Long-read genome assembly of the insect model organism *Tribolium castaneum* reveals extensive propagation of satellite DNA long arrays in gene-rich regions

Volarić Marin[#], Despot-Slade Evelin[#], Mravinac Brankica, Meštrović Nevenka*

Ruđer Bošković Institute, Zagreb, Croatia

[#]These authors contributed equally to this work.

* Correspondence: nevenka@irb.hr

Abstract

Eukaryotic genomes are replete with satellite DNAs (satDNAs), large stretches of tandemly repeated sequences which are mostly underrepresented in genome assemblies. As a result, we know little about their spread, movement, and rearrangement across the genome or about their potential impact on the genome. Here, we have combined Oxford Nanopore long-read sequencing with a reference-guided assembly approach to generate a new, high-quality TcasONT genome assembly of the model beetle *Tribolium castaneum*. The new TcasONT is enriched by 50 Mb in the repetitive genome part. Therefore, we used the enhanced assembly to conduct global and in-depth analyses of ten abundant non-(peri)centromeric satDNAs, Cast1-Cast9. We found a large fraction of long arrays embedded in gene-rich regions common for almost all Cast1-Cast9 satDNAs and a negative correlation between the abundance of these satDNAs and transposable elements. Contrary to accepted opinion, our finding clearly speaks to the ability of satDNAs to expand in euchromatic regions and also demonstrate the flexibility of gene-rich regions to tolerate and accommodate satDNA sequences. In addition, we find that suppressed recombination on X chromosomes has less effect on the spread of satDNAs in euchromatin, but rather tolerates amplification of satDNAs into longer arrays. Our phylogenetic analyses show the extensive spread of Cast1-Cast9 satDNAs, characterized by distribution along the entire chromosome length and between (non)homologous chromosomes. This observed pattern points out very efficient mechanisms of propagation satDNAs on the whole genome scale including gene-rich regions. In conclusion, the presence of such a large portion of satDNAs with extensive propagation and amplification potential in gene-rich regions of the *T. castaneum* genome can strongly influence both gene expression and the dynamics of genome evolution.

Background

Satellite DNAs (satDNAs) are the most abundant and rapidly evolving non-coding DNA of all eukaryotic genomes. In contrast to other groups of repetitive elements such as interspersed transposable elements (TE), which are scattered throughout the genome, satDNAs are characterized by monomer sequences tandemly arranged into long arrays. Previous studies indicate preferential localization of satDNAs in the (peri)centromeric heterochromatin (Plohl et al., 2014). Heterochromatin represents a highly compact gene-poor specialized chromatin structure that is distinct from the gene-rich euchromatin. SatDNAs in (peri)centromeric heterochromatin have a long history of investigation including many plant and animal species, and studies have suggested satDNA role in chromatin packaging, centromere formation/maintenance, and centromere evolution (Aldrup-MacDonald et al., 2016; Blattes et al., 2006; Melters et al., 2013). SatDNAs are also linked to chromosome mis-segregation (Aldrup-MacDonald et al., 2016), disease phenotypes (Bersani et al., 2015), and reproductive isolation between species (Ferree & Prasad, 2012). In addition to standard satDNA-rich regions such as (peri)centromeric heterochromatin, recent studies indicated examples of short arrays of (peri)centromeric satDNAs also dispersed in euchromatic regions. For example, (peri)centromeric satDNA families in humans, *Drosophila* and the beetle *Tribolium* can also be dispersed within euchromatin in the form of clustered or short single repeat arrays (Brajković et al., 2012; Feliciello et al., 2020; Kuhn et al., 2012). In *Drosophila* X-specific (peri)centromeric 1688 satDNAs dispersed in euchromatic regions have a role in dosage compensation through up-regulation of X-linked genes in *Drosophila* males (Joshi & Meller, 2017; Menon et al., 2014). The euchromatic counterpart of (peri)centromeric DNA in the beetle *Tribolium* modulates the local chromatin environment upon heat stress changing the expression of neighboring genes (Feliciello et al., 2015). Although there is clear evidence that satDNAs exist outside the (peri)centromere and that some roles have been assigned to them, understanding of their organization, evolutionary dynamics, and disclosure of the molecular mechanisms driving their spread, movement, and rearrangement in euchromatin is still very limited. The key requirement that would help to understand the evolution and function of these sequences is their detection at the genome level and genome-wide studies of satDNAs in various model organisms. One of the main reasons for the current lack of global and in-depth studies of satDNAs in euchromatin is certainly the fact that satDNA regions are the most difficult part of the genome to sequence and assemble, and therefore they are underrepresented or even absent in the best genome assemblies (reviewed in (Lower et al., 2018; Miga, 2015)). However, with the improvement of long-read sequencing technology, especially Oxford Nanopore platforms, it has recently been demonstrated that genomic loci made of satDNAs can be efficiently assembled (Miga et al., 2020).

Moreover, with the exception of the Y chromosome, end-to-end maps of all human chromosomes, spanning highly repetitive (peri)centromeric and telomeric regions, have recently been obtained (Nurk et al., 2022).

The beetle *Tribolium castaneum* is a worldwide pest of stored products and the representative of the most species-rich animal order on Earth, the coleopterans. *T. castaneum* has become one of the most important models in the field of evolution, physiology and development of insects because its development is more representative for insects compared to *Drosophila* (Campbell et al., 2022; Klingler, 2004). The *T. castaneum* genome has been sequenced, annotated and a first version of genome assembly was available in 2008 (Richards et al., 2008). Structural genome analyses revealed large amounts (>30%) of different classes of repetitive DNA in the assembled part of the *T. castaneum* genome. Among them, the (peri)centromeric heterochromatic satDNA TCAST, which was drastically underrepresented in the assembled genome, dominates on all *T. castaneum* chromosomes comprising up to 17% of the genome (Gržan et al., 2020; Ugarković et al., 1996). All *T. castaneum* chromosomes are characterized by large blocks of (peri)centromeric heterochromatin, while no prominent heterochromatic blocks could be detected cytogenetically on chromosome arms (Wang et al., 2008). Regarding the possible presence of satDNAs in euchromatin of *T. castaneum*, our previous study on the *T. castaneum* reference genome assembly disclosed that, in addition to short arrays of counterpart of (peri)centromeric DNA in euchromatin regions (Brajković et al., 2012), *T. castaneum* genome abounds with many different and unrelated satDNAs distributed out of (peri)centromeric regions (Pavlek et al., 2015). Nine satDNAs, Cast1-Cast9, represent the most prominent non-pericentromeric satDNA portion which constitutes ∼0.3 % of the genome assembly. Unlike of the most euchromatic satDNAs, found for example in *Drosophila*, which have short monomer units, Cast1-Cas9 represent "classical" satellite DNAs based on ∼170 and ∼340 bp long monomers. The experimentally determined genome content of Cast1-Cast9 satDNAs has showed abundance of approximately ∼4.3%, which is ∼10X higher than the analyzed representativeness in the assembled genome (Pavlek et al., 2015). The genome version from 2008 was based on a Sanger 7x draft assembly and was anchored to Linkage Groups (LGs) (Richards et al., 2008). Recently, the genome assembly was upgraded to a new version, Tcas5.2, which was improved by using large-distance jumping libraries and BioNano Genomics optical mapping, bringing also a new official gene set OGS3 (Herndon et al., 2020). Nevertheless, our previous satDNA analyses showed that satDNA genome fractions were highly underrepresented in Tcast5.2.

Therefore, we considered that for a deep genome-wide analysis of dominant satDNAs distributed outside the (peri)centromere, Cast1-Cast9, the first prerequisite is the availability of the assembly enriched in the

repetitive sequences. For that reason, we used Oxford Nanopore Technology (ONT) long-read sequencing and applied reference-guided assembly strategy using the reference assembly Tcas5.2 to produce an improved genome assembly of *T. castaneum*. This approach has been very successful in our study to achieve our ultimate goal: filling the gaps and elongating the repetitive regions, especially the satDNAs in the new TcasONT assembly. The new nanopore-based assembly expanded the repetitive part of the genome by 50.8 Mb. In addition to enrichment in the repetitive part, the analyses of the completeness of the predicted gene model also increased compared to the Tcas5.2 assembly. Further, we used the TcasONT assembly provided here, as a platform for global and in-depth analyses of the dominant satDNA fraction located in euchromatic chromosomal arms. Specifically, we developed a workflow for Cast1-Cast9 detection in TcasONT assembly and subsequent analysis of their chromosomal distribution and organization, phylogenetic analysis of their arrays as well as analysis of arrays' environment.

Results

1. Chromosome-scale ONT-genome assembly

The size of the reference *T. castaneum* genome assembly (Tcas5.2) is 165,9 Mb (Herndon et al., 2020). Removing of placeholders and sequencing gaps results in a genome assembly size of 136 Mb. Considering that the experimentally estimated genome size is 204 Mb (Brown et al., 1990), it is clear that 68 Mb of the genome is missing from the Tcas5.2 reference assembly. In order to improve the existing assembly, especially in repetitive genome parts, we employed Oxford Nanopore Technologies' (ONT) long-read sequencing platform. In total, 121 Gb of nanopore long reads were generated. Incorporating only those reads that had a mean quality score of >7, a total of 89 Gb representing genome coverage of 436X, were used for the new genome assembly (Suppl Table 1 Nanoplot). Selected reads were up to 500 kb in length with a mean read length of 15.8 kb and mean quality score per read of 12.8. Main steps of genome assembly process are presented in Suppl Figure 1. First, nanopore reads were divided into two fractions based on their length, short reads (<20kb; total of 52.7Gb; 258X genome coverage), and long reads (>20kb; total of 36.3 Gb; 178X genome coverage). Next, long reads (>20kb) were used to assemble the initial assembly using Canu (Koren et al., 2017). This initial Canu assembly had 1479 contigs with total length of 321 Mb and a N50 of 835.5 kb. (Suppl Table 2, left Canu statistics). The longest contig was 16.4 Mb in length. Our initial Canu assembly was approximately 117 Mb larger than the previously experimentally estimated genome size of 204 MB (Alvarez-Fuster et al., 1991). To verify the genome size and genome repetitiveness *in silico*, the k-mer frequencies of the corrected reads generated during the Canu assembly process using the FindGSE and CovEST programs were calculated. FindGSE estimated the genome to be 204 Mb, while CovEST RE showed only slightly higher estimation of 208 Mb (Supp Table 3 genome size), which is fully consistent with experimental estimation. The repeat ratio calculated by FindGSE was 27% of genome confirming the previously observed high genome repetitiveness (Wang et al., 2008). In terms of repetitiveness, previous studies showed that up to 17% of the *T. castaneum* genome is (peri)centromeric satDNA TCAST (Gržan et al., 2020; Ugarković et al., 1996). Considering the facts that abundant tandem repeats might be problematic for proper assembly (Tørresen et al., 2019), we assumed that the additional 117 Mb in our Canu assembly (Supp table 2) were primarily generated by gene-poor and highly repetitive (peri)centromeric TCAST arrays. To reduce the suspected negative effects of TCAST satDNA on the genome assembly process, we performed an additional filtering step. In this step, the contigs that did not contain at least 1000 bp of unique gene-coding sequence were removed, resulting in successful filtering of 471 out of 1479 total contigs (Suppl Figure 1).

