DNA Research, 2015, 22(6), 387–401
doi: 10.1093/dnares/dsv021
Advance Access Publication Date: 1 October 2015
Full Paper

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022

Full Paper

Genome-wide analysis of tandem repeats in *Tribolium castaneum* genome reveals abundant and highly dynamic tandem repeat families with satellite DNA features in euchromatic chromosomal arms

Martina Pavlek[1], Yevgeniy Gelfand[2], Miroslav Plohl[1], and Nevenka Meštrović[1,*]

[1]Ruđer Bošković Institute, Bijenička 54, Zagreb HR-10002, Croatia, and [2]Laboratory for Biocomputing and Informatics, Boston University, Boston, MA 02215, USA

*To whom correspondence should be addressed. Tel. +385 1-4571-273. Fax. +385 1-4561-177. E-mail: nevenka@irb.hr

Edited by Prof. Hiroyuki Toh

Received 30 June 2015; Accepted 26 August 2015

Abstract

Although satellite DNAs are well-explored components of heterochromatin and centromeres, little is known about emergence, dispersal and possible impact of comparably structured tandem repeats (TRs) on the genome-wide scale. Our bioinformatics analysis of assembled *Tribolium castaneum* genome disclosed significant contribution of TRs in euchromatic chromosomal arms and clear predominance of satellite DNA-typical 170 bp monomers in arrays of ≥5 repeats. By applying different experimental approaches, we revealed that the nine most prominent TR families Cast1–Cast9 extracted from the assembly comprise ~4.3% of the entire genome and reside almost exclusively in euchromatic regions. Among them, seven families that build ~3.9% of the genome are based on ~170 and ~340 bp long monomers. Results of phylogenetic analyses of 2500 monomers originating from these families show high-sequence dynamics, evident by extensive exchanges between arrays on non-homologous chromosomes. In addition, our analysis shows that concerted evolution acts more efficiently on longer than on shorter arrays. Efficient genome-wide distribution of nine TR families implies the role of transposition only in expansion of the most dispersed family, and involvement of other mechanisms is anticipated. Despite similarities in sequence features, FISH experiments indicate high-level compartmentalization of centromeric and euchromatic tandem repeats.

Key words: Tandem repeats, euchromatic regions, evolutionary trends, transposition, *Tribolium castaneum*

1. Introduction

Eukaryotes typically display high proportions of repetitive elements that outmatch 50% of the genome. Among these repetitive elements, satellite DNAs (satDNAs) are a class of diverse tandemly repeated DNA sequences that build long arrays located in heterochromatin and often represent the most abundant genome fraction.[1] They are usually highly prevalent at and around centromeres, which are regions with suppressed recombination.[2] Centromeric satDNAs change rapidly during evolution despite their conserved function at the centromeric

© The Author 2015. Published by Oxford University Press on behalf of Kazusa DNA Research Institute. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

locus.[3] In addition to extreme diversity in nucleotide sequences between species, centromeric satDNAs are typically characterized by sequential arrangement of monomers in the form of long arrays and by preferential monomer length corresponding to the size of nucleosomal DNA.[4]

Our current knowledge about satDNA evolution is mostly based on studies of centromeric satDNAs. SatDNAs evolve in concert as a result of molecular drive, a process by which mutations are homogenized throughout a family of monomers in a genome, and are fixed in a population.[5] Theoretical models predict unequal crossing-over and gene conversion as the most widespread mechanisms involved in dynamics of satDNAs evolution.[2,6,7] Studies on human centromeric alpha satDNA show that these mechanisms act more efficiently within arrays than between arrays, and decrease progressively between arrays on homologous and on heterologous chromosomes.[8] Internal tandem repeats (TRs) found in some transposable elements (TEs) boost them to propagate satellite sequences throughout the genome.[9,10] In addition, recent results strongly support the idea that rolling circle amplification and reinsertion of extrachromosomal circles can be important for efficient dispersion of satellite arrays through a genome.[11]

Genome-wide annotation and study of TR-rich regions from assemblies of complex genomes represent a challenge. Due to the long arrays composed of nearly identical monomers, these genome fractions remain the most poorly mapped in assembled genomes. Moreover, centromeric regions are commonly omitted from the assembly of a genome. Comprehensive bioinformatics analysis of large arrays of TRs (>10 kb) located in the euchromatic part of the human genome showed a wide range of monomer size variations, from several nucleotides to several kilobases.[12] Among them, one of the largest non-centromeric arrays is 600 kb to 1.7 Mbp long, being located on human chromosome 8 and composed of 12 kb long monomers. Bioinformatics analysis of two mouse whole-genome shotgun assemblies revealed eight new satDNAs including some chromosome-specific satDNA subfamilies, which serve as a unique chromosome bar code.[13] Concerning evolutionary dynamics, comparative analysis of the most abundant satDNA in *D. melanogaster*, known as 1.688 satDNA, shows differential rates of concerted evolution in distinct genomic regions, from euchromatin to centromeric heterochromatin.[14] In addition to variations in chromosomal distribution, a novel alignment-free algorithm applied to Human Satellite 3 estimates extreme array size variation (7–98 Mb) of the HSat3 between individuals on the Y chromosome, thus confirming considerable satDNA array size polymorphisms.[15] Recent reports on satDNA impact on euchromatic genome suggests roles in modulation of gene regulation,[16,17] in disease-associated gene mutations[18] as well as in accumulation of differences within genes between human and chimpanzee.[19]

Tribolium castaneum is considered to be the most important model organism after *Drosophila* in studies of insect development, population genetics, and comparative genomics. Its genome is also the first coleopteran genome to be sequenced.[20] The genomes of *Tribolium* beetles are characterized by massive blocks of one or two abundant species-specific satDNAs localized in centromeric regions.[21–25] In *T. castaneum*, two related subfamilies of TCAST satDNA, estimated to comprise up to 35% of the whole genome, encompass the centromeric heterochromatin.[26] Furthermore, the two types of TCAST-like elements are found dispersed within euchromatin. The first represents TCAST satellite-like elements in the form of short arrays (up to tetramers), whereas the second consists of TCAST-like elements embedded within a complex unit similar to DNA transposon.[27]

Three complementary approaches, used for *de novo* genome-wide identification of repetitive DNA, recovered >30% of repetitive DNA in the assembled part of the *T. castaneum* genome.[28] All *T. castaneum* chromosomes are characterized by large blocks of heterochromatin surrounding centromeres, while no prominent heterochromatic blocks could be detected cytogenetically on chromosome arms.[29] Although centromeric satDNA is underrepresented in the assembled genome of *T. castaneum* analyses performed by Wang et al.[28] revealed significant portion of TRs with monomer over 100 bp in length. These results indicate the existence of a number of satDNA candidate sequences distributed outside the centromeric regions, in euchromatic chromosomal arms. The availability of a genome assembly based on whole-genome shotgun (WGS) reads as well as on Fosmid and BAC end-sequences offers a good platform for genome-wide study of TRs. A similar study of evolutionary trends of 1.688 repetitive DNA has already been performed for the *Drosophila* genome[14] where sequencing and assembly approaches were used in a way similar to that in *T. castaneum*.[20]

In this work, combining bioinformatics and experimental approaches, we identified and studied content, distribution, and structural features of TRs in the *T. castaneum* genome. Given the general lack of knowledge about TRs in euchromatic regions, we further focused on the nine most abundant TR families detected in the assembled genome. FISH experiments confirmed their almost exclusive localization in euchromatic chromosomal arms, while phylogenetic analyses revealed their highly dynamic evolution, particularly evident in their extensive exchanges between non-homologous chromosomes. Our results also suggest that in addition to other mechanisms, transposition may play an important role in the efficient spread of the most expanded TR family.

2. Materials and methods

2.1 Satellite DNA database construction and phylogenetic analyses

The *T. castaneum* (Tcas3.0) genome was assembled using the gold-standard Sanger assembly strategy with the benefit of extensive genetic maps obtained by the Tribolium Genome Sequencing Consortium.[20] More detailed, high-quality sequence reads were produced from WGS sequencing libraries of ∼3 and 6 kb in pUC18 subclones, as well as additional reads from fosmids (40 kb) and BACs (130 kb). These reads were assembled using the Atlas suite of assembly tools.[20] The assembled genome of *T. castanem* was downloaded in the fasta format from the web page ftp://ftp.bioinformatics.ksu.edu/pub/BeetleBase/3.0/ (The third version of the assembly: Tcas_3.0) in the form of 10 chromosomes. The sequence of each chromosome was uploaded to Tandem Repeats Database (TRDB),[30] https://tandem.bu.edu/cgibin/trdb/trdb.exe and analysed using Tandem Repeats Finder (TRF) algorithm.[31] All chromosomes have been processed with TRF using alignment parameters 2, 7, 7 for match, mismatch, and indels, respectively, and a minimum alignment score of 50. The range of the period size 100–500 bp was selected to exclude the micro/minisatellite fraction. The initial raw TRF output included TR arrays with overlapping genome coordinates. Using the redundancy elimination tool in TRDB and additional manual elimination, arrays with the shorter monomer units were selected for further analyses. Arrays with unspecified (N) nucleotides in both flanking regions were removed. Analysis of monomer length trends in extracted arrays was performed using the filtering option for copy number in arrays. For further analyses arrays with ≥5 monomers were selected. They were merged and analysed with the clustering tool integrated in TRDB. Conditions were as follows: *P*-value excluded (set to 0), cut-off value set at 70%, heuristical and DUST algorithm excluded, PAM algorithm included with default

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022

values (0.7 and 0.3). The result of these analyses was formation of clusters, i.e. groups of arrays that represent TR families. Monomer sequences of all arrays from selected clusters were downloaded in the fasta format. Left and right flanking regions (extending up to 4000 bp) of all selected clusters were also downloaded. Multiple sequence alignments of monomers were done using Clustal W for each TR family. Alignments without truncated monomers from the beginning and the end of the array were used for phylogenetic analyses. The Lasergene software package v.7.0.0 (DnaStar) was used in dot plot analyses and PCR primer design. Monomer sequence variability was analysed using DnaSP v.4.10.9.[32] Maximum likelihood (ML) trees based on Clustal W alignment were obtained with the PhyML 3.0 software[33] using best-fit models calculated by the jModelTest 2.1.3.[34] Due to the large number of monomers, branch support was evaluated using the approximate likelihood ratio test.[35] Trees were displayed with FigTree v1.40 and adjusted in Corel11 software. All new TR families were blasted against the NCBI GenBank Database and Repbase[36] to check similarity with published sequences. Sequence editing, selection of restriction enzymes (REs) and MUSCLE alignments of flanking regions were performed using the Geneious 5.4.3 program.[37] In order to extract TCAST variants from unassembled reads, TRF and clustering analyses were done on 2153 unassembled reads (ftp://ftp.bioinformatics.ksu.edu/pub/BeetleBase/3.0/) under the TRF parameters described above. The consensus sequences of five TCAST subfamilies were constructed based on multiple alignments of all extracted variants (available upon request). Recombinant clones with monomers or dimers of satellite DNA were sequenced by the Macrogen Europe Laboratory (Amsterdam, The Netherlands). Monomers of 9 TR families were deposited in EMBL databank under Accession Numbers: Cast1 (KP846079-KP846566), Cast2 (KP846568-KP847079), Cast3 (KP847080-KP847309), Cast4 (KP847310-KP847735), Cast5 (KP847736-KP848112), Cast6 (KP848113-KP848269), Cast7 (KP848352-KP848270), Cast8 (KP848353-KP848472), Cast9 (KP848473-KP848536) and consensus sequence of new TCAST subfamilies (KR046220-KR046222).

