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 genome assembly (TcasONT) of the model beetle *Tribolium castaneum*. The new TcasONT is enriched by 50 Mb in the repetitive genome part, which is as much as 1/4 of *T. castaneum* genome size. Therefore, we used the enhanced assembly to conduct global and in-depth analyses of abundant non-(peri)centromeric satDNAs, Cast1-Cast9. We found a large proportion of long arrays of Cast1-Cast9 satDNAs embedded in gene-rich regions. In addition, occurrence of Cast1-Cast9 satDNAs and transposable elements often did not overlap. Contrary to accepted opinion, our finding clearly speaks to the ability of satDNAs to expand in gene-rich regions and also demonstrate the flexibility of gene-rich regions to tolerate and accommodate satDNA sequences. In addition, we found that suppressed recombination on X chromosomes has less effect on the spread of Cast1-Cast9 satDNAs, but rather tolerates amplification of these satDNAs into longer arrays. Our results also show the extensive spread of Cast1-Cast9 satDNAs, characterized by distribution along the entire chromosome length and between different chromosomes. This observed pattern points out very efficient mechanisms of propagation of satDNAs on the whole genome scale including gene-rich regions. 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.

31

32 ## Background

33 Satellite DNAs (satDNAs) are the most abundant and rapidly evolving non-coding DNA of all eukaryotic
34 genomes. In contrast to other groups of repetitive elements such as interspersed transposable elements
35 (TE), which are scattered throughout the genome, satDNAs are characterized by monomer sequences
36 tandemly arranged into long arrays. Previous studies on many animal and plant species indicate
37 preferential localization of satDNAs in the (peri)centromeric heterochromatin [1]. Heterochromatin
38 represents a highly compact gene-poor specialized chromatin structure that is distinct from the gene-rich
39 euchromatin. SatDNAs in (peri)centromeric heterochromatin have a long history of investigation including
40 many plant and animal species, and studies have suggested a role of satDNA in chromatin packaging,
41 centromere formation/maintenance, and centromere evolution [2–4]. SatDNAs are also linked to
42 chromosome mis-segregation [3], disease phenotypes [5], and reproductive isolation between species [6].
43 In addition to standard satDNA-rich regions such as (peri)centromeric heterochromatin, recent studies
44 revealed examples where short arrays of (peri)centromeric satDNAs also spread to euchromatic regions.
45 For example, (peri)centromeric satDNA families in humans, *Drosophila* and the beetle *T. castaneum* can
46 also be dispersed within euchromatin in the form of clustered or short single repeat arrays [7–9]. In
47 *Drosophila* X-specific (peri)centromeric 1688 satDNAs dispersed in euchromatic regions have a role in
48 dosage compensation through up-regulation of X-linked genes in *Drosophila* males [10,11]. The
49 euchromatic counterpart of (peri)centromeric DNA in the beetle *T. castaneum* modulates the local
50 chromatin environment upon heat stress changing the expression of neighboring genes [12]. Although
51 there is clear evidence that satDNAs exist outside the (peri)centromere and that some roles have been
52 assigned to them, understanding of their organization, evolutionary dynamics, and disclosure of the
53 molecular mechanisms driving their spread, movement, and rearrangement in euchromatin is still very
54 limited. The key requirement that would help to understand the evolution and function of these
55 sequences is their detection at the genome level and genome-wide studies of satDNAs in various model
56 organisms. One of the main reasons for the current lack of global and in-depth studies of satDNAs in
57 euchromatin is certainly the fact that satDNA regions are the most difficult part of the genome to
58 sequence and assemble, and therefore they are underrepresented or absent even in the best genome
59 assemblies (reviewed in [13,14]). However, with the improvement of long-read sequencing technology,
60 especially Oxford Nanopore platform, it has recently been demonstrated that genomic loci made of

61 satDNAs can be efficiently assembled [15]. Moreover, end-to-end maps of all human chromosomes,
62 spanning highly repetitive (peri)centromeric and telomeric regions, have recently been obtained [16].
63 The beetle *T. castaneum* is a worldwide pest of stored products and the representative of the most
64 species-rich animal order on Earth, the coleopterans. *T. castaneum* has become one of the most important
65 models in the field of evolution, physiology and development of insects because its development is more
66 representative for insects compared to *Drosophila* [17,18]. The *T. castaneum* genome has been
67 sequenced, annotated and a first version of genome assembly was available in 2008 [19]. Structural
68 genome analyses revealed large amounts (>30%) of different classes of repetitive DNA in the assembled
69 part of the *T. castaneum* genome. Among them, the (peri)centromeric heterochromatic satDNA TCAST,
70 which was drastically underrepresented in the assembled genome, dominates on all *T. castaneum*
71 chromosomes comprising up to 17% of the genome [20,21]. All *T. castaneum* chromosomes are
72 characterized by large blocks of (peri)centromeric heterochromatin, while no prominent heterochromatic
73 blocks could be detected cytogenetically on chromosome arms [22]. Regarding the possible presence of
74 satDNAs in euchromatin of *T. castaneum*, our previous study on the *T. castaneum* reference genome
75 assembly disclosed that, in addition to short arrays of counterpart of (peri)centromeric DNA in
76 euchromatin regions [9], *T. castaneum* genome abounds with many different and unrelated satDNAs
77 distributed out of (peri)centromeric regions [23]. Nine satDNAs, Cast1-Cast9, represent the most
78 prominent non-pericentromeric satDNA portion which constitutes ∼0.3 % of the genome assembly.
79 Unlike of the most euchromatic satDNAs, found for example in *Drosophila*, which have short monomer
80 units, Cast1-Cas9 represent "classical" satellite DNAs based on ∼170 and ∼340 bp long monomers. The
81 experimentally determined genome content of Cast1-Cast9 satDNAs showed abundance of approximately
82 ∼4.3%, which is ∼10X higher than the analyzed representativeness in the assembled genome [23]. The
83 genome version from 2008 was based on a Sanger 7x draft assembly and was anchored to Linkage Groups
84 (LGs) [19]. Recently, the genome assembly was upgraded to a new version, Tcas5.2, which was improved
85 by using large-distance jumping libraries and BioNano Genomics optical mapping, bringing also a new
86 official gene set OGS3 [24]. Nevertheless, our previous satDNA analyses showed that satDNA genome
87 fractions were highly underrepresented in Tcas5.2 as well.
88 We considered that for a deep genome-wide analysis of dominant satDNAs distributed outside the
89 (peri)centromere, the first prerequisite is the availability of the assembly enriched in the repetitive
90 sequences. For that reason, we used Oxford Nanopore Technology (ONT) long-read sequencing and
91 applied a reference-guided assembly strategy using the reference assembly Tcas5.2 to produce an
92 improved genome assembly of *T. castaneum*. This approach allowed us 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 (TcasONT) 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 the TcasONT assembly and subsequent analysis of their chromosomal distribution, organization, evolutionary trends as well as analysis of satDNAs' environment.

122

123

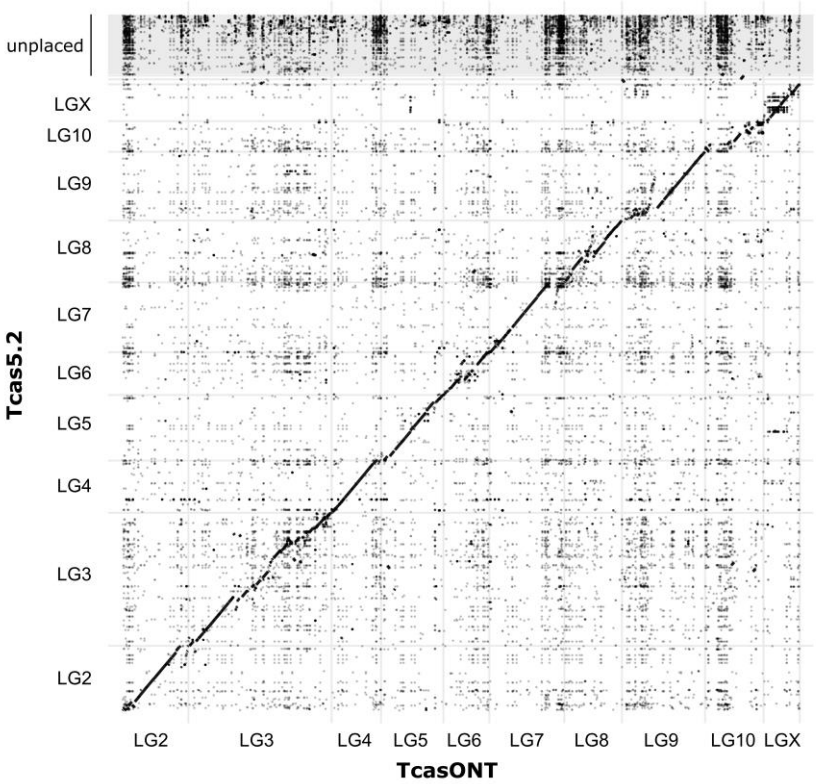
124 # Results

125 **1. Chromosome-scale ONT-genome assembly**

126 The size of the reference *T. castaneum* genome assembly (Tcas5.2) is 165,9 Mb [24]. Removing of
127 placeholders and sequencing gaps results in a genome assembly size of 136 Mb. Considering that the
128 experimentally estimated genome size is 204 Mb [25], it is clear that 68 Mb of the genome is missing from
129 the Tcas5.2 reference assembly. In order to improve the existing assembly, especially in repetitive genome
130 parts, we employed Oxford Nanopore Technologies' (ONT) long-read sequencing platform. In total, 121
131 Gb of nanopore long reads were generated. Incorporating only those reads that had a mean quality score
132 of >7, a total of 89 Gb were used for the new genome assembly representing genome coverage of 436X
133 (Suppl Table 1 Nanoplot). Selected reads were up to 500 kb in length with a mean read length of 15.8 kb
134 and mean quality score per read of 12.8. Main steps of genome assembly process are presented in Suppl
135 Figure 1. First, nanopore reads were divided into two fractions based on their length, short reads (<20kb;
136 total of 52.7Gb; 258X genome coverage), and long reads (>20kb; total of 36.3 Gb; 178X genome coverage).
137 Next, long reads (>20kb) were used for the initial assembly using Canu [26]. This initial Canu assembly
138 had 1479 contigs with total length of 321 Mb and a N50 of 835.5 kb. (Suppl Table 2, left Canu statistics).
139 The longest contig was 16.4 Mb in length. Our initial Canu assembly was approximately 117 Mb larger
140 than the previously experimentally estimated genome size of 204 MB [27]. To verify the genome size and
141 genome repetitiveness *in silico*, the k-mer frequencies of the corrected reads generated during the Canu
142 assembly process using the FindGSE and CovEST programs were calculated. FindGSE estimated the
143 genome to be 204 Mb, while CovEST RE showed only slightly higher estimation of 208 Mb (Supp Table 3
144 genome size), which is fully consistent with experimental estimation. The repeat ratio calculated by
145 FindGSE was 27% of genome confirming the previously observed high genome repetitiveness [22]. In
146 terms of repetitiveness, previous studies showed that up to 17% of the *T. castaneum* genome is
147 (peri)centromeric satDNA TCAST [20,21]. Considering that abundant tandem repeats might be
148 problematic for proper assembly [28], we assumed that the additional 117 Mb in our Canu assembly (Supp
149 table 2) were primarily generated by gene-poor and highly repetitive (peri)centromeric TCAST arrays. To
150 reduce the suspected negative effects of TCAST satDNA on the genome assembly process, we performed
151 an additional filtering step. In this step, the contigs that did not contain at least 1000 bp of unique gene-

