Identification of a genome-specific repetitive element in the *Gossypium* D genome

Hejun Lu[1,2*], Xinglei Cui[1*], Yanyan Zhao[1], Richard Odongo Magwanga[1,3], Pengcheng Li[1], Xiaoyan Cai[1], Zhongli Zhou[1], Xingxing Wang[1], Yuling Liu[4], Yanchao Xu[1], Yuqing Hou[1], Renhai Peng[4], Kunbo Wang[1,5+], Fang Liu[1+]

[1] Research Base of Tarium University, State Key Laboratory of Cotton Biology, Institute of Cotton Research of Chinese Academy of Agricultural Science, Anyang, Henan, 455000, China
[2] Gembloux Agro-Bio Tech, University of Liège, Gembloux, 5030, Belgium
[3] School of Biological and Physical Sciences (SBPS), Jaramogi Oginga Odinga University of Science and Technology (JOOUST), P.O Box 210-40601, Bondo-Kenya
[4] Anyang Institute of Technology, Anyang, Henan, 455000, China
[5] Tarium University, Alar, Xinjiang. 843300, China
[*]These authors contributed equally to this work.
[+]Correspondence: Kunbo Wang, E-mail: wkbcri@163.com; Fang Liu, liufcri@163.com

Abstract: The activity of genome-specific repetitive sequences is the main cause of genome variation between *Gossypium* A and D genomes. Through comparative analysis of the two genomes, we retrieved a repetitive element termed *ICRd* motif, which appears frequently in the diploid *Gossypium raimondii* (D_5) genome but rarely in the diploid *Gossypium arboreum* (A_2) genome. We further explored the existence of the *ICRd* motif in chromosomes of *G. raimondii*, *G. arboreum*, and two tetraploid (AADD) cotton species, *Gossypium hirsutum* and *Gossypium barbadense*, by fluorescence *in situ* hybridization (FISH), and observed that the *ICRd* motif exists in the D_5 and D-subgenomes but not in the A_2 and A-subgenomes. The *ICRd* motif comprises two components, a variable tandem repeat (TR) region and a conservative sequence (CS). The two constituents each have hundreds of repeats that evenly distribute across 13 chromosomes of the D_5 genome. The *ICRd* motif (and its repeats) was revealed as the common conservative region harbored by ancient Long Terminal Repeat Retrotransposons. Identification and investigation of the *ICRd* motif promotes the study of A and D genome differences, facilitates research on *Gossypium* genome evolution, and provides assistance to subgenome identification and genome assembling.

Keywords: *Gossypium*; cotton plant, D genome; repetitive element, genome-specific; fluorescence *in situ* hybridization (FISH); evolution

1. Introduction

Repetitive DNA sequences are common in eukaryotic genomes, and account for a large fraction of the host genome (Ibarra-Laclette et al., 2013). Their amount is highly correlated with the size of the host genome (Feschotte, 2008). Repetitive DNA is divided into two major groups based on their structures: tandem repeats and interspersed repeats (Jurka et al., 2005). Tandem repeats are known as simple sequence repeat (SSR), and include micro-satellites, mini-satellites, and satellites (Jeffreys, et al., 1990). Interspersed repeats are also referred to as transposable elements (TEs) or transposons.

After the first TE of Ac/Ds was reported in maize (McClintock, 1950; Brink & Williams, 1973; Goldschmidt, 2002), further TEs have been identified in many eukaryotic species (Munoz-Lopez & Garcia-Perez, 2010). There are thousands of different TE families in plants, which display extreme diversity (Sanmiguel & Bennetzen, 1998; Bennetzen, 2005; Morgante, 2006). Finnegan first proposed a TE classification system, which includes two classes based on their transposition mechanisms, viz., those mediated by RNA (Retrotransposons) and those by DNA (DNA transposons) (Bowen & Jordan, 2002; Wessler, 2006; Arkhipova, 2018). Wicker unified TEs nomenclature and classification by applying mechanistic and enzymatic criteria (Wicker et al., 2007). TEs play important roles in the genome through diverse ways, such as variation in intron size (Deutsch & Long, 1999; Zhang et al., 2011; Koonin, Csuros & Rogozin, 2013), segmental duplication (Del Pozo & Ramirez-Parra, 2015), transfer of organelle DNA to the nucleus (Adams & Palmer, 2003), expansion/contraction of tandem repeats, and illegitimate recombination

46 (Finnegan, 1989; Koike, Nakai & Takagi, 2002). Long Terminal Repeat Retrotransposons (LTR-TEs), which are
47 usually scattered throughout genomes, are the most abundant TE type and can cause genome expansion over a short
48 evolutionary period particularly in plants (Piegu et al., 2006). The investigation of genome-specific TE is an
49 efficient approach to studying species formation and genome evolution (Dong et al., 2018).