2.2 DNA isolation, cloning and sequencing

A *T. castaneum* culture (laboratory strain) was obtained from the Central Science Laboratory (Sand Hutton, York, UK). Insects were maintained on flour and kept in glass jars at room temperature, in a laboratory at the Ruđer Bošković Institute. Genomic DNA was isolated from adults by standard phenol–chloroform extraction. Primers were constructed based on the consensus sequences of each TR families as well as on the R66-like flanking region. Primer sequences and their positions on monomer consensus sequences are indicated in Supplementary Fig. S4 and Table S1. The reaction mixture consisted of a reaction buffer, 1.5 mM MgCl2, 0.2 mM dNTPs, 0.5 U GoTaq DNA polymerase (Promega), 0.4 mM of each primer and 20 ng of genomic DNA. The PCR cycling parameters used were as follows: 2 min initial denaturation at 94°C, followed by 30 cycles of: 95°C for 30 s, 55°C for 30 s and 72°C for 1 min. Final extension was at 72°C for 10 min. PCR products were purified using the QIAquick PCR Purification Kit (Qiagen). PCR products were ligated in a pGEM T-Easy vector (Promega) and transformed in *Escherichia coli* DH5a-competent cells (Invitrogene). Recombinant clones with monomers or dimers of satellite DNA were sequenced by the Macrogen Europe Laboratory (Amsterdam, The Netherlands).

2.3 Southern and dot blot analyses

Standard procedures were used for restriction endonuclease digestions, electrophoresis and transfer to nylon membranes. For genomic Southern hybridization analysis, 8 μg of genomic DNA was digested with REs which cut once in most monomer sequences and REs with recognition sites only in some monomers; HinfI and HaeIII (Cast1), EcoRI and HaeIII (Cast2), HaeIII and HinfI (Cast4), RsaI and DraI (Cast 5) and HaeIII and HinfI (Cast7). Digested DNA was separated by electrophoresis in a 0.8% agarose, denatured and DNA transferred to Hybond N+ membrane (Amersham). Cloned satellite monomers labelled with biotin-16-dUTP by PCR were used as hybridization probes. Hybridizations were performed overnight under moderate stringency conditions (65°C) in buffer containing 250 mM Na_2HPO_4 (pH 7.2), 20% SDS, 1 mM EDTA, 0.5% blocking reagent and 50 ng/ml of the probe. Post-hybridization washes were done in 20 mM Na_2HPO_4/1 mM EDTA/1% SDS at a temperature 2°C lower than the hybridization temperature. Chemiluminescent detection was carried out using the alkaline phosphatase substrate CDP-Star (Roche Applied Science). The abundance of TR families was estimated by quantitative dot-blot analysis using a series of genomic DNA dilutions ranging from 50 to 200 ng. Satellite monomers, excised from a plasmid, were dot-blotted in the range between 0.05 and 1 ng, and used as a calibration curve.

2.4 Chromosome preparations and two-colour fluorescence *in situ* hybridization

Male gonads were isolated from adults and chromosome spreads were prepared by the 'squash' technique as described previously.[38] Detailed mapping of nine new TR families on chromosomes in meiotic pro metaphase was not possible due to extremely rare observation of this phase. The condensation state in this phase also causes lower FISH sensitivity especially in the case of low copy families. Technical difficulties were overcome by using chromosomes in mitotic pro metaphase which enabled detection of centromere regions together with FISH signals of newly detected TR families. Two-colour FISH was carried out to determine the positions of new TR families related to centromeric regions. A TCAST-specific probe was generated by nick translation labelling of cloned dimer sequences with Cy3-dUTP using the Nick Translation Mix (Roche Applied Science). FISH probes for TR families detected in this work (Cast1–Cast9) were obtained by PCR labelling of cloned monomers with biotin-16-dUTP (Roche Applied Science) by using M13 forward and reverse primers (Invitrogen). In order to investigate *in situ* localization of Cast5 arrays with respect to their flanking regions, two-colour FISH was used. Probes were generated by PCR labelling of cloned Cast5 with Cy3-dUTP and R66-like flanking sequences with biotin-16-dUTP. Hybridizations were performed for 18 h at 37°C in a solution containing 60% formamide, 2× SSC, 10% dextran sulphate, 20 mM sodium phosphate, and 10 ng/μl of each probe. Post-hybridization washes were done in 50% formamide/2× SSC at 37°C. Biotin-labelled probes were detected with fluorescein avidin D and biotinylated antiavidin D (Vector Laboratories). The chromosomes were counter-stained with DAPI 4′6-diamidino-2-phenylindole (Invitrogen) and analysed with appropriate filters on an Olympus BX51 epifluorescence microscope equipped with an Olympus DP70 digital camera system. Merging of images was performed using Adobe Photoshop CS5 Extended Version 12.0 software.

3. Results

3.1 Distribution and monomer length of TRs in the assembled *T. castaneum* genome

In order to identify features and distribution of TRs in the output of the assembled *T. castaneum* genome, we analysed all 10 assembled

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022

chromosomes individually using the TRF software. Two thousand nine hundred and sixty arrays of TRs comprising a total length of 3.25 Mb were obtained. These repeats constitute 2.1% of the 156 Mb long *T. castaneum* genome assembly. To determine whether TRs identified in our study might show any preferential genome localization with respect to the repeat copy number, distributions of short (<5 monomers) and long (≥5 monomers) arrays were analysed separately and presented along 10 *T. castaneum* chromosomes (Fig. 1). The 300 kb long uncaptured gaps between scaffolds were taken into account to mark dominant gaps in the assembled genome. The abundance of short and long TR arrays is evidently higher in CH3, CH6, CH8, CH9 and CH10 than in other chromosomes. The obtained results are in accordance with the previous study of Wang et al.,[28] where RepeatScout and TEpipe tools revealed higher accumulation of the repetitive class named HighA and of TEs in the same chromosomal subset. Here, short TR arrays (<5 monomers) showed almost uniform distribution along the assembled chromosomes (Fig. 1). Interestingly, long arrays (≥5 monomers) displayed a higher tendency to reside in the euchromatic chromosomal compartment, being less represented in the putatively heterochromatic domains proposed by Wang et al.[28] Although the observed trend of long array distribution could be affected by gaps in the assembly of tandemly repeated sequences, marked uncaptured gaps do not seem to be more frequent in putative heterochromatic domains than in euchromatic chromosomal segments. Declared heterochromatic domains in the assembled genome represent chromosomal regions with high proportion of HighA and TE repetitive classes. Taking into account that the HighA sequence library includes 30% of TEs obtained by the TEpipe algorithm, and only a small fraction (~6%) of TRs obtained by TRF,[28] we conclude that HighA is mostly built of dispersed repetitive sequences, for example, non-autonomous TEs.

To explore trends of monomer length in extracted arrays and a possible correlation with copy number of monomers, all 2960 arrays were subdivided into three classes: arrays with ≤2 monomers (634 arrays), 3–4 monomers (1563 arrays) and ≥5 monomers (763 arrays). Each class was analysed separately (Fig. 2). Arrays with only two repeat units are predominantly built of monomers having a length between 100 and 180 bp, while the number of array drops with increased monomer size (Fig. 2A). Arrays with 3–4 monomers as well as arrays ≥5 are predominantly built of 160–180 bp long monomers, while the relative contribution of arrays with <160 bp long monomers drops significantly (Fig. 2B and C). In the same time, the profile of arrays is enriched with 200–220 and 320–340 bp long monomers (Fig. 2B and C). Furthermore, a dramatic decrease in the number of long arrays is evident when monomer length exceeds 340 bp (Fig. 2C).

3.2 Revision of centromeric TCAST satDNA

Centromeric TCAST satDNA was identified experimentally.[21,26] Bioinformatic identification of repetitive DNA by RepeatScout revealed only 0.3% of TCAST satDNA in the assembled genome,[28] while the majority of sequenced TCAST satDNA was retained in unassembled reads. We extracted TCAST monomers from unassembled reads to define all sequence variants (data not shown). The alignment of all monomers revealed five subfamilies of TCAST satDNA with mutual sequence similarity of 70% and monomer length variation from 332 to 384 bp (Fig. 3). Two subfamilies (subf1 and subf3) were previously described,[21,26] while others are identified in this work for the first time. Dramatic monomer length variation is mainly due to insertion/deletion events in one region of extracted monomers. In order to explore the distribution of this satDNA in the assembled genome, we mapped TCAST satDNA on assembled chromosomes by BLAST search using consensus sequences of all subfamilies as queries (red spots in Fig. 1). We determined 130 short arrays of mainly 1–2 monomers, distributed randomly along chromosomes.

Due to small chromosome sizes (only 1–3 µm), previous FISH analysis of TCAST satDNA had very poor resolution.[21] To evaluate the centromeric localization of all TCAST subfamilies as well as to reexamine the karyotype with respect to the centromeric region, we performed individual FISH analyses with monomers specific for each of five subfamilies. Given that all FISH analyses show identical localization of TCAST subfamilies, we present only one of them in Fig. 4a. *Tribolium castaneum* has a $2n = 20$ complement of chromosomes and a meioformula 9+XYp.[29] Detailed cytogenetic analyses confirmed localization of TCAST subfamilies to centromeric regions of all chromosomes without any significant signals outside of centromeric regions. It must be noted that short TCAST arrays dispersed along the assembled genome are evidently below the detection level by FISH. FISH analyses also enabled identification of Yp, two metacentric chromosomes CH2 and CH3, and the largest telocentric chromosome CH4 (Fig. 4a). The remaining chromosomes are mostly small telocentrics.