152 coding sequence were removed, resulting in successful filtering of 471 out of 1479 total contigs (Suppl
153 Figure 1).
154 These 471 sorted contigs were further used to achieve a new chromosome-scale assembly using
155 reference-guided approach. The latest improved *T. castaneum* genome assembly Tcas5.2 [24] was used
156 as the reference for organizing these contigs into chromosomes. Two reasons were crucial for the
157 selection of this approach: 1) the availability of the high-quality *T. castane*um reference genome, Tcas5.2,
158 and 2) assembly approach that allows the input sequence to be ordered based on the mapping quality of
159 the reference sequence. Given that the reference genome Tcas5.2 contained 3669 unresolved gaps with
160 total N (ambiguous nucleotides) size of 11 Mb and mean gap size of 3125 Kb, the first step of reference-
161 guided assembly approach included gap filling of Tcas5.2 with TGS-GapCloser using 8.6 Gb of Canu
162 corrected larger reads (length >20 kb). TGS-GapCloser was efficient in filling most of the gaps, so in total
163 3607 gaps (98.3 %) were closed which increased the genome size by 10.5 Mb (Suppl table 4). The gap-
164 filled Tcas5.2. reference genome assembly was further used as a template for orientating 471 Canu-
165 contigs into the 10 chromosomes using RagTag software tool [29]. A total of 244 contigs were
166 unambiguously assigned to a reference Tcas5.2. genome representing a new ONT-based genome
167 assembly, named TcasONT (Supply table 5). The rest (227 contigs) did not accurately map on the
168 assembly, remaining assigned as unassembled sequences (Supp table 6). To further validate structure of
169 omitted contigs, we mapped them with highly abundant (peri)centromeric TCAST satDNAs and found that
170 around half of unplaced sequences (14 Mb) are indeed mostly covered (>50 %) with (Supp table 6). To
171 gain insight into the actual representation of these regions, we mapped the TcasONT assembly with TCAST
172 satDNA and obtained 9446 monomers, accounting for 3.6 Mb (1.7%) of the genome.
173 To improve genome accuracy, the TcasONT genome assembly was polished using Racon with 56.3 Gb of
174 raw short nanopore reads (<20 kb) which have been excluded from the initial Canu-assembly but passed
175 the quality check. The final polished TcasONT has a size of 225.9 Mb, with 191 Mb assembled in ten
176 chromosomes and the remainder in unassembled contigs (Suppl table 7). Compared to the reference
177 Tcas5.2, the new TcasONT assembly shows a 53 Mb increase of total chromosome length (189 Mb in
178 contrast to 136 Mb genome without gaps or Ns, Suppl Table 7). Comparing TcasONT and Tcas5.2
179 chromosomes, the elongation of chromosome lengths ranges from 10.6 to 56.3 %, indicating a significant
180 improvement of genome assembly continuity at the chromosome level (Supp table 8).
181 To measure completeness and quality of the new assembly, we compared the TcasONT assembly to the
182 reference Tcas5.2 using dot plot analyses. The result revealed high levels of macrosynteny and collinearity
183 across all chromosomes with high sequence identity (Figure 1). Furthermore, we calculated that 88% of

184 the total Tcas5.2 unplaced contig sequence length is now placed within chromosomes of the new TcasONT
185 assembly.
186



187
188 Figure 1. Dot-plot representation of TcasONT vs. Tcas5.2 assembly comparison. The horizontal axis
189 corresponds to the intervals along the TcasONT assembly and the vertical axis corresponds to the intervals
190 along the Tcas5.2 assembly. LG1-LG10 represent chromosomes, while unplaced contigs contain
191 sequences in Tcas5.2. that were not associated with any chromosome. Dots closest to the diagonal line
192 indicate co-linearity between the two assemblies.
193
194
195 We next assessed the completeness of the TcasONT assembly by comparative BUSCO analyses between
196 TcasONT and Tcas5.2 using insect universal orthologs from the odb10 database (Figure 2). Of the 1367
197 genes, we identified 1329, which corresponds to 97.2 % of BUSCO single-copy completeness. Further, 17
198 genes were duplicated, 13 were fragmented and only 8 genes were missing. A considerable increase in
199 number of complete BUSCOs (32 genes) can be observed in the TcasONT assembly compared to Tcas5.2.
200 In addition to BUSCO analysis, we "lifted" genes from the Tcas5.2 annotation onto the TcasONT assembly
201 using the Liftoff package and assessed their final numbers (Suppl Table 9). Only 48 out of 14467 genes
202 (0.3%) were left unmapped, mostly representing genes with no known biological functions.

203

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

204

205

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

209

210 **2. Repeat identification in the TcasONT assembly**

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

218

219

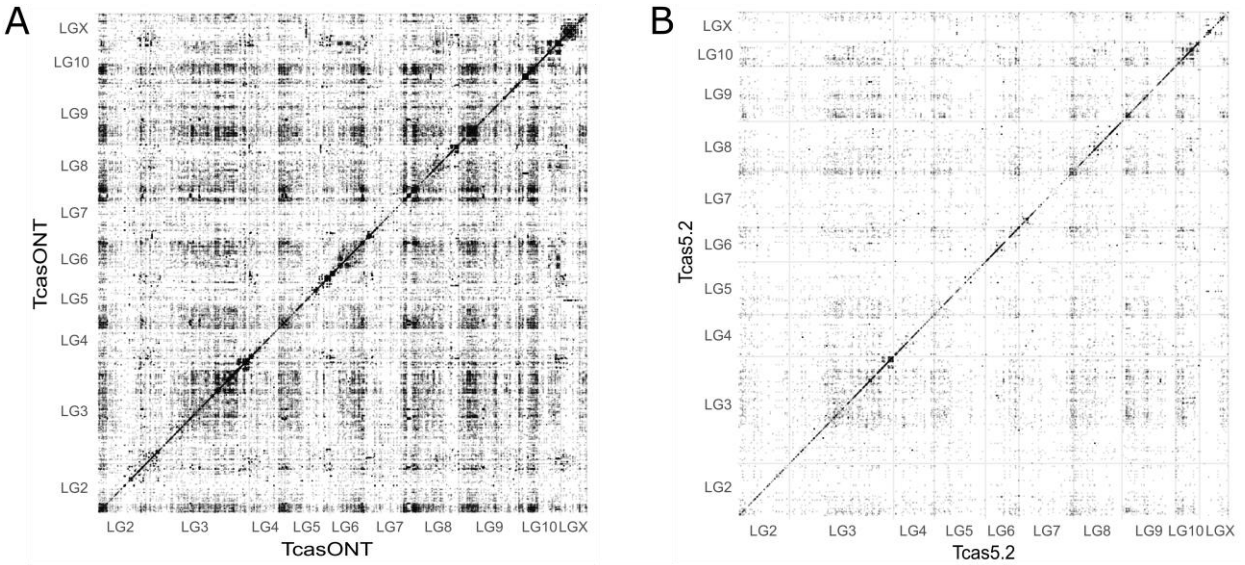


Figure 2. The whole genome-to-genome dot-plots analyses of A) TcasONT, and B) Tcas5.2 assemblies. Each dot represents a region of ≥1000 bp long region which is mapped in its entirety to another part of the genome. The dot density is proportional to the number of highly similar regions.

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* [23]. 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 TcasONT 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).

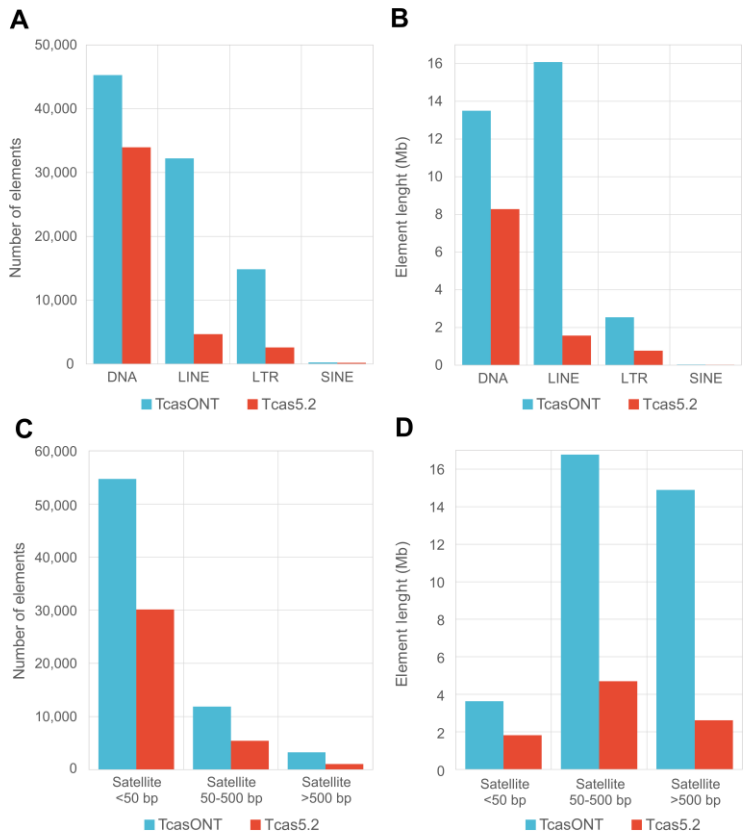


Figure 3. Comparison of main classes of transposable elements (TEs) found by RepeatMasker and tandem repeats (TRs) found by TRF in the TcasONT and Tcas5.2 assemblies. The height of the bars indicates **A)** total number of TEs, **B)** cumulative length of TEs**, C)** total number of TRs, and **D)** cumulative 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 previously disclosed abundant satDNA sequences outside the (peri)centromeres, Cast1-Cast9 (Pavlek et al, 2015).