50   *Gossypium,* a genus of flowering plants from which cotton is harvested, diverged from the common ancestor
51 with *Theobroma cacao* approximately 33.7 million years ago (MYA) (Wang et al., 2012). *Gossypium* comprises
52 eight diploid (2n=2x=26) genomic groups: A, B, C, D, E, F, G, K, and one allotetraploid (2n=4x=52) genomic
53 group: AD (Wang, Wendel & Hua, 2018). *Gossypium* species are good subjects for research on polyploidization,
54 genomic organization and genome-size variation because of their high genome diversity: from the smallest New
55 World D genome with an average of 885 Mb to the Australian K-genome with an average of 2576 Mb (Hendrix &
56 Stewart, 2005). The accumulation of different lineage-specific TEs was thought to be responsible for the variation in
57 genome size in *Gossypium* genomic groups (Hawkins et al., 2006; Lu et al., 2018). Of the eight genomic groups, the
58 A and D groups are the main ones investigated in cotton genomics research (Du et al., 2018). *Gossypium hirsutum*,
59 the major cultivated cotton species, is known to have originated from the progenitors of *G. arboreum* (A_2) and *G.*
60 *raimondii* (D_5) (Paterson et al., 2012). The key phenotype difference between *G. arboreum* and *G. raimondii* is the
61 production of spinnable fibers in the former but not the latter. In terms of the genomics, the former has a genome
62 size of 1,746 Mb/1C, which is about two times that of the latter (885 Mb/1C) (Hendrix & Stewart, 2005). Genome
63 sequencing showed that the difference in the numbers of protein-coding genes between the A (41,330) and D
64 (37,505) genomes is not obvious, while the lineage-specific TE content is the main reason for the size gap between
65 the A and D genome (Li et al., 2015; Du et al., 2018). Moreover, Wang, Huang & Zhu (2016) suggested that the
66 transposable elements play an important role during cotton genome evolution and fiber cell development. Thus,
67 research on the lineage-specific repetitive sequences between A and D genomes is a meaningful path to investigate
68 speciation dynamics.

69   Fluorescence *in situ* hybridization (FISH) is a versatile tool to visualize the distribution of certain DNA
70 sequences in chromosomes and plays a vital role in cytogenetic research. In tetraploid cotton, FISH has played a key
71 role in cytological experiments that have contributed to the understanding the allotetraploid event. FISH with DNA
72 segments harboring dispersed repeats has identified genome-specific repeats between the A and D genome, and
73 showed that some A genome repeats appear to have spread to the D genome (Hanson, Zhao et al. 1998; Zhao, Si et
74 al., 1998). Although the repetitive DNA fragments are more common in the A than in the D genome, one tandem
75 repeat family (B77) has been well-characterized from the D Chromosome (Zhao, Ji et al. 1998). Recently, more
76 repetitive sequences were observed with FISH in the cotton genome after construction of a cotton cytogenetic map
77 (Cui, Liu et al., 2015; Liu, Peng et al., 2016). Lu et al. (2018) suggested that *CICR* was an important contributor to
78 the size gap between the A and D genome. The identification and localization of these repetitive sequences benefit
79 genome assembly and facilitate understanding of the mechanism of genome evolution.

80   The D genomic group represents a diverse group of diploids that diverged from a branch of A, B, C, E, F, G,
81 and K genomic groups about 5-10 MYA (Senchina et al., 2003). Although the D genome has the smallest size of all
82 *Gossypium* species, this study has revealed the presence of a set of repeat elements with high proliferation, which is
83 absent in the A genome. The discovery and characterization of these novel repetitive elements provides components
84 for a repetitive sequences database and new insight into *Gossypium* evolution.

85 **2. Materials and methods**

86 **2.1 Plant materials**

87   Cotton plants were obtained from the National Wild Cotton Nursery in Hainan Island, China, sponsored by the
88 Institute of Cotton Research of Chinese Academy of Agricultural Sciences (ICR-CAAS). They were also conserved
89 in the greenhouse at ICR-CAAS' headquarters in Anyang City, Henan Province, China. The DNA and cells came
90 from specimens listed in Table 1, which is based on the latest nomenclature of *Gossypium* species (Wang, Wendel &
91 Hua, 2018).

