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

Ines Vlahovic\(^1,1\), Matko Glunjčić\(^1,1,*\), Marija Rosandić\(^2\), Durdica Ugarković\(^3\), and Vladimir Paar\(^1,2\)

\(^1\)Faculty of Science, University of Zagreb, Zagreb, Croatia
\(^2\)Croatian Academy of Sciences and Arts, Zagreb, Croatia
\(^3\)Division of Molecular Biology, Institute Ruder Bošković, Zagreb, Croatia

\*These authors contributed equally to this work.

**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://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 TCAS 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/dgene, 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; Rosandic 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 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 (Vlahovic et al. 2015, unpublished data), 5mer in chimpanzee, 14mer in orangutan and 3mer in macaque (Rosandic et al. 2013). In assemblies of chromosomes of lower primates several pronounced 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), 5mer in owl monkey and 12mer in marmoset (Sujivattanarat 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 Pholeuson prosperinai a 532-bp dimer HOR unit composed of two types of 266 bp monomers was found (Pons et al. 2003). In Pimelia radula ascendens 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 brevicomis and Tribolium madens (Ugarkovic 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. brevicomis, T. madens and T. audax (Ugarkovic 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 (Ugarkovic 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 (Feliciello 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 (Rosandic et al. 2008).
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) (Rosandic` et al. 2003; Paar et al. 2005, 2011a, 2011b; Gluncˇic´ 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 (Gluncˇic´ and Paar 2012). The GRM algorithm is an extension of KSA (Key String Algorithm) framework (Gluncˇic´ 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 deleted (fig. 4C). Such HOR pattern, where some constituting 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).
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 $S_2$ (369 bp), (starting at position 2884), is highly homologous to monomer copy $S_3$ (starting at position 3614) and between them is the monomer copy $A_3$ (361 bp). Thus, the distance between the first nucleotides in the $S_2$ and $S_3$ 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 and peaks at 694, 733, 1067, 1427, 1741, 2473, 3126, and 3168 bp (H) represent GRM signature of riddled 4mer HOR (table 2).

**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).
Divergence among homologous monomers of types m1, m3, and m4 are 4.66 ± 2.02%, 3.92 ± 3.87%, and 5.61 ± 1.74%, while divergence of copies of this three types of monomers is 18.51 ± 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 3 0 0t o4 0 0 b p(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 ± 0.33%. Furthermore, monomers m3 (~314 bp) approximately correspond to a segment 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 3 0 0t o4 0 0 b p(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 ± 0.33%. Furthermore, monomers m3 (~314 bp) approximately correspond to a segment of

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Illustration of Origin of Pronounced HOR-signature GRM Peaks in Figure 3F as Distances between Start Positions of Approximately Homologous Monomers in Smer HOR Array from GG695826.1 (fig. 4C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance between start of monomers of the same type (bp)</td>
<td>HOR-signature peak (bp)</td>
</tr>
<tr>
<td>S1/S2 = 369 + 361 = 730</td>
<td>730</td>
</tr>
<tr>
<td>S2/S3 = 369 + 332 + 362 + 369 + 362 = 1794</td>
<td>1793</td>
</tr>
<tr>
<td>2/S2 = 363 + 369 + 329 = 1051</td>
<td>1052</td>
</tr>
<tr>
<td>S5/S2 = 370 + 363 + 369 + 329 + 362 + 369 + 361 = 2523</td>
<td>2523</td>
</tr>
</tbody>
</table>

Labeling of monomer copies: see figure 4C. Example: S1 denotes the second m5-type monomer (counting from top to bottom) in m5 column in figure 4C (HOR copy No. 3); S5 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 ± 2.02%, 3.92 ± 3.87%, and 5.61 ± 1.74%, while divergence of copies of this three types of monomers is 18.51 ± 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 ± 0.33%. Furthermore, monomers m3 (~314 bp) approximately correspond to a segment 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 ± 0.33%. Furthermore, monomers m3 (~314 bp) approximately correspond to a segment 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).
sequences obtained in this way, the one, which is closest to DS497953.1 using the key string AACCATAA. Of monomer reference sequence is obtained by GRM segmentation of reference sequence for Tcast-360 monomers. Slightly different monomers: supplementary table S1b, Supplementary Material online. Monomer zontal bar representing this monomer and labeled I). Consensus mono-

The monomer 42 (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 monomer 42 is denoted by m1, m2, m3, and m4 (top enumeration of columns). The constituent monomers (fig. 5).

Average divergence between copies within each monomer type is low (~1–3%), while divergence between monomer 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 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
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).
38bp segment at the end of type 1 consensus is highly homologous to 38bp 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 ~720bp monomers (type 1, consensus sequence: Supplementary Material online) and one ~2191bp 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 122bp monomers (type 1, consensus sequence: supplementary table S1h, Supplementary Material online) and one ~1108bp monomer (type 2). This HOR can also be characterized as weak.
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 fascin-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 (Janzén 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.
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. freemanii, 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

**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. 6A 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. 6A HOR copy No 2) is located in the second intron of visual system homeobox protein gene (VSX gen, LOC655933). The third HOR copy (fig. 6A HOR copy No 3) is located within second intron of mRNA transcription factor AP-4 gene (LOC656099). The fourth and the fifth HOR copies (fig. 6A 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 dacshund homolog 1 gene (LOC652934). The seventh HOR is located 367 bp from 3’ end of leucine rich repeat protein soc2 homolog gene (LOC103312590) 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.
Table 4
List of HORs in *Tribolium castaneum* Genome Identified in This Work

<table>
<thead>
<tr>
<th>HOR</th>
<th>Constituting monomers</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>5mer</td>
<td>Five Tcast-360</td>
<td>GGE95826.1</td>
</tr>
<tr>
<td>4mer</td>
<td>Three Tcast-360 + one -1067bp</td>
<td>D549795.1</td>
</tr>
<tr>
<td>4mer</td>
<td>Three -170bp + one -941bp</td>
<td>CM000279.1</td>
</tr>
<tr>
<td>4mer</td>
<td>Three -170bp + one -941bp</td>
<td>CM000277.2</td>
</tr>
<tr>
<td>4mer</td>
<td>Four -170bp</td>
<td>CM000277.2</td>
</tr>
<tr>
<td>6mer</td>
<td>Five -311bp + one 567bp</td>
<td>GGE94292.1</td>
</tr>
<tr>
<td>6mer</td>
<td>Five -720bp + one -2191bp</td>
<td>GGE94249.1</td>
</tr>
<tr>
<td>6mer</td>
<td>Five 122bp + one -1108bp</td>
<td>GGE95437.1</td>
</tr>
</tbody>
</table>

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 (Feliciello 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**


Associate editor: Maria Costantini