Long-read genome assembly of the insect model organism *Tribolium castaneum* reveals spread of satellite DNA in gene-rich regions by recurrent burst events

Volarić Marin[#], Despot-Slade Evelin[#], Veseljak Damira, 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. Here we combined Nanopore long-read sequencing with a reference-guided assembly approach, to generate an improved, high-quality genome assembly TcasONT of the model beetle *Tribolium castaneum*. Enriched by 45 Mb in the repetitive part, the new assembly comprises almost the entire genome sequence. We used the enhanced assembly to conduct global and in-depth analyses of abundant euchromatic satDNAs, Cast1-Cast9. Contrary to the commonly adopted view, our finding showed the extensive spread of satDNAs in gene-rich regions, including long arrays. The results of the principal component analysis of monomers and sequence similarity relationships between satDNA arrays, revealed an occurrence of recent satDNAs array exchange between different chromosomes. We proposed a scenario of their genome dynamics characterized by repeated bursts of satDNAs spreading through euchromatin, followed by a process of elongation and homogenization of arrays. We also found that suppressed recombination on the X chromosome has no significant effect on the spread of satDNAs, but rather tolerates the amplification of these satDNAs into longer arrays. Analyses of arrays' neighboring regions showed a tendency of one Cast satDNA to be associated with transposon-like elements. Using 2D electrophoresis followed by Southern blotting, we proved Cast satDNAs presence in the fraction of extrachromosomal circular DNA (eccDNA). We point to two mechanisms that enable the said satDNA spread to occur: transposition by transposable elements and insertion mediated by eccDNA. The presence of such a large proportion of satDNA in gene-rich regions inevitably gives rise to speculation about their possible influence on gene expression.

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, which are scattered throughout the genome, satDNAs are characterized by monomer sequences tandemly arranged into long arrays. Previous studies on many animal and plant species indicate preferential localization of satDNAs in the (peri)centromeric heterochromatin [1]. 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 a role of satDNA in chromatin packaging, centromere formation/maintenance, and centromere evolution [2–4]. SatDNAs are also linked to chromosome mis-segregation [3], disease phenotypes [5], and reproductive isolation between species [6]. In addition to standard satDNA-rich regions such as (peri)centromeric heterochromatin, recent studies revealed examples where short arrays of (peri)centromeric satDNAs also spread to euchromatic regions. For example, (peri)centromeric satDNA families in humans, *Drosophila* and the beetle *Tribolium castaneum* can also be dispersed within euchromatin in the form of clustered or short repeat arrays [7–9]. In *Drosophila,* X-specific (peri)centromeric 1688 satDNA repeats dispersed in euchromatic regions have a role in dosage compensation through up-regulation of X-linked genes in *Drosophila* males [10,11]. The euchromatic counterpart of (peri)centromeric DNA in the beetle *T. castaneum* modulates the local chromatin environment upon heat stress changing the expression of neighboring genes [12]. Although there is clear evidence that satDNAs exist outside the (peri)centromere and have been assigned some roles, the understanding of their organization, their evolutionary dynamics and the disclosure of the molecular mechanisms that drive 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 absent even in the best genome assemblies (reviewed in [13,14]). However, with the improvement of long-read sequencing technology, especially Oxford Nanopore platform, it has recently been demonstrated that genomic loci made of satDNAs can be efficiently assembled [15].

Moreover, end-to-end maps of all human chromosomes, spanning highly repetitive (peri)centromeric and telomeric regions, have recently been obtained [16].

The beetle *T. 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* [17,18]. The *T. castaneum* genome has been sequenced, annotated and a first version of genome assembly was available in 2008 [19]. 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 dominates on all *T. castaneum* chromosomes comprising up to 17% of the genome [20,21], was drastically underrepresented in the assembled genome. 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 [22]. Regarding the possible presence of satDNAs in euchromatin of *T. castaneum*, our previous study on the *T. castaneum* reference genome assembly disclosed that *T. castaneum* genome abounds with many different and unrelated satDNAs distributed out of (peri)centromeric regions [23]. In addition, there are also short arrays of counterpart of (peri)centromeric DNA in euchromatin regions [9]. Nine satDNAs, Cast1-Cast9, represent the most prominent non-pericentromeric satDNA portion, which was mapped to ~0.3 % of the genome assembly at that time. In *Drosophila*, most euchromatic satDNAs feature short monomer units [24]. However, in *T. castaneum*, Cast1-Cast9 represent "classical" satellite DNAs with ~170 and ~340 bp long monomers. The experimentally determined genome content of Cast1-Cast9 satDNAs showed abundance of approximately ~4.3%, which is ~10X higher than the analyzed representativeness in the assembled genome [23]. The genome version from 2008 was based on a Sanger 7x draft assembly and was anchored to ten Linkage Groups (LGs) [19]. 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 [25]. Nevertheless, our previous satDNA analyses showed that satDNA genome fractions were highly underrepresented in Tcas5.2 as well. We considered that for a deep genome-wide analysis of dominant satDNAs distributed outside the (peri)centromere, the first prerequisite would be the availability of the assembly enriched in the repetitive sequences.

In the present work, we used Oxford Nanopore Technology (ONT) long-read sequencing and applied a reference-guided assembly strategy using the Tcas5.2 reference assembly to create an improved

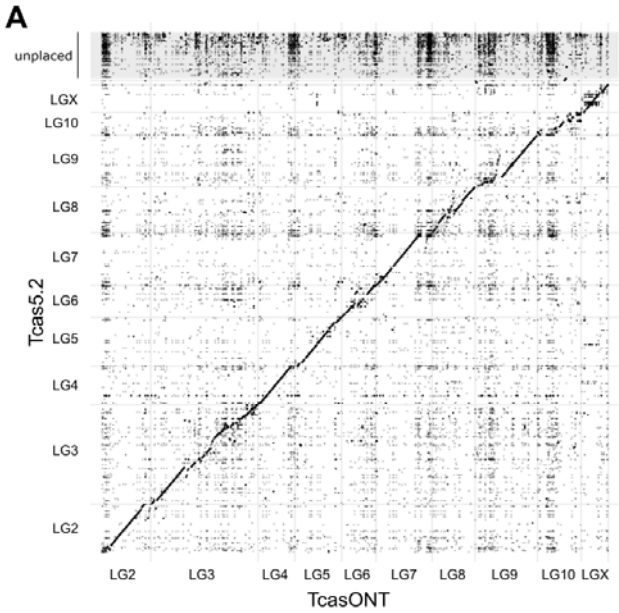
genome assembly of *T. castaneum*. We managed to fill gaps and elongate the repetitive regions, especially the satDNAs in the new TcasONT assembly. Furthermore, we used the TcasONT assembly presented here as a platform for global and in-depth analyses of the dominant satDNA fraction in euchromatin. Our results indicate that satDNAs spread widely in euchromatic regions, including gene-rich regions. We found that these euchromatic satDNAs undergo multiple expansion events that occur on a large scale and in extensive mixed arrangements between different chromosomes. We have proposed two mechanisms for these revealed genome dynamics of these satDNAs: spread by an insertion mediated by extrachromosomal circular DNA and, to a lesser extent, spread by transposon elements (TE).

Results

1. Chromosome-scale ONT-genome assembly and repetitive elements identification The size of the reference *T. castaneum* genome assembly (Tcas5.2) is 165,9 Mb [25]. Removing placeholders and sequencing gaps results in a genome assembly size of 136 Mb. Considering that the experimentally estimated genome size is 204 Mb [26], 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. 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, Suppl Table 1). Next, long reads (>20kb) were used for the initial assembly using Canu [27]. 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 [28]. 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), which is fully consistent with experimental estimation. The repeat ratio calculated by FindGSE was 27% of genome confirming the previously observed high genome repetitiveness [22]. Considering that (peri)centromeric satDNA TCAST, with 17% genome abundance, might be problematic for proper assembly [29], we assumed that the additional 117 Mb in our Canu assembly (Suppl Table 2) were primarily generated by gene-poor and highly repetitive 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 a new chromosome-scale assembly using reference-guided approach. The latest improved *T. castaneum* genome assembly Tcas5.2 [25] was used

154 as the reference for organizing these contigs into chromosomes. Two reasons were crucial for the
155 selection of this approach: 1) the availability of the high-quality *T. castane*um reference genome,
156 Tcas5.2, and 2) assembly approach that allows the input sequence to be ordered based on the mapping
157 quality of the reference sequence. Given that the reference genome Tcas5.2 contained 3669 unresolved
158 gaps with total size of 11 Mb and mean gap size of 3125 Kb, the first step was included gap filling of
159 Tcas5.2 with TGS-GapCloser using 8.6 Gb of Canu corrected larger reads (length >30 kb). TGS-GapCloser
160 was efficient in filling most of the gaps, so in total 3607 gaps (98.3 %) were closed which increased the
161 genome size by 10.5 Mb (Suppl table 4). The gap-filled Tcas5.2. reference genome assembly was further
162 used as a template for orientating 471 Canu-contigs into the 10 chromosomes using RagTag software
163 tool [30]. A total of 244 contigs were unambiguously assigned to the reference Tcas5.2. genome
164 resulting in a new ONT-based genome assembly named TcasONT (Suppl Table 5). The rest (227 contigs)
165 did not accurately map on the assembly, remaining assigned as unassembled sequences (Supp Table 6).
166 To further validate structure of omitted contigs, we mapped them with highly abundant
167 (peri)centromeric TCAST satDNAs and found that around half of unplaced sequences (14 Mb) are indeed
168 mostly covered (>50 %) with TCAST satDNAs (Suppl Table 6). To gain insight into the actual
169 representation of (peri)centromeric TCAST, we annotated the TcasONT assembly with TCAST satDNA
170 and obtained 9446 monomers, accounting for 3.6 Mb (1.7%) of the genome.
171 The final polished TcasONT has a size of 225.9 Mb, with 191 Mb assembled in ten chromosomes and the
172 remainder in unassembled contigs (Suppl Table 7). Compared to the reference Tcas5.2, the new
173 TcasONT assembly shows a 45 Mb increase of total chromosome length (Suppl Table 7). Comparing
174 TcasONT and Tcas5.2 chromosomes, the elongation of chromosome lengths ranges from 10.6 to 56.3 %,
175 indicating a significant improvement of genome assembly continuity at the chromosome level (Suppl
176 Table 8).
177 The result of completeness and quality of the new assembly revealed high levels of macrosynteny and
178 collinearity across all chromosomes with high sequence identity between TcasONT assembly and the
179 reference Tcas5.2 (Figure 1A). Furthermore, we calculated that 88% of the total Tcas5.2 unplaced contig
180 sequence length is now placed within chromosomes of the new TcasONT assembly.
181