92   The repeat elements were characterized in the *G. raimondii* genome (Paterson et al., 2012), and compared to
93 genomes in other *Gossypium* genomes, viz., *G. arboreum* (Li, Fan et al. 2014), *G. hirsutum* (AD)$_1$ (BGI (Li, Fan et
94 al., 2015), NBI (Zhang, Hu et al., 2015), HAU (Wang, Tu et al., 2019), ZJU (Hu, Chen et al., 2019)), *G. barbadense*
95 (AD)$_2$ (HAUv1 (Yuan et al., 2015), CAS (Liu et al., 2015), HAUv2 (Wang, Tu et al., 2019), and ZJU (Hu, Chen et
96 al., 2019)). All genome data was downloaded from Cottongen (https://www.cottongen.org/), except the (AD)$_2$-CAS
97 which was obtained from GenBank under PRJNA251673.

98 **2.2 Characterization of the repetitive element and bioinformatics analysis**

99   BLASTN (v2.6.0) (Camacho, Coulouris et al., 2009) was used to identify repeat elements in the genomes of
100 the plant material, and in the genomes stored in the databases. We used a threshold of greater than or equal to 80%
101 matching ratio and an 80% similarity following the 80-80 rule suggested by Wicker et al. (2007). The tandem
102 repeats (TRs) were identified with Tandem Repeats Finder (v4.09) (Benson, 1999). We used Perl script for batch
103 extracting sequences from the genome (Doc. S1). Sequence alignments were obtained from MUSCLE (v3.81)
104 (Edgar, 2004). The Unipro UGENE (v1.31) was used to present the alignments and train consensus sequences
105 (Okonechnikovet al. 2012). The inner enzyme annotation was obtained by online CD-search in NCBI (Marchler-
106 Bauer et al., 2017). GIRI Repbase (Chen et al., 2007) were queried for annotation. RepeatMasker (v4.07) was used
107 to annotate the insertions and estimate the proportion of repetitive sequences in genomes
108 (http://www.repeatmasker.org).
109   Flanking LTRs of LTR-TEs were identified with LTRharvest (v1.5.8) (Ellinghaus, Kurtz & Willhoeft, 2008).
110 Subsequently, Vmatch (v2.3.0) was used to cluster the LTRs (Kurtz, 2003). The divergence time of the LTR-TEs
111 was estimated using the formula T = d/2r, where r represents a substitution rate of 1.3×10^{-8} per site per year (Ma &
112 Bennetzen, 2004), and d represents the distances of paired LTRs, which was calculated based on the Kimura two-
113 parameter (Kimura, 1980). The insertions of the repeat elements were obtained based on the BLASTN result, and
114 the LTR-TE and Coding sequence (CDS) information was obtained from genome annotation (Paterson et al., 2012),
115 which were illustrated by the ggplot2 R package (Wickham 2016) with a sliding 500 kb window for LTR-TE and
116 CDS. The synteny blocks of the homologous segments were shown by a Perl script (Doc. S1) based on the BLASTN
117 results.

118 **2.3 Fluorescence in situ hybridization (FISH)**

119   A probe was designed with the PCR product of the *ICRd* motif, which was obtained from the forward primer:
120 TTCTATTTTATCCATCGTTATG, reverse: GGAGATAGGATTTGTTGCT; and followed the amplification
121 procedure: firstly, 95°C for 5 min of pre-degeneration; then, 30 cycles at 95°C for 30 s, 52°C for 30 s, and 72°C for
122 2 min. The final extension was done at 72°C for 6 min. Composition of the reaction mix used the following: gDNA
123 (~5 µg/ml), primers (~0.8 µM), PCR Master Mix (Thermo), and H$_2$O. The gDNA was extracted from the leaves of
124 the cotton plants (Table 1). The probe was purified and labeled with digoxigenin-dUTP via nick translation,
125 according to manufacturer's instructions (Roche Diagnostics, USA). Mitotic chromosome preparation and FISH
126 procedures were conducted using a modified protocol (Wang et al., 2001).