3.3 Characterization of the most abundant TR families detected in the assembled genome

In order to explore the most abundant TR families in the assembled *T. castaneum* genome we further focused on arrays with ≥5 monomers obtained in the TRF analysis output. This cutoff level was also selected to prevent blurring the results of phylogenetic analyses on selected most abundant TR families, that would be caused by a large number of branches in phylogenetic trees derived from monomers in shorter arrays. This analysis recovered 763 arrays with ≥5 TRs, cumulatively making up 1.63 Mb that constitute 1.04% of the assembled genome. Extracted arrays were clustered using tools implemented in the TRDB.[30] Sequence identity >70% was selected as the clustering requirement in order to keep the identity level similar to that detected within the majority of satDNA families.[39] Using this parameter, 371 arrays were further classified into 56 clusters; 23 containing 3–42 arrays, while others were those with only two related arrays. In order to explore in detail the most prominent TR families, we selected nine largest clusters which were represented with at least two arrays on at least two chromosomes. These criteria were set to enable comparative studies of monomers from different arrays located on a single chromosome as well as comparisons of arrays among chromosomes.

Nine selected clusters comprise about two-thirds of arrays with ≥5 monomers (Table 1, clustering results are available on the TRDB website upon request). Monomers belonging to each cluster were further exported and named with letters and numbers related to the chromosome number, genomic position and position in the particular array. Each cluster thus represents the most abundant TR families hereafter indicated as Cast1–Cast9. Monomers were aligned with ClustalW (Supplementary Fig. S1) and structural characteristics of the nine TR families are summarized in Table 1. In agreement with the estimated correlation between monomer size and array length, 5 of 9 families are based on ~170 while two have ~340 bp long monomers. The remaining two families have monomers of ~110 and 210 bp, respectively. The number of tandemly repeated monomers in the obtained arrays is up to 54. The nucleotide sequences of all families show high AT content (≥60%) and 9–28% nucleotide diversity within the family. The abundance of newly detected TRs in the assembled genome ranges from 0.006% up to 0.075% (9–117 kb). We also observed that periodicity of AT tracts is a prominent feature of all analysed TRs (Supplementary Fig. S2).

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022

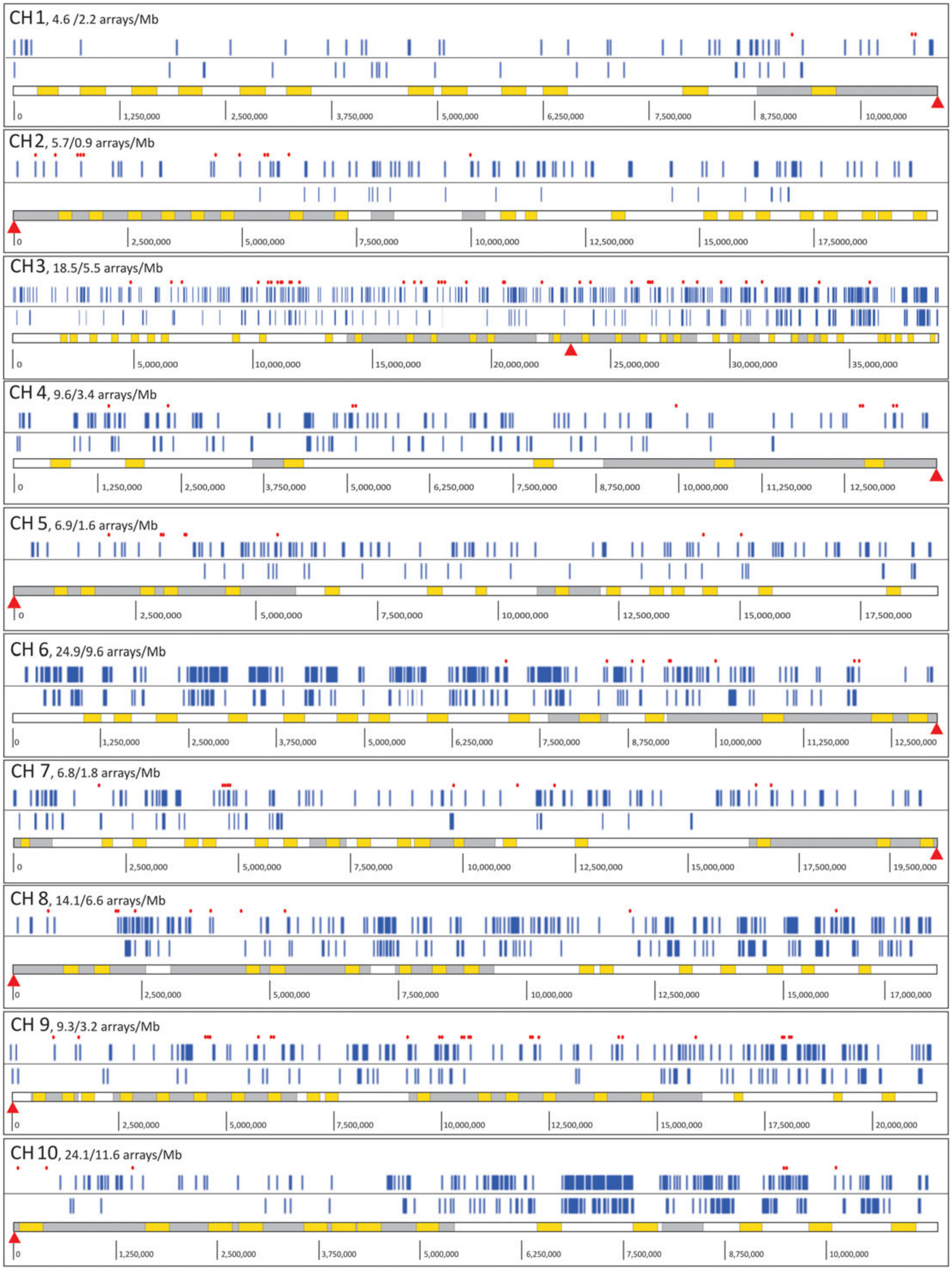


Figure 1. Genomic distribution of arrays of TRs on *T. castaneum* assembled chromosomes (CH1–CH10). Vertical bars represent short arrays (<5 monomers/array, upper line) and long arrays (≥5 monomers/array, lower line). The actual number of arrays per Mb for short and long arrays is indicated above each chromosome. Red dots correspond to centromeric TCAST satDNA found in the assembled genome. Horizontal bar represents putative euchromatin (white) and heterochromatin (HighA domain, grey) regions as identified in Wang et al.[28] Locations of the 300 kb placeholders were included to define uncaptured gaps (yellow). Red triangles indicate assumed position of the centromere and large blocks of centromeric heterochromatin based on our FISH analyses (Fig. 4), chromosome banding[29] and HighA domain (putative heterochromatin) defined in Wang et al.[28]

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022

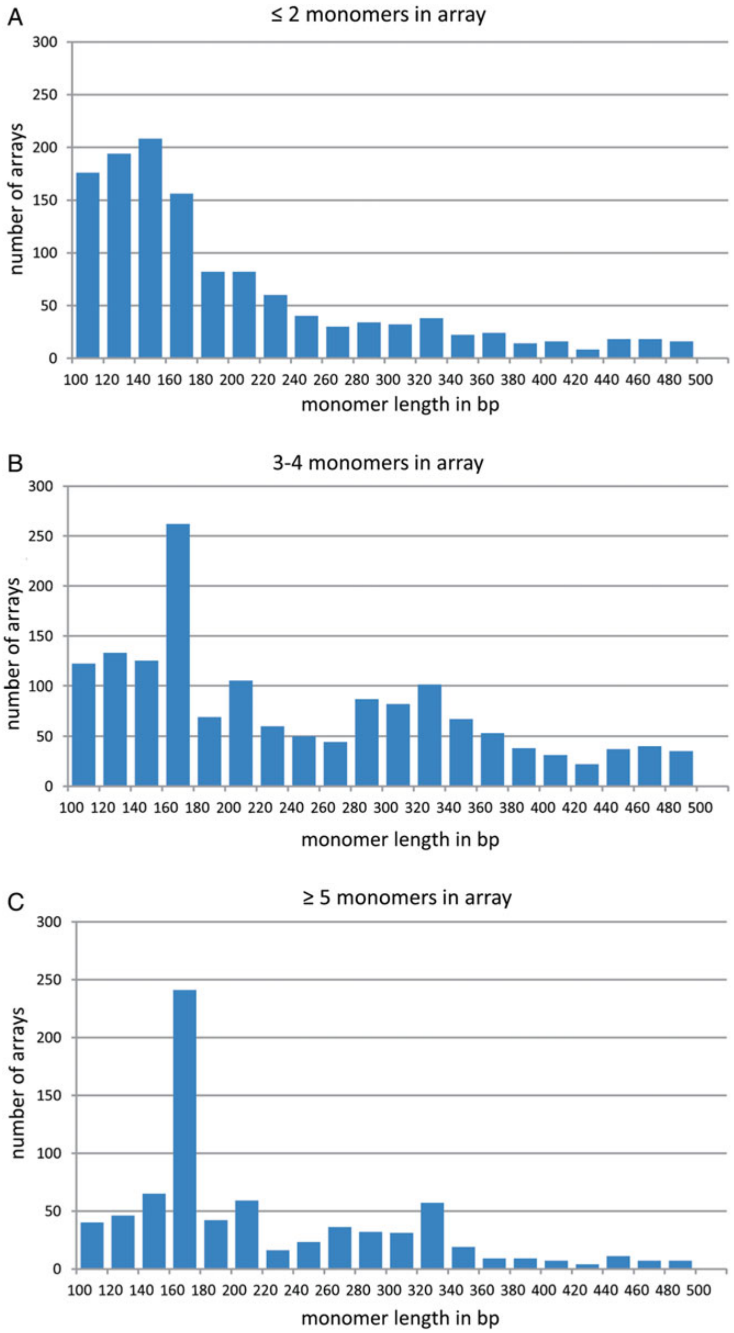
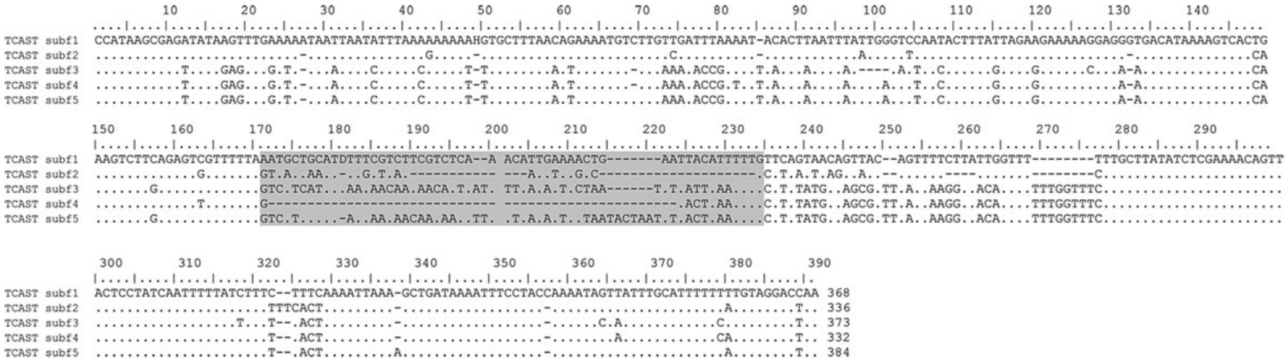


Figure 2. Correlation of monomer number in extracted TRF arrays and monomer length. Plotted is number of arrays as a function of monomer length for arrays with ≤2 monomers (A) 3–4 monomers (B) and ≥5 monomers (C). This figure is available in black and white in print and in colour at *DNA Research* online.