B

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

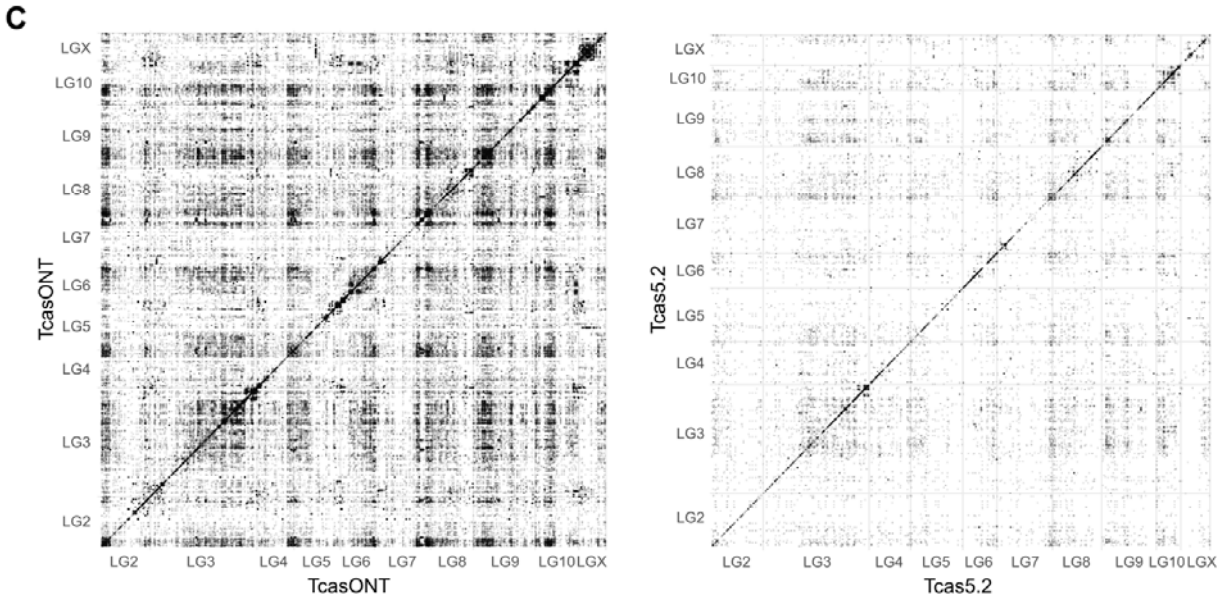
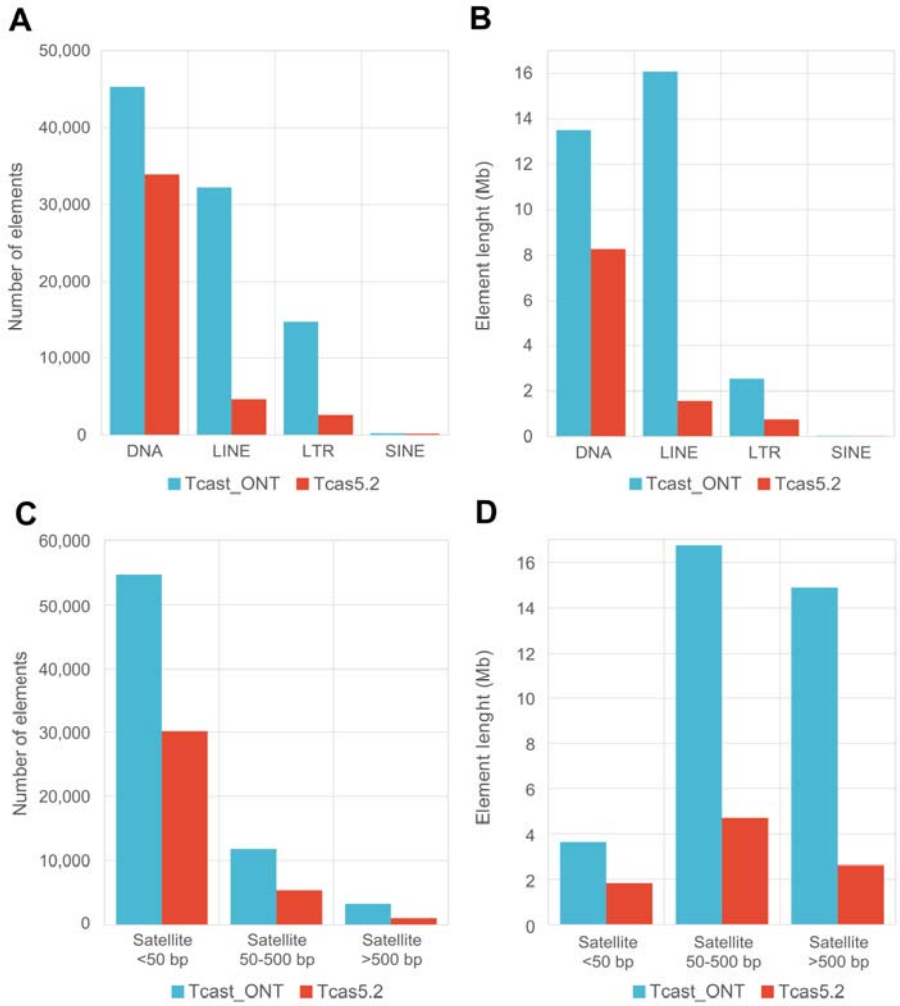


Figure 1. Assessment of *T. castaneum* assemblies. **A** Dot-plot representation of TcasONT vs. Tcas5.2 assembly comparison. The horizontal axis corresponds to the intervals along the TcasONT assembly, and the vertical axis corresponds to the intervals along the Tcas5.2 assembly. LG2-LG10, including LGX 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. **B** 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. **C** The whole genome-to-genome dot-plots analyses of TcasONT (left), and Tcas5.2 (right) 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.

195 We next assessed the gene completeness of the TcasONT assembly by comparative BUSCO analyses
196 between TcasONT and Tcas5.2 using insect universal orthologs from the odb10 database (Figure 1B). Of
197 the 1367 genes, we identified 1329, which corresponds to 97.2 % of BUSCO single-copy completeness.
198 Further, 17 genes were duplicated, 13 were fragmented and only 8 genes were missing. A considerable
199 increase in number of complete BUSCOs (32 genes) can be observed in the TcasONT assembly compared
200 to Tcas5.2. In addition to BUSCO analysis, we "lifted" genes from the Tcas5.2 annotations onto the
201 TcasONT assembly using the Liftoff package and assessed their final numbers (Suppl Table 9). Only 48
202 out of 14467 genes (0.3%) were left unmapped, mostly representing genes with no known biological
203 functions.
204 To evaluate the improvement in the representation of the repetitive genome portion in the new
205 TcasONT assembly compared to Tcas5.2, a self-similarity dot-plot analyses were performed using >70%
206 sequence identity as a criterion. The comparison of dot-plots revealed that the TcasONT assembly has a
207 higher level of self-similarity with medium to large dark blocks representing regions of repetitive
208 sequences in comparison to Tcas5.2 assembly (Figure 1C). Quantified, the number of self-similar
209 sequences was 44,671 in the TcasONT genome assembly, while in Tcas5.2 it was only 2,202, which
210 means that the new TcasONT assembly has about 20-fold higher amount of repetitive genome fraction.
211 To define the improvement specifically in transposable elements (TEs) content, the TEs in both Tcas5.2
212 and TcasONT assemblies were annotated using RepBase database and RepeatMasker. The TEs were
213 classified into 4 main types, including DNA elements, long interspersed nuclear elements (LINEs), long
214 terminal repeats (LTRs) and short interspersed nuclear elements (SINEs). The TcasONT assembly
215 revealed an increase of 11,297 DNA TEs, 12,268 LTR TEs and 27,553 LINE elements (Figure 2A and Suppl
216 Table 10) compared to Tcas5.2. Next, we investigated abundance of these elements in both genomes.
217 Among the four TE categories, in terms of cumulative sequence occupancy there is a notable increase in
218 TcasONT compared to Tcas5.2 for three of them; DNA TEs (5.2 Mb), LINE (14.5 Mb) and LTR (1.8 Mb)
219 (Figure 2B and Suppl Table 10). Taken together, compared to Tcas5.2, the TcasONT assembly is enriched
220 with 21.5 Mb of TEs (Suppl Table 10).
221 Additionally, we assessed the content of another class of repetitive DNA sequences, tandem repeats
222 (TRs), detecting them by the Tandem Repeat Finder (TRF), which was successfully used in previous
223 studies of *T. castaneum* [23]. Detected TRs were classified into 3 groups based on their monomer length
224 (<50bp, 50-500bp, and >500bp). The analysis revealed a total of 35.3 Mb TRs in the TcasONT assembly
225 compared to 9.1 Mb in Tcas5.2 (Suppl Table 10), indicating a significant improvement in TRs content. On
226 closer inspection, the results show that the number of TRs elements detected in the TcasONT assembly

227 was doubled in all three classes (Figure 2C), while the 50-500 bp and >500 bp TRs, which represent
228 "classical" satDNAs, contributed significantly to the cumulative genome length (31.7 Mb in TcasONT vs
229 7.3Mb in Tcas5.2) (Figure 2D).



230

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

235

236 From all above, the 45 Mb size difference between the Tcas5.2 and TcasONT genome assemblies is
237 mainly due to the enrichment of repetitive regions (45.8 Mb; Suppl Table 10). The repetitive fractions
238 that were most enriched in the TcasONT assembly were TEs (21.5 Mb) and TRs (26.2 Mb). If we consider
239 that the (peri)centromeric TCAST satDNA accounts for only 3.6 Mb, it is evident that the TcasONT

240 assembly has an increase of 22.6 Mb in TRs outside of (peri)centromeric regions which mainly includes
241 an increase in “classical” satDNAs with >50 bp long monomer units.
242

243 **2. Characterization of Cast1-Cast9 satDNAs in TcasONT assembly**

244 Given the demonstrated improvement in the genome assembly, especially in its repetitive parts, the
245 new TcasONT assembly represents an excellent platform for in-deep analyses of previously disclosed
246 abundant satDNA sequences outside the (peri)centromeres, Cast1-Cast9 [23].
247 The first step, essential for downstream analyses, was the mapping of the Cast1-Cast9 arrays in the
248 TcasONT assembly. Because of the well-known variability of monomer sequences within a satDNA
249 family, it was important to establish sequence similarity and sequence coverage parameters, which
250 would ensure the detection of the vast majority of Cast1-Cast9 monomers in the TcasONT assembly. For
251 this reason, the BLAST search of the raw Nanopore data was performed with the consensus sequences of
252 the Cast1-Cast9 monomers using two parameters, sequence coverage and similarity. The results were
253 shown as density plots where color intensity correlated with number of monomers of a particular Cast
254 satDNA (Suppl Figure 2). The majority of Cast satDNAs show the highest intensity, i.e. monomer
255 aggregation in the area corresponding to >75% of sequence coverage and >70% of percent identity.
256 However, the pattern with two areas of high density observed in Cast2 and Cast4 is a consequence of the
257 mutual similarity in one part of sequences between these satDNAs (Suppl Figure 3). Despite their partial
258 similarity they were still considered as individual satDNAs. Considering density plots for all Cast satDNAs,
259 the parameters sequence coverage > 70% and percent identity > 70% were used to ensure mapping of
260 most monomers of all 9 satDNAs (Cast1-Cast9).
261 To disclose the organization of the Cast1-Cast9 arrays, we analyzed the relationship between monomer
262 distance in the arrays for each Cast (Suppl Figure 4). More specifically, we wanted to determine whether
263 the arrays were organized in tandem continuously (consisting exclusively of monomers of a particular
264 Cast satDNA) or had mixed tandem organization (with a different sequence incorporated into the
265 arrays). The results showed that arrays of most satDNAs show no steep increases in the curve, pattern
266 representing a typical satDNA array organization consisting of continuously arranged monomeric
267 variants. In contrast, Cast2, Cast5 and Cast7 exhibited a disturbed tandem contiguity demonstrated by
268 sharp increases in average array length at certain distances. The sharpest increase is visible in Cast2,
269 indicating the presence of a specific sequence within the arrays that required further investigation. In
270 addition to monomeric Cast2 arrays, we found Cast2 monomers also organized as a part of a new,