127 **3. Results**

128 **3.1. One specific repetitive sequence in the *Gossypium* D$_5$ genome**

129   We performed BLAST to query all of the interspersed repetitive sequences of *G. raimondii* (Paterson et al.,
130 2012) with the whole genome of *G. arboreum* (A$_2$) (Li et al., 2014). One segment in the *G. raimondii* (D$_5$) genome
131 (Chr05: 50639971-50641791) was filtered out and recognized as D$_5$ genome-specific. This sequence repeats
132 frequently and is spread over 13 chromosomes of the D$_5$ genome (Supplementary Table 1), while it is absent from
133 the A$_2$ genome. Searches in Repbase (Chen et al., 2007) and NCBI found no related annotation and LTRharvest
134 (Ellinghaus, Kurtz & Willhoeft, 2008) and a CD-search (Marchler-Bauer et al., 2017) revealed it is neither LTR nor
135 a coding sequence.

136   Manual inspection revealed the structure of the genome-specific sequence as having two constituents: a tandem
137   repeats array (referred as TR hereafter) composed of 133 bp basic units, and an unknown conservative sequence
138   (referred as CS hereafter) (Figure 1). Based on this feature, we identified 72 sequences in total from the D_5 genome
139   with RepeatMasker (Supplementary Table 2), referred to here as the *ICRd* motif following our previous work (Lu et
140   al., 2018). Among the 72 *ICRd* motifs, the TRs are length-variable having 2-20 times of basic units (Figure 2a),
141   while the CSs are stable and have an average size ~ 860 bp.
142   To verify the genome specificity and chromosome distribution of the *ICRd* motif, we used the PCR product of
143   the *ICRd* motif from *G. raimondii* to design the probe for FISH on the mitotic chromosomes of diploid A_2 and D_5,
144   and tetraploid *G. hirsutum* ($(AD)_1$) and *G. barbadense* ($(AD)_2$). The probe generated bright signals covering all the
145   chromosomes of the D_5 and D-subgenome, but no signals on the A_2 and A-subgenome (Figure 3). These cytogenetic
146   inspections were in accordance with the genomic comparative analysis and further revealed that the *ICRd* motif is a
147   genome-specific and highly repetitive element in the D_5 genome, as well as in the D-subgenome of tetraploid cotton.

3.2 LTR-TEs inserted with the *ICRd* motif

149   We compared the insertion loci of 72 *ICRd* motifs with the whole genome repeats annotation (gff file) of the
150   D_5 genome (Paterson et al., 2012) and found that each of the motifs is one-to-one harbored within the 72 LTR-TEs
151   (Supplementary Table 3), which meant the former is the inner part of the latter.
152   We extracted the 72 LTR-TEs sequences from the D_5 genome and parsed their structure, which showed all
153   sequences are incomplete, lacking either enzyme or flanking LTRs, the required elements for an intact LTR-TE
154   (Wicker et al., 2007). A consensus accumulation histogram obtained from aligning all of these LTR-TEs together
155   (Supplementary Figure 1) showed these TEs to have a vast sequence variation and a single conservative region
156   representing the insertion region of the *ICRd* motif (Figure 4). The *ICRd* motif appears to be more stable than other
157   parts of the TEs along with degradation and evolution. This stability implies the importance of *ICRd* motif to the
158   TEs.
159   Of the 72 LTR-TEs, 25 were identified as having paired flanking LTRs, and were used to represent the
160   classification and evolution of these TEs. The LTR cluster results showed that, except for two TEs having similar
161   LTR regions, the other 23 TEs are totally different from each other, indicating that they do not belong to the same
162   family based on the LTR similarity rules (Wicker et al., 2007). The estimated active date curve of these TEs –
163   almost all prior to 10 MYA and peaking at ~30 MYA (Figure 5) – shows the peak is close to the time that *G.
164   raimondii* and *T. cacao* diverged approximately 33.7 MYA (Wang et al., 2012), far earlier than the putative
165   divergence time of the *Gossypium* A and D genomes (Wendel & Cronn, 2001). These results indicate that these
166   LTR-TEs are ancient and potentially contributed to speciation of *Gossypium*.