259 The first step, essential for downstream analyses, was the mapping of the Cast1-Cast9 arrays in the
260 TcasONT assembly. Because of the well-known variability of monomer sequence within a satDNA family,
261 it was important to establish sequence similarity and sequence coverage parameters, which would ensure
262 the detection of the vast majority of Cast1-Cast9 monomers in the TcasONT assembly. For that reason,
263 the BLAST search of nanopore raw reads using consensus sequences of Cast1-Cast9 monomers [23] were
264 performed. The results were shown as density plots where color intensity correlated with number of
265 monomers of a particular Cast satDNA (Suppl Figure 2). The majority of Cast satDNAs show highest
266 intensity, i.e. monomer aggregation in the area corresponding to >75% of sequence coverage and >70%
267 of percent identity. However, in Cast2 and Cast4, two areas with high density of which one area is
268 characterized by lower coverage/identity parameters were detected. This pattern is a consequence of the
269 mutual similarity in one part of consensus sequences between Cast2 and Cast4 (Suppl Figure 3). Despite
270 their partial similarity they were still considered as individual satDNAs. In addition, the only satDNA that
271 in BLAST analysis shows increased monomer variability is Cast7, evident by the increased density in the
272 lower sequence identity area. Therefore, to map the vast majority of arrays from all 9 satDNAs (Cast1-
273 Cast9) in the TcasONT assembly, the slightly lower parameters of sequence coverage >70% and percent
274 identity >70% were used.
275 To disclose the organization of the Cast1-Cast9 arrays, we analyzed the relationship between monomer
276 distance in the arrays for each Cast (Suppl Figure 4.). More specifically, we wanted to determine whether
277 the arrays were organized in tandem continuously (consisting exclusively of monomers of a particular Cast
278 satDNA) or had mixed tandem organization (with a different sequence incorporated into the arrays). The
279 results showed that, arrays of most satDNAs show continuity of the curve, which depicts a typical satDNA
280 array organization consisting of continuously arranged monomeric variants. In contrast, Cast2, Cast5 and
281 Cast7 exhibited a disturbed tandem contiguity demonstrated by sharp increases in average array length
282 at certain distances. Most prominently, Cast2 showed a sharp increase in probability, indicating the
283 presence of a specific sequence within the arrays that required further investigation. In addition to
284 monomeric Cast2 arrays, we found Cast2 monomers also organized as a part of a new, longer repeat unit
285 of approximately 1270 bp in length (Suppl Fig. 5A and B). This new family of tandem repeats is termed
286 Cast2', and in the downstream analysis these two types of Cast2 arrays were examined separately. A more
287 detailed introspection of Cast5 arrays showed that the observed profile is the result of sporadic
288 occurrence of previously described R66-like sequences [23] dispersed within continuous Cast5 arrays
289 (Suppl Fig. 5C). In addition to R66-like sequence which turned out to be randomly distributed within Cast5
290 arrays, R140-like sequences were found exclusively at the end of some Cast5 arrays. In addition, analyses

291 of Cast7 arrays also reveal a mixed organization where Cast7 arrays often occur in association with
292 (peri)centromeric TCAST satDNA (Suppl Fig. 5D).
293 Considering the difficulty of correctly assembling satDNA sequences, the next question was whether the
294 array lengths and arrangement of the Cast1-Cast9 satDNAs in the TcasONT assembly correspond to the
295 real situation of the genomic loci containing the Cast1-Cast9 repeats. This is because satDNAs can lead to
296 assembly collapse, where the number of monomer units (array length) found in a genome assembly could
297 be less than the actual number. Based on the assumption that raw reads provide a realistic picture of what
298 is originally present in the genome because they do not undergo an assembly process, we performed a
299 comparative analysis of Cast1-Cast9 including new Cast2' array pattern across the individual reads and
300 across the new genome assembly. For this analysis, it was crucial to define optimal parameters for array
301 detection to avoid fragmentation due to short gaps. Therefore, considering monomer length and mixed
302 array organization of specific Cast satDNAs such as Cast2, Cast5 and Cast7, the best window length that
303 ensures detection of the maximum number of arrays was evaluated for each Cast satDNA. The analysis
304 showed that a minimum of three consecutive monomer units should be a criterion for correct array
305 characterization for each Cast satDNA. The array patterns between the datasets of individual reads and
306 the TcasONT assembly show remarkable similarity for all Cast satDNAs (Suppl Figure 6).
307 To detect enhancement of Cast1-Cast9 regions in TcasONT genome assembly, we compared both their
308 content and array length with Tcas5.2 (Table 2). In terms of overall enrichment in Cast satDNA portion,
309 the TcasONT assembly shows an increase of 7.2 Mb in the representation of analyzed Cast satDNAs
310 compared with Tcas5.2. Among them, Cast1, Cast2' and Cast5 emerged as particularly prominent satDNAs
311 with an increase of 0.7 Mb, 2.8 Mb and 2.6 Mb, respectively. The total content of Cast satDNAs in the new
312 TcasONT genome was 8.8 Mb (4.6% of the genome), which is remarkably close to the previous
313 experimental estimate of the genome abundance for these satDNAs (> 4.3% in [23]). Among them, Cast2'
314 and Cast5 showed the highest genome abundance with 1.5% for each. Except for Cast8 and Cast9 satDNAs
315 with low abundance (<0.1%), most Cast satDNAs showed moderate genome abundance in the range of
316 0.1-0.4%. The results also show an exceptional increase in the number of arrays in the TcasONT assembly
317 for Cast2', Cast5 and Cast7. In Cast2', for example, the number of arrays has increased by more than 9X.
318 In addition to the number of arrays, average increase in array length (expressed as mean values) can also
319 be observed in the TcasONT genome assembly for all Cast satDNAs. A significant increase in average array
320 length of up to 3X was observed for Cast2', Cast5 and Cast6 satDNAs compared to Tcas5.2. The length of
321 an average array for Cast satDNAs in the TcasONT assembly ranges from 1.2 to 31 kb. The highest average
322 array size of ~31 kb has been found for Cast6, while Cast1 has the longest array of ~112 kb.

323

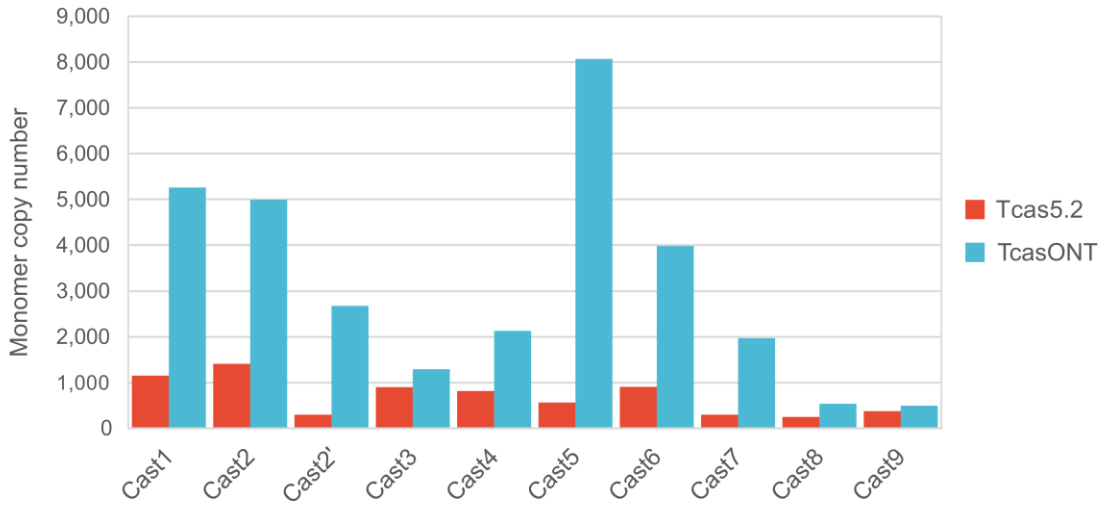
324 Table 2. Statistics of Cast1-9 arrays detected in the Tcas5.2 and TcasONT genome assemblies.

	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

325

326 The graph of Cast1-Cast9 monomer copy number in TcasONT compared to Tcas5.2 also clearly shows the
327 presence of significant enrichment (2-10X) in the majority of satDNAs analyzed (Figure 4). From the above,
328 it is clear that the new TcasONT assembly contains the representative sample of Cast arrays and can be
329 taken as a credible platform for further genome-wide analysis of Cast satDNAs.

330



331

332 Figure 4. Comparison of monomer copy numbers for Cast satDNAs annotated in the reference Tcas5.2
333 and the TcasONT genome assemblies (see also Suppl Table 11). The assemblies were mapped with Cast1-
334 Cast9 satDNA monomeric consensus sequences published previously [23].

335

336 **4. Genome organization and evolution of Cast1-Cast9 satDNAs**

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

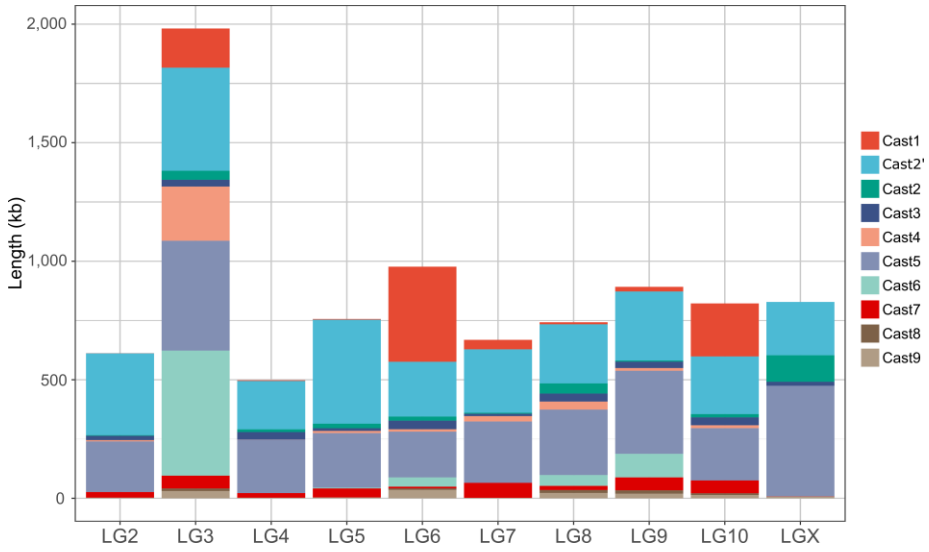


Figure 5. Distribution of Cast satDNAs across different chromosomes in the TcasONT genome assembly. Different colors represent specific Cast satDNA across ten chromosomes of *T. castaneum*. Monomers within the genome assembly were annotated based on a query percentage identity >70% and query coverage >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 [20] 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 show a tendency to cluster in distal chromosomal regions.

The distribution of Cast1-Cast9 satDNAs on chromosomes compared with genes and transposon elements (TEs) was also analyzed (Figure 6). Interestingly, Cast1-Cast9 satDNAs are often part of gene rich regions and rarely overlap with TE (the most prominent regions with these features are highlighted in Figure 6). The only coincidence of the occurrence of Cast satDNA and TEs was found in the (peri)centromeric regions, which correlates with the occurrence of Cast7 satDNA.