271 longer repeat unit of approximately 1270 bp in length (Suppl Figure 5A-B). This new family of tandem
272 repeats is termed Cast2', and in the downstream analysis these two types of Cast2 arrays were
273 examined separately. A more detailed introspection of Cast5 arrays showed that the observed profile is
274 the result of occurrence of previously described R66-like sequences [23] dispersed within continuous
275 Cast5 arrays (Suppl Figure 5C). In addition, analyses of Cast7 arrays also reveal a mixed organization
276 where Cast7 monomers often occur in association with (peri)centromeric TCAST satDNA but with low
277 sequence length and similarity (Suppl Figure 4 and 5D).
278 Considering the difficulty of correctly assembling satDNA sequences, the next question was whether the
279 array lengths and arrangement of the Cast1-Cast9 satDNAs in the TcasONT assembly correspond to the
280 real situation of the genomic loci containing the Cast1-Cast9 repeats. Namely, repetitiveness of satDNAs
281 can lead to assembly collapse, which can result in the number of monomer units (array length) in a
282 genome assembly being lower than in the real genome. Given that raw reads provide a realistic picture
283 of what is originally present in the genome as they do not undergo an assembly process, we performed
284 a comparative analysis of Cast1-Cast9 across the individual reads and across the new genome assembly.
285 For this analysis, it was crucial to define optimal parameters for array detection to avoid fragmentation
286 due to short gaps. Therefore, considering monomer length and mixed array organization of specific Cast
287 satDNAs such as Cast2, Cast5 and Cast7, the best window length that ensures detection of the maximum
288 number of arrays was evaluated for each Cast satDNA. The analysis showed that a minimum of three
289 consecutive monomer units should be a criterion for correct array characterization for each Cast
290 satDNA. The array patterns between the datasets of individual reads and the TcasONT assembly show
291 remarkable similarity for almost all Cast satDNAs (Suppl Figure 6). The only exception is Cast6, whose
292 pattern indicates a suppression of the representation of arrays in the assembly regardless of size.
293 To quantify the enhanced detection of Cast1-Cast9 regions in TcasONT genome assembly, we compared
294 both their content and array length with Tcas5.2 (Table 2). In terms of overall enrichment in Cast satDNA
295 portion, the TcasONT assembly shows an increase of 7.2 Mb in the representation of analyzed Cast
296 satDNAs compared with Tcas5.2. Among them, Cast1, Cast2' and Cast5 emerged as particularly
297 prominent satDNAs with an increase of 0.7 Mb, 2.8 Mb and 2.6 Mb, respectively. The total content of
298 Cast satDNAs in the new TcasONT genome was 8.8 Mb (4.6% of the genome), which is remarkably close
299 to the previous experimental estimate of the genome abundance for these satDNAs (>4.3% in [23]).
300 Among them, Cast2' and Cast5 showed the highest genome abundance with 1.5% for each, Cast8 and
301 Cast9 the lowest (<0.1%) while the rest remained in the moderate genome abundance level (0.1-0.4%).
302 The results also show an exceptional increase in the number of arrays in the TcasONT assembly for

303 Cast2', Cast5 and Cast7. In Cast2', for example, the number of arrays has increased by more than 9X. In
304 addition to the number of arrays, average increase in array length (expressed as mean values) can also
305 be observed in the TcasONT genome assembly for all Cast satDNAs. A significant increase in average
306 array length of up to 3X was observed for Cast2', Cast5 and Cast6 satDNAs compared to Tcas5.2. The
307 length of an average array for Cast satDNAs in the TcasONT assembly ranges from 1.2 to 31 kb. The
308 highest average array size of ~31 kb has been found for Cast6, while Cast1 has the longest array of ~112
309 kb.

310

311 Table 1. 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

312

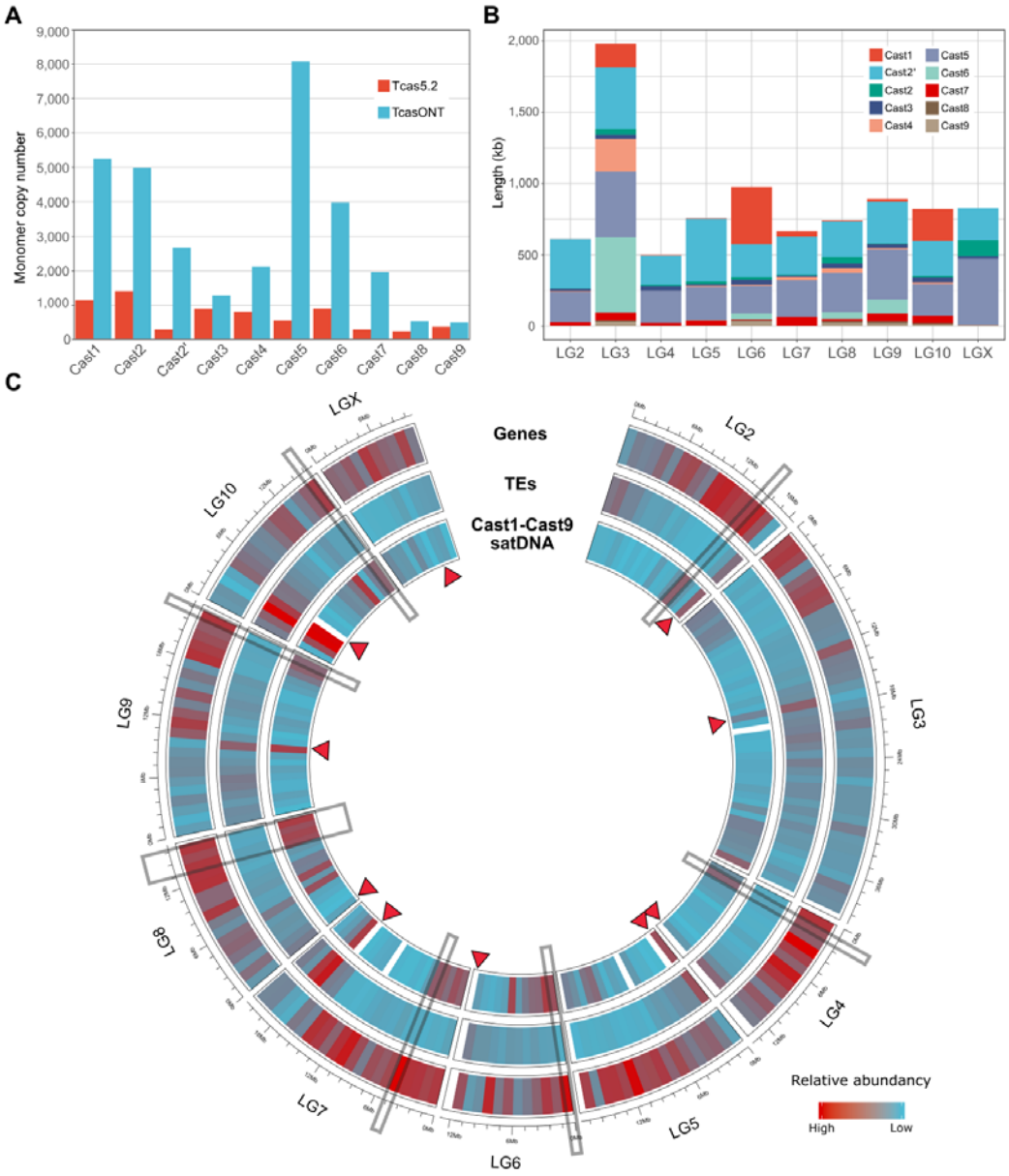
313 The graph of Cast1-Cast9 monomer copy number in TcasONT compared to Tcas5.2 also clearly shows
314 the presence of significant enrichment (2-10X) in the majority of satDNAs analyzed (Figure 3A). From the
315 above, the new TcasONT assembly contains the representative sample of Cast arrays and can be taken
316 as a credible platform for further genome-wide analysis of Cast satDNAs.

317

318 **3. Cast1-Cast9 satDNAs chromosome distribution and genomic environment**

319 The distribution of Cast1-Cast9 including Cast2' satDNAs on *T. castaneum* chromosomes was further
320 investigated by mapping all Cast1-Cast9 monomers on the assembled TcasONT linkage groups. The two
321 main patterns were observed. The first pattern is characterized by Cast satDNAs located on a subset of

322 chromosomes (Figure 3B). For example, Cast1 dominates on LG3, LG6 and LG10, while it is
323 underrepresented on other chromosomes. Similarly, Cast6 dominates on LG3, but is also present, albeit
324 to a much lesser extent, on LG6, LG8 and LG9. In contrast, the other Cast satDNAs such as Cast2', Cast5
325 and Cast7 show a more uniform pattern characterized by similar content on almost all chromosomes
326 (Figure 3B).



327

328 **Figure 3.** Distribution evaluation of Cast1-Cast9 satDNA in TcasONT assembly. **A** Comparison of
329 monomer copy numbers for Cast satDNAs annotated in the reference Tcas5.2 and the TcasONT genome
330 assemblies (see also Suppl Table 11). The assemblies were mapped with Cast1-Cast9 satDNA monomeric
331 consensus sequences published previously [23]. **B** Comparison of monomer copy numbers for Cast
332 satDNAs annotated in the reference Tcas5.2 and the TcasONT genome assemblies (see also Suppl Table

333 11). The assemblies were mapped with Cast1-Cast9 satDNA monomeric consensus sequences published
334 previously [23]. **C** Circular representation of the TcasONT genome assembly with annotated Cast1-Cast9
335 satDNAs, genes and transposable elements (TEs). The boxes indicate the most prominent regions where
336 the accumulation of Cast satDNAs correlates positively with gene density and negatively with the TEs
337 density. The position of the (peri)centromere on each chromosome is indicated by a red arrow.

338 Next, we analyzed the distribution of Cast1-Cast9 along chromosomal length (Suppl Figure 7). Since it is
339 already known that TCAST builds up huge blocks of (peri)centromeric heterochromatin [20], the
340 presumed localization of the (peri)centromere on each chromosome is marked. Detailed examination of
341 Cast1-Cast9 distribution disclosed an exceptional spread of Cast satDNAs along the chromosomes with
342 no tendency to cluster in proximal regions of the (peri)centromere. Only Cast7 occurs preferentially in
343 regions enclosed by (peri)centromeric satDNA TCAST. In addition, some Cast satDNAs show a tendency
344 to cluster in distal chromosomal regions.

345 The distribution of Cast1-Cast9 satDNAs on chromosomes compared with genes and transposon
346 elements (TEs) was also analyzed (Figure 3C). Interestingly, Cast1-Cast9 satDNAs are often part of gene-
347 rich regions and rarely overlap with TEs (the most prominent regions with these features are highlighted
348 in Figure 3C). The only coincidence of the occurrence of Cast satDNA and TEs was found in the
349 (peri)centromeric regions, which corresponds to the occurrence of Cast7 satDNA.