3.3 Abundant constituents of the *ICRd* motif in the D_5 genome

168   To further analyze the genomic features of the *ICRd* motif, we separately investigated the content and
169   distribution of its two constituents (TR and CS) in the D_5 genome (Figure 6a). In total 350 *TR* insertions were
170   detected (Supplementary Table 2). Insertions varied in length (due to the unit repeating at different times) between
171   2–21, but mainly 2–10 times the basic unit length (Figure 2b). The longest TR insertion in D_5 ($D_5$03: 25689303–
172   25697234) was an extraordinary 61 units up to 8 kb; how it was formed is unknown. On the other hand, a total of
173   463 CSs were found (Supplementary Table 2). Combining the analyses of the insertion loci of the two constituents,
174   we found 72 TRs and 72 CSs constituting the *ICRd* motifs (Figure 1).
175   Further analysis showed that the TR and CS were evenly distributed on the chromosomes based on an χ^2 test,
176   with the number of insertions being proportional to the size of the chromosome [TR insertions, $\chi^2 = 5.56$ (df = 12, P
177   > 0.9); CS insertions, $\chi^2 = 9.08$ (df = 12, P > 0.5)]. The even distributions meant that the CS and *TR* are possible
178   ancient repetitive sequences that have evolved along with the chromosomes. Previous *G. raimondii* genome
179   sequencing work reported that most TEs in *G. raimondii* are deletion derivatives lacking the domains that are
180   typically necessary for transposition and that only 3% of LTR base pairs derived from full-length LTR-TEs

181 (Paterson et al., 2012). We show that hundreds of constituents of the *ICRd* motif in D_5 are potentially the fragments
182 produced from the ancient LTR-TEs.

3.4 Disappearance of the *ICRd* motif from *Gossypium*

184 Aiming to observe the disappearance of the *ICRd* motif in the newly formed *Gossypium* A genome, we
185 selected two homologous segments from the highly collinear Chromosome 1 of *G. raimondii* ($D_5$01) and *G.*
186 *arboreum* ($A_2$01) (Li et al., 2014), respectively. The segment from Chromosome 1 of *G. raimondii* ($D_5$01) harbored
187 one *ICRd* motif and its homologous segment from $A_2$01 was obtained from homologous SSR markers
188 (Supplementary Table 4). The illustration of the syntenic block of the two segments showed that the *ICRd* motif
189 together with its host LTR-TE were lost on the $A_2$01 segment, while their up- and downstream conservative regions
190 remained (Figure 7). In the upstream, we observed two insertions rich in repeat sequences especially on the $A_2$01
191 segment (Supplementary Table 4), which was potentially due to a recent TE expanding event happening in the A
192 genome (Lu et al., 2018). Thus, we observed that the *ICRd* motifs and host LTR-TEs were lost from the genome
193 with the recent formation of the A genome (Wendel & Cronn, 2001; Wendel, Flagel & Adams, 2012), but remained
194 in the D genome despite mass damage accumulation.

3.5 Distributions of *ICRd* motifs in tetraploid cotton

196 Tetraploid cotton, *G. hirsutum* and *G. barbadense,* are the major cultivated fiber-producing cotton species.
197 Research on the genome of these two species is an important approach to improving cotton yield and quality.
198 However, due to the large number of homologous segments between A and D-subgenomes, the tetraploid cotton
199 genome assemblage has been a great challenge to molecular geneticists (Bowers et al., 2003; Chen et al., 2007b).
200 Through high-throughput sequencing methods, two versions of the *G. hirsutum* genome assembly (($AD)_1$-BGI (Li et
201 al., 2015), ($AD)_1$-NBI (Zhang et al., 2015)), and two *G. barbadense* versions ($AD)_2$-HAU (Yuan et al., 2015) and
202 ($AD)_2$-CAS (Liu et al., 2015)) were completed in 2015. With the advance of sequencing techniques, the tetraploid
203 genome assemblies were improved in quality (Hu, Chen et al., 2019, Wang, Tu et al., 2019). However, to benefit
204 research in the post-genome era, such as facilitating molecular breeding of cotton, suitable evaluation is needed to
205 provide accurate reference data. Application of the lineage-specific repetitive element and the *ICRd* motifs are
206 important tools in evaluating the quality of the genome assembly of tetraploid cotton.

207 To observe the assembling quality of the *ICRd* motif in tetraploid genomes, we queried it with BLAST in all
208 published tetraploid cotton genomes, including four versions of *G. hirsutum* (($AD)_1$) and four versions of *G.*
209 *barbadense* (($AD)_2$) (Table 2). In the case of ($AD)_1$, the two recently published (Hu, Chen et al. 2019; Wang, Tu et
210 al. 2019) versions and the previous NBI version were in agreement with our FISH inspection results, viz., that the
211 *ICRd* motifs only generated the signals on the D-subgenome chromosomes (Figure 3). However, the BGI version (Li
212 et al. 2015) is inconsistent with the FISH results in that the *ICRd* motif was misassembled into the A-subgenome.
213 For the ($AD)_2$ assemblies, the two newly published (Hu, Chen et al. 2019; Wang, Tu et al. 2019) and CAS versions
214 were better assembled than the HAUv1 version. The HAUv1 showed the number of matches in the chromosome-
215 unassembled scaffolds, while the HAUv2 has improved quality (Supplementary Table 5). This means that with
216 advances in genome sequencing techniques, tetraploid genomes can be more precisely assembled though the
217 existence of homologous segments from At and Dt.