These 471 sorted contigs were further used to achieve new chromosome-scale assembly using reference-guided approach. The latest improved *T. castaneum* genome assembly Tcas5.2 (Herndon et al., 2020) was used as the reference for organizing these contigs into chromosomes. Two reasons were crucial for the selection of this approach: 1) the availability of the high-quality *T. castane*um reference genome, Tcas5.2, and 2) assembly approach that allows the input sequence not to change, but rather to be ordered based on the mapping quality of the reference sequence. Given that the reference genome Tcast5.2 contained 3669 unresolved gaps with total N (ambiguous nucleotides) size of 11 Mb and mean gap size of 3125 Kb, the first step of reference-guided assembly approach included gap filling of Tcas5.2 with TGS-GapCloser using 8.6 Gb of Canu corrected larger reads (length >20 kb). TGS-GapCloser was efficient in filling most of the gaps, so in total 3607 gaps (98.3 %) were closed which increased the genome size by 10.5 Mb (Suppl table 4). The gap-filled Tcas5.2. reference genome assembly was further used as a template for orientating 471 Canu-contigs into the 10 chromosomes using RagTag software tool (Alonge et al., 2021). A total of 244 contigs were unambiguously assigned to a reference Tcas5.2. genome representing a new ONT-based genome assembly, named TcasONT (Supply table 5). The rest (227 contigs) did not accurately map on the assembly, remaining assigned as unassembled sequences (Supp table 6). To further validate structure of omitted contigs, we mapped them with highly abundant (peri)centromeric TCAST satDNAs and found that around half of unplaced sequences (14 Mb) are indeed mostly covered (>50 %) with this repetitive sequence (Supp table 6). It is evident that filtered contigs and the unassembled genome part are enriched in TCAST satDNA. To gain insight into the actual representation of these regions, we mapped the TcasONT assembly with TCAST satDNA. The results showed that 9446 monomers were mapped, accounting for 3.6 Mb (1.7%) of the genome.

To improve genome accuracy, the TcasONT genome assembly was polished using Racon with 56.3 Gb of raw short nanopore reads (<20 kb) which have been excluded from the initial Canu-assembly but passed the quality check. The final polished TcasONT assembly has a size of 225.9 Mb, with 191 Mb assembled in ten chromosomes and the remainder in unassembled contigs (Suppl table 7). Compared to the reference Tcas5.2, the new TcasONT assembly shows a 53 Mb increase of total chromosome length (189 Mb in contrast to 136 Mb genome without gaps or Ns, Suppl Table 7). Comparing TcasONT and Tcas5.2 chromosomes, the elongation of chromosome lengths ranges from 10.6 to 56.3 %, indicating a significant improvement of genome assembly continuity at the chromosome level (Supp table 8).

To measure completeness and quality of the new assembly, we compared the TcasONT assembly to the reference Tcas5.2 using dot plot analyses. The result revealed high levels of macrosynteny and collinearity across all chromosomes with high sequence identity (Figure 1). Furthermore, we calculated that 88% of

the total Tcas5.2 unplaced contig sequence length is now placed within chromosomes of the new TcasONT assembly.

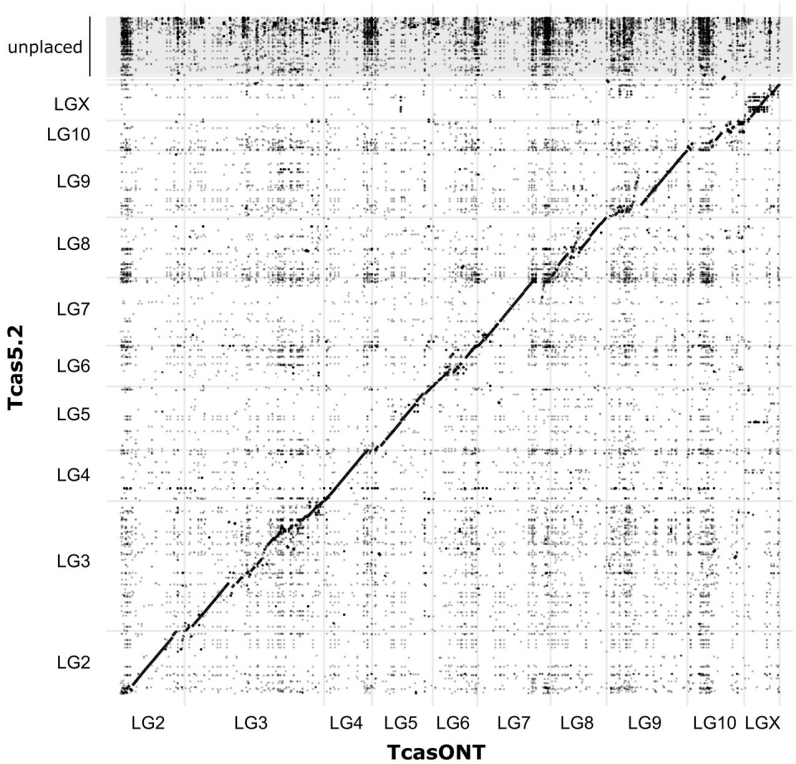

Figure 1. Dot-plot representation of TcasONT vs. Tcas5.2 assembly comparison. The horizontal axis corresponds to the intervals along the TcasONT assembly and the vertical axis corresponds to the intervals along the Tcas5.2 assembly. LG1-LG10 represent chromosomes, while unplaced contigs contain sequences in Tcas5.2. that were not associated with any chromosome. Dots closest to the diagonal line indicate co-linearity between the two assemblies.

We next assessed the completeness of the TcasONT assembly by comparative BUSCO analyses between TcasONT and Tcas5.2 using insect universal orthologs, odb10 database (Figure 2). Of the 1367 genes, we identified 1329, which corresponds to 97.2 % of BUSCO single-copy completeness. Further, 17 genes were duplicated, 13 were fragmented and only 8 genes were missing. A considerable increase in number of complete BUSCOs (32 genes) can be observed in the TcasONT assembly compared to Tcas5.2. In addition to BUSCO analysis, we “lifted” genes from the Tcas5.2 annotation onto the TcastONT assembly using the Liftoff package and assessed their final numbers (Suppl Table 9). Only 48 out of 14467 genes (0.3%) were left unmapped, mostly representing genes with no known biological functions.

Genes (N=1367)	**Tcas5.2**	**TcasONT**
Complete	1297	1329
Duplicated	6	17
Fragmented	4	13
Missing	60	8

Table 1. Gene completeness assessment based on BUSCO analysis of Tcas5.2 and TcasONT assemblies using insect universal orthologs. The analysis was expressed as absolute numbers for complete and single-copy, complete and duplicated, fragmented and missing genes.

2. Repeat identification in the TcasONT assembly

To evaluate the improvement in the representation of the repetitive genome portion in the new TcasONT assembly compared to Tcas5.2, a self-similarity dot-plot analyses were performed using >70% sequence identity as a criterion (Figure 3). The comparison of dot-plots revealed that the TcasONT assembly has a higher level of self-similarity with medium to large dark blocks representing regions of repetitive sequences in comparison to Tcas5.2 assembly (Figure 2). Quantified, the number of self-similar sequences was 44,671 in the TcasONT genome assembly, while in Tcas5.2 it was only 2,202, which means that the new TcasONT assembly has about 20-fold higher proportion of repetitive genome fraction.

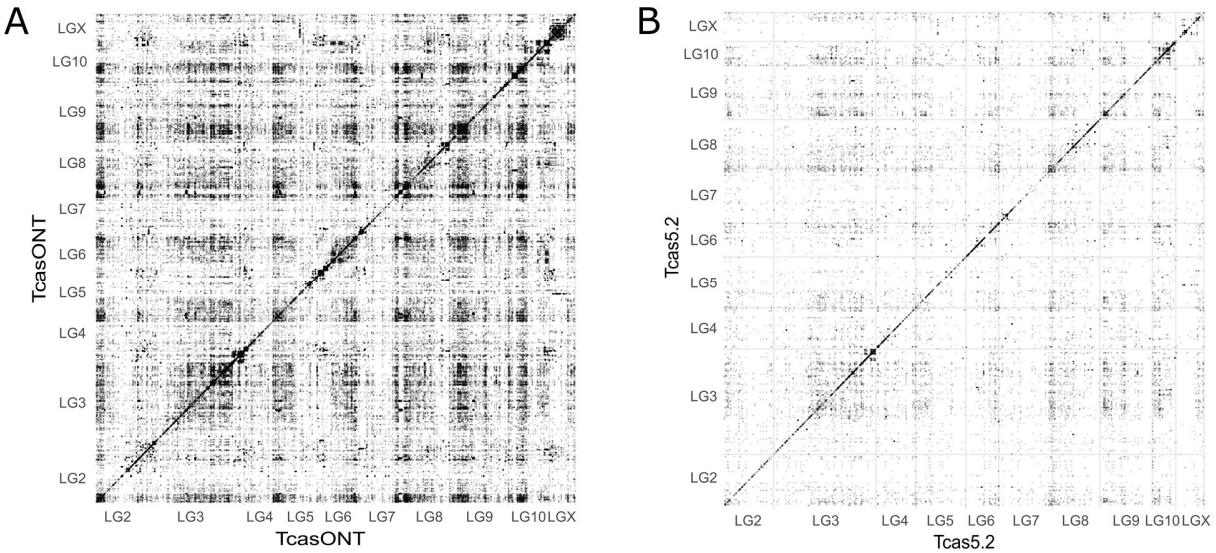

Figure 2. The whole genome-to-genome dot-plots analyses of A) TcasONT, and B) Tcas5.2 assemblies. Each dot represents a region of ≥1000 bp long region which is mapped in its entirety to another part of the genome. The dot density is proportional to the number of highly similar regions.

To define the improvement specifically in transposable elements (TEs) content, the TEs in both Tcas5.2 and TcasONT assemblies were annotated using RepBase database. The TEs were classified into 4 main types, including DNA elements, long interspersed nuclear elements (LINEs), long terminal repeats (LTRs) and short interspersed nuclear elements (SINEs). The TcasONT assembly revealed an increase of 11,297 DNA TE, 12,268 LTR TE and 27,553 LINE elements (Figure 3A and Suppl Table 10) compared to the Tcas5.2. Next, we investigated abundance of these elements in both genomes. Among the four TE categories, in terms of cumulative sequence occupancy there is a notable increase in TcasONT compared to Tcas5.2 for three of them; DNA TE (5.2 Mb), LINE (14.5 Mb) and LTR (1.8 Mb) (Figure 3B and Suppl Table 10). Taken together, the TcasONT assembly is enriched with 21.5 Mb of TE elements, compared to Tcas5.2. (Suppl Table 10).

Next, we assessed the content of another class of repetitive DNA sequences, tandem repeats (TRs), detecting them by the Tandem Repeat Finder (TRF), which was successfully used in previous studies of *T. castaneum* (Pavlek et al., 2015). Detected TRs were classified into 3 groups based on their monomer length (<50bp, 50-500bp, and >500bp). The analysis revealed a total of 35.3 Mb TRs in the TcasONT assembly compared to 9.1 Mb in Tcas5.2 (Suppl Table 10), indicating a significant improvement in TRs content. On closer inspection, the results show that the number of TRs elements detected in the TcastONT assembly was doubled in all three classes (Figure 3C), while the 50-500 bp and >500 bp TRs, which represent “classical” satDNAs, contributed significantly to the cumulative genome length (31.5 Mb in TcasONT vs 7.3Mb in Tcas5.2) (Figure 3D).

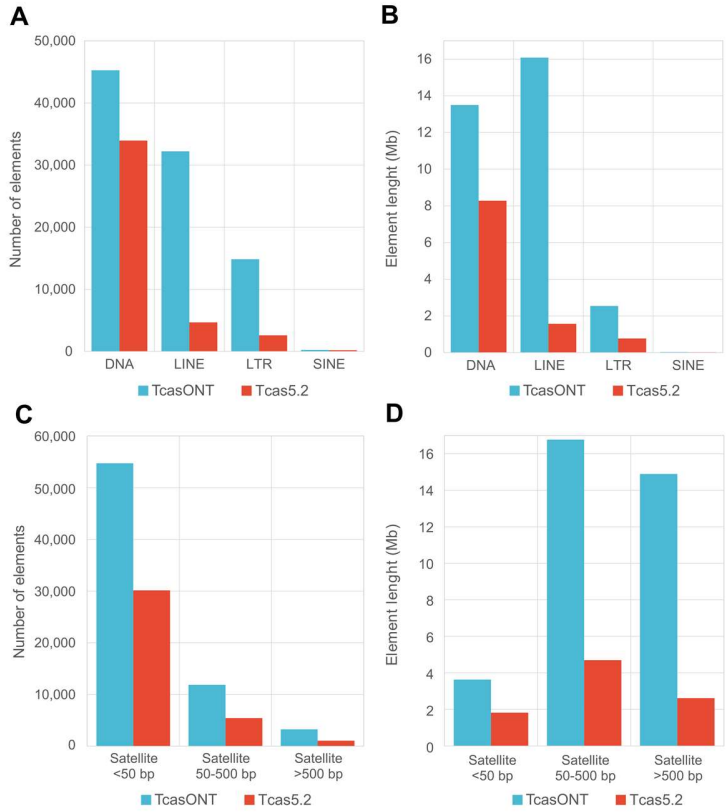

Figure 3. Comparison of main classes of transposable elements (TEs) found by RepeatMasker and tandem repeats (TRs) found by TRF in the TcasONT and Tcas5.2 assemblies. The height of the bars indicates **A)** total number of TEs, **B)** cumulative length of TEs**, C)** total number of TRs, and **D)** cumulative length of TRs.