350 For a more detailed analysis of Cast1-Cast9 satDNAs arrays, 50 kb flanking regions surrounding Cast1-
351 Cast9 arrays were examined for the presence of genes and TEs. More specifically, all genes in the
352 flanking regions of Cast1-Cast9 arrays were counted, and the base gene content of the whole genome
353 assembly was calculated and used to assess the relative gene density in the vicinity of Cast arrays (Figure
354 4A). From the scaled values based on the total number of arrays, it is evident that Cast satDNAs have
355 median gene content values that are above the median of the genome, except for Cast1, Cast6 and
356 Cast7. Moreover, the arrays Cast2', Cast3, Cast5 and Cast8 are surrounded by a significantly larger
357 number of genes (Kolmogorov-Smirnov test, $p<0.01$) than the rest of the genome (Suppl Table 12). In
358 addition, these Cast satDNAs have distributions above 3rd quartile range of the genome indicating the
359 presence of arrays embedded in extremely gene-rich regions. The only Cast satDNA that has a lower
360 number of adjacent genes is Cast7 (Supple Table 12, Figure 4A). This correlates with the observation that
361 Cast7 arrays occur in an intermingled organization with (peri)centromeric TCAST satDNA (Suppl Figure
362 7). In contrast, analyses of Cast satDNAs distribution in relation to TEs showed a significantly lower
363 number of TEs near Cast satDNA arrays compared to the whole genome (Kolmogorov-Smirnov $p<0.01$)
364 (Figure 4B).

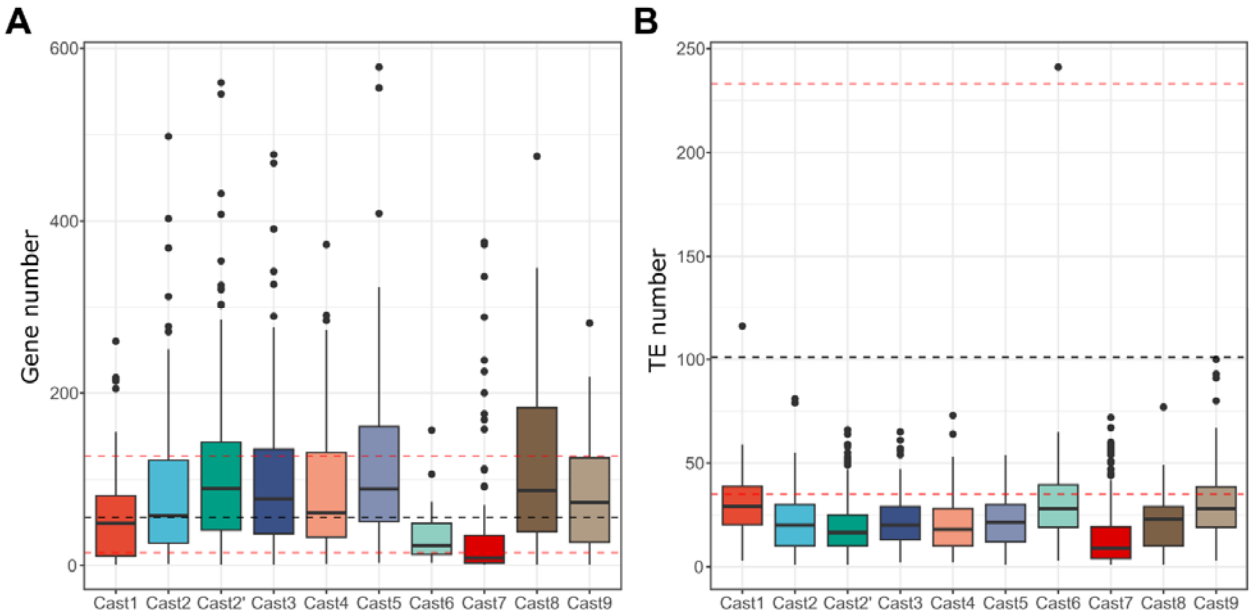


Figure 4. Distribution of gene and transposable element counts in vicinity of Cast arrays. **A** The number of genes was calculated based on the distribution of gene content in the assembled genome with the black line representing the median number of genes in regions of same size (100 kb windows, 5kb sliding frame) and the red dotted lines represent the first and third quartiles (Q1 and Q3) of the genome wide distribution. Results of the Kolmogorov-Smirnov tests are represented in Suppl Table 12. **B** The number of transposable elements was calculated using the same window size as for the genes, with black and red lines representing the same values. All Cast arrays had significantly (Kolmogorov-Smirnov $p<0.01$) less transposable elements in their vicinity when compared to the genome wide distribution, highlighting the fact that they do not overlap with TE presence.

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 Suppl Figure 8. Interestingly, most Cast satDNAs have the same profile with the highest gene density being associated with 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). In the examined 50kb windows around satDNA monomers, gene counts show similar trends (Suppl Figure 8), without distinct graph patterns, meaning that the arrays are uniformly embedded in gene-rich regions.

388 **4. Genomic dynamics and mechanisms of propagation of Cast satDNAs**

389

390 We considered that analyzes based on the comparison of the mutual variability of monomers and
391 arrays, together with their position on the chromosomes, can be informative for monitoring the
392 genomic dynamics of Cast satDNAs. Taking into account that the data set (monomers and arrays) was
393 large and the overall variability of monomers within the family relatively low (Suppl Figure 2), we noted
394 that phylogenetic analyzes of a huge number of similar sequences would not be sufficiently sensitive to
395 disclose trends in Cast satDNAs genome dynamics. Therefore, we decided to perform the principal
396 component analysis (PCA) of monomer alignments, and sequence similarity relationships between the
397 arrays for each of Cast satDNAs (Figure 5A and C). First, a genome-wide database of Cast monomers was
398 created and annotated with their actual chromosomal positions. PCA results obtained this way showed a
399 similar trend for almost all Cast satDNAs characterized by a scattered pattern of monomers. This is
400 particularly evident for Cast1, Cast2, Cast2', Cast3 and Cast4 satDNAs, where regions of extremely high
401 density of monomers from different chromosomes become visible. The only exception with a pattern
402 characterized by an accumulation of monomers from the same chromosome was observed in Cast6,
403 satDNAs with long homogenized arrays.

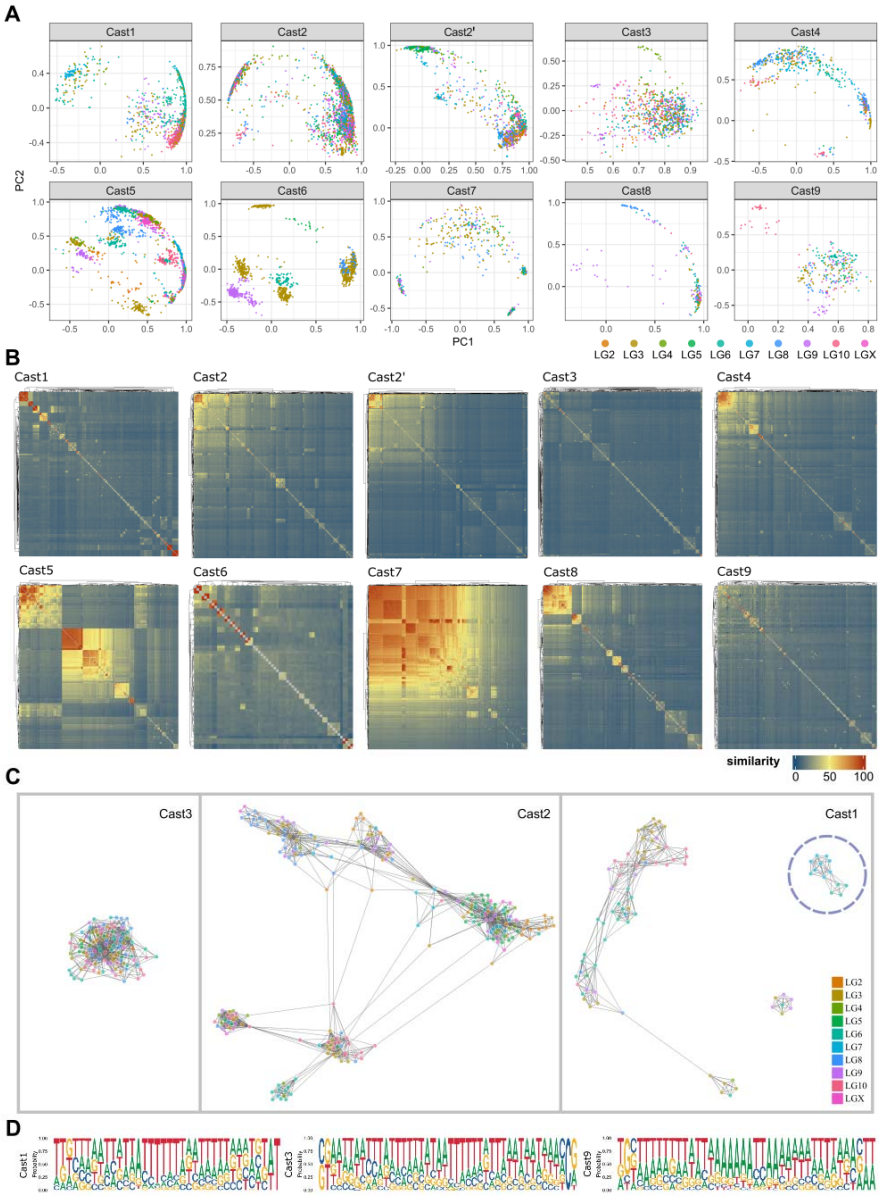
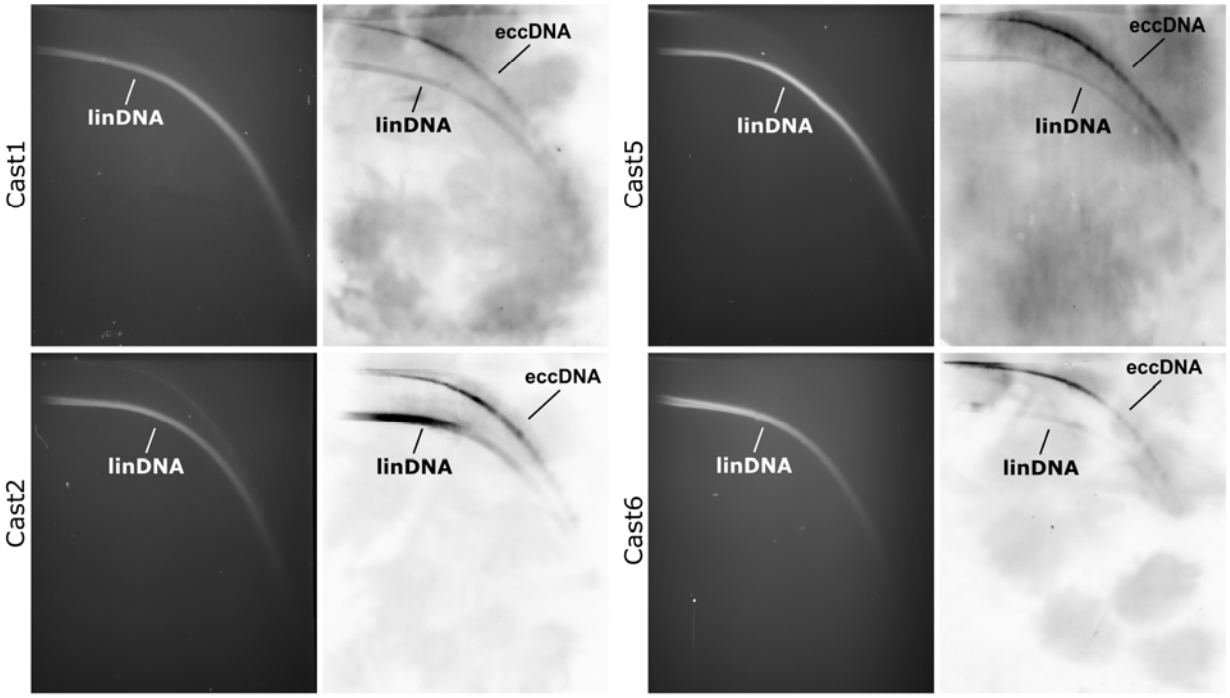