4. Discussion

4.1 Identification of *ICRd* motif and *Gossypium* evolution

220 TEs have played an important function in *Gossypium* speciation and the accumulation of different genomic-
221 specific TEs were thought to be responsible for genome-size variation in *Gossypium* (Hawkins et al., 2006).
222 Through FISH inspection, some A genome-specific repetitive elements have been well identified and characterized
223 (Liu et al., 2016), but similar work in the D genome have been rare; this may be because the genome-specific
224 repetitive sequences in the A genome are much more numerous than in the D genome (Liu et al., 2018). However, in

225 the present study, starting with comparative genomic data, we have screened out one kind of specific sequence in the
226 D genome, and subsequently, we have identified and characterized TEs.
227   The TEs harboring the *ICRd* motif showed an ancient active date of much earlier than 10 MYA, while the time
228 of divergence of the A and D genomes from the common ancestor is estimated to have occurred 5-10 MYA (Grover
229 et al., 2004). Thus the *ICRd* motifs have existed in the ancestor of A and D genome, while disappeared along with
230 the formation of the A genome. Previous researchers have considered that the accumulation of lineage-specific TEs,
231 which is thought to be responsible for the variation of *Gossypium* genomes (Hawkins et al., 2006), and the LTR-TE
232 activities after 5 MYA mainly contributed to the two-fold size difference of the A and D genomes (Zhang et al.,
233 2015). Based on our analysis, we presumed that as in the activity of new repetitive sequences the extinction of
234 ancient repetitive sequences, such as the disappearance of the *ICRd* motif in the A genome, also contributed
235 significantly to genome evolution. Through FISH, we observed that the *ICRd* motifs were only distributed in D-
236 subgenome chromosomes, and the results were in agreement with a previous study which reported that the TE have
237 proliferated in the progenitor genomes but were retained after allopolyploid formation in the D-subgenome (Zhang
238 et al., 2015).

239 **4.2 *ICRd* motif support cytogenetic markers for tetraploid cotton**

240   The identification of the *ICRd* motif provides a new subgenome marker for the accurate assembling of
241 tetraploid cotton (Chen et al., 2007). Chromosome identification is the foundation of plant genetics, evolution and
242 genomics research (Saranga, 2007; Xie et al., 2012). Although many species have been sequenced, the rapid
243 identification of the subgenome is still useful in applied research. FISH has been used as a reliable cytological
244 technique for chromosome identification in many species (Wang, Guo & Zhang, 2007), but has only been used
245 recently for the identification of cotton chromosomes (Gan et al., 2012). In the present study, the identified *ICRd*
246 motifs can be used as a new cytological marker in *Gossypium*, especially in tetraploids. Further, the repetitive
247 sequence probes are easier and more successfully detected than other probes. Several similar markers have been
248 reported (Liu et al., 2016). The addition of these new cytological markers will enrich the marker database for
249 chromosome identification and facilitate cotton genomic studies.
250   Eukaryotic genomes have a high proportion of TEs and these TEs make eukaryotic genome assembly much
251 more difficult than simple genome assembly (Treangen & Salzberg, 2012). Many reported genome sequences have
252 gaps because of the high proportion of TEs (Adams et al., 2000). Allopolyploid genomes are especially difficult to
253 assemble homologous fragments from subgenomes (Chen et al., 2007). Incorrect assembling of the genomes leads to
254 ambiguity in research which, in turn, produces biases and errors when interpreting results (Adams et al., 2000). The
255 repetitive sequences analysis in this work were screened out from the whole genome comparison, we characterize
256 the distribution feature on referenced genome assembly, moreover, FISH observation on chromosomes of somatic
257 cell verified the lineage-specific feature. Combining FISH with genome-specific repeat segments is a convenient and
258 practical approach to observe chromosome differences, in addition to assisting polyploid genome assembling, and
259 evaluating assembling accuracy. With the progress of genome sequencing and assembling, genome assembly will
260 become increasingly more precise and convincing, and it is likely that the latter published tetraploid genome will
261 adopt the BioNano and Hi-C approaches (Hu, Chen et al., 2019, Wang, Tu et al., 2019) and improve the
262 identification of homologous segments from subgenomes. The improved tetraploid cotton genome assemblies were
263 consistent with FISH, which provides a reference for researchers deciding which genomes to adopt in their research.