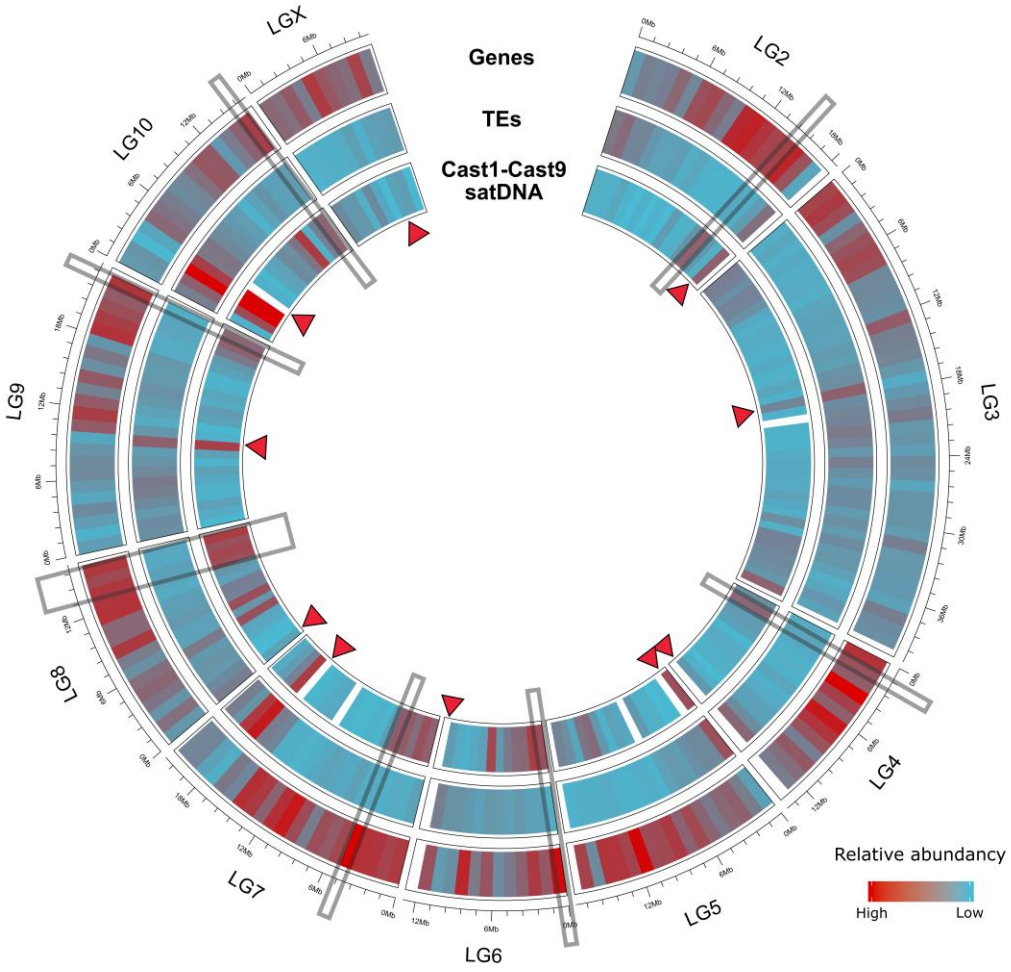
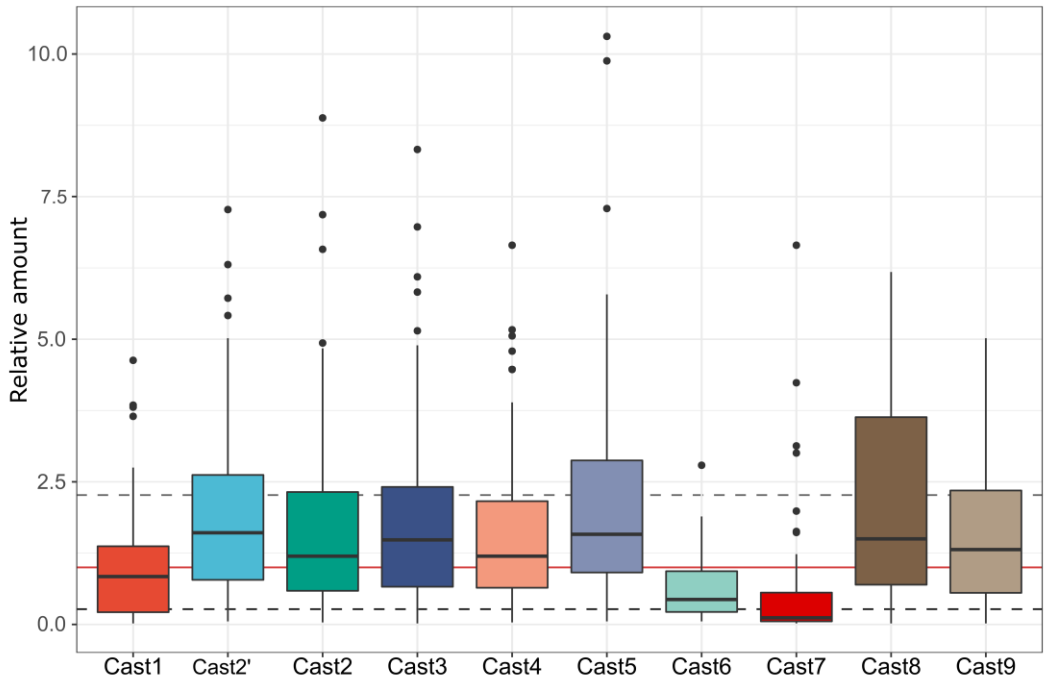


Figure 6. Circular representation of the TcasONT genome assembly with annotated Cast1-Cast9 satDNAs, genes and transposable elements (TEs). The boxes indicate the most prominent regions where the accumulation of Cast satDNAs correlates positively with gene density and negatively with the TEs density. The position of the (peri)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

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



383

384 Figure 7. Distribution of exon counts in vicinity of Cast arrays. The relative amount of exons was calculated
385 based on the distribution of exon content in the assembled genome with the red line representing the
386 median number of exons in regions of same size (100 kb) and the dotted line represents the first and third
387 quartiles (Q1 and Q3) of the distribution.
388
389 Next, we aimed to determine if there is a correlation between the arrays' length and the density of nearby
390 genes. For this purpose, Cast1-Cast9 arrays were divided into three classes according to their length:
391 (small arrays (<1000kb), moderate arrays (1000-10 000 kb) and long arrays (>10 000kb) and gene density
392 was calculated. The results for each Cast satDNAs are presented in Supplementary Figure 8. Interestingly,
393 most Cast satDNAs have the same profile with most gene occurring associated with arrays of intermediate
394 length (1000-10 000 bp). In addition, a significant number of genes were observed in the Cast5 arrays,
395 especially near large arrays (>10 kb). Furthermore, the raw exon numbers in the 50kb windows show no
396 discernible pattern, suggesting that the arrays are immersed in regions that exhibit uniform gene
397 presentation.
398 To better understand the evolution of Cast satDNAs, a genome-wide database of Cast monomers was
399 created and annotated with respect to their actual chromosomal positions. In addition, principal

400 component analysis (PCA) of monomer alignments was performed separately for each Cast satDNA
401 (Figure 6). The PCA analyses revealed a similar trend for almost all Cast satDNAs, which was characterized
402 by a scattered pattern of monomers with rare chromosome-specific clusters. These results argue for a
403 significant level of recent array exchange of Cast satDNAs between different chromosomes. This is
404 particularly evident for Cast1, Cast2, Cast2', Cast3, and Cast4 satDNAs, where regions of extremely high
405 density of monomers from different chromosomes are visible. A notable clustering of monomers from the
406 same chromosome can be observed only in Cast6. Additional analysis showed that the clustered
407 monomers mostly originate either from the same arrays or from arrays located nearby on the
408 chromosome. From the above, it is evident that the propagation of Cast satDNAs is dynamic and extensive
409 process both on the inter and intra chromosomal level.
410
411
412