Figure 5. A Principal component analysis (PCA) of aligned Cast1–Cast9 satDNAs monomer distance matrices. Dots represent a monomer unit. Monomers were colored according to the chromosome of origin. The most prominent and recent exchange events between different chromosomes are visible in regions with high density of points with different colors. **B** 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. **C** Graph networks of Cast3, Cast2 and Cast1 arrays based on their sequence similarity relationship. Cast 3 shows a cluster containing all arrays. Cast2 shows a pattern with six different clusters containing arrays mixed from different chromosomes. Cast1 shows no significant clustering of arrays, except small divergent clusters. **D** Sequence logo of junction regions of Cast1, Cast3 and Cast9 arrays in which an enrichment of polyT and polyA tracts can be recognized.

417 Second, after a database of Cast arrays with corresponding chromosome annotation having been
418 created, comparative analyzes of the arrays were performed based on their mutual sequence variability
419 for each Cast satDNAs. Sequence similarity relationships between the arrays for each of Cast satDNAs
420 are represented by graph networks (Figure 5C, Suppl Data), except for Cast6 and Cast7 which were not
421 analyzed due to the small number of arrays. The distance and interconnectivity between dots correlate
422 with the sequence similarity between arrays, with closer dots indicating higher similarity of monomers
423 between different arrays. Our assumption is that the relationship between the arrays reflects their
424 spread throughout genome. Three graphs, Cast1, Cast2, and Cast3 were selected as illustrative examples
425 (Figure 5C), while the other Cast satDNAs arrays networks are presented in Suppl Data. The Cast3 shows
426 the pattern of array dynamics with a dominant cluster containing relatively closely related arrays from
427 all or almost all chromosomes, and the similar pattern applies to Cast4, Cast8 and Cast9. In addition to
428 the dominant mixed clusters, Cast4, Cast8 and Cast9 also have some small distant clusters (up to 10
429 arrays) from the same chromosome. The second pattern includes Cast2 with six relatively distant array
430 clusters with related arrays from different chromosomes, indicating multiple genome events of intense
431 array expansion during Cast2 evolution. Interestingly, only one of them represents intrachromosomal
432 expansion, while the rest show very extensive interchromosomal exchange. A similar pattern with
433 several distant clusters containing related sequences from different chromosomes is also observed in
434 Cast2' and Cast5. Finally, the Cast1 network shows clusters with higher distance between arrays, with
435 some array sets completely separated due to their sequence divergence. The Cast1 network also shows
436 two separate subgroups of sequences representing different variants, one of them (Figure 5C, labelled)
437 directly connected to the transposon element, Polytron (Suppl Figure 9).
438 Interestingly, these three different patterns of Cast satDNA propagation events correlate with the
439 average lengths of the arrays. For example, satDNAs for which only one expansion event can be
440 observed (Cast3, Cast4, Cast8 and Cast9) have a relatively short array length (mostly around 4000 bp).
441 SatDNAs with several expansion events, as seen in Cast2, Cast2' and Cast5, have a moderate array
442 length of about 15000 bp. Finally, Cast1, for which no recent expansion centers were observed, intends
443 to have very long arrays (up to 112kb).
444 To gain insight into the spread of Cast1-Cast9 arrays throughout the genome, the surrounding regions of
445 the Cast arrays were also analyzed. To address this, regions of 2 kb in length upstream and downstream
446 of individual Cast arrays were extracted. The extracted regions were aligned using the MAFFT program
447 and the sequence similarity values were calculated from the alignment for each Cast satDNA. The results
448 showed that of the ten Cast satDNAs analyzed, only Cast5 and Cast7 show similarity of surrounding

449 regions for most arrays (Figure 5B). A detailed analysis of the surrounding regions for Cast5 arrays
450 revealed two dominant regions consisting of sequences with high similarity. One of them is the R66-like
451 sequence, found as a scattered sequence in the Cast5 arrays (see Suppl Figure 5C) being also located at
452 one end. On the other side of most Cast5 arrays the R140-like sequence was found. Detailed analysis
453 showed that out of 150 Cast5 arrays, 2/3 arrays had R140-like sequences and 1/3 arrays had R66-like
454 sequences at the ends of the arrays. Likewise, Cast7 arrays showed that most of them are surrounded by
455 the (peri)centromeric TCAST (Suppl Figure 5D). The rest of Cast satDNAs showed only partial similarities
456 of the surrounding regions noticed only for a smaller number of arrays. (for example, Figure 5B, Cast1
457 panel). For example, a high-similarity subset of Cast1 arrays showed that those nine arrays originate
458 from the same chromosome (LG7), having the same transposon-like sequence at the end of arrays
459 (Suppl Figure 9).
460 For proper examination of the precise Cast array junction regions, their edges need to be precisely
461 defined. Since the monomers at the ends of the arrays have a higher variability due to the less efficient
462 recombination at the ends of arrays (reviewed in [14]), we developed a k-mer similarity-based approach.
463 The results of the analyzes did not reveal any part of the monomer that would be favored at the ends of
464 the Cast arrays (data not shown). However, the analyzes of 20 bp sequence motifs near the array
465 boundaries showed that Cast1, Cast3, and Cast9 have regions in the vicinity of arrays with significant
466 poly A/poly T tracts, whereas the other Cast satDNAs do not share any common motif (Figure 5D, Suppl
467 Figure 10).
468 Considering the observed extensive inter- and intrachromosomal exchanges of all Cast satDNAs, one of
469 the possible mechanisms that influence this phenomenon could be the insertion of satDNA arrays
470 mediated by extrachromosomal circular DNA (eccDNA). To test the presence of Cast satDNAs in the
471 eccDNA fraction, we performed two-dimensional (2D) agarose gel electrophoresis followed by Southern
472 blot hybridization. The hybridization probes were developed for the most abundant Cast satDNAs,
473 Cast1, Cast2', Cast5 and Cast6 (0.3-1.5% of the genome), and the eccDNA molecules containing the
474 examined satDNAs were indeed detected (Figure 6).

475



476

477 **Figure 6.** Detection of extrachromosomal circular DNA molecules containing *T. castaneum* Cast1, Cast2,
478 Cast5 and Cast6 satDNA. Results of *T. castaneum* genomic DNA separation in 2D electrophoresis are
479 shown on the left, while Southern blot hybridization with specific probes for each Cast element is shown
480 on the right. Signals obtained by Southern blots confirm the presence of specific Cast satDNA in
481 chromosomal linear double-stranded DNA (linDNA), but also in extrachromosomal circular DNA
482 (eccDNA) molecules.

483 In general, chromosomal regions with suppressed or without recombination tend to accumulate
484 satDNAs [31]. Since sex chromosomes represent regions with suppressed recombination, we examined
485 the number and length of Cast1-Cast9 arrays on the X chromosome to compared it with autosomal
486 chromosomes. This type of analysis would be even more informative using the Y chromosome, which is
487 mostly non-recombining, but unfortunately, the Y chromosome was not available in the former Tcas5.2
488 nor in the new TcasONT assembly. Mapping of Cast1-Cast9 arrays to chromosomes disclosed that the X
489 chromosome did not have a significantly higher average number of Cast1-Cast9 arrays per Mb,
490 compared to the others (Supp Table 13). The comparison of the sequence variability of Cast satDNA
491 monomers from X chromosome in contrast to autosomes did not show significant difference. However,
492 the analysis of Cast1-Cast9 array length showed a statistically significant longer length of arrays on the X
493 chromosome compared to the others (Wilcoxon test, $p<0.05$) (Figure 7). This trend is especially
494 prominent in Cast 2, Cast2', Cast5 and Cast9, where the average length of the arrays is increased even
495 10X at the X chromosome (e.g. Cast2') (Suppl Figure 11).

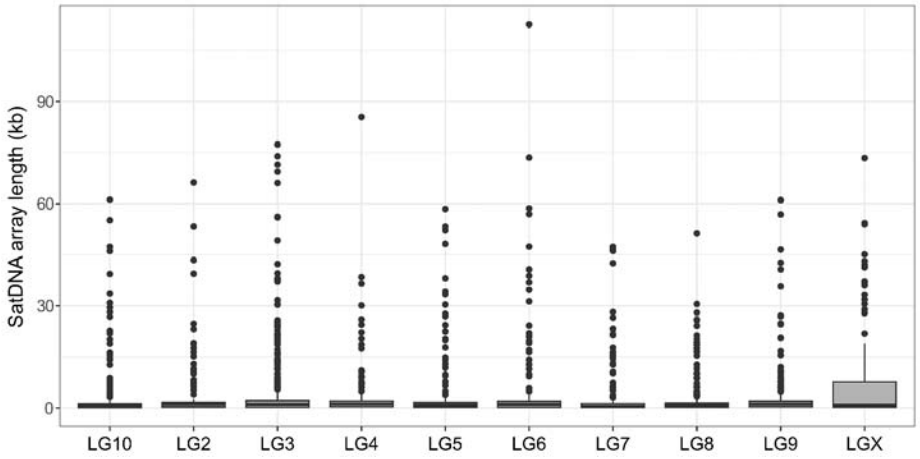


Figure 7. Distribution of Cast1-Cast9 satDNA array lengths across chromosomes, LG2-LG10 including X chromosome (LGX).