264 **5. Conclusions**

265   We identified and characterized a new type of repetitive sequence termed *ICRd* motif in the *Gossypium* D
266 genome. The motifs are interspersed in 13 chromosomes of the D genome, but absent in the A genome, and retained
267 in D-subgenome in tetraploid cotton. We analyzed their structure, genomic distribution, affiliation, and evolution,
268 which revealed a conserved region harbored in ancient LTR-TEs. The identification and characterization of the
269 *ICRd* motif provided new insight into understanding TE evolution along with the formation of cotton genomes as
270 well as providing a convenient and practical tool to distinguish the A and D genome subsets of the tetraploid cotton

271 genome assembly. The *ICRd* motif has a novel structure and affiliation; how the structure was formed and what
272 function the *ICRd* motif plays in LTR-TEs would be valuable areas for future research.

273 **Supplementary materials:**, Figure S1: Supplementary Figure 1. The whole alignment of the 72 LTR-TEs, Table S1: Blast of the
274 1.8 kb sequences in *G. raimondii* genome, Table S2: The *ICRd* motifs and their constituents, Table S3: The structures of the
275 LTR-TEs harboring the *ICRd* motif, Table S4: Information on the two homologous segments, Table S5: Blast results of the ICRd
276 motif with tetraploid cotton.

277 **Acknowledgments:** We are indebted to Dr Syed Shan-e-Ali Zaidi of the University of Liège, Belgium, for his guidance in
278 analysis and interpretation of the data

279 References

280 Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA,
281   Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell
282   MD, Zhang Q, Chen LX, Brandon RC, Rogers YHC, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C,
283   Baxter EG, Helt G, Nelson CR, Gabor Miklos GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C,
284   Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos P V.,
285   Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC,
286   Busam DA, Butler H, Cadieu E, Center A, Chandra I, Michael Cherry J, Cawley S, Dahlke C, Davenport LB,
287   Davies P, de Pablos B, Delcher A, Deng Z, Deslattes Mays A, Dew I, Dietz SM, Dodson K, Doup LE,
288   Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S,
289   Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Harley
290   Gorrell J, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D,
291   Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA,
292   Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J,
293   Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina N
294   V., Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL,
295   Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri
296   V, Reese MG, Reinert K, Remington K, Saunders RDC, Scheeler F, Shen H, Christopher Shue B, Siden-
297   Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas
298   R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM,
299   Weissenbach J, Williams SM, Woodage T, Worley KC, Wu D, Yang S, Alison Yao Q, Ye J, Yeh RF, Zaveri
300   JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO,
301   Gibbs RA, Myers EW, Rubin GM, Craig Venter J. 2000. The genome sequence of Drosophila melanogaster.
302   *Science* 287:2185–2195. DOI: 10.1126/science.287.5461.2185.

303 Adams KL, Palmer JD. 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus.
304   *Molecular phylogenetics and evolution* 29:380–95.

305 Arkhipova IR. 2018. Neutral theory, transposable elements, and eukaryotic genome evolution. *Molecular Biology*
306   *and Evolution* 35:1332–1337. DOI: 10.1093/molbev/msy083.

307 Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. *Current*
308   *Opinion in Genetics and Development* 15:621–627. DOI: 10.1016/j.gde.2005.09.010.

309 Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research* 27:573–
310   580. DOI: 10.1093/nar/27.2.573.

311 Bowen NJ, Jordan IK. 2002. Transposable elements and the evolution of eukaryotic complexity. *Current Issues in*
312   *Molecular Biology* 4:65–76.

313 Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic
314   analysis of chromosomal duplication events. *Nature* 422:433–438. DOI: 10.1038/nature01521.

315 Brink RA, Williams E. 1973. Mutable R-navajo alleles of cyclic origin in maize. *Genetics* 73:273–296.

316 Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer and T. L. Madden (2009). "BLAST+:
317   architecture and applications." BMC bioinformatics 10(1): 421. DOI: 10.1186/1471-2105-10-421.