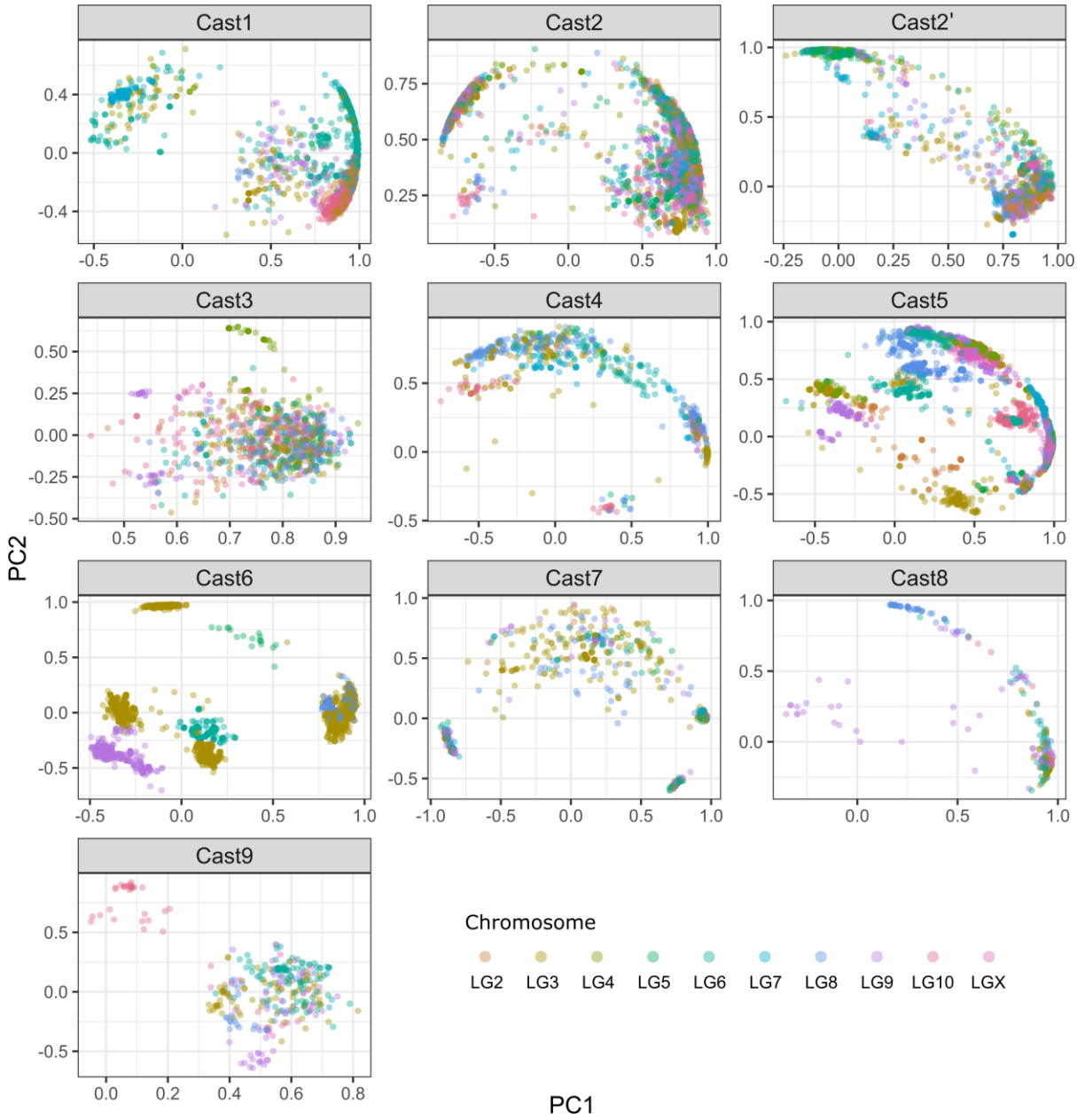
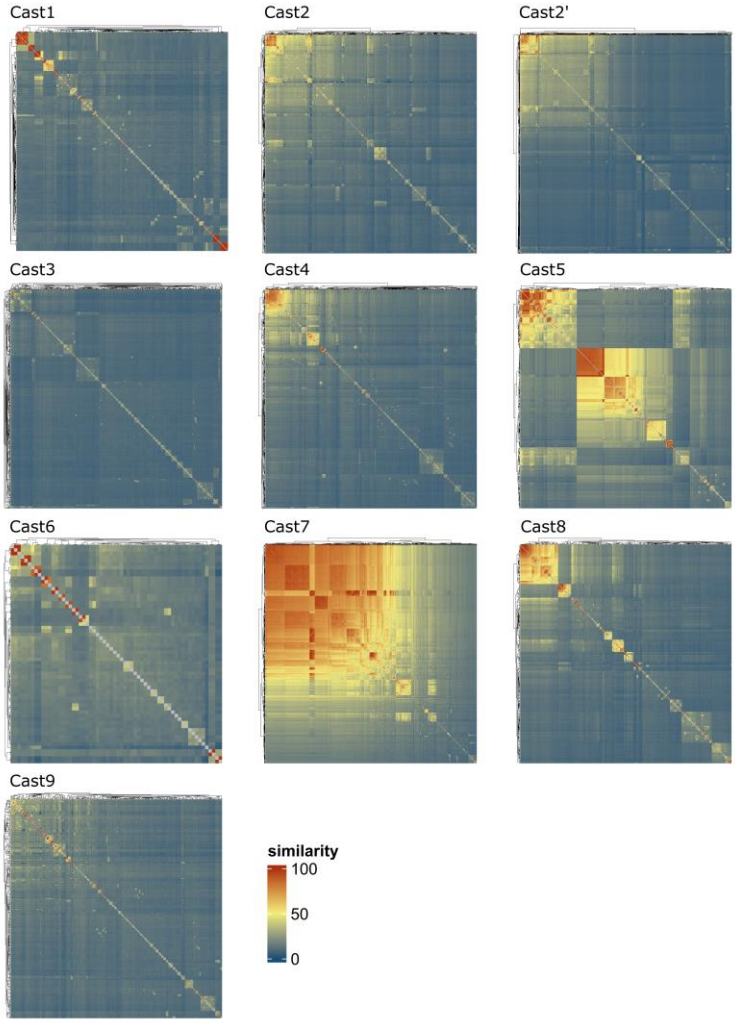


Figure 8. Principal component analysis (PCA) of aligned Cast1–Cast9 satDNAs monomer distance matrices. Dots represent a monomer unit. Monomers were colored according to the chromosome of origin. The most prominent and recent exchange events between different chromosome is visible in regions with high density of points with different colors.

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

423 alignment. The results show that of the ten Cast satDNAs analyzed, only Cast5 and Cast7 show significant
424 similarity of surrounding regions in most arrays (Figure 9). Detailed analysis of Cast5 arrays surrounding
425 regions showed that previously detected R66-like (1 kb in length) were incorporated as dispersed
426 sequence in the Cast5 arrays and R140-like sequences were found at both ends of Cast5 arrays (Suppl Fig.
427 5C). In addition to Cast5, most Cast7 arrays were arranged with (peri)centromeric TCAST (Suppl Fig 5D.
428 The rest of Cast satDNAs showed only partial similarities of the surrounding regions characterized for
429 some subsets of arrays. For example, detailed analyzes of the subset of Cast1 arrays with similarity of
430 surrounding regions showed that all these arrays originate from the same chromosome (LG7), and had
431 the same transposon-like sequence on one side (Suppl Fig 9). To identify putative sequence motifs
432 enriched near junctions of Cast arrays we generated the MEME dataset for each Cast satDNAs by
433 extracting 500 bases on each side of all arrays. In addition to the similarities already identified above in
434 the surrounding regions (e.g. Cast5, part of Cast1), we could not identify any motifs enriched near the
435 junction regions that apply to the majority of the arrays in the specific Cast satDNA.
436

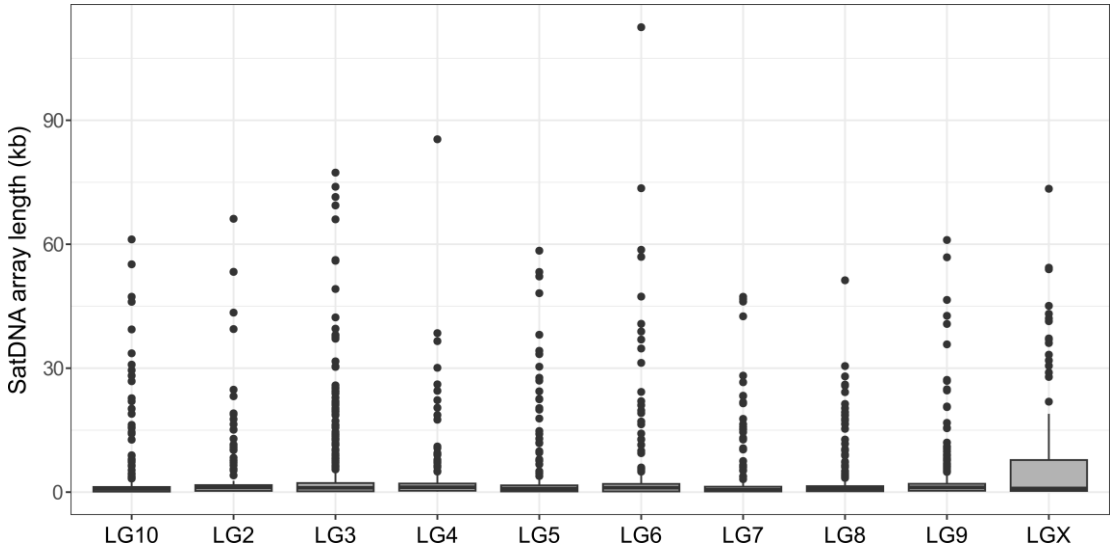


437
438 Figure 9. Heatmap visualization of sequence similarity of 2 kb regions around Cast1-Cast9 arrays
439 throughout the TcasONT assembly. The legend indicates that the sequence similarity of the surrounding
440 regions increases, with colors going from blue to red.

441

442 In general, chromosomal regions with suppressed or without recombination tend to accumulate satDNAs
443 [30]. Because sex chromosomes represent such regions, we examined the number and length of Cast1-
444 Cast9 arrays on the X chromosome and compared them with other chromosomes. This type of analysis
445 would be even more informative using the Y chromosome, which is mostly non-recombining but
446 unfortunately, the Y chromosome was not available because neither the former Tcas5.2 nor the new
447 TcasONT assembly contain the Y chromosome. Mapping of Cast1-Cast9 arrays to chromosomes disclosed
448 that the X chromosome did not have a significantly higher average number of Cast1-Cast9 arrays per Mb
449 compared to the others (Supp table 13). However, in contrast to number of arrays per chromosome, the

450 analysis of Cast1-Cast9 array length showed a statistically significant longer length of arrays on the X
451 chromosome compared to the others (Wilcoxon test, pval<0.05) (Figure 10.)



452

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

455

456 ## Discussion

457 In our study, a high-quality *T. castaneum* genome assembly was generated at the chromosome level by
458 combining nanopore long-read sequencing and reference-guided approach. The reference-guided
459 approach has recently been successfully used to generate genome assemblies in many species (reviewed
460 in [31]). Therefore, in our attempt to improve the *T. castaneum* assembly, especially in the repetitive
461 genome part, we used the latest version of the *T. castaneum* assembly Tcas5.2 [24] as a reference. The
462 new TcasONT assembly generated by this approach was found to be nearly 53 Mb larger than the Tcas5.2
463 reference genome with an extension of chromosome lengths in the range of 10.6-56.3%. In addition, the
464 TcasONT assembly has a higher completeness of genes than the last version of the reference Tcas5.2,
465 which underlines the credibility of the gene representation. It is also important to note that the TcasONT
466 genome assembly lacks only 13 Mb of the estimated *T. castaneum* genome sequence of 204 Mb,
467 previously determined experimentally [27] and also *in silico* in this study. We assume that the missing 13
468 Mb could be attributed primarily to (peri)centromeric regions, as the gene-poor and highly repetitive
469 (peri)centromeric satDNA TCAST regions [21]. These highly repetitive regions proved to be problematic
470 for correct assembly and thus were largely excluded from the TcasONT assembly process. This is not
471 unusual that despite significant advances in sequencing technologies and assembly methods,
472 (peri)centromeric regions remains still incomplete in most even nanopore based assemblies [32]. The only
473 exceptions are the recently published Telomere-to-Telomere assemblies of prominent model organisms
474 such as human and *Arabidopsis* that contain the complete (peri)centromere (Nurk et al., 2022; Altemose
475 et al., 2022; Naish et al., 2021).