502 ## Discussion

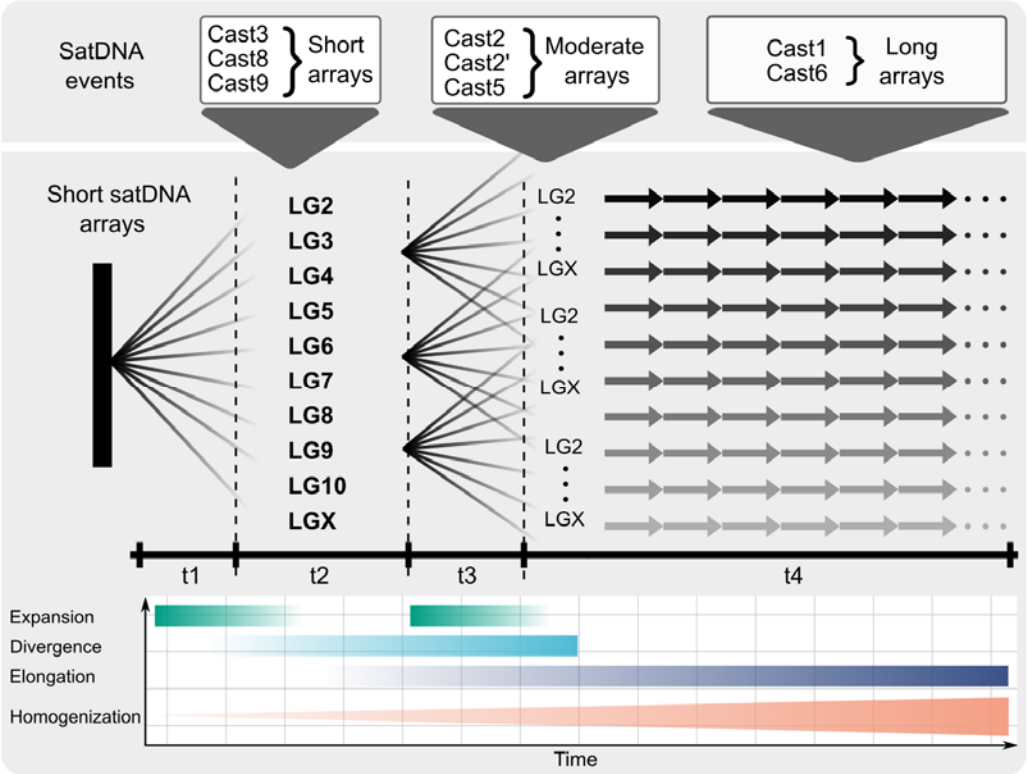
503 Despite significant advances in sequencing technologies and plethora bioinformatics tools, satDNAs are
504 still the most difficult part of a genome to assemble. The most illustrative example of the problem with
505 satDNA in assembly process using long-read sequencing is found in the recently published work on
506 holocentric nematodes [32] . Although these genomes are small, the assemblies were quite fragmented,
507 and the authors pointed out that abundant and scattered satDNAs occupying holocentromeres are the
508 main problem preventing assembly at the chromosome level [32,33]. It is clear that the assembly of
509 satDNAs-rich genomes is challenging, even when long-read sequencing is used. Previous studies [22,23]
510 indicated that *T. castaneum* is one of the species with a considerable amount of diverse repetitive
511 sequences, and therefore it is not surprising that the official reference assembly, although of good
512 quality in the coding part, was significantly deficient in long-range assembly of repetitive DNAs.
513 The aim of our research was a thorough investigation of satDNAs in *T. castaneum* euchromatin, and for
514 this purpose we needed an assembly of reliable quality in repetitive regions. Unfortunately, the official
515 assembly Tcas5.2 did not fulfill this criterion, and we decided to improve the *T. castaneum* assembly,
516 focusing on upgrading the difficult-to-assemble repetitive DNA regions. To this goal, a high-quality *T.*
517 *castaneum* genome assembly at the chromosome level was generated by combining nanopore long-
518 read sequencing and a reference-guided approach. The new TcasONT assembly is nearly 45 Mb larger
519 than the latest version of the Tcas5.2 reference genome. The achieved better completeness of genes
520 compared to the reference Tcas5.2 underlines the credibility of TcasONT for gene representation. The
521 TcasONT assembled chromosomes lack only 13 Mb of the estimated *T. castaneum* genome sequence of
522 204 Mb, previously determined experimentally [28] and also *in silico* in our study. The missing 13 Mb
523 could primarily be attributed to (peri)centromeric regions, due to assembly-impeding highly repetitive
524 TCAST satDNA regions [21]. Despite significant advances in sequencing technologies and assembly
525 methods, it is not unusual that (peri)centromeric regions remain still incomplete even in most nanopore-
526 based assemblies [34]. The only exceptions are the recently published Telomere-to-Telomere assemblies
527 of prominent model organisms such as human and *Arabidopsis* that contain the complete
528 (peri)centromere [16,35,36]
529 In addition to enrichment in the gene regions, TcasONT assembly proved to be highly representative for
530 the study of the repetitive regions. Namely, comparison of the assemblies showed a 20-fold enrichment
531 in the repetitive part in the TcasONT assembly compared to Tcas5.2. In TcasONT, the representation of
532 TEs has also been increased, both in terms of the number of elements and the abundance. However,
533 the increase is particularly significant for tandem repeats (TRs) with a monomer unit length of >50 bp,

534 which are declared as “classical” satDNA. The enrichment of the proportion of TEs and TRs totals 47.8
535 Mb compared to Tcas5.2. Our study proves that the nanopore sequencing approach enables the
536 resolution of previously difficult-to-assemble regions, most of which consist of highly repetitive
537 sequences such as satDNAs and TEs. It is also important to note that TcasONT is larger than the previous
538 version by 25%, which means that a substantial portion of the genome is now available for further
539 research.
540 The assembly process often does not assemble the satDNAs correctly, resulting in collapsed arrays [29],
541 but our comparative analyses of the raw data and the TcasONT assembly show that the TcasONT
542 assembly is highly credible in terms of representativeness of Cast satDNAs. The TcasONT assembly
543 showed Cast satDNAs abundance of 8.8 Mb, which accounts for 4.6% of the genome, being consistent
544 with their experimentally confirmed genome abundance [23]. The previous assembly Tcas5.2 contained
545 only 1/10 of experimentally estimated amount of Cast satDNAs. Next, we used the TcasONT assembly
546 provided here as a source for a comprehensive and in-depth analysis of the ten most dominant satDNAs
547 in euchromatin, Cast1-Cas9, previously characterized in [23]. Our genomic analysis of these satDNAs
548 included their distribution, correlations with genes and TEs, genomic dynamics, and possible
549 mechanisms of spread in euchromatin.
550 Interestingly, we found that these ten "classical" satDNAs (Cast1-Cast9, including a new version of Cast2,
551 Cast2') do occur in the form of long tandem arrays in euchromatin, a region in which the accumulation
552 of abundant satDNAs is not as common as in the (peri)centromeric heterochromatin. Moreover, some
553 Cast satDNAs arrays show a very high average length (Cast6 with 31kb) and maximum length (Cast1 with
554 112kb). The question arises whether the presence of such long satDNA arrays in regions that are not
555 considered suitable for the accumulation of satDNAs, is indeed a distinctive feature of the *T. castaneum*
556 genome. We suspect that the presence of relatively numerous and abundant satDNAs in euchromatin is
557 not a unique feature of the *T. castaneum* genome, but that the absence of satDNAs in these regions in
558 some other genome assemblies is a consequence of significant gaps caused by assembling difficulties.
559 We therefore believe that the strategy of nanopore sequencing, combined with reference-guided
560 assembly approach, ensures an enviable representation of satDNAs, which should be crucial for
561 comprehensive and in-depth satDNA studies in other species.
562 Furthermore, one of the important questions that arose from the observed organization, characterized
563 by long arrays in non-pericentromeric regions, is whether Cast1-Cast9 satDNAs were truly euchromatic.
564 Indeed, it is possible that satDNAs, even if located outside the (peri)centromeric region, could be
565 localized in gene-poor regions populated by, for example, transposons or other types of noncoding DNA

566 [37]. However, our gene density analyses showed that the surrounding regions of the arrays of almost
567 all Cast satDNAs correlated positively with the gene-rich regions compared to the average gene density
568 in the genome. Moreover, some satDNA arrays were surrounded by extremely gene-rich regions. It is
569 also interesting that these regions are not characteristic of short arrays as expected, but long satDNA
570 arrays (>10kb) are also found near the genes.
571 It has not been seen before that satDNAs in euchromatin, including those with long arrays, were mostly
572 integrated into an environment rich in genes and poor in transposons. Contrary to expectations, we
573 found that the occurrence of Cast1-Cast9 satDNAs and TEs often do not overlap. This fact clearly speaks
574 of the ability of Cast satDNA to invade euchromatic regions and indicates the flexibility of euchromatin,
575 i.e., gene-rich regions, to tolerate and accommodate these sequences. Previous models support the fact
576 that the accumulation of satDNA is characteristic of regions where recombination is suppressed, such as
577 (peri)centromeric heterochromatin [38]. In addition, a commonly accepted view is that genome
578 recombination events ultimately result in array loss, and therefore satDNAs mainly persist in
579 heterochromatic regions (reviewed in [14]). However, from our study, it appears that gene-rich
580 euchromatin, where recombination is expected to occur, may tolerate the accumulation of satDNA.
581 Further, comparative analysis of Cast satDNA arrays between the X and autosomal chromosomes
582 showed that suppressed recombination in the X chromosome stimulates the propagation of satDNAs
583 into longer arrays. However, it has no significant impact on the number of arrays and sequence
584 variability. We propose that recombination in the euchromatin does not prevent the satDNA extensive
585 spread but discourages arrays prolongation. This impact on arrays' length in euchromatin is probably
586 due to rearrangement potential of satDNAs elongation, which can affect neighboring regions.
587 Considering that previous studies have mainly reported short satDNA arrays in euchromatin regions, our
588 finding of "classical" satDNAs, widely present across the euchromatin, in the form of long arrays in gene-
589 rich regions, provides new insights into the genomic distribution of satDNAs. Previous studies were
590 based on euchromatic satDNA with short arrays, suggesting that these satDNAs may play roles in gene
591 regulation by acting as "evolutionary tuning knobs" [35], regulating chromatin [9,12], and facilitating X
592 chromosome recognition/dosage compensation [11]. SatDNAs located in euchromatic regions could also
593 regulate gene expression by modulating local chromatin structure, or via transcripts derived from the
594 repeats. For example, contractions of the human subtelomeric satellite D4Z4 alter the chromatin state
595 of nearby genes, causing muscular dystrophy [40]. Recent studies disclosed that expression of
596 euchromatic satDNA-derived transcripts control embryonal development in the mosquito via sequence-
597 specific gene silencing [41]. Being in the gene-rich regions, epigenetic histone marks of satDNAs in

598 euchromatin could also influence surrounding gene regions. For example, human genome-wide analysis
599 of satDNAs in euchromatin has shown their association with repressive histone mark H3K9me3,
600 suggesting that they may affect the expression of neighboring genes [42]. In addition to the direct
601 impact on genes, large-scale rearrangements that involve the long Cast1-Cast9 arrays scattered
602 throughout the genome, are also highly probable. The ability of euchromatic satDNA to undergo high
603 evolutionary turnover may also contribute to the rapid change of gene landscape across the genome,
604 and consequently on gene function [43]. In summary, the presence of such a substantial portion of
605 Cast1-Cast9 satDNAs in euchromatin regions of the *T. castaneum* genome can strongly influence both
606 gene expression and the dynamics of genome evolution.
607 Another important emerging issue, given the widespread distribution of Cast1-Cast9 satDNAs in
608 euchromatin and their putative impact on the genome evolution, is the way in which these satDNAs are
609 propagated. While the mechanisms of TE propagation are well known and depend on the type of
610 transposon elements [44], the propagation of satDNA, especially in euchromatin regions, is still quite a
611 mystery. Given that we know little of the genome dynamics and molecular mechanisms that shape the
612 distribution of euchromatic satDNA on the genome scale, we carried out analysis of the distribution
613 pattern, genome dynamics as well as analysis of surrounding and junction regions of Cast1-Cast9
614 satDNAs. The results mainly show arrays dispersed on all chromosomes and along their entire
615 chromosome length without regional preference. Interestingly, the results of the sequence similarity
616 relationships between arrays analyzed at the genome level for each Cast satDNAs, showed grouping of
617 Cast satDNAs in three characteristic patterns (shown in Figure 5C). If we assume that the observed
618 patterns of satDNA behavior are events captured at a specific and distinct point of "satDNA life" in the
619 genome, here we propose a course of events that can explain the genome dynamics of euchromatic
620 satDNAs over time. The timeline of possible events is presented in a simplified form, as shown in Figure
621 8. The first event could be extensive expansion from one center, when short arrays dramatically spread
622 and shuffled among different chromosomes (t1). After expansion, the arrays of the satDNAs diverged in
623 sequence due to their different localization and reduced effect of homogenization. At the same time, if
624 arrays are located in a stimulating environment for the propagation, their arrays become lengthened
625 (t2). After this phase, some of the arrays, probably short ones, could be roused and form new expansion
626 centers, characterized again by a rapid spread to different chromosomes (t3). After many expansion
627 events, satDNA could enter a dormant state in terms of spreading but continue to extend arrays (t4). In
628 a favorable environment, the arrays can be subjected to extensive elongation and homogenization. It is
629 difficult to determine what causes a particular sequence to become a promoter of dispersal. However,

630 the observed pattern of euchromatic satDNA genome dynamics suggests an efficient mechanism of their
631 extensive and efficient spread on different chromosomes.