BLAST search of both GenBank and RepBase with the consensus monomer sequence of each TR family as a query did not reveal significant similarity with any other sequence. However, local BLAST search with new TR families on repetitive classes obtained previously by RepeatScout[28] recovered their significant homology with Cast4 and Cast5 (Supplementary Fig. S3). In particular, high homology is shown between the Cast4 family and 7 repetitive elements previously identified by RepeatScout which mostly represent dispersed dimers of Cast4 family as well as parts of monomers associated with different flanking regions. Similarly, three RepeatScout-defined elements show homology with Cast5. They are composed of Cast5 monomer and flanking regions (for details, see below).

In addition to bioinformatic predictions of abundance and genome distribution, the newly found TR families were also examined by FISH, Southern analyses and dot-blot experiments (Fig. 4 and Table 1). Individual TR family members were amplified with specific primers derived from consensus monomer sequences obtained in this work (Supplementary Fig. S4). Cloned and sequenced monomers or dimers were further used as a template for the probe in hybridization experiments. Taking into account that genome segments rich in TRs remain poorly assembled and underestimated in outputs of genome projects, dot-blot hybridization was performed for each *in silico* detected TR family. Cast1 and Cast5 are the most abundant, each comprising >1% of the genome (Table 1). Cast2, Cast4 and Cast6

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022



```
                    10        20        30        40        50        60        70        80        90       100       110       120       130       140
            ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....
TCAST subf1 CCATAAGCGAGATATAAGTTTGAAAAATAATTAATATTTAAAAAAAAAHGTGCTTTAACAGAAAATGTCTTGTTGATTTAAAAT-ACACTTAATTTATTGGGTCCAATACTTTATTAGAAGAAAAAGGAGGGTGACATAAAAGTCACTG
TCAST subf2 ..........................-...............G.....-........................C..........-............A.....T...........................-...............CA
TCAST subf3 ...........T....GAG...G.T.-...A....C.....C.....T-T........A.T.......-...AAA.ACCG....T.A...A....A.----.A.T..C......G....G......C...A-A..............CA
TCAST subf4 ...........T....GAG...G.T.-...A....C.....C.....T-T........A.T.......-...AAA.ACCG.T..T.A...A....A....A...T..C......G....G..........A-A..............CA
TCAST subf5 ...........T....GAG...G.T.-...A....C.....C.....T-T........A.T...........AAA.ACCG....T.A...A....A....A...T..C......G....G..........A-A..............CA

            150       160       170       180       190       200       210       220       230       240       250       260       270       280       290
            |....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....
TCAST subf1 AAGTCTTCAGAGTCGTTTTTAAATGCTGCATDTTTCGTCTTCGTCTCA--A ACATTGAAAACTG-------AATTACATTTTTGTTCAGTAACAGTTAC--AGTTTTCTTATTGGTTT--------TTTGCTTATATCTCGAAAACAGTT
TCAST subf2 .............G.......GT.A..AA..-..G.T.A.----------- ---A..T..G.C--------------------.C.T.A.T.AG..A..--......----.......--------C.......................
TCAST subf3 .......G.............GTC.TCAT...AA.AACAA.AACA.T.AT. TT.A.A.T.CTAA------T.T.ATT.AA....C.T.TATG..AGCG.TT.A..AAGG..ACA....TTTGGTTTC.......................
TCAST subf4 .............T.......G----------------------------- ----------------------.ACT.AA....C.T.TATG..AGCG.TT.A..AAGG..ACA....TTTGGTTTC.......................
TCAST subf5 .......G.............GTC.T.....-A..AA.AACAA.AA..TT. .T.A.A.T..TAATACTAAT.T.ACT.AA....C.T.TATG..AGCG.TT.A..AAGG..ACA....TTTGGTTTC.......................

            300       310       320       330       340       350       360       370       380       390
            |....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|.
TCAST subf1 ACTCCTATCAATTTTTATCTTTC--TTTCAAAATTAAA-GCTGATAAAATTTCCTACCAAAATAGTTATTTGCATTTTTTTTGTAGGACCAA 368
TCAST subf2 ......................TTTCACT.........-..................-......................A........T.. 336
TCAST subf3 ..................T...T--.ACT.........-..................-......C.A............C.........T.. 373
TCAST subf4 ......................T--.ACT.........-..................-........A............CA........T.. 332
TCAST subf5 ......................T--.ACT.........A..................-......................A........T.. 384
```

Figure 3. Alignment of consensus monomer of five TCAST satDNA subfamilies from unassembled reads of *T. castaneum* genome. Consensus sequences of subfamilies are derived according to the majority principle in alignment of variants from unassembled reads. Monomer length variation in the central region is highlighted in grey.

are less represented, each making up 0.5% of the genome, while other TR families comprise ~0.2% of the genome. In total, according to the results of dot-blot assay, estimated abundance of all analysed TR families is >4% of the genome. The experimentally determined genome content of these TR families is ~10 times higher than the representativeness in the assembled genome. For further support of abundance, we also analysed the presence of these TR families in *T. castaneum* unassembled reads, and these estimations confirmed the highest relative proportion of Cast1, Cast2, Cast5 and Cast6 (data not shown). It is important to observe that among the experimentally estimated contribution of TR families that make >4% of the genome, the most abundant (>2.7%) are those based on ~170 bp monomer, while families with ~340 bp monomer comprise >1.2% of the genome.

To validate the TR profile of the sequence sets generated *in silico*, Southern blot hybridization analyses were carried out on the most prominent TR families (Fig. 4). Genomic DNA was digested using REs which cut once in the majority of monomer sequences as well as with REs having a recognition site only in some monomers. In addition to a strong signal of predicted monomer size, typical satellite ladder pattern was observed in all hybridization analyses.

In order to detect chromosomal localization of Cast1–Cast9 families relative to centromeric regions, two-colour FISH was done combining each detected TR family and the centromeric TCAST satDNA. FISH analyses on chromosomes in mitotic pro metaphase enabled detection of centromeric regions together with signals of newly detected TR families (Fig. 4). As already shown by dot-blot experiments, signals obtained after FISH with the most abundant Cast1 and Cast5 were significantly stronger relative to other Cast families. Interestingly, FISH results show localization of all nine TR families almost exclusively on non-centromeric/euchromatic chromosome parts. Only a few overlapping signals with TCAST could be observed in centromeric regions, particularly in the case of Cast5.

3.4 Study of TR family evolution in euchromatic regions

In order to assess evolutionary trends of dominant TR families located outside of centromeric regions in *T. castaneum*, we examined phylogenetic relationships between their monomers (Supplementary Fig. S1). Two thousand five hundred monomers originating from the nine selected families (Cast1–Cast9) were included in the analyses. Annotation with respect to their actual position allows detailed identification of phylogenetic relationships among monomers within and between arrays and chromosomes (Supplementary Fig. S5). Simplified forms of ML trees with marked subgroups which show recent exchange events within a chromosome and between non-homologs are presented in Fig. 5. In all ML trees, the monomer groups show a significant level of exchange events which occurred between non-homologous chromosomes.

The ML tree of Cast1 shows three groups of monomers originating from distant arrays (>20 kb) on homologous chromosomes and a recent exchange between non-homologous chromosomes (Fig. 5A). However, according to our FISH results, it can be concluded that the Cast1 family makes very long arrays which are not represented in the assembled genome.

The Cast2 phylogenetic tree displays two distinct branches with monomers originating from arrays located on 6 or 7 non-homologous chromosomes (Fig. 5B). This dispersion pattern together with short branches grouped in subclusters indicates a relatively recent expansion of arrays between non-homologous chromosomes which occurred in two independent events. In addition, monomers of this family form one chromosome-specific cluster composed of monomers originating from three distantly located arrays (separated by at least 20 kb). In addition, the Cast2 tree reveals exchange events between the sex-chromosome (X, marked as CH1) and autosomes.

The Cast3 family is characterized by low monomer divergence (11%; Table 1) and short arrays mainly comprised of 5–7 monomers (Fig. 5C). Cast3 monomers derived from the majority of short arrays show a scattered distribution in the ML tree (see symbols on the tree), while those from longer arrays (up to 19 monomers) group together.

The Cast4 family is characterized by significant monomer divergence (21%; Table 1) observed even among monomers within arrays, as evident by long branches in the tree. Nevertheless, monomers from the same array tend to group together (Fig. 5D). Two modes of chromosome-specific clustering are distinctive: (i) arrays are located close to each other (<20 kb) and probably homogenized together, and (ii) arrays are distant (>20 kb) and show intra-chromosomal exchanges.

Regardless the lowest overall monomer sequence divergence (9%), Cast5 shows clear grouping of monomers originating from the same arrays (Fig. 5E). In addition, monomers from the same chromosome show a tendency to group together even when arrays are positioned at distant locations. The tree also suggests recent inter-chromosomal exchange in a fraction of monomers. Cast7, Cast8 and Cast9 trees are not shown here because they follow similar evolutionary trends as for the TR families described above.