476 In addition to enrichment in the gene regions, TcasONT genome proved to be representative for the study
477 of the repetitive regions. In that line, comparative studies of the assemblies show a 20-fold enrichment in
478 the repetitive part in the TcasONT assembly compared to Tcas5.2. Transposons have increased in both
479 number of elements and abundance, which is particularly evident in the case of LINE elements.
480 Considering all transposon classes, an increase of as much as 21.5 Mb was observed in the TcasONT
481 genome. When all classes of tandem repeats (TRs) were considered the TcasONT genome assembly
482 enlargement was up to 26.3 Mb. The increase is particularly significant for TRs with a monomer length of
483 >50bp, which are usually declared as “classical” satDNA. The enrichment of the proportion of transposons
484 and TRs totaling 47.8 Mb compared to Tcas5.2 indicates that the TcasONT genome provides an excellent
485 platform for the analysis of different repetitive fractions of the genome. Our study shows that nanopore
486 sequencing approach enable resolution of previously assembly-impeding regions, most of which consist
487 of highly repetitive sequences such as satDNAs and transposable elements. It is also important to note

488 that TcasONT is larger than the previous version by 25%, which means that a large portion of the genome
489 is now available for further research.
490 In this work, we used the TcasONT genome assembly provided here as a source for a detailed analysis of
491 the most dominant satDNAs outside the (peri)centromere, Cast1-Cas9, previously characterized in [23].
492 Given that our previous study of these satDNAs was limited because it was based on a portion of Cast1-
493 Cast9 which was found in the Tcas3.0 reference genome, representing less than 1/10 of the actual Cast1-
494 Cast9 genome fraction. The availability of the TcasONT genome assembly provided the opportunity for a
495 comprehensive analysis of Cast1-Cast9 satDNAs. The assembly process often does not correctly assemble
496 the satDNAs leading to collapsed arrays (Tørresen et al., 2019). However, our comparative analyses of
497 Cast1-Cast9 satDNAs between raw data and TcasONT assembly show credibility of the TcasONT assembly
498 in terms of representativeness of these sequences. The TcasONT assembly also showed abundance of Cast
499 satDNAs of 8.8 Mb, which accounts for 4.6% of the genome and it is consistent with their experimentally
500 confirmed genome abundance [23]. These two features, have provided the basis for a deep and credible
501 platform for genome-wide analyses of Cast1-Cast9 satDNAs. Interestingly we found that these ten most
502 abundant and "classical" satDNAs (Cast1-Cast9, including a new version of Cast2, Cast2') occur in the form
503 of long tandem arrays. Considering that they are located outside the (peri)centromere where
504 accumulation of satDNAs is not common, they show a very high average and maximum) length of arrays
505 (Cast6 with 31kb and Cast1 with 112kb respectively). The question arises whether the presence of such
506 long satDNA arrays in regions considered not suitable for the accumulation of satDNAs is a distinctive
507 feature of the *T. castaneum* genome, or whether the absence of satDNA arrays in these regions in other
508 analyzed genomes is a consequence of significant gaps in the assemblies. Given that the previous high-
509 quality genome assembly Tcas5.2, based on a library BAC and improved by optical mapping, contained
510 only 1/10 of experimentally estimated amount of Cast satDNAs, and that their content in the TcasONT
511 assembly almost completely approached their real amount, we conclude that the strategy of nanopore
512 sequencing combined with reference-guided assembly approach ensures the enviable representation of
513 satDNAs.
514 Furthermore, one of the important questions that arose from the observed organization characterized by
515 long arrays localized in non-pericentromeric regions that were assumed to be euchromatic is whether
516 Cast1-Cast9 satDNAs are truly euchromatic satDNAs. Although it has been previously shown
517 experimentally that these satDNAs are mostly located outside the (peri)centromeric regions, we believe
518 that this fact alone cannot classify them as euchromatic. Indeed, it is possible that satDNAs, even if located
519 outside the (peri)centromeric region, could be localized in gene-poor regions populated by, for example,

520 transposons or other types of noncoding DNA [34]. For this reason, we performed the analysis of
521 surrounding regions of Cast satDNA arrays in terms of gene density and compared them with the average
522 gene density in the whole genome. Interestingly, the results showed that the surrounding regions of
523 almost all Cast satDNAs correlated positively with the gene regions. In addition, some satDNA arrays were
524 surrounded by extremely gene-rich regions. It is also important to note that the surrounding gene-rich
525 regions are not characteristic for short arrays as would be expected but the gene-rich regions are also
526 found surrounding the >10kb arrays. Because it is known that satDNAs can inhabit regions rich in
527 transposons, another dominant repetitive DNA, we also analyzed the distribution of Cast1-Cast9 satDNAs
528 compared to transposons. Contrary to expectations, we found that the occurrence of Cast1-Cast9 satDNAs
529 and TEs often did not overlap. To summarize, Cast1-Cast9 satDNAs including their long arrays, were mostly
530 integrated into an environment rich in genes and poor in transposons. This fact clearly speaks to the ability
531 of Cast satDNA to invade euchromatic regions and also indicates the flexibility of euchromatin, i.e., gene-
532 rich regions, to tolerate and accommodate these sequences. Previous models support the fact that the
533 accumulation of satDNA is characteristic of regions where recombination is repressed, such as
534 (peri)centromeric heterochromatin (Charlesworth et al., 2005). This is thought to be due to the potentially
535 deleterious effect of satDNAs, which can affect neighboring regions because of its high turnover. However,
536 from our study, it appears that euchromatin where recombination is expected to occur may have some
537 tolerance to the accumulation of satDNA even in the form of long arrays. In order to test how potentially
538 suppressed recombination influences the dynamics of Cast1-Cast9 propagation, we performed a
539 comparative analysis of arrays between autosomal chromosomes and X chromosome. In that line, our
540 analyses of Cast1-Cast9 showed that the X chromosome had a similar number of loci with Cast1-Cast9
541 satDNAs in comparison with autosomes, but arrays were longer in the X chromosome than in the
542 autosomes. These results suggest that suppressed recombination on X chromosomes has less effect on
543 the spread of satDNAs in euchromatin, but rather tolerate propagation of satDNAs into longer arrays.
544 Considering that previous studies have mostly reported short arrays in euchromatin regions, our finding
545 that there are "classical" satDNAs widely distributed across the genome, even in the form of long arrays
546 in gene-rich regions, is something completely new. Previous studies based on euchromatic satDNA in a
547 form of short arrays suggest that these satDNAs may play roles in gene regulation by acting as
548 “evolutionary tuning knobs” [35] regulating chromatin [9,12], and facilitating X chromosome
549 recognition/dosage compensation [11]. SatDNAs located in euchromatic regions also could regulate gene
550 expression by modulating local chromatin structure, or via transcripts derived from the repeats. For
551 example, contractions of the human subtelomeric satellite D4Z4 alter the chromatin state of nearby genes

552 causing muscular dystrophy [36]. Recent studies disclosed that expression of euchromatic satDNA-derived
553 transcripts control embryonal development in the mosquito via sequence-specific gene silencing [37].
554 Being located in the gene-rich regions, epigenetic histone marks of satDNAs in euchromatin could also
555 have an effect on surrounding gene regions. For example, human genome-wide analysis of satDNAs
556 located in euchromatin, has shown that they are associated with repressive histone mark H3K9me3,
557 suggesting that they may affect the expression of neighboring genes as repressive H3K9me3 encompasses
558 surrounding regions in addition to satDNA arrays [38]. In addition to the direct impact on genes, large-
559 scale rearrangements that involve the observed long Cast1-Cast9 arrays scattered throughout the genome
560 are also highly probable. The ability of euchromatic satDNA to undergo high evolutionary turnover may
561 also contribute to the rapid change of gene landscape across the genome and consequently on gene
562 function [39]. In summary, the presence of such a large portion of Cast1-Cast9 satDNAs in euchromatin
563 regions of the *T. castaneum* genome can strongly influence both gene expression and the dynamics of
564 genome evolution.
565 Another important question that arose in view of the extensive distribution of Cast1-Cast9 satDNAs
566 through the *T. castaneum* genome is how these satDNAs are propagated. While the mechanisms of TE
567 propagation are relatively well known and depend on the type of transposon elements [40], the
568 propagation of satDNA especially in euchromatin regions is still quite a mystery. Given that we know little
569 about the evolutionary dynamics and molecular mechanisms that shape the distribution of euchromatic
570 satDNA on the genome scale, we carried out analysis of the distribution pattern, evolutionary trends as
571 well as analysis of surrounding and junction regions of euchromatic Cast1-Cast9 satDNAs. The results
572 mainly show a scattered pattern characterized by arrays dispersed on all chromosomes and along their
573 entire chromosome length without regional preference. Interestingly, most Cast satDNAs showed
574 relatively recent and extensive propagation between different chromosomes. Observed pattern of
575 chromosomal distribution of Cast satDNAs indicates efficient mechanisms responsible for their
576 propagation. Although satDNAs were considered less mobile in contrast to TEs, which are able to move
577 extensively throughout the genome ultimately forming interspersed repeats, our study showed that
578 satDNAs have the same potential to spread extensively throughout the genome. To determine whether
579 some of them are passively spreading as part of some other sequence, analyzes of neighbor regions were
580 made. Apart from the already known associations of Cast7 with TCAST major satDNAs and Cast5 with
581 transposon-like sequences, other Cast satDNAs showed association with a common sequence only in a
582 small fraction of the arrays. To date, there are three mechanisms that have been proposed to be involved
583 in the propagation of satDNA through the euchromatin (i) dispersion in a form of short arrays integrated

584 as central repeats of non-autonomous transposons elements [41]; (ii) the spread of satDNAs across long
585 physical distances in euchromatin through extrachromosomal circular DNAs (eccDNAs) [42]; (iii) interlocus
586 gene conversion events via 3-D interactions between loci in interphase nucleus [43]. There are quite a few
587 examples where dispersed satDNAs have been found in the form of short arrays integrated as central
588 repeats of transposons elements [41,44,45]. Moreover, tandem repeats within Tetris transposon DNA
589 have been reported to be the basis for the generation of long satellite arrays in *Drosophila virilis* that
590 eventually lose transposon features in adjacent regions [41]. Considering that most of the Cast5 arrays
591 have common adjacent regions, one of which shows low similarities with the Mariner element, it is to be
592 assumed that this Cast5 satDNA owes its expansion to this element, similar to the origin of satDNA in
593 *Drosophila* through mediation by Tetris element. However, Cast satDNAs did not show a tendency to
594 associate with other mobile elements but show very extensive propagation indicating an efficient self-
595 propagation mechanism that operates at the inter- and intra-chromosome level. Mechanisms that have
596 recently been highlighted as possible drivers in the dispersal of satDNAs in *Drosophila* euchromatin are
597 eccDNA reintegration and interlocus gene conversion [39]. However, a detailed analysis of the junction
598 regions of Cast satDNAs arrays did not indicate which, if any, of the two remaining proposed mechanisms
599 could be involved in this extensive propagation of Cast satDNAs in the *T. castaneum* genome.
600 In conclusion, this work aims to (1) provide a high-quality reference genome of the insect model organism
601 *T. castaneum*, TcasONT, (2) take advantage of the new *T. castaneum* assembly to perform a deep genome-
602 wide analysis of dominant euchromatic satDNAs (3) and provide a highly reliable data point for future
603 comparative analyses of repetitive genome portion among related species. First, we have produced the
604 most contiguous genome assembly of *T. castaneum* to date with significant improvement in the
605 representation of the repetitive genome portion by Oxford Nanopore long-read sequencing. Our study
606 demonstrates that Nanopore sequencing approach has the potential to overcome the limitations of short
607 reads and enable resolution of previously assembly-impeding regions, most of which consist of highly
608 repetitive sequences such as satDNAs and transposable elements. Our reference genome sequences have
609 great potential to be important genomic resources for future research of both genes and various
610 repetitive fractions of the genome. Second, our comprehensive analyses of the abundant satDNAs
611 embedded in euchromatin regions of the *T. castaneum* genome show their massive proliferation of a wide
612 range, implying efficient mechanisms of their spread and amplification in gene-rich regions. Contrary to
613 the previously generally accepted view that the coexistence of gene-rich regions and abundant satDNA is
614 precluded, our study raises the question to which extent this hypothesis is actually based on assemblies

615 that generally show a lack of satDNA sequences not only in highly repetitive (peri)centromeric regions,
616 but also in satDNAs embedded in gene-rich regions.
617
618
619

620 ## Materials and methods

621 Insects sample and DNA preparation

622 Laboratory cultures of the red flour beetle *T. castaneum*, highly cultured strain Georgia 2 (GA2), were
623 routinely reared in whole wheat flour with the addition of whole rye flour and oats under conditions
624 favoring faster multiplication (32 °C and 70% relative humidity) in the dark. Selection of the different life
625 stages of the insects was performed by sieving through a 0.71-mm sieve and picking the beetles
626 individually with tweezers. Larvae and pupae were collected to achieve the most efficient DNA isolation
627 efficiency. The collected amount of beetles had to be sufficient to perform the isolation with 200 mg of
628 the starting material for pupae and 500 mg for larvae as starting sample for each DNA isolation. DNA
629 isolation was performed followed the protocol described in [46] .Briefly, in order to isolate DNA of
630 sufficient quality and quantity, nuclei were first isolated using a nuclei isolation buffer (NIB), followed by
631 lysis , and finally transferred to a Qiagen Genomic Tip 100/G column. Pure DNA was spooled from eluates
632 and precipitated in isopropanol. In total, we have performed 8 of described isolations, 4 for larvae and 4
633 for pupae.