From all above, the 53 Mb size difference between the Tcas5.2 and TcasONT genome assemblies is mainly due to the enrichment of repetitive regions (50.8 Mb; 95.8%; Suppl Table 10). The repetitive fractions that were most enriched in the TcasONT assembly were TEs (21.5 Mb) and TRs (26.2 Mb). If we consider that the (peri)centromeric TCAST satDNA accounts for only 3.6 Mb, it is evident that the TcasONT assembly has an increase of 22.6 in TRs outside of (peri)centromeric regions which mainly includes an increase in "classical" satDNAs with >50bp long monomer units.

3. Characterization of Cast1-Cast9 satDNAs in TcasONT assembly

Given the demonstrated improvement in the genome assembly, especially in its repetitive parts, the new TcasONT assembly represents a very good platform for in-deep analyses of satDNA sequences outside the

(peri)centromeres. TcasONT thus served as a crucial basis to achieve the final goal of the work, which was the comprehensive analysis of Cast1-Cast9, abundant satDNAs localized outside the (peri)centromeric regions.

The first step, essential for downstream analyses, was the mapping of the Cast1-Cast9 arrays in the TcasONT assembly. Because of the well-known variability of monomer sequence within a satDNA family, it was important to establish sequence similarity and sequence coverage parameters, which would ensure the detection of the vast majority of Cast1-Cast9 monomers in the TcasONT assembly. For that reason, the BLAST search of nanopore raw reads using consensus sequences of Cast1-Cast9 monomers (Pavlek et al., 2015) were performed. The results were shown as density plots where color intensity correlates with number of monomers of a particular Cast satDNA (Suppl Figure 2). The majority of Cast satDNAs show highest intensity, i.e. monomer aggregation in the area corresponding to >75% of sequence coverage and >70% of percent identity. However, in Cast2 and Cast4, the appearance of two areas with high density of which one area is characterized by lower coverage/identity parameters can be observed. This pattern is a consequence of the mutual similarity in one part of consensus sequences between Cast2 and Cast4 (Suppl Figure 3). Despite their partial similarity they were still considered as iIndividual satDNAs. In addition, the only satDNA that in BLAST analysis shows increased monomer variability is Cast7, evident by the increased density in the lower sequence identity area. Therefore, to map the vast majority of monomers from all 9 satDNAs (Cast1-Cast9) from the TcastONT assembly, the slightly lower parameters of sequence coverage >70% and percent identity >70% were used.

To disclose the organization of the Cast1-Cast9 arrays on the whole genome scale, we analyzed the relationship between monomer distance in the arrays for each Cast (Suppl Figure 4.) More specifically, we wanted to determine whether the arrays were organized in tandem continuously (consisting exclusively of monomers of a particular Cast satDNA) or had mixed tandem organization (with a different sequence incorporated into the arrays). The results showed that, most satDNAs show continuity of the curve, which depicts a typical satDNA array organization consisting of continuously arranged monomeric variants. In contrast, Cast2, Cast5 and Cast7 exhibited a disturbed tandem contiguity demonstrated by sharp increases in average array length at certain distances. Most prominently, Cast2 showed an extremely sharp and uniform increase in average array length, indicating the presence of a specific sequence within the arrays that required further investigation. In addition to monomeric Cast2 arrays, we found Cast2 monomers also organized as a part of a new, longer repeat unit of approximately 1270 bp in length (Suppl Fig. 5A and B). This new family of tandem repeats is termed Cast2', and in the downstream analysis these two types of Cast2 arrays were examined separately. A more detailed introspection of Cast5 arrays showed that the

observed profile is the result of sporadic occurrence of previously described R66-like sequences (Pavlek et al., 2015) dispersed within continuous Cast5 arrays (Suppl Fig. 5C). In addition to R66-like sequence which turned out to be randomly distributed within Cast5 arrays, R140-like sequences were found exclusively at the end of some Cast5 arrays. In addition, analyses of Cast7 arrays also reveal a mixed organization where Cast7 arrays often occur in association with (peri)centromeric TCAST satDNA (Suppl Fig. 5D).

Considering the difficulty of correctly assembling satDNA sequences, the next question was whether the array lengths and arrangement of the Cast1-Cast9 satDNAs in the TcasONT assembly correspond to the real situation of the genomic loci containing the Cast1-Cast9 repeats. This is because satDNAs can lead to assembly collapse, where the number of monomer units (array length) found in a genome assembly could be less than the actual number. Based on the assumption that raw reads provide a realistic picture of what is originally present in the genome because they do not undergo assembly process, we performed a comparative analysis of Cast1-Cast9 including new Cast2' array pattern across the individual reads and across the new genome assembly. For this analysis, it was crucial to define optimal parameters for arrays detection to avoid fragmentation due to short gaps. Therefore, considering monomer length and mixed array organization of specific Cast satDNAs such as Cast2, Cast5 and Cast7, the best window length that ensures detection of the maximum number of arrays was evaluated for each Cast satDNA. The analysis showed that a minimum of three consecutive monomer units should be a criterion for correct array characterization for each Cast satDNA. The array patterns between the datasets of individual reads and the TcasONT assembly show remarkable similarity for all Cast satDNAs (Suppl Figure 6). The only exception is Cast6, where there is an apparent loss of array portion in TcasONT assembly compared to individual raw reads, leading to the assumption that Cast6 has a larger portion of longer arrays that are not fully represented in TcasONT. The results of this analyses indicate that the TcasONT assembly contains the representative sample of Cast arrays and can be taken as a credible platform for further analysis of Cast satDNAs.

To detect enhancement of Cast1-Cast9 regions in TcasONT genome assembly, we compared both their content and array length with Tcas5.2 (Table 2). In terms of overall enrichment in Cast satDNA portion, the TcasONT assembly shows an increase of 7.2 Mb in the representation of analyzed Cast satDNAs compared with Tcas5.2. Among them, Cast1, interCast2 and Cast5 emerged as particularly prominent

satDNAs with an increase of 0.7 Mb, 2.8 Mb and 2.6 Mb, respectively. The total content of Cast satDNAs in the new TcasONT genome was 8.8 Mb (4.6% of the genome), which is remarkably close to the previous experimental estimate of the genome abundance for these satDNAs (> 4.3% in (Pavlek et al., 2015)). Among them, Cast2' and Cast5 showed the highest genome abundance with 1.5% for each. Except for Cast8 and Cast9 satDNAs with low abundance (<0.1%), most Cast satDNAs showed moderate genome abundance in the range of 0.1-0.4%. The results also show an exceptional increase in the number of arrays in the TcasONT assembly for Cast2', Cast5 and Cast7. In Cast2', for example, the number of arrays has increased by more than 9X. In addition to the number of arrays, average increase in array length (expressed as mean values) can also be observed in the TcasONT genome assembly for all Cast satDNAs. A significant increase in average array length of up to 3X was observed for interCast2, Cast5 and Cast6 satDNAs compared to Tcas5.2. The length of an average array for Cast satDNAs in the TcasONT assembly ranges from 1.2 to 31 kb. The highest average array size of ~31 kb has been found for Cast6, while Cast1 has the longest array of ~112 kb.

Table 2. Statistics of Cast1-9 arrays detected in the Tcas5.2 and TcasONT genome assemblies.

	Tcas5.2					**TcasONT**					
	Array number	Total length (bp)	Array mean (bp)	Maximal array length (bp)	genome abundance (%)	Array number	Total length (bp)	Array mean (bp)	Maximal array length (bp)	genome abundance (%)	Total length difference (bp)
Cast1	71	206,931	2,915	61,324	0.140	106	865,816	8,168	112,523	0.452	658,885
Cast2	121	178,456	1,475	17,682	0.121	77	267,482	3,474	54,320	0.140	89,026
Cast2'	29	117,847	4,064	21,938	0.080	258	2,928,418	11,350	66,157	1.529	2,810,571
Cast3	173	220,599	1,275	4,173	0.149	172	246,560	1,433	10,679	0.129	25,961
Cast4	100	147,755	1,478	9,103	0.100	88	334,031	3,796	56,181	0.174	186,276
Cast5	66	258,820	3,922	12,025	0.175	148	2,897,522	19,578	85,403	1.512	2,638,702
Cast6	17	166,879	9,816	43,796	0.113	23	715,260	31,098	77,335	0.373	548,381
Cast7	86	82,188	956	3,729	0.056	273	332,266	1,217	6,028	0.173	250,078
Cast8	32	38,553	1,205	4,280	0.026	24	51,621	2,151	6,463	0.027	13,068
Cast9	117	145,832	1,246	4,575	0.099	98	139,574	1,424	5,958	0.073	-6,258
Total	812	1,563,860			1.059	1267	8,778,550			4.582	7,214,690

The graph of Cast1-Cast9 monomer copy number in TcasONT compared to Tcas5.2 also clearly shows the presence of significant enrichment (2-10X) in the majority of satDNAs analyzed (Figure 4). The exception were Cast3 and Cast9 satDNAs whose content did not change significantly. From the above, it is clear that

the new TcasONT assembly shows a much more credible representation of Cast satDNAs than the reference Tcas5.2.

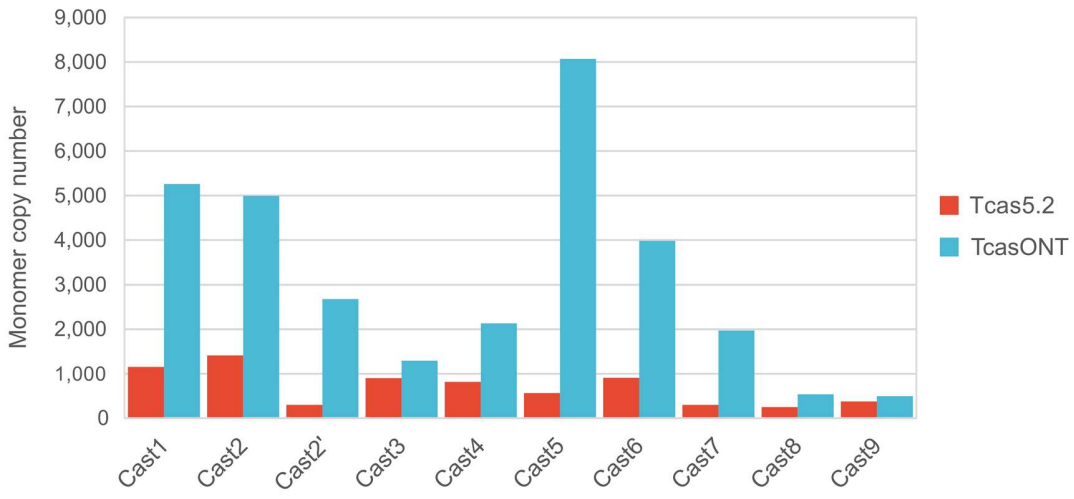

Figure 4. Comparison of monomer copy numbers for Cast satDNAs annotated in the reference Tcas5.2 and the TcasONT genome assemblies (see also Suppl Table 11). The assemblies were mapped with Cast1-Cast9 satDNA monomeric consensus sequences published previously (Pavlek et al., 2015).

4. Genome organization and evolution of Cast1-Cast9 satDNAs

The distribution of Cast1-Cast9 including Cast2' satDNAs on *T. castaneum* chromosomes was further investigated by mapping all Cast1-Cast9 monomers on the assembled TcasONT linkage groups. The two main patterns were observed. The first pattern is characterized by Cast satDNAs located on a subset of chromosomes (Figure 5). For example, Cast1 dominates on LG3, LG6 and LG10, while it is underrepresented on other chromosomes. Similarly, Cast6 dominates on LG3, but is also present, albeit to a much lesser extent, on LG6, LG8 and LG9. In contrast, the other Cast satDNAs such as Cast2', Cast5 and Cast7 show a more uniform pattern characterized by similar content on almost all chromosomes (Figure 5).