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022

Table 1. Structural features of euchromatic TR families

TR families	Number of arrays	Number of arrays per chromosome										Max. number of monomers per array	Avarage number of monomers per array	Nucleotide diversity (Pi) of monomers in cluster ± standard deviation	The monomer length (bp)	AT content	Number of monomers	Total repeat family length (kb)	% of assembled genome	% of genome estimated by dot blot
		CH1	CH2	CH3	CH4	CH5	CH6	CH7	CH8	CH9	CH10									
Cast1	46	0	1	13	1	1	8	3	4	10	5	31	10.63	0.20143 ± 0.00320	166–173	66.2	489	83	0.053	>1
Cast2	42	6	1	4	5	4	4	1	10	1	3	33	12.19	0.12138 ± 0.00303	166–172	72.3	512	87	0.056	0.5
Cast3	35	2	1	3	6	3	7	2	5	2	4	17	6.57	0.11734 ± 0.006	205–219	74.8	230	48	0.031	0.2
Cast4	33	0	1	5	0	2	3	8	8	4	0	31	12.91	0.21765 ± 0.00296	168–176	69.7	426	73	0.047	0.5
Cast5	30	0	1	6	3	3	5	1	7	1	3	28	11.6	0.09620 ± 0.00182	270–338	73.1	384	117	0.075	>1
Cast6	6	0	0	4	0	1	0	0	1	0	0	54	21.17	0.16874 ± 0.00385	179–181	66	157	28	0.018	0.5
Cast7	10	0	0	4	1	2	1	0	2	0	0	12	8.3	0.28310 ± 0.00129	109–114	68.9	83	9	0.006	0.2
Cast8	10	0	0	2	0	0	1	0	1	4	2	24	12	0.16321 ± 0.00733	161–167	67.3	120	20	0.013	0.2
Cast9	10	0	0	2	0	1	3	0	1	0	3	10	6.4	0.25299 ± 0.00481	311–346	72.6	64	21	0.013	0.2
Total	222																2501	486	0.312	>4.3

Interestingly, phylogenetic analysis of Cast6 (Fig. 5F) revealed that its long arrays are predominantly located on the chromosome CH3. Monomers in arrays evidence significant intra-chromosomal exchange. An almost exclusive localization of the Cast6 family on CH3 was also confirmed by FISH with the Cast6 monomer as a probe on the meiotic pro metaphase plate (Fig. 5G). The coherence between sequence data for Cast6 from the assembled genome and FISH experiments further confirms the authenticity of the genome assembly.

Although each family shows its own pattern in the phylogenetic tree, common features can be deduced. The presented tree topologies show predominant clustering of monomers derived from the same array. This is particularly true for monomers derived from long arrays while those from short arrays generally do not show any consistent grouping. A general observation deduced from all extracted TR families is that monomers in arrays positioned on the same chromosome do not group with a frequency significantly higher than monomers located on non-homologous chromosomes. In support of this, but with exception of Cast6, all of the TR families show grouping of monomers from arrays located on nearly all non-homologous chromosomes, thus suggesting extensive inter-chromosomal exchanges.

3.5 Possible mechanisms of propagation

In order to investigate putative propagation mechanisms for the analysed TR families we first wanted to check if they are just passively carried by expansion of other DNA segments in the genome. If this is so, it would be expected that array flanking regions or at least some of them are mutually homologous. To address this question, we generated 4 kb long left and right array flanking regions from the TRF output and compared them separately for each family. Our results show that among the nine TR families, only Cast2 and Cast5 display homologies in their flanking regions. The Cast2 family show homology only in a small number of left flanking regions (7 of 36 arrays), while the vast majority of Cast5 arrays (22 of 28) could be grouped according to similarities in left and right flanking regions. Flanking regions of Cast5 are further similar to R66 and R140 repetitive sequences (Supplementary Fig. S3) extracted previously from the sequenced *T. castaneum* genome by RepeatScout analyses.[28] Detailed analyses of R66 and R140 show that both sequences are composed of Cast5 monomer parts together with parts of the array flanking region. Alignments of R66 and R140 (without the monomer part) with the flanking sequences of Cast5 family are presented in Supplementary Fig. S6. In both cases, flanking regions show homology in 1 kb long segments, although R140-like flanking regions in some variants extend up to 2 kb. Significant variations were detected in positions of both left and right junctions of flanking sequences with respect to the TR array. A search of RepBase with R66-like and R140-like Cast5 flanking regions as a query shows stretch of 140 bp with a high homology (84%) of R140-like sequence to the non-autonomous Tc1/Mariner transposon defined in *T. castaneum*. This result suggests possibly a transposon property of Cast5 flanking regions (Supplementary Fig. S6). In support, two-colour FISH with Cast5 monomer and centromeric satDNA (Fig. 4F) indicates higher expansion of Cast5 family in comparison with all other families.

A schematic presentation of flanking regions and arrays of Cast5 family members with respect to monomer composition and orientation is presented in Fig. 6A. The same orientation of monomers relative to flanking regions can be observed in almost all arrays, while flanking modules and monomers together are positioned in both directions with respect to the orientation of genomic sequences.

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022

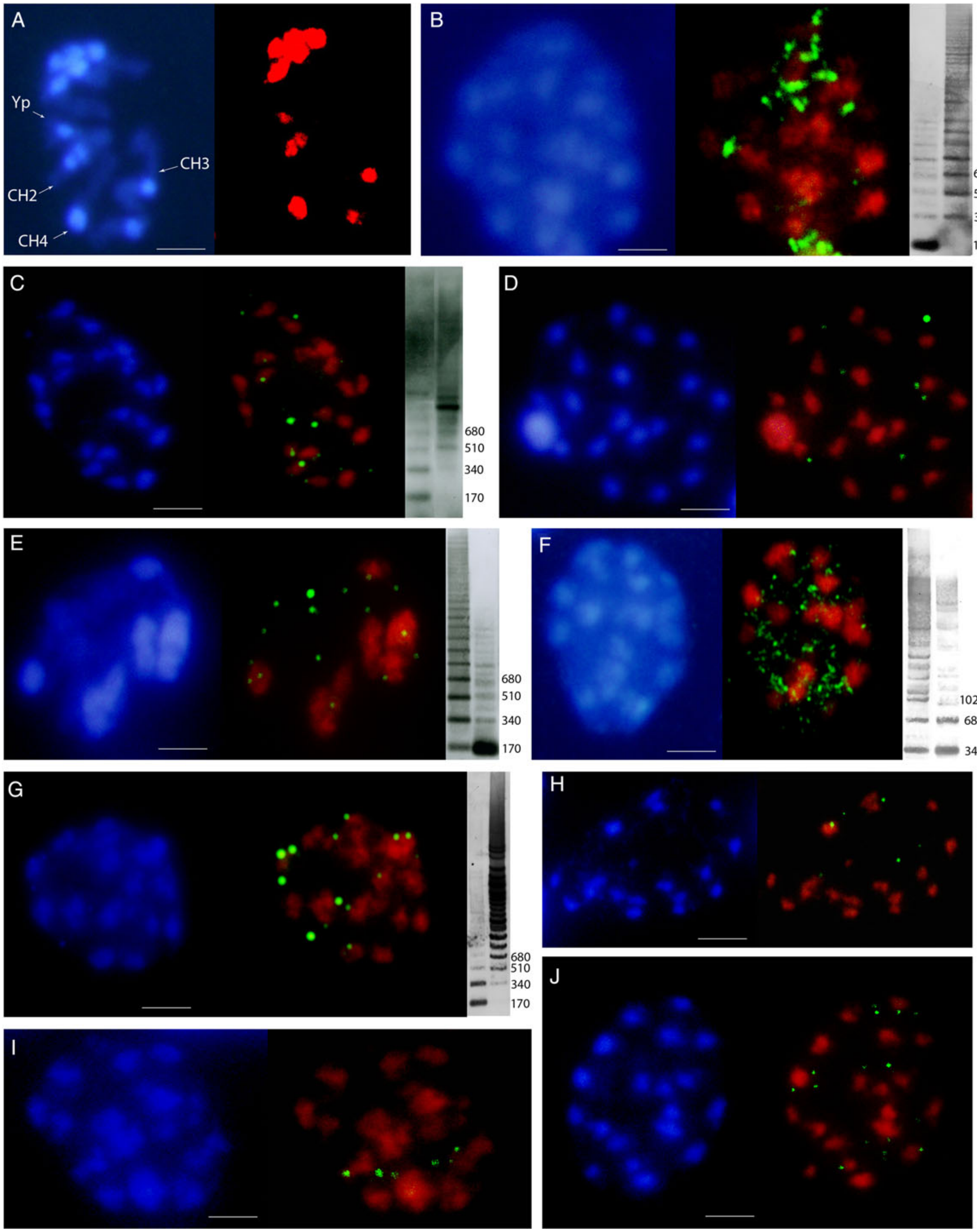
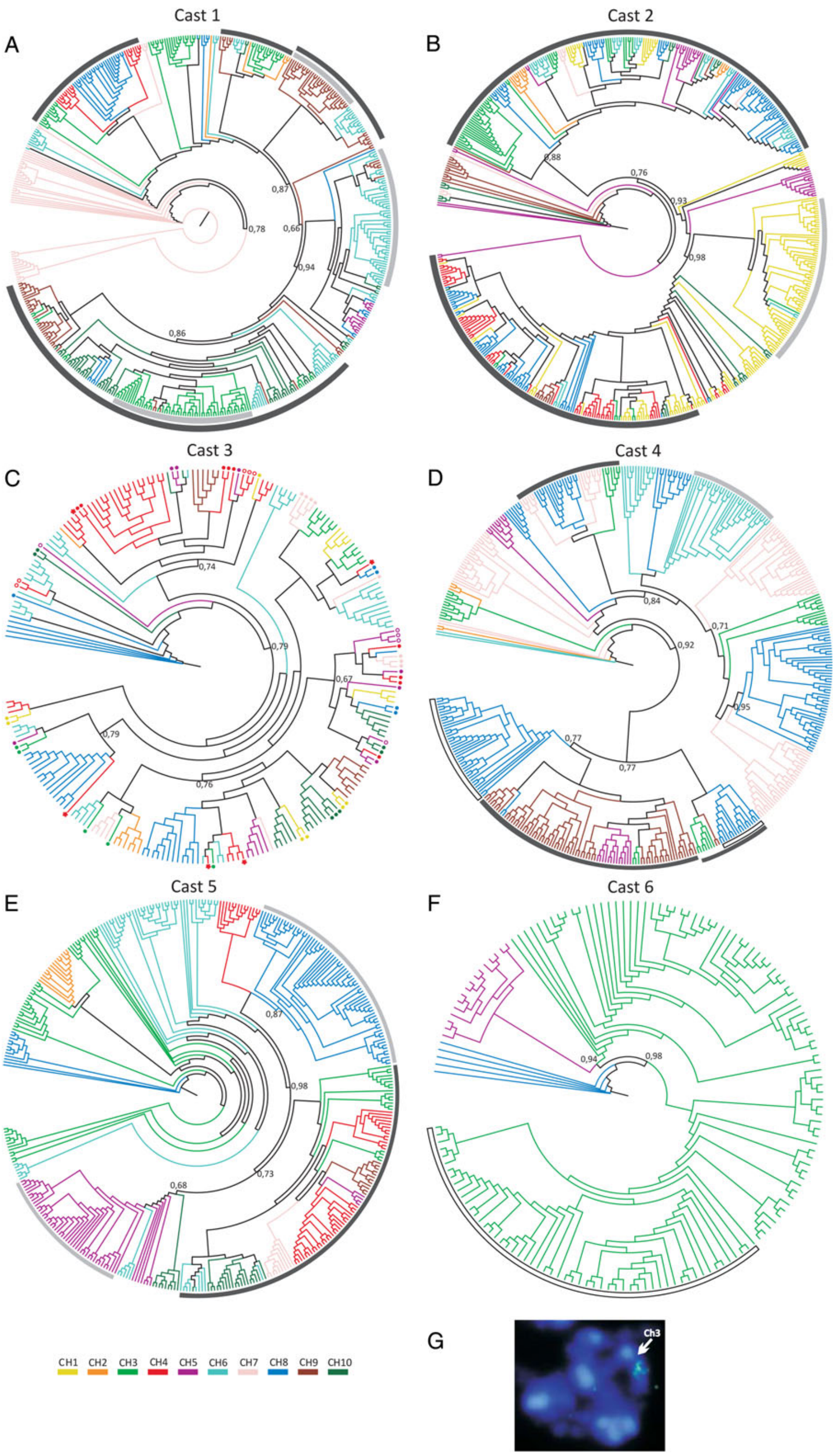


