tandemly repeated DNA families in the mouse genome. *BMC Genomics* 12:531.
- Lee C, Wevrick R, Fisher RB, Ferguson-Smith MA, Lin CC. 1997. Human centromeric DNAs. *Hum. Genet.* 100:291–304.
- Manuelidis L. 1978. Complex and simple sequences in human repeated DNAs. *Chromosoma* 66:1–21.
- Modi WS, Ivanov S, Gallagher DS. 2004. Concerted evolution and higher-order repeat structure of the 1.709 (satellite IV) family in bovids. *J Mol Evol.* 58:460–465.
- Mravinac B, Ugarković Đ, Franjević D, Plohl M. 2005. Long inversely oriented subunits form a complex monomer of *Tribolium brevicornis* satellite DNA. *J Mol Evol.* 60:513–525.
- Mravinac B, Plohl M. 2007. Satellite DNA junctions identify the potential origin of new repetitive elements in the beetle *Tribolium madens*. *Gene* 394:45–52.
- Mravinac B, Plohl M. 2010. Parallelism in evolution of highly repetitive DNAs in sibling species. *Mol Biol Evol.* 27:1857–1867.
- Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* 48:443–453.
- Noonan JP, McCallion AS. 2010. Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet.* 11:1–23.
- Paar V, et al. 2005. ColorHOR – novel graphical algorithm for fast scan of alpha satellite higher-order repeats and HOR annotation for GenBank sequence of human genome. *Bioinformatics* 21:846–852.
- Paar V, et al. 2011a. Large tandem, higher order repeats and regularly dispersed repeat units contribute substantially to divergence between human and chimpanzee Y chromosomes. *J Mol Evol.* 72:34–55.
- Paar V, Glunčić M, Rosandić M, Basar I, Vlahović I. 2011b. Intragenic higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees. *Mol Biol Evol.* 28:1877–1892.
- Palomeque T, Muñoz-López M, Carrillo JA, Lorite P. 2005. Characterization and evolutionary dynamics of a complex family of satellite DNA in the leaf beetle *Chrysolina carnifex* (Coleoptera, Chrysomelidae). *Chromosome Res.* 13:795–807.
- Palomeque T, Lorite P. 2008. Satellite DNA in insects: a review. *Heredity* 100:564–573.
- Pennachio LA, Rubin EM. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet.* 2:100–109.
- Pertile MD, Graham AN, Choo KHA, Kalitsis P. 2009. Rapid evolution of mouse Y centromere repeat DNA belies recent sequence stability. *Genome Res.* 19:2202–2213.
- Pezer Z, Brajković J, Feliciello J, Ugarković Đ. 2012. Satellite DNA-methylated effects on genome regulation. *Genome Dyn* 7:153–169.
- Pons J, Juan C, Petitpierre E. 2002. Higher-order organization and compartmentalization of satellite DNA PIM357 in species of the coleopteran genus *Pimelia*. *Chromosome Res.* 10:597–606.
- Pons J, Bucur R, Vogler AP. 2003. Higher-order repeats in the satellite DNA of the cave beetle *Pholeuon proserpinae glaciale* (Coleoptera: Cholevidae). *Hereditas* 139:28–34.
- Richards S, et al. 2008. The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452:949–955.
- Rosandić M, Paar V, Basar I. 2003. Key-string segmentation algorithm and higher-order repeat 16mer (54 copies) in human alpha satellite DNA in chromosome 7. *J Theor Biol.* 221:29–37.
- Rosandić M, et al. 2006. CENP-B box and pl $\alpha$  sequence distribution in human alpha satellite higher-order repeats (HOR). *Chromosome Res.* 14:735–753.
- Rosandić M, Glunčić M, Paar V, Basar I. 2008. The role of alphoid higher order repeats (HORs) in the centromere folding. *J Theor Biol.* 254:555–560.
- Rosandić M, Glunčić M, Paar V. 2013. Start/stop codon like trinucleotides extensions in primate alpha satellites. *J Theor Biol.* 317:301–309.
- Roth S, Hartenstein V. 2008. Development of *Tribolium castaneum*. *Dev Genes Evol.* 218:115–118.
- Rudd MK, Willard HF. 2004. Analysis of the centromeric regions of the human genome assembly. *Trends Genet.* 20:529–533.
- Rudd MA, Wray GA, Willard H. 2006. The evolutionary dynamics of alpha-satellite. *Genome Res.* 16:88–96.
- Schueler MG, Sullivan B. 2006. Structural and functional dynamics of human centromeric heterochromatin. *Annu Rev Genom Hum Genet.* 7:301–313.
- Sokoloff A. 1972. The biology of *Tribolium* with special emphasis on genetic aspects. Oxford: Oxford University Press.
- Sujjwattananat P, et al. 2015. Higher-order repeat structure in alpha satellite DNA occurs in New World monkeys and is not confined to hominoids. *Sci Rep* 5:10315.
- Terada S, Hirai Y, Hirai H, Koga A. 2013. Higher-order repeat structure in alpha satellite DNA is an attribute of hominoids rather than hominids. *J Hum Genet.* 58:752–754.
- Thakurela S, Sahu SK, Garding A, Tiwari VK. 2015. Dynamics and function of distal regulatory elements during neurogenesis and neuroplasticity. *Genome Res.* 25:1309–1324.
- Tomilin NV. 2008. Regulation of mammalian gene expression by retroelements and non-coding tandem repeats. *Bioessays* 30:338–348.
- Tyler-Smith C, Brown WRA. 1987. Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J Mol Biol.* 195:457–470.
- Ugarković Đ. 2005. Functional elements residing within satellite DNAs. *EMBO Rep* 6:1035–1039.
- Ugarković Đ, Podnar M, Plohl M. 1996a. Satellite DNA of the red flour beetle *Tribolium castaneum* – comparative study of satellites from the genus *Tribolium*. *Mol Biol Evol.* 13:1059–1066.
- Ugarković Đ, Durajlija S, Plohl M. 1996b. Evolution of *T. madens* (Insecta, Coleoptera) satellite DNA through DNA inversion and insertion. *J Mol Evol.* 42:350–358.
- Ugarković Đ. 2009. Centromere-competent DNA: structure and evolution. *Prog Mol Subcell Biol.* 48:53–76.
- Wang S, Lorenzen MD, Beeman RW, Brown SJ. 2008. Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. *Genome Biol.* 9:R61.
- Warburton PE, Willard HF. 1996. Evolution of centromeric alpha satellite DNA: molecular organization within and between human and primate chromosomes. In: Jackson M, Strachan T, Dover G, editors. *Human genome evolution*. Oxford: BIOS. p. 121–145.
- Warburton PE, et al. 2008. Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics* 9:533.
- Willard HF, Wayne JS. 1987. Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* 3:192–198.

Associate editor: Maria Costantini