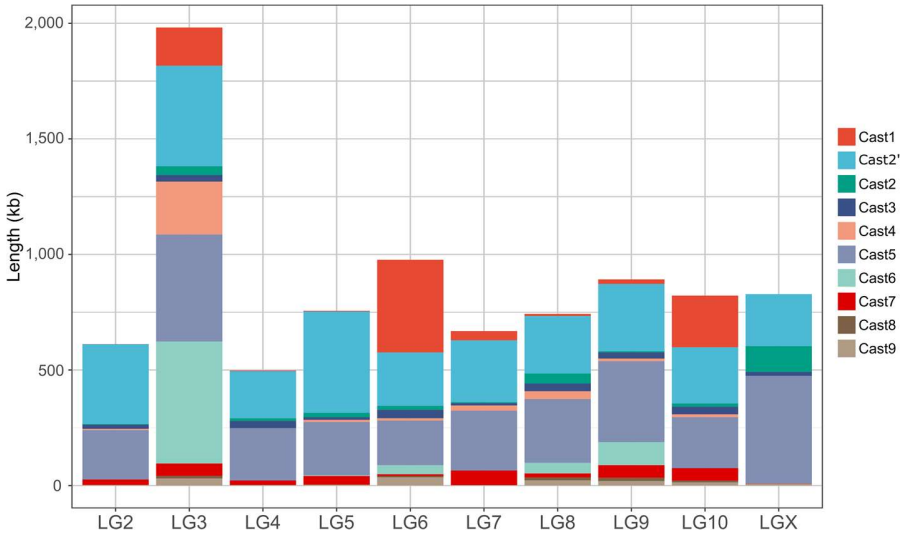

Figure 5. Distribution of Cast satDNAs across different chromosomes in the TcasONT genome assembly. Different colors represent specific Cast satDNA across ten chromosomes of *T. castaneum*. Monomers within the genome assembly were annotated based on a query percentage identity >70% and query coverage >75% for all Cast satDNAs.

Next, we analyzed the distribution of Cast1-Cast9 along chromosomal length (Suppl Figure 7). Since it is already known that TCAST builds up huge blocks of (peri)centromeric heterochromatin (Gržan et al., 2020) the presumed localization of the (peri)centromere on each chromosome is marked. Detailed examination of Cast1-Cast9 distribution disclosed an exceptional spread of Cast satDNAs along the chromosomes with no tendency to cluster in proximal regions of the (peri)centromere. Only Cast7 occurs preferentially in regions enclosed by (peri)centromeric satDNA TCAST. In addition, some Cast satDNAs such as Cast2, Cast2', Cast3 and Cast4 show a tendency to cluster in distal chromosomal regions.

The distribution of Cast1-Cast9 satDNAs on chromosomes compared with transposon elements (TEs) and genes was also analyzed (Figure 6). Interestingly, the distribution of Cast1-Cast9 satDNAs often coincides with regions of higher gene density, while it is inversely correlated with the distribution of TEs (the most prominent regions with these features are highlighted in Figure 6).

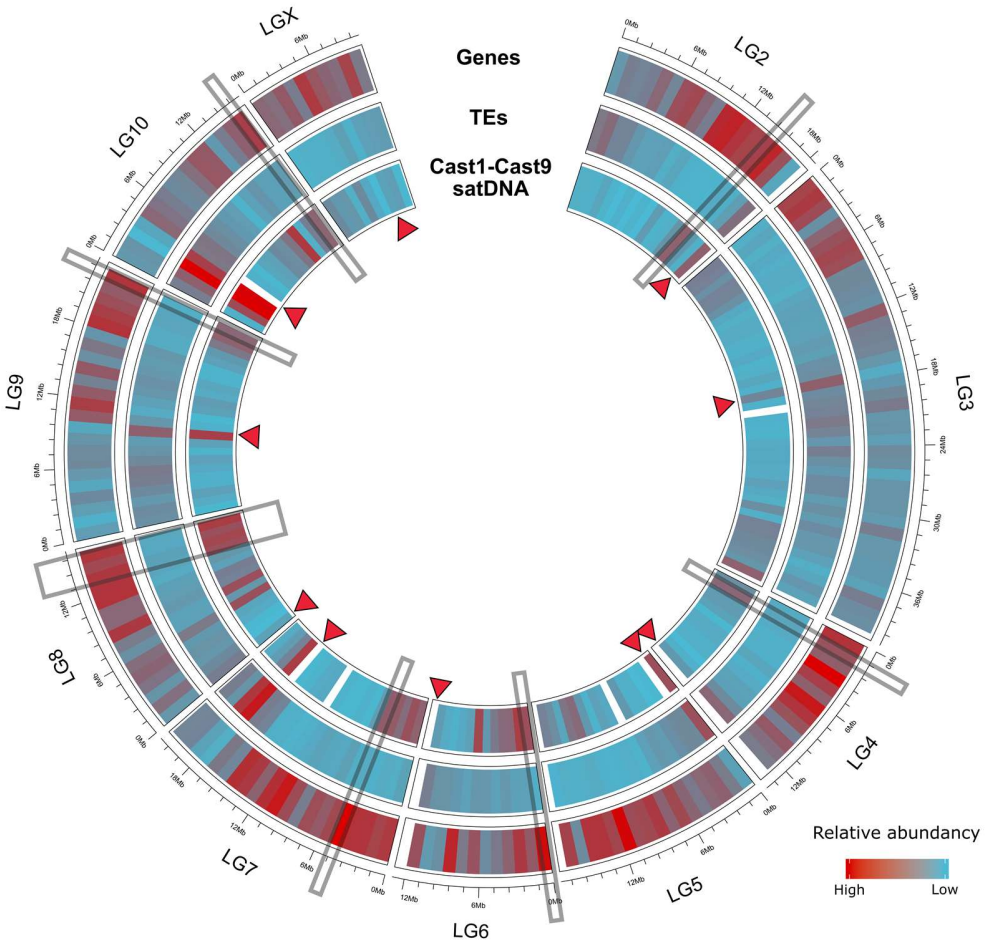

Figure 6. Circular representation of the TcasONT genome assembly with annotated Cast1-Cast9 satDNAs, genes and transposable elements (TEs). The boxes indicate the most prominent regions where the accumulation of Cast satDNAs correlates positively with gene density and negatively with the TEs density. The position of the centromere on each chromosome is indicated by a red arrow.

For a more detailed correlation analysis of Cast1-Cast9 satDNAs and genes, 50 kb flanking regions surrounding Cast1-Cast9 arrays were examined for the presence of genes. More specifically, all exons in the flanking regions of Cast1-Cast9 arrays were counted, and the base exon content of the whole genome assembly was calculated and used to assess the relative exon density in the vicinity of Cast arrays (Figure 7). From the scaled values based on the total number of arrays, it is evident that Cast satDNAs have median exon content values that are above the median of the genome, except for Cast1, Cast6 and Cast7. Moreover, the arrays Cast2', Cast3, Cast5 and Cast8 are surrounded by a significantly larger number of genes (Kolmogorov-Smirnov test, $p<0.01$) than the rest of the genome (Suppl Table 12). In addition, these Cast satDNAs have distributions above 3rd quartile range of the genome indicating the presence of several

arrays surrounded by extremely gene-rich regions. The only Cast satDNA that has significantly lower number of adjacent genes is Cast7 (Supple Table 12, Figure 7). This correlates with the observation that Cast7 arrays occur in an intermingled organization with (peri)centromeric TCAST satDNA (Suppl Figure 7).

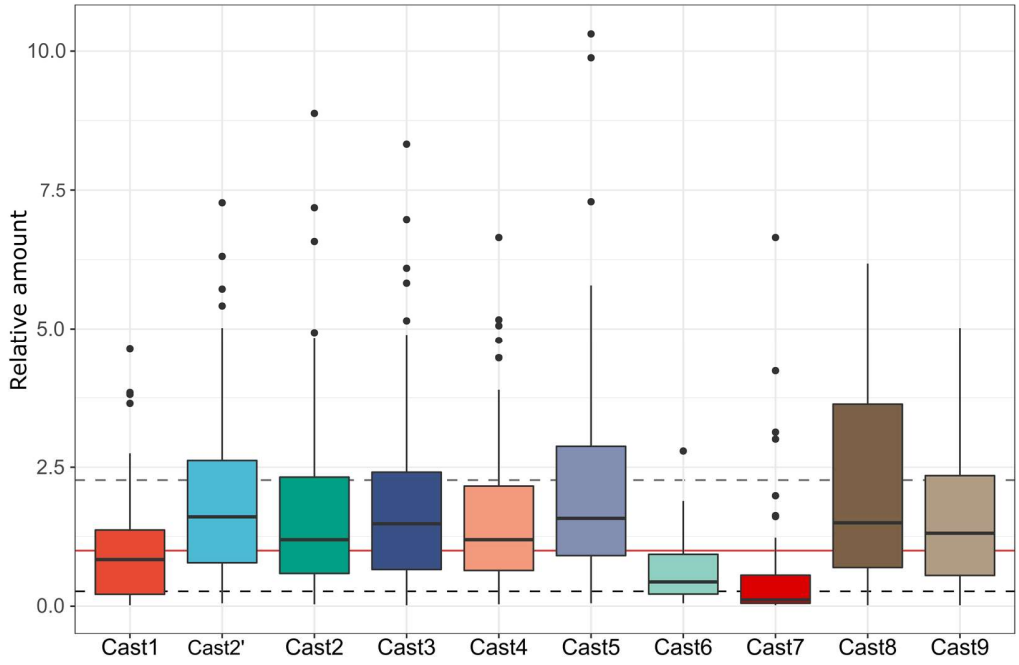

Figure 7. Distribution of exon counts in vicinity of Cast arrays. The relative amount of exons was calculated based on the distribution of exon content in the assembled genome with the red line representing the median number of exons in regions of same size (100 kb) and the dotted line represents the first and third quartiles (Q1 and Q3) of the distribution.

Next, we aimed to determine if there is a correlation between the arrays' length and the density of nearby genes. For this purpose, Cast1-Cast9 arrays were divided into three classes according to their length: (small arrays (<1000kb), moderate arrays (1000-10 000 kb) and long arrays (>10 000kb) and gene density was calculated. The results for each Cast satDNAs are presented in Supplementary Figure 8. Interestingly, most Cast satDNAs have the same profile with most gene occurring in arrays of intermediate length (1000-10 000 bp). In addition, a significant number of genes were observed in the Cast5 arrays, especially near large arrays (>10 kb). Furthermore, the raw exon numbers in the 50kb windows show no discernible pattern, suggesting that the arrays are immersed in regions that exhibit uniform gene presentation.

To better understand the evolution of the Cast satDNAs, a detailed identification of phylogenetic relationships among arrays was performed. First, we created a genome-wide database of Cast arrays that

were annotated with respect to their actual chromosomal positions. Considering that the monomers within the arrays were found to be homogenized, a consensus monomer sequence was generated for each array. Phylogenetic trees were generated separately for each Cast satDNA, and simplified forms of the trees with the consensuses of the arrays colored according to the chromosome of origin are shown in Figure 8. The trees show significant recent intra-chromosomal exchange events. This recent expansion within a chromosome is particularly significant for Cast1, Cast4 and Cast7. However, in addition to intra-chromosomal dispersion, all Cast satDNAs show very dynamic and recent exchange between non-homologs. The Cast2, Cast3, and Cast5 satDNAs exhibit significant levels of recent array exchange between non-homologous chromosomes (the most prominent branches with these features have been marked with asterisks in Figure 8).