Figure 4. Fluorescence *in situ* hybridization of centromeric TCAST satDNA and TR families determined in this work by TRF analysis. Chromosomes are counter-stained with DAPI. FISH showing centromeric TCAST satDNA (TCAST subf3 as a probe; red signals) on *T. castaneum* chromosomes in meiotic pro metaphase (A). Arrows point to chromosomes CH2, CH3, CH4 and Yp. Chromosomes were named according to the karyotype analysis provided by Stuart and Mocelin.[29] Two-coloured FISH performed on chromosomes in mitotic pro metaphase show localization of new TR families (green): Cast1 (B), Cast2 (C), Cast3 (D), Cast4 (E), Cast5 (F), Cast6 (G), Cast7 (H), Cast8 (I) and Cast9 (J) with respect to centromeric regions marked with TCAST satDNA (red). The bar represents 1 μm. Aside to chromosome spreads are shown Southern blot analyses of genomic DNA digested with REs and hybridized with Cast1 (B), Cast2 (C), Cast4 (E), Cast 5 (F) and Cast6 (G). Only TR families with >0.5% of genomic DNA are presented.

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022



Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022

To investigate the extent of genomic co-localization of Cast5 repeats and flanking regions, we further performed two-colour FISH experiment with probes specific for the flanking sequence similar to R66 and for the Cast5 monomer (Fig. 6B). The results obtained on meiotic pro metaphase chromosome spreads produced mostly co-localized hybridization signals, although individual signals corresponding to the flanking region as well as to TR family alone can be seen. This result is in accordance with bioinformatic analysis of Cast5 arrays in which some flanking sequence diverges significantly from R66-like.

Given that the other analysed TR families do not show any regularity in flanking regions, we inspected junction position in monomer sequences in order to asses if there is any preferential monomer part that might be linked with the transition site. Inspection of these junctions did not reveal any evident correlation between the sequence and junction position (data not shown).

4. Discussion

A global survey of TRs throughout the entire genome is important for correct annotation of these sequences in genome assemblies, and also to improve insights into their evolutionary dynamics and mechanisms of emergence and expansion. The availability of a whole-genome assembly of *T. castaneum* mapped to chromosomes was the most important prerequisite in addressing these issues, making possible examination of TR evolution at the chromosomal and at the repeat-array level. To reveal the distribution profile and structural features of TRs as well as to build a database of highly abundant TRs, we first applied TRF analysis to the *T. castaneum* genome assembly, and focused on arrays with repeats in the range of 100–500 bp. The total amount of detected TRs is 1.63 Mb, which constitutes ~1.04% of the assembled genome. Arrays are not uniformly distributed between chromosomes, showing higher density on five chromosomes. This result is in agreement with a previous report of overall repeat families and TE classes in the assembled genome.[28] However, our combined FISH and bioinformatics studies of TCAST, the most abundant *T. castaneum* satDNA,[26] confirmed the complete absence of centromeric regions in the genome assembly. The distribution profile of TCAST shows dominant localization in centromeric regions and arrays with only a few copies outside of centromeric regions. Another centromeric satDNA from *T. castaneum*, TCAST2, is similar to TCAST, and found in diverse euchromatic locations only in the monomeric form.[40] Distribution profiles with large blocks of satDNA in centromeric regions and short arrays of centromeric satDNA (up to five monomers) located in the euchromatin were also defined for 1.688 and Rsp satDNAs in the *Drosophila melanogaster* genome.[14,41]

In contrast to the centromeric region with dominant satDNA sequences, domains rich in dispearsed repeats and proclaimed as putative heterochromatin regions in Wang et al. [28] flank large blocks of centromeric heterochromatin in *T. castaneum*. A similar architecture with alpha satDNAs predominantly located in centromeric regions and non-LTR transposons which colonize pericentromeric regions is observed in the human genome.[42,43]

Regardless of insignificant presence of TCAST and TCAST2 outside of centromeric regions, the distribution profile of the entire TR content indicates their significant contribution in the assembled genome. Interestingly, in addition to the arrays composed of several repeats (<5 copies), there is a significant number of arrays based on ≥5 repeats.

Moreover, FISH and quantitative analysis of the new, most abundant TR families (Cast1–Cast9) characterized in the assembled genome established them as euchromatic and recovered their existence in significantly higher amounts than predicted by the assembled genome data. This was particularly obvious in Cast1 FISH analysis, where strong signals suggest the presence of long unassembled arrays almost exclusively located outside of centromeric regions. The estimated genome size of 204 Mb is 44 Mb larger than the assembled genome sequence. Based on our results, it can be assumed that the dominant fraction of unassembled genome constitute highly abundant TRs, including centromeric satDNAs TCAST and euchromatic TR families Cast1–Cast9.

Structural analysis of overall TRs in the assembled *T. castaneum* genome revealed a correlation between the monomer length and the number of monomers in arrays. There is an obvious predominance of ~170 bp monomer length if the number of monomers in an array increases. In agreement with the experimentally estimated contribution, the most abundant TR families detected in euchromatic regions have the monomer lengths of ~170 bp and of its double length. Similarly, all five subfamilies of centromeric TCAST satDNA are based on monomers of ~360 bp. The preferential length of repeat units in abundant TR families (both euchromatic and centromeric) can be linked to requirements for efficient DNA packing of long arrays in chromatin structures. It is well known that chromatin in eukaryotic genomes is organized into the nucleosome with 1.67 turns of DNA around the histone octamer (147 bp) plus the linker DNA with a variable distance (10–70 bp).[44] Similar rules govern the human and plant centromeres where the periodicity of centromeric nucleosomes (CenH3) is found to be in accord with the satDNA monomer length,[45–47] being predominantly of ~170 bp and of its double length.[4] The phenomenon observed in this work can be thus explained by the length of DNA wrapped around 1 or 2 nucleosomes as a requirement that may facilitate regular phasing of nucleosomes, and it can be a necessary prerequisite for dramatic amplification of TRs in both, centromeric and euchromatic regions. It has also been proposed for centromeric satDNAs that the monomer length longer than two nucleosomes is rare because longer sequences are unlikely to stabilize nucleosomes.[3,46] Our overall analysis of TRs in the *T. castaneum* assembled genome is in accordance with this idea, showing that the proportion of arrays built of monomers >380 bp dramatically decreases.

In addition to preferential monomer length in the nine TR families structural analyses further revealed a periodic distribution of A or T tracts (4–10 nucleotides), in the same manner as observed in TCAST and other centromeric satDNAs found in several other *Tribolium* species.[22] The periodic appearance of AA/TT dimers along eukaryotic DNA sequences, including those in centromeric regions, promote this nucleotide sequence pattern as a feature that may facilitate nucleosome formation.[45,46] In this regard, periodically distributed tracts of As

Figure 5. Phylogenetic relationships of monomers of Cast1–Cast6 families presented in ML trees (A–F). Terminal branches are coloured according to the chromosome of origin. Monomers originating from a single array and grouped in a single branch of the tree are not specifically marked. Dominant groups which display putative recent exchange events are indicated with arches. Black arches mark monomers that group together in the tree, although they originate from arrays located on non-homologous chromosomes. Grey arches represent monomers in chromosome-specific arrays distant one from the other (>20 kb). Blank arches indicate monomers located in dominant chromosome-specific arrays positioned <20 kb apart. Symbols indicate monomers originating from the same array but dispersed in the tree. The branch support values are indicated at major branch points. (G) FISH of Cast6 family on *T. castaneum* chromosomes. Arrow points to the CH3 chromosome.