634 Library preparation and sequencing

635 Oxford Nanopore library was prepared using SQK-LSK110 library preparation kit. Preparation was
636 performed according to the manufacturer's instructions with modifications described below. Prior to
637 sequencing DNA was sheared 30X times using a 30-gauge needle and purified of short fragments using
638 the Short Read Eliminator (SRE) XS kit (Circulomics) according to the manufacturers protocol. The sheared
639 and purified DNA was then used in library preparation protocols in amounts 2X larger than the ONT
640 protocol recommendation. Additionally, all recommended elution and incubation times were extended
641 to twice the length. These steps were performed to avoid library loss and to increase final library
642 concentration sufficient to load multiple MinION flow cell (final concentrations of the library were >100
643 ng/µL allowing multiple loads from a single library). The flow cells were washed and reloaded 2-5 times
644 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
645 a total cumulative output of 89.9 GB and an N50 of 20.1 kb. The software used for sequencing was the
646 Oxford Nanopore MinKnow 20.10.3.

647

648

649 Genome assembly

650 Data obtained from all sequencing runs was basecalled using Oxford Nanopore basecalling software
651 Guppy 5.0.1 with parameters as listed in the Supplementary table 14. The basecalled reads were then
652 used in the assembly process, using Canu v2.2 [26] with the parameters listed in the Supplementary table
653 14. According to Canu documentation the parameters were adjusted for high repetitiveness of the
654 genome [22] and the high observed AT content in the reference Tas5.2 genome assembly [24]. The reads
655 were filtered to >20 kb in order to reduce the computational intensity needed to perform the assembly,
656 given the amount of data and the relatively small genome of the beetle. Canu assembly was performed
657 using the resources of computer cluster Isabella based in University Computing Centre (SRCE), University
658 of Zagreb.

659 Filling the gaps

660 To successfully arrange the Canu contigs into chromosomes based on the Tcas5.2 (GCF_000002335.3)
661 assembly, pre-existing gaps in the Tcas5.2 assembly needed to be bridged. This was done using the TGS-
662 Gapcloser software [47] with default for gap-filling settings (Supplementary table 14) using corrected
663 reads generated as the output of the Canu pipeline. Gap filling was used to bridge small and medium gaps
664 in the Tcas5.2 assembly to prevent that Nanopore-based long contigs be interrupted due to discontinuity
665 of the previous assembly.

666 Contig orientation

667 After filling the gaps in the Tcas5.2 assembly chromosomes, the RagTag collection of software tools [29]
668 was used. The contigs obtained with Canu were used as the query sequence and the gap-filled Tcas5.2
669 assembly was used as the reference sequence with "scaffold" parameter (Supplementary table 14).
670 RagTag uses the reference sequence (gap filled Tcas5.2 assembly), as a guide to which it aligns and places
671 the input sequences (Canu contigs) to create the new assembly and fill the gaps in the subject assembly
672 with new successfully aligned contigs. Since the highly repetitive regions are in majority lacking from the
673 Tcas5.2 assembly, this approach allowed to incorporate a new region of repetitive elements in the
674 TcasONT assembly. RagTag maps the high confidence genomic regions on chromosomes and then, if a
675 contig ends or begins with a repetitive region it is placed in the gap giving us insight into previously
676 unknown regions. The output of the RagTag tool is the unpolished assembly that is used as a template for
677 polishing (Supplementary Figure 1).

678

679 Polishing

680 To improve the quality of the assembly and reduce the error ratio, correction of Canu contigs was
681 performed. To polish the assembly, two separate rounds of RACON [48] polishing were performed using
682 reads excluded from the initial assembly (<20 kb). These short reads used for polishing involved significant
683 addition of genomic region information as this excluded data was ~50 Gb. Polishing was performed
684 according to the instructions within the RACON documentation. The RACON workflow consists of mapping
685 the reads onto the assembled genome with minimap2 [49] and then using the information about the
686 mapped reference reads to perform polishing. Produced assembly, named TcasONT was used for
687 downstream analysis.

688 BUSCO analysis

689 Benchmarking Single Copy orthologs (BUSCO) [50] analysis were performed using the BUSCO v5.0.0.
690 module on the Galaxy web platform (usegalaxy.org). Same settings, as listed in the Supplementary table
691 14 were repeated for all assembly validations.

692 Repeat detection

693 RepeatMasker is a widely used tool for finding and masking different repeat elements within a given target
694 sequence [51]. RepeatMasker was used to get the GFF/GTF formatted data with the position and
695 orientation of classified RepBase repeat elements. This was a base for the information about the quantity,
696 size, and distribution of different elements within examined genome assemblies. Assemblies were
697 annotated with different elements using the RepeatMasker program on the Galaxy web platform
698 (usegalaxy.org) with different repeat elements from the latest RepBase database (RELEASE 20181026) of
699 repeats and the "Hexapoda" species listing for clade specific repeats. We also rerun RepeatMasker for the
700 Tcas5.2 assembly to update the repeat annotations, as it was created using previous version of RepBase
701 database.

702 For quantifying of three satDNAs classes (with defined monomer length >50; 50-500; >500bp) in
703 TcasONT and Tcas5.2 assembly, Tandem Repeat Finder (TRF) program [52] was used with default
704 parameters.

705 Gene annotation

706 In order to map genes in Canu contigs (filtering) and in TcasONT assembly, we used the novel LiftOff
707 package [53] using gene annotations from the Tcas5.2 assembly. The Liftoff first maps the newly created

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

713 Identification of Cast1-Cast9 satDNAs

714 Annotation of satellite repeats within the genomes was performed using the standalone NCBI's BLAST
715 algorithm, and the interface to the R programming language package metablastr [54]. The subject
716 sequences were the different analyzed assemblies (Tcas5.2 and TcasONT) while the queries were the
717 previously characterized Cast1-Cast9 [23]. Post-analysis included filtering BLAST hits in the genome to
718 discover the Cast1-Cast9 trends and arrays.

719 Analysis of Cast1-Cast9 satDNA arrays

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

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

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

733 Analysis of gene content in vicinity of Cast1-Cast9 arrays

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

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

740 Analysis of Cast1-Cast9 junction regions (MEME)

741 In order to find particular motifs surrounding Cast1-9 arrays we used source code compiled version of
742 MEME-5.5.1. [55]. We performed MEME motif identification in 100 bp, 200 bp and 500 bp windows
743 around particular arrays of interest, resulting motifs were investigated and in both quantity and
744 localization.

745 Multiple sequence alignment and tree construction

746 We used MAFFT [56] to create multiple sequence alignments of Cast1-9 monomers in the assembly. After
747 alignment we used the "F81" genetic distance evolutionary model from the package "ape"[57] in R to
748 generate genetic distance matrices which were subsequently used in the PCA analysis. PCA was performed
749 using the FactoMineR [58] package PCA function and first 2 dimensions for each satDNA were visualized
750 using *ggplot2*.

751 Visualizations and calculations

752 All plots and calculations were produced in R using custom made data processing notebooks. Apart from
753 the standard libraries, we also used the *circilize* package [59] to create the circular visualization plots of
754 global genome patterns. To create the complex heatmaps used in the characterization of neighboring
755 regions similarity we used the *ComplexHeatmap* package [60].

756 Data avaliability

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

761

762 ## References

763

764 1. Plohl M, Meštrović N, Mravinac B. Centromere identity from the DNA point of view. *Chromosoma*.
765 2014; doi: 10.1007/s00412-014-0462-0.

766 2. Blattes R, Monod C, Susbielle G, Cuvier O, Wu JH, Hsieh TS, et al.. Displacement of D1, HP1 and
767 topoisomerase II from satellite heterochromatin by a specific polyamide. *EMBO J*. 2006; doi:
768 10.1038/sj.emboj.7601125.

769 3. Aldrup-MacDonald ME, Kuo ME, Sullivan LL, Chew K, Sullivan BA. Genomic variation within alpha
770 satellite DNA influences centromere location on human chromosomes with metastable epialleles.
771 *Genome Res*. 2016; doi: 10.1101/gr.206706.116.

772 4. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, et al.. Comparative analysis of tandem
773 repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol*. 2013;
774 doi: 10.1186/gb-2013-14-1-r10.

775 5. Bersani F, Lee E, Kharchenko P V., Xu AW, Liu M, Xega K, et al.. Pericentromeric satellite repeat
776 expansions through RNA-derived DNA intermediates in cancer. *Proc Natl Acad Sci U S A*. 2015; doi:
777 10.1073/pnas.1518008112.

778 6. Ferree PM, Prasad S. How Can Satellite DNA Divergence Cause Reproductive Isolation? Let Us Count
779 the Chromosomal Ways. *Genet Res Int*. 2012; doi: 10.1155/2012/430136.

780 7. Feliciello I, Pezer Ž, Kordiš D, Madarić BB, Ugarković D. Evolutionary history of alpha satellite DNA
781 repeats dispersed within human genome euchromatin. *Genome Biol Evol*. 2020; doi:
782 10.1093/gbe/evaa224.

783 8. Kuhn GCS, Küttler H, Moreira-Filho O, Heslop-Harrison JS. The 1.688 repetitive DNA of *Drosophila*:
784 Concerted evolution at different genomic scales and association with genes. *Mol Biol Evol*. 2012; doi:
785 10.1093/molbev/msr173.

786 9. Brajković J, Feliciello I, Bruvo-MadWarić B, Ugarković DW. Satellite DNA-like elements associated with
787 genes within euchromatin of the beetle *Tribolium castaneum*. *G3 Genes, Genomes, Genet*. 2012; doi:
788 10.1534/g3.112.003467.