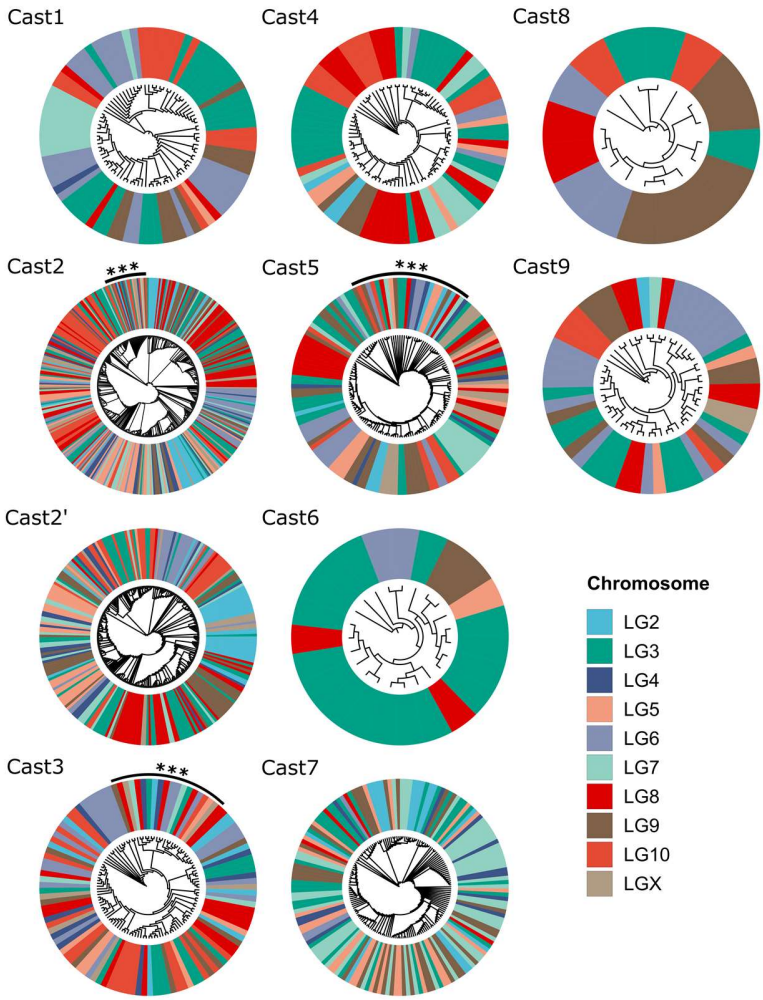

Figure 8. Simplified representation of phylogenetic trees of array consensus sequences for Cast1–Cast9 satDNAs. Terminal branches are represented by strip colored according to the chromosome of origin. The most prominent and recent exchange between non-homologous chromosomes is marked with asterisks.

To gain insight into the spread of Cast1-Cast9 arrays throughout the genome the surrounding regions of the Cast arrays were analyzed. To address this, regions of 2 kb in length upstream and downstream of individual Cast arrays were extracted and grouped according to Cast satDNA. Then the regions were aligned using the MAFFT program and the mutual sequence similarity values were calculated from the alignment. The results show that of the ten Cast satDNAs analyzed, only Cast5 and Cast7 show significant similarity of surrounding regions in most arrays (Figure 9). Detailed analysis of Cast5 arrays surrounding regions showed that previously detected R66-like (1 kb in length) were incorporated as dispersed sequence in the Cast5 arrays and R140-like sequences were found at both ends of Cast5 arrays (Suppl Fig. 5C). In addition to Cast5, most Cast7 arrays were arranged with (peri)centromeric TCAST (Suppl Fig 5D.

The rest of Cast satDNAs showed only partial similarities of the surrounding regions characterized for some subsets of arrays. For example, detailed analyzes of the subset of Cast1 arrays which showed significant mutual similarity of surrounding regions showed that all these arrays originate from the same chromosome (LG7), and had the same transposon-like sequence on one side (Suppl Fig 9). To identify putative sequence motifs enriched near junctions of Cast arrays we generated the MEME dataset for each Cast satDNAs by extracting 500 bases on each side of all arrays. In addition to the similarities already identified above in the surrounding regions (e.g. Cast5, part of Cast1), we could not identify any motifs enriched near the junction regions that apply to the majority of the arrays in the specific Cast satDNA.

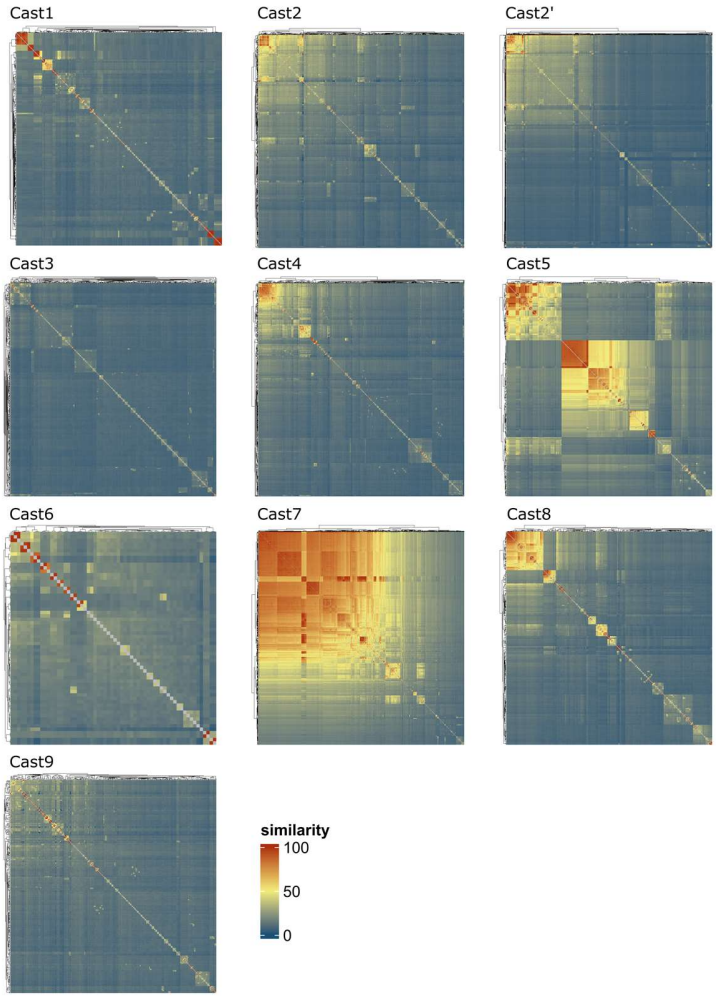

Figure 9. Heatmap visualization of sequence similarity of 2 kb regions around Cast1-Cast9 arrays throughout the TcasONT assembly. The legend indicates that the sequence similarity of the surrounding regions increases, with colors going from blue to red.

In general, chromosomal regions with suppressed or without recombination tend to accumulate satDNAs (Stephan, 1986). Because sex chromosomes represent such regions, we examined the number and length of Cast1-Cast9 arrays on the X chromosome and compared them with other chromosomes. This type of analysis would be even more informative using the Y chromosome, which is mostly non-recombining but unfortunately, the Y chromosome was not available because neither the former Tcas5.2 nor the new TcasONT assembly contain the Y chromosome. Mapping of Cast1-Cast9 arrays to chromosomes disclosed that the X chromosome did not have a significantly higher average number of Cast1-Cast9 arrays per Mb compared to the others (Supp table 13). However, in contrast to number of arrays per chromosome, the analysis of Cast1-Cast9 array length showed a statistically significant longer length of arrays on the X chromosome compared to the others (Wilcoxon test, pval<0.05) (Figure 10.)

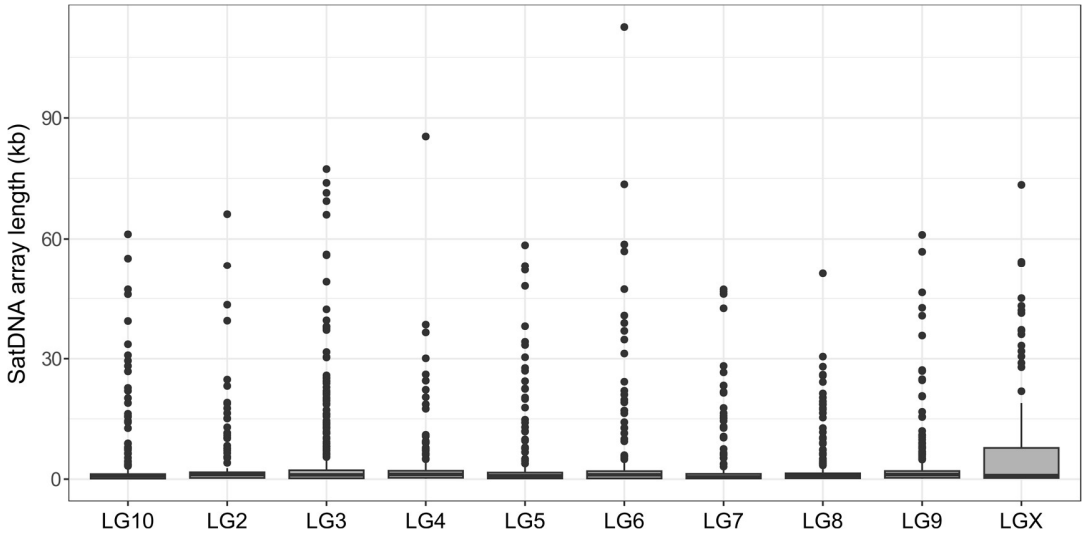

Figure 10. Distribution of Cast1-Cast9 satDNA array lengths across chromosomes, LG2-LG10 including X chromosome (LGX).

Discussion

In our study, a high-quality *T. castaneum* genome assembly was generated at the chromosome level by combining nanopore long-read sequencing and reference-guided approach. The reference-guided approach has recently been successfully used to generate genome assemblies in many species (reviewed in (Lischer & Shimizu, 2017)). Therefore, in our attempt to improve the *T. castaneum* assembly, especially in the repetitive genome part, we used the latest version of the *T. castaneum* assembly Tcas5.2 (Herndon et al., 2020) as a reference. The new TcasONT assembly generated by this approach proved to be enhanced by almost 53 Mb compared to the Tcas5.2 reference genome. In addition to genome increase, the TcasONT assembly also showed significant improvements of genome assembly continuity on the chromosome-level with the elongation of chromosome lengths in the range of 10.6-56.3 %. The TcasONT genome assembly shows collinearity and macrosynteny to the Tcas5.2 assembly. In addition, the TcasONT assembly has a higher completeness of genes than the last version of the reference Tcas5.2, which underlines the credibility of the gene representation. Namely, while 60 genes were missing in Tcast5.2 out of the 1367 insect orthologs analyzed, only 8 were missing in the TcasONT assembly. It is also important to note that the TcasONT genome assembly lacks only 13 Mb of the estimated *T. castaneum* genome sequence of 204 Mb, previously determined experimentally (Alvarez-Fuster et al., 1991) and also *in silico* in this study. We assume that the missing 13 Mb could be attributed primarily to (peri)centromeric regions, as the gene-poor and highly repetitive (peri)centromeric satDNA TCAST regions, which are estimated to comprise 17% of the *T. castaneum* genome (Ugarković et al., 1996). These highly repetitive regions proved to be problematic for correct assembly and thus were largely excluded from the TcasONT assembly process. Consequently, (peri)centromeric regions make up 1.7% of the TcasONT assembly, and they were mainly present in the unassembled reads. This is not unusual that despite significant advances in sequencing technologies and assembly methods, pericentromeric regions remains still incomplete in most even nanopore based assemblies (Talbert & Henikoff, 2022). The only exceptions are the recently published assemblies of prominent model organisms such as human and *Arabidopsis* that contain the complete (peri)centromere (Nurk et al., 2022; Altemose et al., 2022; Naish et al., 2021).

In addition to enrichment in the gene regions, TcasONT genome proved to be representative for the study of the repetitive regions. In that line, comparative studies of the assemblies show a 20-fold enrichment in the repetitive part in the TcasONT assembly compared to Tcas5.2. Transposons have increased in both number of elements and abundance, which is particularly evident in the case of LINE elements, which show even a 90% increase in abundance. Considering all transposon classes, an increase of as much as 21.5 Mb was observed in the TcasONT genome. The TcasONT genome assembly enlargement was up to

26.3 Mb when all classes of tandem repeats (TRs) were considered. The increase is particularly significant for TRs with a monomer length of >50bp, which are usually declared as “classical” satDNA. The enrichment of the proportion of transposons and TRs totaling 47.8 Mb compared to Tcas5.2 indicates that the TcasONT genome provides an excellent platform for the analysis of different repetitive fractions of the genome. Our study shows that nanopore sequencing approach have the potential to overcome the limitations of short reads and enable resolution of previously assembly-impeding regions, most of which consist of highly repetitive sequences such as satDNAs and transposable elements.