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022

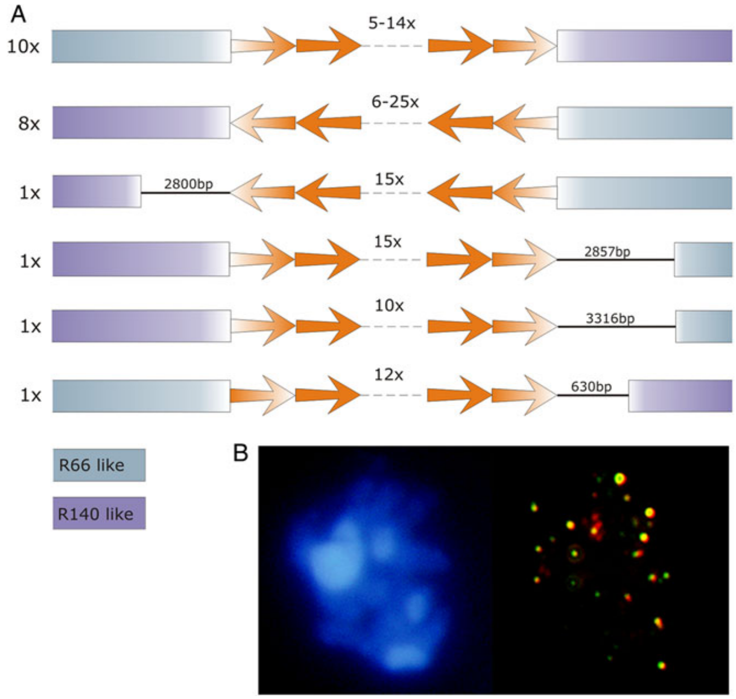


Figure 6. Schematic presentation of flanking regions composed of R66-like and R140-like sequences associated with arrays of Cast5 (A). Two-colour FISH of R66-like (green) and Cast5 (red) probes hybridized on *T. castaneum* chromosomes (B).

and/or Ts, present in many centromeric satDNAs of tenebrionid beetles were shown to define the sequence-induced curvature of the DNA helix axis and could facilitate the tight packing of DNA in centromeric heterochromatin.[48,49] We hypothesize that, similar to the centromeric regions, monomer length and other sequence features (e.g. AT tracts) could be equally important for expansion of TRs in euchromatic regions and may be linked to the formation of hypothetical micro-heterochromatic regions embedded within the euchromatic chromosomal arms. Regardless of their similar structural features, our analyses show a high level of chromosomal compartmentalization of TRs in the *T. castaneum* genome, where prominent arrays of Cast1–Cast9 families are located almost exclusively in non-centromeric, i.e. euchromatic domains, in contrast to centromeric distribution of TCAST satDNA. Clear chromosomal compartmentalization of centromeric and different euchromatic TRs might suggest the existence of some additional requirements which could be important for expansion of TRs in different domains, for example, in functional centromeres.[50]

Within this study, we also focused on the nine most prominent non-centromeric TR families (Cast1–Cast9) in order to define evolutionary trends and mechanisms of TR dispersion throughout the genome. Phylogenetic analyses of monomers show a similar evolutionary scheme for all analysed TR families. Clustering of repeats from the same array is mainly observed in longer arrays, while monomers originating from arrays built of 5–7 monomers are often found dispersed in phylogenetic trees. These data suggest that the efficiency of homogenization mechanisms and concerted evolution in euchromatic regions depends on array length. In support of this, early computer simulations showed that homogenization mechanisms work better in long arrays, but poorly in those containing only few repeats.[7] In addition, our results show that monomers originating from the same chromosome do not group with significantly higher frequency than monomers located on heterologous chromosomes. Moreover, phylogenetic analyses recovered extensive exchanges between non-homologous chromosomes in almost all analysed TR families. This trend is particularly obvious in Cast2, Cast3, Cast4 and Cast5 families, where dominant clusters include arrays originating from all chromosomes. The presence of several subclusters originating from non-homologous chromosomes in almost each analysed tree allows us to propose that these families were spread in several rounds of inter-chromosomal exchange and subsequent amplification. Genome-wide expansion events suggest efficient mechanism(s) of TR propagation in euchromatic genome regions. Previous studies on evolutionary trends involving TRs were focused on centromeric satDNAs. Studies in plants and human thus showed higher sequence divergence between satDNAs located on different chromosomes than within a chromosome, and revealed the preferential occurrence of sequence exchange and homogenization at the intra-chromosomal level.[51,52] The most prominent example is human alpha satDNA whose higher order repeat units show chromosome-specific differences in monomer composition and length.[8] However, this homogenization pattern is not universal, and TRs in subtelomeric regions share a high degree of sequence identity despite being located on non-homologous chromosomes.[53,54] It has been postulated that exchanges between non-homologous chromosome ends occur during meiotic prophase, when all chromatids are interconnected.[53] Similarly, enhanced efficiency in the spread of centromeric satDNA between non-homologous chromosomes noticed in many *Tribolium* species could be facilitated by a bouquet formation which occurs during the first meiotic division.[25,55]

To assess mechanisms underlying the spread of TR sequences throughout the *T. castaneum* genome, we analysed TR array flanking regions, and recovered homologous flanking regions which resemble TEs only for the extremely dispersed Cast5 family. This suggests that repetitive sequences of the Cast5 family were initially distributed by a certain transposition activity and additionally amplified in longer arrays. SatDNA transposition has been advocated in some studies as

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022

alternative mechanism in the evolutionary dynamics of human centromeric region[56] Similarly, enrichment in TRs derived from LTRs in *Zea mays* centromeres raises the possibility that centromeric satDNA can be renewed or replaced by novel satDNA repeats derived from retrotransposons.[57] In *T. castaneum*, TCAST-like elements have been identified within complex units that resemble a DNA transposon and some of them were inserted into introns.[27]

It has also been proposed that the mechanisms of dispersion of TR families in human may be related to duplication of large segments in which arrays are embedded.[12] Flanking regions of other analysed non-centromeric TR families in *T. castaneum* do not display mutual homology thus eliminating mechanisms of segmental duplication in the spread of these sequences. In addition, their junctions and flanking regions do not show any specific feature in the form of inverted repeats or motifs which could act as mediators in the mechanism of dispersion. However, taking into account that the *T. castaneum* genome is as AT-rich (67% A+T) as analysed TR families (66–74% A + T), we suspect that rolling circle amplification could be the mechanism of dispersion, while regions of micro-homologies in the form of AT tracts could be promoters of insertions at different locations. Recent data suggest rolling circle replication and reinsertion of extrachromosomal circular DNA as the mechanism of TR propagation in various organisms, including humans.[11,58]

It has been proposed that recombination in centromeric regions is suppressed to prevent the deleterious effects of crossing over between megabase-sized arrays.[59] However, the extremely efficient distribution and rapid expansion of different TR families found in euchromatic regions can induce intensive rearrangements between diverse genome loci and thus contribute to genomic instability. In addition, TRs dispersed throughout the genome have exceptional potential to evolve independently, creating lineage-specific changes in the structure, sequence, or chromosomal localization, and thus generating incompatibilities between populations/species. Such rapid changes in genomes have been implicated in the post-zygotic isolation of several *Drosophila* species demonstrating the critical role of satDNA in hybrid incompatibility.[60,61] Recent data based on analysis of TR profiles in multiple populations of *D. melanogaster* show significant differentiation of many analysed simple TRs at the population level.[62]

Recent studies indicated that TRs can have remarkable effects on the euchromatic part of the genome. For example, length variation in multimegabase stretches of satDNA repeats of the *Drosophila* Y chromosome could be the major source of epigenetic variation which can modulate gene expression and cause variable phenotypes including differences in immune response.[63] Our analysis of non-centromeric TRs in *T. castaneum* shows extremely efficient propagation in the euchromatic regions and suggest that they could be important factors in the modulation of gene expression. Their impact could be the result of direct insertion of arrays in introns/regulatory elements of genes, or through the formation of micro-heterochromatic environment in the vicinity of genes which may lead to the position effect variegation, and thus modulating the gene expression.

In conclusion, using a combination of bioinformatics and experimental approaches, we have delineated structural and evolutionary trends of TRs in euchromatic regions. The euchromatic chromosomal regions of the *T. castaneum* genome are replete with different highly abundant TRs which are prone to amplify into long arrays if they meet the preferential monomer length. In contrast to similar structural features between centromeric and euchromatic TRs, these two categories show clear chromosomal compartmentalization, suggesting that additional requirements may be imposed on sequences belonging to different regions. The observed evolutionary pattern of euchromatic TR families suggests intensive impact of concerted evolution on longer arrays, while homogenization of short arrays remains limited. Evolutionary trends suggest that efficient inter-chromosomal exchanges were followed by amplification on almost all chromosomes. Our results suggest efficient dispersion of nine TR families through the euchromatin regions and possible role of a transposition-like mechanism only in the case of the most expanded family. We propose that recombination between homologous TRs dispersed among euchromatic sequences can affect both expression of coding information as well as lead to deleterious chromosomal rearrangements.

Authors' contributions

N.M. conceived the study, designed the experiments and analyses, and wrote the manuscript. M. Pavlek performed experiments and analyses, and participated in data interpretation. Y.G. participated in TRF analyses. M. Plohl participated in data interpretation, and critical revision of the manuscript. All authors read, and approved the manuscript.

Acknowledgments

The authors thank Brankica Mravinac for critical reading of the manuscript. We are indebted to the colleagues from the Department of Molecular Biology (Faculty of Science, University of Zagreb) for using their fluorescence microscope. We thank Mary Sopta for critical reading of the manuscript.

Supplementary Data

Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by Research Fund of Republic of Croatia, project no. 098-0982913-2756 and by funds received from the European Union's Seventh Framework Programme for Research and Technological Development under grant agreement No 316289 - InnoMol, FP7-REGPOT-2012-2013-1. Funding to pay the Open Access publication charges for this article was provided by Research Fund of Republic of Croatia, project no. 098-0982913-2756.

References

1. Plohl, M., Meštrović, N. and Mravinac, B. 2012, Satellite DNA evolution, In: Garrido-Ramos, M. (ed.), *Repetitive DNA. Genome Dyn 7*. Karger Publishers, Basel, pp. 126–52.
2. Talbert, P.B. and Henikoff, S. 2010, Centromeres convert but don't cross, *PLoS Biol.*, **8**, 1–5.
3. Melters, D.P., Bradnam, K.R., Young, H.A., et al. 2013, Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution, *Genome Biol.*, **14**, R10.
4. Heslop-Harrison, J.S.P. and Schwarzacher, T. 2013, Nucleosomes and centromeric DNA packaging, *Proc. Natl. Acad. Sci. USA*, **110**, 1–2.
5. Dover, G.A. 1986, Molecular drive in multigene families: how biological novelties arise, spread and are assimilated, *Trends Genet.*, **2**, 159–65.
6. Mahtani, M.M. and Willard, H.F. 1998, Physical and genetic mapping of the human×chromosome centromere: repression of recombination, *Genome Res.*, **8**, 100–10.
7. Smith, G.P. 1976, Evolution of repeated DNA sequences by unequal crossover, *Science*, **191**, 528–35.
8. Rudd, M.K., Wray, G.A. and Willard, H.F. 2006, The evolutionary dynamics of alpha-satellite, *Genome Res.*, **16**, 88–96.
9. Dias, G.B., Svartman, M., Delprat, A., Ruiz, A. and Kuhn, G.C.S. 2014, Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of *Drosophila virilis*, *Genome Biol. Evol.*, **6**, 1302–13.