789 10. Menon DU, Coarfa C, Xiao W, Gunaratne PH, Meller VH. siRNAs from an X-linked satellite repeat
790 promote X-chromosome recognition in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A*. National
791 Academy of Sciences; 2014; doi: 10.1073/pnas.1410534111.

792 11. Joshi SS, Meller VH. Satellite Repeats Identify X Chromatin for Dosage Compensation in *Drosophila*
793 *melanogaster* Males. *Curr Biol*. Elsevier Ltd.; 2017; doi: 10.1016/j.cub.2017.03.078.

794 12. Feliciello I, Akrap I, Ugarković Đ. Satellite DNA Modulates Gene Expression in the Beetle *Tribolium*
795 *castaneum* after Heat Stress. *PLoS Genet*. 2015; doi: 10.1371/journal.pgen.1005466.

796 13. Miga KH. Completing the human genome: the progress and challenge of satellite DNA assembly.
797 *Chromosom Res*. Springer Netherlands; 2015; doi: 10.1007/s10577-015-9488-2.

798 14. Lower SS, McGurk MP, Clark AG, Barbash DA. Satellite DNA evolution: old ideas, new approaches.
799 *Curr Opin Genet Dev*. Elsevier Ltd; 2018; doi: 10.1016/j.gde.2018.03.003.

15. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al.. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*. Springer US; 2020; doi: 10.1038/s41586-020-2547-7.

16. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, et al.. The complete sequence of a human genome. *Science (80-)*. American Association for the Advancement of Science; 2022; doi: 10.1126/science.abj6987.

17. Klingler M. Tribolium. *Curr Biol*. 2004; doi: 10.1016/j.cub.2004.08.004.

18. Campbell JF, Athanassiou CG, Hagstrum DW, Zhu KY. *Tribolium castaneum*: A Model Insect for Fundamental and Applied Research. *https://doi.org/101146/annurev-ento-080921-075157*. Annual Reviews; 2022; doi: 10.1146/ANNUREV-ENTO-080921-075157.

19. Richards S, Gibbs RA, Weinstock GM, Brown S, Denell R, Beeman RW, et al.. The genome of the model beetle and pest *Tribolium castaneum*. *Nature*. 2008; doi: 10.1038/nature06784.

20. Gržan T, Despot-Slade E, Meštrović N, Plohl M, Mravinac B. CenH3 distribution reveals extended centromeres in the model beetle *Tribolium castaneum*. Malik HS, editor. *PLoS Genet*. 2020; doi: 10.1371/journal.pgen.1009115.

21. Ugarković D, Podnar M, Plohl M. Satellite DNA of the red flour beetle *Tribolium castaneum* - Comparative study of satellites from the genus *Tribolium*. *Mol Biol Evol*. 1996; doi: 10.1093/oxfordjournals.molbev.a025668.

22. Wang S, Lorenzen MD, Beeman RW, Brown SJ. Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. *Genome Biol*. 2008; doi: 10.1186/gb-2008-9-3-r61.

23. Pavlek M, Gelfand Y, Plohl M, Meštrović N. 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 Res*. Oxford Academic; 2015; doi: 10.1093/dnares/dsv021.

24. Herndon N, Shelton J, Gerischer L, Ioannidis P, Ninova M, Dönitz J, et al.. Enhanced genome assembly and a new official gene set for *Tribolium castaneum*. *BMC Genomics*. BMC Genomics; 2020; doi: 10.1186/s12864-019-6394-6.

25. Brown SJ, Henry JK, Black IV WC, Denell RE. Molecular genetic manipulation of the red flour beetle: Genome organization and cloning of a ribosomal protein gene. *Insect Biochem*. 1990; doi: 10.1016/0020-1790(90)90011-I.

26. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res*. 2017; doi: 10.1101/gr.215087.116.

27. Alvarez-Fuster A, Juan C, Petitpierre E. Genome size in *Tribolium* flour-beetles: Inter-and intraspecific variation. *Genet Res*. 1991; doi: 10.1017/S0016672300029542.

28. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, et al.. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res*. Oxford University Press; 2019; doi: 10.1093/nar/gkz841.

29. Alonge M, Lebeigle L, Kirsche M, Aganezov S, Wang X, Lippman ZB, et al.. Automated assembly

839 scaffolding elevates a new tomato system for high-throughput genome editing. *bioRxiv*.
840 :2021.11.18.4691352021;

841 30. Stephan W. Nonlinear Phenomena in the Evolution of Satellite. *Berichte der Bunsengesellschaft für*
842 *Phys Chemie*. 1986; doi: 10.1002/bbpc.19860901119.

843 31. Lischer HEL, Shimizu KK. Reference-guided de novo assembly approach improves genome
844 reconstruction for related species. *BMC Bioinformatics*. 2017; doi: 10.1186/s12859-017-1911-6.

845 32. Talbert PB, Henikoff S. The genetics and epigenetics of satellite centromeres. *Genome Res*. 2022; doi:
846 10.1101/gr.275351.121.

847 33. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, et al.. The genetic and
848 epigenetic landscape of the *Arabidopsis* centromeres. *Science (80-)*. 2021; doi:
849 10.1126/science.abi7489.

850 34. Meštrović N, Mravinac B, Pavlek M, Vojvoda-Zeljko T, Šatović E, Plohl M. Structural and functional
851 liaisons between transposable elements and satellite DNAs. *Chromosom Res*. 2015; doi:
852 10.1007/s10577-015-9483-7.

853 35. King DG, Soller M, Kashi Y. Evolutionary tuning knobs. *Endeavour*. 1997; doi: 10.1016/S0160-
854 9327(97)01005-3.

855 36. Zeng W, De Greef JC, Chen YY, Chien R, Kong X, Gregson HC, et al.. Specific loss of histone H3 lysine 9
856 trimethylation and HP1γ/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral
857 dystrophy (FSHD). *PLoS Genet*. 2009; doi: 10.1371/journal.pgen.1000559.

858 37. Halbach R, Miesen P, Joosten J, Taşköprü E, Rondeel I, Pennings B, et al.. A satellite repeat-derived
859 piRNA controls embryonic development of *Aedes*. *Nature*. Springer US; 2020; doi: 10.1038/s41586-020-
860 2159-2.

861 38. Vojvoda Zeljko T, Ugarković Đ, Pezer Ž. Differential enrichment of H3K9me3 at annotated satellite
862 DNA repeats in human cell lines and during fetal development in mouse. *Epigenetics and Chromatin*.
863 BioMed Central; 2021; doi: 10.1186/s13072-021-00423-6.

864 39. Sproul JS, Khost DE, Eickbush DG, Negm S, Wei X, Wong I, et al.. Dynamic evolution of euchromatic
865 satellites on the x chromosome in *Drosophila melanogaster* and the simulans clade. *Mol Biol Evol*. 2020;
866 doi: 10.1093/MOLBEV/MSAA078.

867 40. Kelleher ES, Barbash DA, Blumenstiel JP. Taming the Turmoil Within: New Insights on the
868 Containment of Transposable Elements. *Trends Genet*. NIH Public Access; 2020; doi:
869 10.1016/J.TIG.2020.04.007.

870 41. Dias GB, Svartman M, Delprat A, Ruiz A, Kuhn GCS. Tetris is a foldback transposon that provided the
871 building blocks for an emerging satellite DNA of *Drosophila virilis*. *Genome Biol Evol*. 2014; doi:
872 10.1093/gbe/evu108.

873 42. Cohen S, Yacobi K, Segal D. Extrachromosomal circular DNA of tandemly repeated genomic
874 sequences in *Drosophila*. *Genome Res*. 2003; doi: 10.1101/gr.907603.

875 43. Lee YCG, Ogiyama Y, Martins NMC, Beliveau BJ, Acevedo D, Wu C ting, et al.. Pericentromeric
876 heterochromatin is hierarchically organized and spatially contacts H3K9me2 islands in euchromatin.
877 *PLoS Genet*. 2020; doi: 10.1371/journal.pgen.1008673.

878 44. Vojvoda Zeljko T, Pavlek M, Meštrović N, Plohl M. Satellite DNA-like repeats are dispersed
879 throughout the genome of the Pacific oyster *Crassostrea gigas* carried by Helentron non-autonomous
880 mobile elements. *Sci Rep*. Nature Publishing Group; 2020; doi: 10.1038/s41598-020-71886-y.

881 45. Macas J, Neumann P. Ogre elements - A distinct group of plant Ty3/gypsy-like retrotransposons.
882 *Gene*. 2007; doi: 10.1016/j.gene.2006.08.007.

883 46. Volarić M, Veseljak D, Mravinac B, Meštrović N, Despot-Slade E. Isolation of high molecular weight
884 DNA from the model beetle *Tribolium* for nanopore sequencing. *Genes (Basel)*. 2021; doi:
885 10.3390/genes12081114.

886 47. Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, et al.. TGS-GapCloser: A fast and accurate gap closer
887 for large genomes with low coverage of error-prone long reads. *Gigascience*. Oxford University Press;
888 2020; doi: 10.1093/gigascience/giaa094.

889 48. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long
890 uncorrected reads. *Genome Res*. 2017; doi: 10.1101/gr.214270.116.

891 49. Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018; doi:
892 10.1093/bioinformatics/bty191.

893 50. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome
894 assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015; doi:
895 10.1093/bioinformatics/btv351.

896 51. Smit, AFA, Hubley, R & Green P. RepeatMasker Open-4.0.

897 52. Benson G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res*. 1999;
898 doi: 10.1093/nar/27.2.573.

899 53. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. *Bioinformatics*. Oxford
900 Academic; 2021; doi: 10.1093/BIOINFORMATICS/BTAA1016.

901 54. Benoit M, Drost HG. A Predictive Approach to Infer the Activity and Natural Variation of
902 Retrotransposon Families in Plants. *Methods Mol Biol*. Humana, New York, NY; 2021; doi: 10.1007/978-
903 1-0716-1134-0_1.

904 55. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in
905 biopolymers. *Proc Int Conf Intell Syst Mol Biol*. 2:28–361994;

906 56. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in
907 performance and usability. *Mol Biol Evol*. 2013; doi: 10.1093/molbev/mst010.

908 57. Paradis E, Schliep K. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in
909 R. *Bioinformatics*. 2019; doi: 10.1093/bioinformatics/bty633.

910 58. Lê S, Josse J, Husson F. FactoMineR : An R Package for Multivariate Analysis. *J Stat Softw*. 2008; doi:
911 10.18637/jss.v025.i01.

912 59. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R.
913 *Bioinformatics*. Oxford Academic; 2014; doi: 10.1093/BIOINFORMATICS/BTU393.

914 60. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional
915 genomic data. *Bioinformatics*. Oxford Academic; 2016; doi: 10.1093/BIOINFORMATICS/BTW313.