In this work, we used the TcastONT genome assembly provided here as a source for a detailed analysis of the most dominant satDNAs outside the (peri)centromere, Cast1-Cas9, previously characterized in (Pavlek et al., 2015). Given that our previous study of these satDNAs was limited because it was based on a portion of Cast1-Cast9 which was found in the Tcas3.0 reference genome, representing less than 1/10 of the actual Cast1-Cast9 genome fraction. The availability of the TcasONT genome assembly provided the opportunity for a comprehensive analysis of Cast1-Cast9 satDNAs. Considering that the assembly process might not assemble the satDNAs correctly and might lead to collapsed arrays (Tørresen et al., 2019), we first performed comparative analyses of the nanopore raw data and the TcasONT assembly to test the credibility of the assembly in terms of representativeness of these sequences. Our results showed that the length distribution of the Cast1-Cast9 arrays in the TcasONT assembly matched that of raw reads, indicating that there were few to no collapsing events in assembling Cast1-Cast9 arrays. In addition to credibility, the content of Cast1-Cast9 satDNA in the TcasONT assembly was also important. The TcasONT assembly showed a representation of Cast satDNAs of 8.8 Mb, which accounts for 4.6% of the genome and it is consistent with their experimentally confirmed genome abundance (Pavlek et al., 2015). These two features, the representativeness and the abundance of Cast satDNAs in the TcasONT assembly, have provided the basis for a deep and credible platform for genome-wide analyses of Cast1-Cast9 satDNAs. Therefore, we first used a new approach to genome-wide characterization of Cast1-Cast9 arrays in the TcasONT assembly. Interestingly we found that these ten most abundant and "classical" satDNAs (Cast1-Cast9, including a new version of Cast2, Cast2') occur in the form of long tandem arrays. Taking into account that they are located outside the (peri)centromere where accumulation of satDNAs is not common, some of them show very high average arrays length (for example, Cast6 with 31kb) and also very long maximum array length (for instance, Cast1 with 112kb). The question arises whether the presence of such long satDNA arrays in regions considered not suitable for the accumulation of satDNAs is a distinctive feature of the *T. castaneum* genome, or whether the absence of saDNA arrays in these regions in other analyzed genomes is a consequence of significant gaps in the assemblies. Given that the

previous high-quality genome assembly Tcas5.2, based on a library BAC and improved by optical mapping, contained only 1/10 of experimentally estimated amount of Cast satDNAs, and that their content in the TcasONT assembly almost completely approached their real amount, we conclude that the strategy of nanopore sequencing combined with reference-guided assembly approach ensures the enviable representation of satDNAs located outside of (peri)centromere.

Furthermore, one of the important questions that arose from the observed organization characterized by long arrays localized in non-pericentromeric regions that were assumed to be euchromatic is whether Cast1-Cast9 satDNAs are truly euchromatic satDNAs. Although it has been previously shown experimentally that these satDNAs are mostly located outside the (peri)centromeric regions, we believe that this fact alone cannot classify them as euchromatic. Indeed, it is possible that satDNAs, even if located outside the (peri)centromeric region, could be localized in gene-poor regions populated by, for example, transposons or other types of noncoding DNA (Meštrović et al., 2015). For this reason, we performed the analysis of surrounding regions of Cast satDNA arrays in terms of gene density and compared them with the average gene density in the whole genome. Interestingly, the results showed that the surrounding regions of almost all Cast satDNAs correlated positively with the gene regions. In addition, the majority of Cast satDNAs were found to have median gene content values that were even slightly higher than the average gene density of the whole genome. In addition, some arrays were surrounded by extremely gene-rich regions. It is also important to note that the surrounding gene-rich regions are not characteristic for short arrays as would be expected but the gene-rich regions are also found surrounding the >10kb arrays. The only exception was Cast7, which was found in association with TCAST satDNA. Because it is known that satDNAs can inhabit regions rich in transposons, another dominant repetitive DNA, we also analyzed the distribution of Cast1-Cast9 satDNAs compared to transposons. We used a database of *T. castaneum* transposons and found a negative correlation between the occurrence of Cast1-Cast9 satDNAs and TEs. To summarize, Cast1-Cast9 satDNAs including their long arrays, were mostly integrated into an environment rich in genes and poor in transposons. This fact clearly speaks to the ability of Cast satDNA to invade euchromatic regions and also indicates the flexibility of euchromatin, i.e., gene-rich regions, to tolerate and accommodate these sequences. Previous models support the fact that the accumulation of satDNA is characteristic of regions where recombination is repressed, such as (peri)centromeric heterochromatin (Charlesworth et al., 2005). This is thought to be due to the potentially deleterious effect of satDNAs, which can affect neighboring regions because of its high turnover. However, from our study, it appears that euchromatin where recombination is expected to occur may have some tolerance to the accumulation of satDNA even in the form of long arrays. In order to test how potentially suppressed

recombination influences the dynamics of Cast1-Cast9 propagation, we performed a comparative analysis of arrays between autosomal chromosomes and X chromosome. In that line, our analyses of Cast1-Cast9 showed that the X chromosome had a similar number of loci with Cast1-Cast9 satDNAs in comparison with autosomes, but arrays were longer in the X chromosome than in the autosomes. These results suggest that suppressed recombination on X chromosomes has less effect on the spread of satDNAs in euchromatin, but rather tolerate propagation of satDNAs into longer arrays.

Considering that previous studies have mostly reported short arrays in euchromatin regions, our finding that there are "classical" satDNAs widely distributed across the genome, even in the form of long arrays in gene-rich regions, is something completely new. Previous studies based on euchromatic satDNA in a form of short arrays suggest that these satDNAs may play roles in gene regulation by acting as “evolutionary tuning knobs” (King et al., 1997) regulating chromatin (Brajković et al., 2012; Feliciello et al., 2015), and facilitating X chromosome recognition/dosage compensation (Joshi & Meller, 2017). SatDNAs located in euchromatic regions also could regulate gene expression by modulating local chromatin structure, or via transcripts derived from the repeats. For example, contractions of the human subtelomeric satellite D4Z4 alter the chromatin state of nearby genes causing muscular dystrophy (Zeng et al., 2009). Recent studies disclosed that expression of euchromatic satDNA-derived transcripts control embryonal development in the mosquito via sequence-specific gene silencing (Halbach et al., 2020). Being located in the gene-rich regions, epigenetic histone marks of satDNAs in euchromatn could also have an effects on surrounding gene regions. For example, human genome-wide analysis of satDNAs located in euchromatin, has shown that they are associated with repressive histone mark H3K9me3, suggesting that they may affect the expression of neighboring genes as repressive H3K9me3 encompasses surrounding regions in addition to satDNA arrays (Vojvoda Zeljko et al., 2021). In addition to direct impact on genes, large-scale rearrangements which could involve observed long Cast1-Cast9 arrays scattered throughout the genome are also highly probable. The ability of euchromatic satDNA to undergo high evolutionary turnover may also contribute to the rapid change of gene landscape across genome and consequently on gene function (Sproul et al., 2020). In summary, the presence of such a large portion of Cast1-Cast9 satDNAs in euchromatin regions of the *T. castaneum* genome can strongly influence both gene expression and the dynamics of genome evolution.

Another important question that arose in view of the extensive distribution of Cast1-Cast9 satDNAs through the *T. castaneum* genome is how these satDNAs are propagated. While the mechanisms of TE propagation are relatively well known and depend on the type of transposon elements (Kelleher et al., 2020), the propagation of satDNA especially in euchromatin regions is still quite a mystery. Given that we

know little about the evolutionary dynamics and molecular mechanisms that shape the distribution of euchromatic satDNA on the genome scale, we carried out analysis of the distribution pattern, phylogenetic analyses as well as analysis of surrounding and junction regions of euchromatic Cast1-Cast9 satDNAs. The observed Cast1-Cast9 satDNAs distribution showed a scattered pattern characterized by arrays dispersed on all chromosomes and along their entire chromosome length without regional preference. The only exception is Cast7, which is mainly located in the proximal pericentromeric regions with the major satDNA TCAST. In accordance with scattered distribution, phylogenetic analysis of Cast satDNA arrays indicate an exceptional expansion of Cast satDNA along the chromosome. However, in addition to this intrachromosomal propagation, phylogenetic analysis also showed relatively recent and extensive dispersion of Cast1-Cast9 satDNA among non-homologous chromosomes. This is particularly characteristic for some Cast satDNAs, such as Cast2, Cast2', Cast3 and Cast5. Observed pattern of distribution of Cast satDNAs characterized by frequent exchange between homologous and non-homologous chromosomes indicates efficient mechanisms responsible for their expansion. Although satDNAs were considered less mobile in contrast to TEs, which are able to move extensively throughout the genome ultimately forming interspersed repeats, our study showed that satDNAs have the same potential to spread through the genome extensively. To determine whether some of them are passively spreading as part of some other sequence, analyzes of neighbor regions were made. Apart from the already known associations of Cast7 with TCAST major satDNAs and Cast 5 with transposon-like sequences, showing that most of their arrays were associated with potential carriers, other Cast satDNAs showed association with a common sequence only in a small fraction of the arrays. To date, there are three mechanisms that have been proposed to be involved in the propagation of satDNA through the euchromatin (i) dispersion in a form of short arrays integrated as central repeats of non-autonomous transposons elements (Dias et al., 2014); (ii) the spread of satDNAs across long physical distances in euchromatin through extrachromosomal circular DNAs (eccDNAs) (Cohen et al., 2003); (iii) interlocus gene conversion events via 3-D interactions between loci in interphase nucleus may also contribute to movement of satDNA (Lee et al., 2020). There are quite a few examples where dispersed satDNAs have been found in the form of short arrays integrated as central repeats of transposons elements (Dias et al., 2014; Macas & Neumann, 2007; Vojvoda Zeljko et al., 2020). Moreover, tandem repeats within Tetris transposon DNA have been reported to be the basis for the generation of long satellite arrays in *Drosophila virilis* that eventually lose transposon features in adjacent regions (Dias et al., 2014). Considering that most of the Cast5 arrays have common adjacent regions, one of which shows low similarities with the Mariner element, it is to be assumed that this Cast satDNA owes its expansion to this

element, similar to the origin of satDNA in *Drosophila* through mediation by Tetris element. Apart from Cast5, and a subpopulation of arrays of some Cast satDNA, the other Cast satDNAs did not show a tendency to associate with other mobile elements but also showed a very extensive propagation that indicates an efficient self-propagation mechanism that operates at the chromosomes level and between chromosomes. Mechanisms that have recently been highlighted as possible drivers in the dispersal of satDNAs in Drosophila euchromatin are eccDNA reintegration and interlocus gene conversion (Sproul et al., 2020). However, a detailed analysis of the junction regions of Cast satDNAs arrays did not indicate which, if any, of the two remaining proposed mechanisms could be involved in this extensive propagation of Cast satDNAs in the *T. castaneum* genome.

In conclusion, this work aims to (1) provide a high-quality reference genome of the insect model organism *T. castaneum*, TcasONT, and (2) take advantage of the new *T. castaneum* assembly to perform a deep genome-wide analysis of dominant satDNAs distributed outside the (peri)centromere. First, we have produced the most contiguous genome assembly of *T. castaneum* to date with significant improvement in the representation of the repetitive genome portion by Oxford Nanopore long-read sequencing. Our study demonstrates that Nanopore sequencing approach has the potential to overcome the limitations of short reads and enable resolution of previously assembly-impeding regions, most of which consist of highly repetitive sequences such as satDNAs and transposable elements. Our reference genome sequences have great potential to be important genomic resources for future research of both genes and various repetitive fractions of the genome. Second, our comprehensive analyses of the abundant satDNAs embedded in euchromatin regions of the *T. castaneum* genome show their massive proliferation of a wide range, implying efficient mechanisms of their spread and amplification in gene-rich regions. Contrary to the previously generally accepted view that the coexistence of gene-rich regions and abundant satDNA is precluded, our study raises the question to which extent this hypothesis is actually based on assemblies that generally show a lack of satDNA sequences not only in highly repetitive (peri)centromeric regions, but also in satDNAs embedded in gene-rich regions.