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022

10. Šatović, E. and Plohl, M. 2013, Tandem repeat-containing MITE elements in the clam *Donax trunculus*, *Genome Biol. Evol.*, **5**, 2549–59.
11. Cohen, S., Agmon, N., Sobol, O. and Segal, D. 2010, Extrachromosomal circles of satellite repeats and 5S ribosomal DNA in human cells, *Mob. DNA*, **1**, 11.
12. Warburton, P.E., Hasson, D., Guillem, F., Lescale, C., Jin, X. and Abrusan, G. 2008, Analysis of the largest tandemly repeated DNA families in the human genome, *BMC Genomics*, **9**, 533.
13. Komissarov, A.S., Gavrilova, E.V., Demin, S.J., Ishov, A.M. and Podgornaya, O.I. 2011, Tandemly repeated DNA families in the mouse genome, *BMC Genomics*, **12**, 531.
14. Kuhn, G.C.S., Küttler, H., Moreira-Filho, O. and Heslop-Harrison, J.S. 2012, The 1. 688 repetitive DNA of *Drosophila*: concerted evolution at different genomic scales and association with genes, *Mol. Biol. Evol.*, **29**, 7–11.
15. Altemose, N., Miga, K.H., Maggioni, M. and Willard, H.F. 2014, Genomic characterization of large heterochromatic gaps in the human genome assembly, *PLoS Comput. Biol.*, **10**, 1–14.
16. Stam, M., Belele, C., Ramakrishna, W., Dorweiler, J.E., Bennetzen, J.L. and Chandler, V.L. 2002, The regulatory regions required for B paramutation and expression are located far upstream of the maize b1 transcribed sequences, *Genetics*, **162**, 917–30.
17. Zhu, Q., Pao, G.M., Huynh, A.M., et al. 2011, BRCA1 tumour suppression occurs via heterochromatin-mediated silencing, *Nature*, **477**, 179–84.
18. Scott, H.S., Kudoh, J., Wattenhofer, M., et al. 2001, Insertion of beta-satellite repeats identifies a transmembrane protease causing both congenital and childhood onset autosomal recessive deafness, *Nat. Genet.*, **27**, 59–63.
19. Paar, V., Gluncić, M., Rosandić, M., Basar, I. and Vlahović, I. 2011, Intragene higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees, *Mol. Biol. Evol.*, **28**, 1877–92.
20. Richards, S., Gibbs, R.A., Weinstock, G.M., et al. 2008, The genome of the model beetle and pest *Tribolium castaneum*, *Nature*, **452**, 949–55.
21. Ugarković, Đ., Podnar, M. and Plohl, M. 1996, Satellite DNA of the red flour beetle *Tribolium castaneum* – comparative study of satellites from the genus Tribolium, *Mol. Biol. Evol.*, **13**, 1059–66.
22. Mravinac, B., Plohl, M. and Ugarković, Đ. 2004, Conserved patterns in the evolution of *Tribolium* satellite DNAs.pdf, *Gene*, **332**, 169–77.
23. Mravinac, B., Ugarković, Đ., Franjević, D. and Plohl, M. 2005, Long inversely oriented subunits form a complex monomer of *Tribolium brevicornis* satellite DNA, *J. Mol. Evol.*, **60**, 513–25.
24. Mravinac, B. and Plohl, M. 2007, Satellite DNA junctions identify the potential origin of new repetitive elements in the beetle *Tribolium madens*, *Gene*, **394**, 45–52.
25. Mravinac, B. and Plohl, M. 2010, Parallelism in evolution of highly repetitive DNAs in sibling species, *Mol. Biol. Evol.*, **27**, 1857–67.
26. Feliciello, I., Chinali, G. and Ugarković, Đ. 2011, Structure and population dynamics of the major satellite DNA in the red flour beetle *Tribolium castaneum*, *Genetica*, **139**, 999–1008.
27. Brajković, J., Feliciello, I., Bruvo-Mađarić, B., et al. 2012, Satellite DNA-like elements associated with genes within euchromatin of the beetle *Tribolium castaneum*, *G3*, **2**, 931–41.
28. Wang, S., Lorenzen, M.D., Beeman, R.W. and Brown, S.J. 2008, Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome, *Genome Biol.*, **9**, 14.
29. Stuart, J. and Mocelin, G. 1995, Cytogenetics of chromosome rearrangements in *Tribolium castaneum*, *Genome*, **38**, 673–80.
30. Gelfand, Y., Rodriguez, A. and Benson, G. 2006, TRDB – the Tandem Repeats Database, *Nucleic Acids Res.*, **35**, D80–7.
31. Benson, G. 1999, Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acids Res.*, **27**, 573–80.
32. Rozas, J. and Rozas, R. 1999, DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis, *Bioinformatics*, **15**, 174–5.
33. Guindon, S. and Gascuel, O. 2003, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, *Syst. Biol.*, **52**, 696–704.
34. Darriba, D., Taboada, G.L., Doallo, R. and Posada, D. 2012, jModelTest 2: more models, new heuristics and parallel computing, *Nat. Methods*, **9**, 772.
35. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. 2010, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, *Syst. Biol.*, **59**, 307–21.
36. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. 2005, Repbase update, a database of eukaryotic repetitive elements, *Cytogenet. Genome Res.*, **110**, 462–7.
37. Drummond, A.J., Ashton, B., Buxton, S., et al. 2011, *Geneious. Version 5.4. Biomatters Ltd.*, Auckland, New Zealand.
38. Juan, C. and Petitpierre, E. 1991, Chromosome number and sex-determining systems in Tenebrionidae (Coleoptera), In: Zunino, M., Belles, X. and Blas, M. (ed.), *Advances in Coleopterology*. AEC, Barcelona, pp. 167–76.
39. Plohl, M., Meštrović, N. and Mravinac, B. 2014, Centromere identity from the DNA point of view, *Chromosoma*, **4**, 313–25.
40. Feliciello, I., Akrap, I., Brajković, J., Zlatar, I. and Ugarković, Đ. 2014, Satellite DNA as a driver of population divergence in the red flour beetle *Tribolium castaneum*, *Genome Biol. Evol.*, **7**, 228–39.
41. Larracuente, A.M. 2014, The organization and evolution of the responder satellite in species of the *Drosophila melanogaster* group: dynamic evolution of a target of meiotic drive, *BMC Evol. Biol.*, **14**, 1–12.
42. Schueler, M.G., Dunn, J.M., Bird, C.P., et al. 2005, Progressive proximal expansion of the primate X chromosome centromere, *Proc. Natl. Acad. Sci. USA*, **102**, 10563–8.
43. Lander, E.S., Linton, L.M., Birren, B., et al. 2001, Initial sequencing and analysis of the human genome, *Nature*, **409**, 860–921.
44. Richmond, T.J. and Davey, C.A. 2003, The structure of DNA in the nucleosome core, *Nature*, **423**, 145–50.
45. Hasson, D., Panchenko, T., Salimian, K.J., et al. 2013, The octamer is the major form of CENP-A nucleosomes at human centromeres, *Nat. Struct. Mol. Biol.*, **20**, 687–95.
46. Zhang, T., Talbert, P.B., Zhang, W., et al. 2013, The CentO satellite confers translational and rotational phasing on cenH3 nucleosomes in rice centromeres, *Proc. Natl. Acad. Sci. USA*, **110**, E4875–83.
47. Levitsky, V.G., Babenko, V.N. and Vershinin, A.V. 2014, The roles of the monomer length and nucleotide context of plant tandem repeats in nucleosome positioning, *J. Biomol. Struct. Dyn.*, **32**, 115–26.
48. Meštrović, N., Mravinac, B., Juan, C., Ugarković, Đ. and Plohl, M. 2000, Comparative study of satellite sequences and phylogeny of five species from the genus *Palorus* (Insecta, Coleoptera), *Genome*, **43**, 776–85.
49. Barceló, F., Gutiérrez, F., Barjau, I. and Portugal, J. 1998, A theoretical perusal of the satellite DNA curvature in tenebrionid beetles, *J. Biomol. Struct. Dyn.*, **16**, 41–50.
50. Hayden, K.E. and Willard, H.F. 2012, Composition and organization of active centromere sequences in complex genomes, *BMC Genomics*, **13**, 324.
51. Lee, H.-R.R., Neumann, P., Macas, J. and Jiang, J. 2006, Transcription and evolutionary dynamics of the centromeric satellite repeat CentO in rice, *Mol. Biol. Evol.*, **23**, 2505–20.
52. Macas, J., Neumann, P., Novák, P. and Jiang, J. 2010, Global sequence characterization of rice centromeric satellite based on oligomer frequency analysis in large-scale sequencing data, *Bioinformatics*, **26**, 2101–8.
53. Ventura, M., Catacchio, C.R., Sajjadian, S., et al. 2012, The evolution of African great ape subtelomeric heterochromatin and the fusion of human chromosome 2, *Genome Res.*, **22**, 1036–49.
54. Macas, J., Navrátilová, A. and Koblížková, A. 2006, Sequence homogenization and chromosomal localization of VicTR-B satellites differ between closely related *Vicia* species, *Chromosoma*, **115**, 437–47.
55. Žinić, S.D., Ugarković, Đ., Cornudella, L. and Plohl, M. 2000, A novel interspersed type of organization of satellite DNAs in *Tribolium madens* heterochromatin, *Chromosome Res.*, **8**, 201–12.
56. Alkan, C., Eichler, E.E., Bailey, J.A., Sahinalp, S.C. and Tüzün, E. 2004, The role of unequal crossover in alpha-satellite DNA evolution: a computational analysis, *J. Comput. Biol.*, **11**, 933–44.
57. Sharma, A., Wolfgruber, T.K. and Presting, G.G. 2013, Tandem repeats derived from centromeric retrotransposons, *BMC Genomics*, **14**, 142.

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022

58. Navrátilová, A., Koblízková, A. and Macas, J. 2008, Survey of extrachromosomal circular DNA derived from plant satellite repeats, *BMC Plant Biol.*, **8**, 90.
59. Malik, H.S. and Henikoff, S. 2009, Major evolutionary transitions in centromere complexity, *Cell*, **138**, 1067–82.
60. 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.
61. Bayes, J.J. and Malik, H.S. 2009, Altered heterochromatin binding by a hybrid sterility protein in *Drosophila* sibling species, *Science*, **326**, 1538–41.
62. Wei, K.H.-C., Grenier, J.K., Barbash, D.A. and Clark, A.G. 2014, Correlated variation and population differentiation in satellite DNA abundance among lines of *Drosophila melanogaster*, *Proc. Natl. Acad. Sci. USA*, **111**, 18793–8.
63. 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. USA*, **107**, 15826–31.

Downloaded from https://academic.oup.com/dnaresearch/article/22/6/387/2462399 by U.S. Department of Justice user on 16 August 2022