632

633 **Figure 8.** The most parsimonious scenario of genome dynamics of euchromatic Cast satDNAs. The
634 multiple expanding lines extending from a single genomic locus represent a sudden burst event in which
635 satDNA spreads on different chromosomes, followed by a second expansion event that leads to further
636 propagation of satDNA. Evidence of such behavior in *T. castaneum* is shown on top of the figures with
637 three different examples of euchromatic satDNA. The timeline depicts the main events in different
638 colors. Expansion is described as multiple sudden bursts denoted by a sudden increase in intensity and
639 subsequently followed by a slow decrease. The divergence event lasts longer, and its effect is more
640 prominent with time. Elongation has an even later onset with a similarly increasing effect over time.
641 Homogenization is a permanent process affecting all satDNAs, however the impact of homogenization is
642 the strongest on the longest arrays.
643

644 Although satDNAs were considered less mobile compared to TEs, which move extensively throughout
645 the genome ultimately forming interspersed repeats, our study showed that satDNAs also have very high
646 capacity to spread throughout the genome. It is common for TEs to spread through the genome in
647 recurrent bursts, which is characterized by the movement of many TEs through the genome within a
648 short evolutionary time [45]. Our results show that euchromatin satDNAs exhibit a similar pattern of
649 genomic dynamics, which includes the cycles of repeated bursts. Although it is assumed that spreading
650 bursts of TEs can occur for no apparent reason, it has been shown that they are often associated with

651 stressful situations such as extreme temperatures, radiation, chemical exposure, or viral infections [45].
652 In a view of a similar expansion pattern with TEs, it can be speculated that euchromatic satDNAs can
653 also enter in expansion cycles triggered by an external cause such as stressful living conditions. Another
654 indication that stress can affect satDNAs in euchromatin of *T. castaneum* is the finding that euchromatin
655 counterparts of pericentromeric TCAST satDNAs are more strongly expressed as response to the heat
656 stress [12].
657 To date, the three mechanisms have been proposed to be involved in the propagation of satDNA
658 through the euchromatin: (i) dispersion in a form of short arrays integrated as central repeats of non-
659 autonomous transposons elements [46]; (ii) the spread of satDNAs across long physical distances in
660 euchromatin through eccDNAs [47]; (iii) interlocus gene conversion events via 3D interactions between
661 loci in interphase nucleus [48].
662 There are a few examples where dispersed satDNAs have been found in the form of short arrays
663 integrated as central repeats of non-autonomous transposons elements [46,49,50]. Moreover, tandem
664 repeats within Tetris transposon DNA have been reported to be the basis for generating long satellite
665 arrays in *Drosophila virilis,* that eventually lose transposon features in adjacent regions [46]. Considering
666 that most of the Cast5 arrays have transposon-like sequences in adjacent regions, one of which shows
667 low similarities with the Mariner element, it is highly probable that Cast5 satDNA owes its expansion to
668 this element. However, other Cast satDNAs, except in some subsets, did not show a tendency to
669 associate with other mobile elements, but show very extensive propagation, indicating an efficient self-
670 propagation mechanism that operates at the inter- and intra-chromosome level. Mechanisms that have
671 recently been highlighted in *Drosophila* as possible drivers of satDNAs spread throughout euchromatin in
672 the X chromosome, are eccDNA reintegration and interlocus gene conversion [43]. To disclose the
673 putative mechanisms involved in the spread of euchromatin of satDNA, we analyzed junction regions.
674 The ends of the arrays, which could represent a critical sequence for the insertion of the array into the
675 environment, did not show a regular pattern, suggesting that gene conversion does not represent
676 mechanisms primarily involved in the spread. We found, however, that some Cast satDNAs have a
677 preferential insertion site in regions characterized by the presence of poly-A/T tracts with
678 microhomology. On the other hand, Cast satDNAs were shown to be present in the eccDNA fraction,
679 suggesting that the spread of Cast satDNA may be mediated by eccDNA. Based on the analyses of
680 neighboring regions, junctions, eccDNA fraction and considering the observed extensive spread of
681 satDNA, we proposed the two mechanisms responsible for euchromatin satDNAs genome dynamics and
682 evolution: transposition and eccDNA insertion.

683 In conclusion, this work has achieved two important accomplishments: (1) a high-quality reference
684 genome TcasONT of the insect model organism *T. castaneum*, 2) an in-deep genome-wide analysis of
685 the dominant euchromatic satDNAs using the new *T. castaneum* assembly. First, by using Oxford
686 Nanopore long-read sequencing, we have produced the most contiguous genome assembly of *T.*
687 *castaneum* to date, with significant improvement in the representation of the repetitive genome portion
688 and making a significant step toward completing the genome assembly of this scientifically and
689 economically important model organism. Our reference genome assembly has a great potential to be
690 important genomic resources for future research of both genes and various repetitive fractions of the
691 genome. Second, using the new assembly, our comprehensive analyses of the satDNAs, embedded in
692 euchromatin regions, revealed their massive spread and proliferation in gene-rich regions, challenging
693 the current hypothesis of a precluded coexistence of genes and abundant satellite DNA. Based on the
694 analyzed dynamics of satDNA spreading through the euchromatin, we suggested the evolutionary
695 scenario characterized by the recurrent burst, implying movement on the different chromosomes within
696 a short evolutionary time. Then, we propose two mechanisms that are most likely involved in efficient
697 propagation of satDNAs in euchromatin: transition by TEs and/or insertion by eccDNA. In summary, such
698 dynamical sequences embedded in euchromatin, which are subject to changes and rearrangements,
699 would have an extraordinary potential for rapid evolution of the genome and consequently of the
700 species itself. This opens a new perspective on satDNAs by considering them as inevitable parts of
701 euchromatin, thus stimulating new research involving transcriptomic and epigenetic studies, which
702 could disclose their role and putative influence on gene content.
703

704 ## Materials and methods

705 Insects sample and DNA preparation

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

718 Library preparation and sequencing

719 Oxford Nanopore library was prepared using SQK-LSK110 library preparation kit. Preparation was
720 performed according to the manufacturer's instructions with modifications described below. Prior to
721 sequencing DNA was sheared 30X times using a 30-gauge needle and purified of short fragments using
722 the Short Read Eliminator (SRE) XS kit (Circulomics) according to the manufacturers protocol. The
723 sheared and purified DNA was then used in library preparation protocols in amounts 2X larger than the
724 ONT protocol recommendation. Additionally, all recommended elution and incubation times were
725 extended to twice the length. These steps were performed to avoid library loss and to increase final
726 library concentration sufficient to load multiple MinION flow cell (final concentrations of the library
727 were >100 ng/µL allowing multiple loads from a single library). The flow cells were washed and reloaded
728 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
729 the data with a total cumulative output of 89.9 GB and an N50 of 20.1 kb. The software used for
730 sequencing was the Oxford Nanopore MinKnow 20.10.3.

731 Genome assembly

732 Data obtained from all sequencing runs was basecalled using Oxford Nanopore basecalling software
733 Guppy 5.0.1 with parameters as listed in the Suppl Table 14. The basecalled reads were then used in the

assembly process, using Canu v2.2 [26] with the parameters listed in the Suppl Table 14. According to Canu documentation the parameters were adjusted for high repetitiveness of the genome [22] and the high observed AT content in the reference Tas5.2 genome assembly [24]. The reads were filtered to >20 kb to reduce the computational intensity needed to perform the assembly, given the amount of data and the relatively small genome of the beetle. Incorporating only those reads that had a mean quality score of >7 and a length >20kb, a total of 44 Gb was used for the new genome assembly representing genome coverage of 224X (Suppl Table 1). Selected reads were up to 500 kb in length with a mean read length of 47 kb and mean quality score per read of 12.8. 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 [47] with default for gap-filling settings (Suppl 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 Nanopore-based long contigs from being 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 [29] 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 (Suppl 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 region 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 (Suppl Figure 1).

Polishing

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

772 BUSCO analysis

773 Benchmarking Single Copy orthologs (BUSCO) [50] analysis was performed using BUSCO v5.0.0. module
774 on the Galaxy web platform (usegalaxy.org). Same settings, as listed in the Suppl Table 14 were
775 repeated for all assembly validations.

776 Repeat detection

777 RepeatMasker is a widely used tool for finding and masking different repeat elements within a given
778 target sequence [51]. RepeatMasker was used to get the GFF/GTF formatted data with the position and
779 orientation of classified RepBase repeat elements. This was a base for the information about the
780 quantity, size, and distribution of different elements within examined genome assemblies. Assemblies
781 were annotated with different elements using the RepeatMasker program on the Galaxy web platform
782 (usegalaxy.org) with different repeat elements from the latest RepBase database (RELEASE 20181026) of
783 repeats and the "Hexapoda" species listing for clade specific repeats. We also rerun RepeatMasker for
784 the Tcas5.2 assembly to update the repeat annotations, as it was created using previous version of
785 RepBase database. For the quantification of three satDNAs classes (with defined monomer length >50;
786 50-500; >500bp) in TcasONT and Tcas5.2 assembly, Tandem Repeat Finder (TRF) program [52] was used
787 with default parameters.

788 Gene annotation

789 In order to map genes in Canu contigs (filtering) and in TcasONT assembly, we used the novel *LiftOff*
790 package [53] using gene annotations from the Tcas5.2 assembly. The Liftoff first maps the newly created
791 TcasONT assembly to the reference Tcas5.2 and then uses these overlaps to accurately align genes from
792 a reference Tcas5.2 assembly to a target TcasONT. This approach was limited to finding possible new
793 genes in the improved version of the TcasONT assembly, but because the Tcas5.2 assembly was

794 annotated based on a large database of RNA-seq reads combined with gene and general trait prediction
795 methods, this approach ensures gene completeness.