Materials and methods

Insects sample and DNA preparation

Laboratory cultures of the red flour beetle *T. castaneum*, highly cultured strain Georgia 2 (GA2), were routinely reared in whole wheat flour with the addition of whole rye flour and oats under conditions favoring faster multiplication (32 °C and 70% relative humidity) in the dark. Selection of the different life stages of the insects was performed by sieving through a 0.71-mm sieve and picking the beetles individually with tweezers. Larvae and pupae were collected to achieve the most efficient DNA isolation efficiency. The collected amount of beetles had to be sufficient to perform the isolation with 200 mg of the starting material for pupae and 500 mg for larvae as starting sample for each DNA isolation. DNA isolation was performed followed the protocol described in (Volarić et al., 2021) .Briefly, in order to isolate DNA of sufficient quality and quantity, nuclei were first isolated using a nuclei isolation buffer (NIB), followed by lysis , and finally transferred to a Qiagen Genomic Tip 100/G column. Pure DNA was spooled from eluates and precipitated in isopropanol. In total, we have performed 8 of described isolations, 4 for larvae and 4 for pupae.

Library preparation and sequencing

Oxford Nanopore library was prepared using SQK-LSK110 library preparation kit. Preparation was performed according to the manufacturer's instructions with modifications described below. Prior to sequencing DNA was sheared 30X times using a 30-gauge needle and purified of short fragments using the Short Read Eliminator (SRE) XS kit (Circulomics) according to the manufacturers protocol. The sheared and purified DNA was then used in library preparation protocols in amounts 2X larger than the ONT protocol recommendation. Additionally, all recommended elution and incubation times were extended to twice the length. These steps were performed to avoid library loss and to increase final library concentration sufficient to load multiple MinION flow cell (final concentrations of the library were >100 ng/µL allowing multiple loads from a single library). The flow cells were washed and reloaded 2-5 times to maximize the output. A total of 7 MinION 10.3.4 and 9.4.1 flow cells were used to obtain the data with a total cumulative output of 89.9 GB and an N50 of 20.1 kb. The software used for sequencing was the Oxford Nanopore MinKnow 20.10.3.

Genome assembly

Data obtained from all sequencing runs was basecalled using Oxford Nanopore basecalling software Guppy 5.0.1 with parameters as listed in the Supplementary table 14. The basecalled reads were then used in the assembly process, using Canu v2.2 (Koren et al., 2017) with the parameters listed in the Supplementary table 14. According to Canu documentation the parameters were adjusted for high repetitiveness of the genome (Wang et al., 2008) and the high observed AT content in the reference Tas5.2 genome assembly (Herndon et al., 2020). The reads were filtered to >20 kb in order to reduce the computational intensity needed to perform the assembly, given the amount of data and the relatively small genome of the beetle. Canu assembly was performed using the resources of computer cluster Isabella based in University Computing Centre (SRCE), University of Zagreb.

Filling the gaps

To successfully arrange the Canu contigs into chromosomes based on the Tcas5.2 (GCF_000002335.3) assembly, pre-existing gaps in the Tcas5.2 assembly needed to be bridged. This was done using the TGS-Gapcloser software (Xu et al., 2020) with default for gap-filling settings (Supplementary table 14) using corrected reads generated as the output of the Canu pipeline. Gap filling was used to bridge small and medium gaps in the Tcas5.2 assembly to prevent that Nanopore-based long contigs be interrupted due to discontinuity of the previous assembly.

Contig orientation

After filling the gaps in the Tcas5.2 assembly chromosomes, the RagTag collection of software tools (Alonge et al., 2021) was used. The contigs obtained with Canu were used as the query sequence and the gap-filled Tcas5.2 assembly was used as the reference sequence with “scaffold” parameter (Supplementary table 14). RagTag uses the reference sequence (gap filled Tcas5.2 assembly), as a guide to which it aligns and places the input sequences (Canu contigs) to create the new assembly and fill the gaps in the subject assembly with new successfully aligned contigs. Since the highly repetitive regions are in majority lacking from the Tcas5.2 assembly, this approach allowed to incorporate a new regions of repetitive elements in the TcasONT assembly. RagTag maps the high confidence genomic regions on chromosomes and then, if a contig ends or begins with a repetitive region it is placed in the gap giving us insight into previously unknown regions. The output of the RagTag tool is the unpolished assembly that is used as a template for polishing (Supplementary Figure 1).

Polishing

To improve the quality of the assembly and reduce the error ratio, correction of Canu contigs was performed. To polish the assembly, two separate rounds of RACON (Vaser et al., 2017) polishing were performed using reads excluded from the initial assembly (<20 kb). These short reads used for polishing involved significant addition of genomic region information as this excluded data was ~50 Gb. Polishing was performed according to the instructions within the RACON documentation. The RACON workflow consists of mapping the reads onto the assembled genome with minimap2 (Li, 2018) and then using the information about the mapped reference reads to perform polishing. Produced assembly, named TcasONT was used for downstream analysis.

BUSCO analysis

Benchmarking Single Copy orthologs (BUSCO) (Simão et al., 2015) analysis were performed using the BUSCO v5.0.0. module on the Galaxy web platform (usegalaxy.org). Same settings, as listed in the Supplementary table 14 were repeated for all assembly validations.

Repeat detection

RepeatMasker is a widely used tool for finding and masking different repeat elements within a given target sequence (Smit, AFA, Hubley, R & Green, 2015). RepeatMasker was used to get the GFF/GTF formatted data with the position and orientation of classified RepBase repeat elements. This was a base for the information about the quantity, size, and distribution of different elements within examined genome assemblies. Assemblies were annotated with different elements using the RepeatMasker program on the Galaxy web platform (usegalaxy.org) with different repeat elements from the latest RepBase database (RELEASE 20181026) of repeats and the “Hexapoda” species listing for clade specific repeats. We also rerun RepeatMasker for the Tcas5.2 assembly to update the repeat annotations, as it was created using previous version of RepBase database.

For quantifying of three satDNAs classes (with defined monomer length >50; 50-500; >500bp) in Tcast_ONT and Tcas5.2 assembly, Tandem Repeat Finder (TRF) program (Benson, 1999) was used with default parameters.

Gene annotation

In order to map genes in Canu contigs (filtering) and in TcasONT assembly, we used the novel LiftOff package (Shumate & Salzberg, 2021) using gene annotations from the Tcas5.2 assembly. The Liftoff first

maps the newly created TcasONT assembly to the reference Tcas5.2 and then uses these overlaps to accurately align genes from a reference Tcas5.2 assembly to a target TcasONT. This approach was limited to finding possible new genes in the improved version of the TcasONT assembly, but because the Tcas5.2 assembly was annotated based on a large database of RNA-seq reads combined with gene and general trait prediction methods, this approach ensures gene completeness.

Identification of Cast1-Cast9 satDNAs

Annotation of satellite repeats within the genomes was performed using the standalone NCBI's BLAST algorithm, and the interface to the R programming language package metablastr (Benoit & Drost, 2021). The subject sequences were the different analyzed assemblies (Tcas5.2 and TcasONT) while the queries were the previously characterized Cast1-Cast9 (Pavlek et al., 2015). Post-analysis included filtering BLAST hits in the genome to discover the Cast1-Cast9 trends and arrays.

Analysis of Cast1-Cast9 satDNA arrays

All Cast1-Cast9 monomers were defined from the BLAST result table, filtered according to parameters described in Supplementary Figure 2. Next, it was crucial to define optimal parameters for satDNA arrays detection to avoid fragmentation due to putative short different sequence(s) incorporated into the arrays. Total arrays of particular Cast satDNA were analysed to evaluate best neighboring window length that ensures the connection of the continuous repeating monomers into one array (Supplementary Figure 3). This approach to array definition was introduced to allow for errors and array insertions and to correctly link all existing monomers of a given satellite as accurately as possible. Afterwards we performed basic filtering in order to define arrays and remove short interspersed monomers using the following parameters:

(array!="Cast2-mix" & width>530) | (array=="Cast2-mix" & width > 2000)

This parameters ensured arrays with 3 or more repeat units for each satDNA. Exception was the Cast2' array (Cast2 monomer interspersed with the newly discovered sequence interCast2) where 3 different length monomers were included, out of which 1100 bp interCast2 is mixed with 170 bp Cast2.

Analysis of gene content in vicinity of Cast1-Cast9 arrays

A region of 50 kb upstream or downstream of the arrays was selected to define gene profiles around the different Cast1-Cast9 arrays. In these 100-kb regions, 100 equal bins of 1 kb size (50 upstream and 50 downstream of a particular array) were created, and the number of exons was counted in each of them.

This allowed gene profiling around different Cast1-Cast9 arrays. Expected exon densities were calculated by computing median, 1Q, and 3Q exon densities across the entire genome in 100-kb windows using a custom R script.

Analysis of Cast1-Cast9 junction regions (MEME)

In order to find particular motifs surrounding Cast1-9 arrays we used source code compiled version of MEME-5.5.1. (Bailey & Elkan, 1994). We performed MEME motif identification in 100 bp, 200 bp and 500 bp windows around particular arrays of interest, resulting motifs were investigated and in both quantity and localization.

Multiple sequence alignment and tree construction

We used MAFFT (Katoh & Standley, 2013) to create multiple sequence alignments to generate both intra-array consensus sequences and the multiple alignment required to generate evolutionary trees. Evolutionary trees were constructed using IQ-TREE (Nguyen et al., 2015) and the maximum-likelihood model with 200 iterations per Cast satDNA monomer.

Visualizations and calculations

All plots and calculations were produced in R using custom made data processing notebooks. Apart from the standard libraries, we also used the *circilize* package (Gu et al., 2014) to create the circular visualization plots of global genome patterns. To create the complex heatmaps used in the characterization of neighboring regions similarity we used the *ComplexHeatmap* package (Gu et al., 2016).

Data avaliability

The *T. castaneum* genome assembly TcasONT has been deposited to the European Nucleotide Archive (ENA) under the BioProject accession PRJEB61413 and assembly accession number GCA_XXXX. The genome annotation presented in this study are available in FigShare at https://doi.org/10.6084/m9.figshare.22683325.v1

References

Aldrup-MacDonald, M. E., Kuo, M. E., Sullivan, L. L., Chew, K., & Sullivan, B. A. (2016). Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Research*, *26*(10), 1301–1311. https://doi.org/10.1101/gr.206706.116

Alonge, M., Lebeigle, L., Kirsche, M., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., & Soyk, S. (2021). Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. *BioRxiv*, 2021.11.18.469135. https://www.biorxiv.org/content/10.1101/2021.11.18.469135v1%0Ahttps://www.biorxiv.org/content/10.1101/2021.11.18.469135v1.abstract

Alvarez-Fuster, A., Juan, C., & Petitpierre, E. (1991). Genome size in *Tribolium* flour-beetles: Inter-and intraspecific variation. *Genetical Research*, *58*(1), 1–5. https://doi.org/10.1017/S0016672300029542

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, *2*, 28–36.