796 Identification of Cast1-Cast9 satDNAs

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

802 Analysis of Cast1-Cast9 satDNA arrays

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

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

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

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

817 A region of 50 kb upstream or downstream of the arrays was selected to define gene profiles around the
818 different Cast1-Cast9 arrays. In these 100-kb regions, 100 equal bins of 1 kb size (50 upstream and 50
819 downstream of a particular array) were created, and the number of exons was counted in each of them.
820 This allowed gene profiling around different Cast1-Cast9 arrays. Expected exon densities were calculated
821 by computing median, 1Q, and 3Q exon densities across the entire genome in 100-kb windows using a
822 custom R script.

823

824

825 Multiple sequence alignment and PCA

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

831 Tandem array region detection

832 To determine exact edges of Cast1-9 arrays in the genome, we employed a specific strategy.
833 Conventional monomer detection strategy uses a hard cutoff, based on monomer similarity, to
834 determine their relative positions. As edges of the arrays often contain degenerate monomers,
835 discovery of small homology and potential junction regions is often very difficult. Therefore, our strategy
836 was based on several steps. First, a database of all monomers of particular satDNA in the genome was
837 created together with a database of all arrays with their flanking regions (500 bp). To encompass all
838 possible variations of satDNA sequences, all k-mers (32 bp) were extracted from monomers. Also, k-
839 mers were extracted based on position in extracted arrays with flanks. For each position in the extended
840 array the k-mer from the monomer database, the closest match based on Hamming distance was found
841 and the score recorded. Next, rolling mean position score was calculated based on the mean of +/- 5
842 position scores. Real edges were determined by finding the minimum and maximum position for each
843 array with distance lower than 5. Based on these new edges, we extracted the surrounding and
844 microhomology regions.

845 Visualizations and calculations

846 All plots and calculations were produced in R using custom made data processing notebooks. Apart from
847 the standard libraries, we also used the *circilize* package [59] to create the circular visualization plots of
848 global genome patterns. To create the complex heatmaps used in the characterization of neighboring
849 regions similarity we used the *ComplexHeatmap* package [60]. Utilizing a graph-based visualization
850 method, we addressed the challenge of low variation among satDNA monomers and their tendency for
851 intra- and interchromosomal exchange and homogenization, as illustrated by the mixing observed in the

852 PCA plots. For each monomer in each array, we found the closest 5 monomers which do not belong to
853 the same array using *dist.dna* function from the *ape* package in R and "F81" genetic distance model.
854 Afterwards, we produced a graph network visualization using the *networkD3* package. Clustered and
855 connected dots in the network represent possible satDNA arrays involved in frequent exchange via
856 some of the proposed mechanisms contrary to the disconnected nodes. Homology of 20 bp regions
857 before and after arrays was visualized using *ggseqlogo* package in R, after alignment with MAFFT.

858 Extrachromosomal circular DNA analysis on two-dimensional agarose gel electrophoresis

859 Two-dimensional agarose gel electrophoresis was performed according to [51], with modifications. Total
860 DNA was isolated from 500 mg *T. castaneum* pupae by standard phenol-chloroform extraction and
861 dissolved in Tris-EDTA buffer, pH=8.0. The concentration was measured with the Qubit 4 fluorometer
862 (Invitrogen). 20 µg of DNA was passed 25 times through a 0.33 mm hypodermic needle to shear the
863 linear DNA. Given that the proportion of linear double-stranded DNA (dsDNA) fragments in the total
864 genomic DNA (gDNA) isolate greatly exceeds the amount of potential extrachromosomal DNA (eccDNA)
865 molecules, we treated the total gDNA isolate with exonuclease V, which cleaves linear dsDNA in both
866 directions starting at both 5' and 3' termini of linear dsDNA. The sheared DNA was digested with
867 exonuclease V (New England Biolabs) overnight at 37 °C to remove as much as possible linear dsDNA
868 fragments and keep the circular DNA intact. The reaction was stopped by adding 11 mM EDTA pH 8.0
869 and incubated at 70 °C for 30 minutes. The DNA was then purified using the Monarch PCR & DNA
870 Cleanup Kit (NEB). The first-dimension electrophoresis was run in 0.7% agarose in 1xTBE buffer at 0.7
871 V/cm for 18h. After electrophoresis, the agarose gel was stained with 1xTBE buffer containing 0.2 µg/ml
872 ethidium bromide. The lane with the separated DNA was cut out and the 1.5% agarose in 1xTBE
873 containing 0.2 µg/ml ethidium bromide was poured around the lane, which was positioned at a 90° angle
874 to the first electrophoresis. The second dimension was run at 4 V/cm for 3 h.

875 Southern blot hybridization

876 To ensure effective transfer of DNA from the agarose gel to the membrane, the gel was rinsed with
877 0.25M HCl for 30 minutes and then with 0.4M NaOH for a further 30 minutes. The DNA was transferred
878 to a positively charged nylon membrane (Roche Life Science) overnight by capillary transfer.
879 Hybridization probes for the Cast1, Cast2, Cast5 and Cast6 satellite DNA were labelled by biotin-16 dUTP
880 (Jena Biosciences) using PCR amplification of the cloned plasmids containing specific satDNA sequences.
881 Amplification was performed with satellite-specific primers (Cast1: 5' AAGTCGGCTACGACTAACCGTTC 3'

882 and 5' TTGCAAATTTGGATTCCGCCCGG 3'; Cast2: 5' TATACGCAAAATGAGCCGC 3' and 5'
883 AAAGTCGTAGAGCAATGCGG 3'; Cast5: 5' GGTGTTGAAAAGTCATAARTTGAGTG 3' and 5'
884 AGAGCCGGTGTACACAACATT 3'; Cast6: 5' CGACGCATGGGTCAATCTAAGACA 3' and 5'
885 ATTCGAAACTTTTCAAAAAAATTGG 3'). Hybridization was performed as described by [23]. Signals were
886 detected using streptavidin-alkaline phosphatase and the chemiluminescent substrate CDP-star (Roche
887 Life Science) and blots were visualised using the Alliance Q9 Mini (Uvitec) imager.

888 Data avaliability

889 The *T. castaneum* genome assembly TcasONT has been deposited to the European Nucleotide Archive
890 (ENA) under the BioProject accession PRJEB61413 and assembly accession number GCA_950066185.
891 The genome annotation and supplementary data presented in this study are available in FigShare at
892 https://doi.org/10.6084/m9.figshare.22683325.

893

894 Competing interest statement

895 The authors declare no competing interests.

896

897 Acknowledgments

898 We thank Raul Horvat for critical reading of the manuscript.

899

900 Funding

901 This work has been fully supported by Croatian Science Foundation under the project IP-2019-04-4910.

902

903

904

905

906

907

908

References

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

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

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

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

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

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

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

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

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

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

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

12. Feliciello I, Akrap I, Ugarković Đ. Satellite DNA Modulates Gene Expression in the Beetle Tribolium

945 castaneum after Heat Stress. *PLoS Genet*. 2015; doi: 10.1371/journal.pgen.1005466.

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

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

950 15. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al.. Telomere-to-telomere
951 assembly of a complete human X chromosome. *Nature*. Springer US; 2020; doi: 10.1038/s41586-020-
952 2547-7.

953 16. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, et al.. The complete sequence of
954 a human genome. *Science (80-)*. American Association for the Advancement of Science; 2022; doi:
955 10.1126/science.abj6987.

956 17. Klingler M, Bucher G. The red flour beetle T. castaneum: elaborate genetic toolkit and unbiased large
957 scale RNAi screening to study insect biology and evolution. *Evodevo*. BioMed Central; 2022; doi:
958 10.1186/s13227-022-00201-9.

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

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

964 20. Gržan T, Despot-Slade E, Meštrović N, Plohl M, Mravinac B. CenH3 distribution reveals extended
965 centromeres in the model beetle *Tribolium castaneum*. Malik HS, editor. *PLoS Genet*. 2020; doi:
966 10.1371/journal.pgen.1009115.

967 21. Ugarković D, Podnar M, Plohl M. Satellite DNA of the red flour beetle *Tribolium castaneum* -
968 Comparative study of satellites from the genus *Tribolium*. *Mol Biol Evol*. 1996; doi:
969 10.1093/oxfordjournals.molbev.a025668.

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

972 23. Pavlek M, Gelfand Y, Plohl M, Meštrović N. Genome-wide analysis of tandem repeats in *Tribolium*
973 *castaneum* genome reveals abundant and highly dynamic tandem repeat families with satellite DNA
974 features in euchromatic chromosomal arms. *DNA Res*. Oxford Academic; 2015; doi:
975 10.1093/dnares/dsv021.

976 24. Flynn JM, Long M, Wing RA, Clark AG, Arkhipova I. Evolutionary Dynamics of Abundant 7-bp
977 Satellites in the Genome of *Drosophila virilis*. *Mol Biol Evol*. 2020; doi: 10.1093/molbev/msaa010.

978 25. Herndon N, Shelton J, Gerischer L, Ioannidis P, Ninova M, Dönitz J, et al.. Enhanced genome assembly
979 and a new official gene set for *Tribolium castaneum*. *BMC Genomics*. BMC Genomics; 2020; doi:
980 10.1186/s12864-019-6394-6.

981 26. Brown SJ, Henry JK, Black IV WC, Denell RE. Molecular genetic manipulation of the red flour beetle:
982 Genome organization and cloning of a ribosomal protein gene. *Insect Biochem*. 1990; doi: 10.1016/0020-
983 1790(90)90011-I.

27. 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.

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

29. 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.

30. Alonge M, Lebeigle L, Kirsche M, Aganezov S, Wang X, Lippman ZB, et al.. Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. *bioRxiv*. :2021.11.18.4691352021;

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

32. Mota APZ, Koutsovoulos GD, Perfus-Barbeoch L, Despot-Slade E, Labadie K, Aury J-M, et al.. Unzipped genome assemblies of polyploid root-knot nematodes reveal unusual and clade-specific telomeric repeats. 2024; doi: 10.1038/s41467-024-44914-y.

33. Despot-Slade E, Mravinac B, Širca S, Castagnone-Sereno P, Plohl M, Meštrović N. The Centromere Histone Is Conserved and Associated with Tandem Repeats Sharing a Conserved 19-bp Box in the Holocentromere of Meloidogyne Nematodes. *Mol Biol Evol*. 2021; doi: 10.1093/molbev/msaa336.

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

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

36. Altemose N, Logsdon GA, Bzikadze A V., Sidhwani P, Langley SA, Caldas G V., et al.. Complete genomic and epigenetic maps of human centromeres. *Science*. NIH Public Access; 2022; doi: 10.1126/SCIENCE.ABL4178.

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

38. Charlesworth D, Charlesworth B, Marais G. Steps in the evolution of heteromorphic sex chromosomes. *Heredity (Edinb)*. 2005; doi: 10.1038/sj.hdy.6800697.

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

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

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

2159-2.

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

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

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

45. Mérel V, Boulesteix M, Fablet M, Vieira C. Transposable Elements in *Drosophila melanogaster*. *Mob DNA II*. Wiley;

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

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

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

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

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

51. Cohen S, Lavi S. Induction of Circles of Heterogeneous Sizes in Carcinogen-Treated Cells: Two-Dimensional Gel Analysis of Circular DNA Molecules. *Mol Cell Biol*. 1996; doi: 10.1128/mcb.16.5.2002.