Benoit, M., & Drost, H. G. (2021). A Predictive Approach to Infer the Activity and Natural Variation of Retrotransposon Families in Plants. *Methods in Molecular Biology*, *2250*, 1–14. https://doi.org/10.1007/978-1-0716-1134-0_1

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Bersani, F., Lee, E., Kharchenko, P. V., Xu, A. W., Liu, M., Xega, K., MacKenzie, O. C., Brannigan, B. W., Wittner, B. S., Jung, H., Ramaswamy, S., Park, P. J., Maheswaran, S., Ting, D. T., & Haber, D. A. (2015). Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(49), 15148–15153. https://doi.org/10.1073/pnas.1518008112

Blattes, R., Monod, C., Susbielle, G., Cuvier, O., Wu, J. H., Hsieh, T. S., Laemmli, U. K., & Käs, E. (2006). Displacement of D1, HP1 and topoisomerase II from satellite heterochromatin by a specific polyamide. *EMBO Journal*, *25*(11), 2397–2408. https://doi.org/10.1038/sj.emboj.7601125

Brajković, J., Feliciello, I., Bruvo-MadWarić, B., & Ugarković, D. W. (2012). Satellite DNA-like elements associated with genes within euchromatin of the beetle *Tribolium castaneum*. *G3: Genes, Genomes, Genetics*, *2*(8), 931–941. https://doi.org/10.1534/g3.112.003467

Brown, S. J., Henry, J. K., Black IV, W. C., & Denell, R. E. (1990). Molecular genetic manipulation of the red flour beetle: Genome organization and cloning of a ribosomal protein gene. *Insect Biochemistry*, *20*(2), 185–193. https://doi.org/10.1016/0020-1790(90)90011-I

Campbell, J. F., Athanassiou, C. G., Hagstrum, D. W., & Zhu, K. Y. (2022). *Tribolium castaneum*: A Model Insect for Fundamental and Applied Research. *Https://Doi.Org/10.1146/Annurev-Ento-080921-075157*, *67*, 347–365. https://doi.org/10.1146/ANNUREV-ENTO-080921-075157

Cohen, S., Yacobi, K., & Segal, D. (2003). Extrachromosomal circular DNA of tandemly repeated genomic sequences in *Drosophila*. *Genome Research*, *13*(6 A), 1133–1145. https://doi.org/10.1101/gr.907603

Dias, G. B., Svartman, M., Delprat, A., Ruiz, A., & Kuhn, G. C. S. (2014). Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of *Drosophila virilis*. *Genome Biology and Evolution*, *6*(6), 1302–1313. https://doi.org/10.1093/gbe/evu108

Feliciello, I., Akrap, I., & Ugarković, Đ. (2015). Satellite DNA Modulates Gene Expression in the Beetle *Tribolium castaneum* after Heat Stress. *PLoS Genetics*, *11*(8), 1–18. https://doi.org/10.1371/journal.pgen.1005466

Feliciello, I., Pezer, Ž., Kordiš, D., Madarić, B. B., & Ugarković, D. (2020). Evolutionary history of alpha satellite DNA repeats dispersed within human genome euchromatin. *Genome Biology and Evolution*, *12*(11), 2125–2138. https://doi.org/10.1093/gbe/evaa224

Ferree, P. M., & Prasad, S. (2012). How Can Satellite DNA Divergence Cause Reproductive Isolation? Let Us Count the Chromosomal Ways. *Genetics Research International*, *2012*(Figure 1), 1–11. https://doi.org/10.1155/2012/430136

Gržan, T., Despot-Slade, E., Meštrović, N., Plohl, M., & Mravinac, B. (2020). CenH3 distribution reveals extended centromeres in the model beetle *Tribolium castaneum*. *PLoS Genetics*, *16*(10), e1009115. https://doi.org/10.1371/journal.pgen.1009115

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/BIOINFORMATICS/BTW313

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize implements and enhances circular visualization in R. *Bioinformatics*, *30*(19), 2811–2812. https://doi.org/10.1093/BIOINFORMATICS/BTU393

Halbach, R., Miesen, P., Joosten, J., Taşköprü, E., Rondeel, I., Pennings, B., Vogels, C. B. F., Merkling, S. H., Koenraadt, C. J., Lambrechts, L., & van Rij, R. P. (2020). A satellite repeat-derived piRNA controls embryonic development of *Aedes*. *Nature*, *580*(7802), 274–277. https://doi.org/10.1038/s41586-020-2159-2

Herndon, N., Shelton, J., Gerischer, L., Ioannidis, P., Ninova, M., Dönitz, J., Waterhouse, R. M., Liang, C., Damm, C., Siemanowski, J., Kitzmann, P., Ulrich, J., Dippel, S., Oberhofer, G., Hu, Y., Schwirz, J., Schacht, M., Lehmann, S., Montino, A., … Bucher, G. (2020). Enhanced genome assembly and a new official gene set for *Tribolium castaneum*. *BMC Genomics*, *21*(1), 47. https://doi.org/10.1186/s12864-019-6394-6

Joshi, S. S., & Meller, V. H. (2017). Satellite Repeats Identify X Chromatin for Dosage Compensation in *Drosophila melanogaster* Males. *Current Biology*, *27*(10), 1393-1402.e2. https://doi.org/10.1016/j.cub.2017.03.078

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kelleher, E. S., Barbash, D. A., & Blumenstiel, J. P. (2020). Taming the Turmoil Within: New Insights on the Containment of Transposable Elements. *Trends in Genetics : TIG*, *36*(7), 474.

https://doi.org/10.1016/J.TIG.2020.04.007

King, D. G., Soller, M., & Kashi, Y. (1997). Evolutionary tuning knobs. *Endeavour*, *21*(1), 36–40. https://doi.org/10.1016/S0160-9327(97)01005-3

Klingler, M. (2004). Tribolium. *Current Biology : CB*, *14*(16), 639–640. https://doi.org/10.1016/j.cub.2004.08.004

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Kuhn, G. C. S., Küttler, H., Moreira-Filho, O., & Heslop-Harrison, J. S. (2012). The 1.688 repetitive DNA of *Drosophila*: Concerted evolution at different genomic scales and association with genes. *Molecular Biology and Evolution*, *29*(1), 7–11. https://doi.org/10.1093/molbev/msr173

Lee, Y. C. G., Ogiyama, Y., Martins, N. M. C., Beliveau, B. J., Acevedo, D., Wu, C. ting, Cavalli, G., & Karpen, G. H. (2020). Pericentromeric heterochromatin is hierarchically organized and spatially contacts H3K9me2 islands in euchromatin. *PLoS Genetics*, *16*(3), 1–27. https://doi.org/10.1371/journal.pgen.1008673

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lischer, H. E. L., & Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, *18*(1), 1–12. https://doi.org/10.1186/s12859-017-1911-6

Lower, S. S., McGurk, M. P., Clark, A. G., & Barbash, D. A. (2018). Satellite DNA evolution: old ideas, new approaches. *Current Opinion in Genetics and Development*, *49*(i), 70–78. https://doi.org/10.1016/j.gde.2018.03.003

Macas, J., & Neumann, P. (2007). Ogre elements - A distinct group of plant Ty3/gypsy-like retrotransposons. *Gene*, *390*(1–2), 108–116. https://doi.org/10.1016/j.gene.2006.08.007

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., Sebra, R., Peluso, P., Eid, J., Rank, D., Garcia, J. F., DeRisi, J. L., Smith, T., Tobias, C., Ross-Ibarra, J., Korf, I., & Chan, S. W. L. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, *14*(1), 1–20. https://doi.org/10.1186/gb-2013-14-1-r10

Menon, D. U., Coarfa, C., Xiao, W., Gunaratne, P. H., & Meller, V. H. (2014). siRNAs from an X-linked satellite repeat promote X-chromosome recognition in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(46), 16460–16465. https://doi.org/10.1073/pnas.1410534111

Meštrović, N., Mravinac, B., Pavlek, M., Vojvoda-Zeljko, T., Šatović, E., & Plohl, M. (2015). Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosome Research*, *23*(3), 583–596. https://doi.org/10.1007/s10577-015-9483-7

Miga, K. H. (2015). Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Research*, *23*(3), 421–426. https://doi.org/10.1007/s10577-015-9488-2

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, V. A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., … Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, *585*(7823), 79–84. https://doi.org/10.1038/s41586-020-2547-7

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A. J., Abramson, B. W., Schmücker, A., Mandáková, T., Jamge, B., Lambing, C., Kuo, P., Yelina, N., Hartwick, N., Colt, K., Smith, L. M., Ton, J., Kakutani, T., Martienssen, R. A., Schneeberger, K., Lysak, M. A., … Henderson, I. R. (2021). The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science*, *374*(6569). https://doi.org/10.1126/science.abi7489

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

Pavlek, M., Gelfand, Y., Plohl, M., & Meštrović, N. (2015). 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. *DNA Research*, *22*(6), 387–401. https://doi.org/10.1093/dnares/dsv021

Plohl, M., Meštrović, N., & Mravinac, B. (2014). Centromere identity from the DNA point of view. *Chromosoma*, *123*(4), 313–325. https://doi.org/10.1007/s00412-014-0462-0

Richards, S., Gibbs, R. A., Weinstock, G. M., Brown, S., Denell, R., Beeman, R. W., Bucher, G., Friedrich, M., Grimmelikhuijzen, C. J. P., Klingler, M., Lorenzen, M., Roth, S., Schröder, R., Tautz, D., Zdobnov, E. M., Muzny, D., Attaway, T., Bell, S., Buhay, C. J., … Grossmann, D. (2008). The genome of the model beetle and pest *Tribolium castaneum*. *Nature*, *452*(7190), 949–955. https://doi.org/10.1038/nature06784

Shumate, A., & Salzberg, S. L. (2021). Liftoff: accurate mapping of gene annotations. *Bioinformatics*, *37*(12), 1639–1643. https://doi.org/10.1093/BIOINFORMATICS/BTAA1016

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smit, AFA, Hubley, R & Green, P. (2015). *RepeatMasker Open-4.0.* http://www.repeatmasker.org

Sproul, J. S., Khost, D. E., Eickbush, D. G., Negm, S., Wei, X., Wong, I., & Larracuente, A. M. (2020). Dynamic evolution of euchromatic satellites on the x chromosome in *Drosophila melanogaster* and the simulans clade. *Molecular Biology and Evolution*, *37*(8), 2241–2256. https://doi.org/10.1093/MOLBEV/MSAA078

Stephan, W. (1986). Nonlinear Phenomena in the Evolution of Satellite. *Berichte Der Bunsengesellschaft Für Physikalische Chemie*, *90*(11), 1029–1034. https://doi.org/10.1002/bbpc.19860901119

Talbert, P. B., & Henikoff, S. (2022). The genetics and epigenetics of satellite centromeres. *Genome Research*, *32*(4), 608–615. https://doi.org/10.1101/gr.275351.121

Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., Anisimova, M., Jakobsen, K. S., & Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, *47*(21), 10994–11006. https://doi.org/10.1093/nar/gkz841

Ugarković, D., Podnar, M., & Plohl, M. (1996). Satellite DNA of the red flour beetle *Tribolium castaneum* - Comparative study of satellites from the genus *Tribolium*. *Molecular Biology and Evolution*, *13*(8), 1059–1066. https://doi.org/10.1093/oxfordjournals.molbev.a025668

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/gr.214270.116

Vojvoda Zeljko, T., Pavlek, M., Meštrović, N., & Plohl, M. (2020). Satellite DNA-like repeats are dispersed throughout the genome of the Pacific oyster *Crassostrea gigas* carried by Helentron non-autonomous mobile elements. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-71886-y

Vojvoda Zeljko, T., Ugarković, Đ., & Pezer, Ž. (2021). Differential enrichment of H3K9me3 at annotated satellite DNA repeats in human cell lines and during fetal development in mouse. *Epigenetics and Chromatin*, *14*(1), 1–16. https://doi.org/10.1186/s13072-021-00423-6

Volarić, M., Veseljak, D., Mravinac, B., Meštrović, N., & Despot-Slade, E. (2021). Isolation of high molecular weight DNA from the model beetle *Tribolium* for nanopore sequencing. *Genes*, *12*(8). https://doi.org/10.3390/genes12081114

Wang, S., Lorenzen, M. D., Beeman, R. W., & Brown, S. J. (2008). Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. *Genome Biology*, *9*(3), 1–14. https://doi.org/10.1186/gb-2008-9-3-r61

Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B. A., Fan, G., Liu, X., Xu, X., Deng, L., & Zhang, Y. (2020). TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience*, *9*(9), 1–11. https://doi.org/10.1093/gigascience/giaa094

Zeng, W., De Greef, J. C., Chen, Y. Y., Chien, R., Kong, X., Gregson, H. C., Winokur, S. T., Pyle, A., Robertson, K. D., Schmiesing, J. A., Kimonis, V. E., Balog, J., Frants, R. R., Ball, A. R., Lock, L. F., Donovan, P. J., Van Der Maarel, S. M., & Yokomori, K. (2009). Specific loss of histone H3 lysine 9 trimethylation and HP1γ/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD). *PLoS Genetics*, *5*(7). https://doi.org/10.1371/journal.pgen.1000559