318 Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Guo W, Chen X, Stelly DM, Rabinowicz PD, Town CD,
319   Arioli T, Brubaker C, Cantrell RG, Lacape J-M, Ulloa M, Chee P, Gingle AR, Haigler CH, Percy R, Saha S,
320   Wilkins T, Wright RJ, Van Deynze A, Zhu Y, Yu S, Abdurakhmonov I, Katageri I, Kumar PA, Mehboob-ur-
321   Rahman, Zafar Y, Yu JZ, Kohel RJ, Wendel JF, Paterson AH. 2007. Toward Sequencing Cotton (Gossypium)
322   Genomes. *Plant Physiology* 145:1303–1310. DOI: 10.1104/pp.107.107672.

323 Del Pozo JC, Ramirez-Parra E. 2015. Whole genome duplications in plants: An overview from Arabidopsis. *Journal
324   of Experimental Botany* 66:6991–7003. DOI: 10.1093/jxb/erv432.

325 Deutsch M, Long M. 1999. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Research* 27:3219–
326   3228. DOI: 10.1093/nar/27.15.3219.

327 Dong G, Shen J, Zhang Q, Wang J, Yu Q, Ming R, Wang K, Zhang J. 2018. Development and Applications of
328   Chromosome-Specific Cytogenetic BAC-FISH Probes in S. spontaneum. *Frontiers in Plant Science* 9. DOI:
329   10.3389/fpls.2018.00218.

330 Du X, Huang G, He S, Yang Z, Sun G, Ma X, Li N, Zhang X, Sun J, Liu M, Jia Y, Pan Z, Gong W, Liu Z, Zhu H,
331   Ma L, Liu F, Yang D, Wang F, Fan W, Gong Q, Peng Z, Wang L, Wang X, Xu S, Shang H, Lu C, Zheng H,
332   Huang S, Lin T, Zhu Y, Li F. 2018. Resequencing of 243 diploid cotton accessions based on an updated A
333   genome identifies the genetic basis of key agronomic traits. *Nature Genetics* 50:796–802. DOI:
334   10.1038/s41588-018-0116-x.

335 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids
336   research* 32:1792–7. DOI: 10.1093/nar/gkh340.

337 Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of
338   LTR retrotransposons. *BMC Bioinformatics* 9. DOI: 10.1186/1471-2105-9-18.

339 Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nature reviews. Genetics*
340   9:397–405. DOI: 10.1038/nrg2337.

341 Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends in Genetics* 5:103–107. DOI:
342   10.1016/0168-9525(89)90039-5.

343 Gan Y, Liu F, Peng R, Wang C, Li S, Zhang X, Wang Y, Wang K. 2012. Individual chromosome identification,
344   chromosomal collinearity and genetic-physical integrated map in Gossypium darwinii and four D genome
345   cotton species revealed by BAC-FISH. *Genes & genetic systems* 87:233–41. DOI: 10.1266/ggs.87.233.

346 Goldschmidt RB. 2002. Marginalia to McClintock's Work on Mutable Loci in Maize. *The American Naturalist*
347   84:437–455. DOI: 10.1086/281640.

348 Grover CE, Kim HR, Wing RA, Paterson AH, Wendel JF. 2004. Incongruent patterns of local and global genome

349   size evolution in cotton. *Genome Research* 14:1474–1482. DOI: 10.1101/gr.2673204.

350   Hanson, R. E., X. p. Zhao, M. N. Islam‐Faridi, A. H. Paterson, M. S. Zwick, C. F. Crane, T. D. McKnight, D. M.
351   Stelly and H. J. J. A. J. o. B. Price (1998). "Evolution of interspersed repetitive elements in Gossypium
352   (Malvaceae)." 85(10): 1364-1368. DOI: 10.2307/2446394.

353   Hawkins JS, Kim HR, Nason JD, Wing RA, Wendel JF. 2006. Differential lineage-specific amplification of
354   transposable elements is responsible for genome size variation in Gossypium. *Genome Research* 16:1252–
355   1261. DOI: 10.1101/gr.5282906.

356   Hendrix B, Stewart JM. 2005. Estimation of the nuclear DNA content of Gossypium species. *Annals of Botany*
357   95:789–797. DOI: 10.1093/aob/mci078.

358   Hu, Y., J. Chen, L. Fang, Z. Zhang, W. Ma, Y. Niu, L. Ju, J. Deng, T. Zhao, J. Lian, K. Baruch, D. Fang, X. Liu, Y.
359   L. Ruan, M. U. Rahman, J. Han, K. Wang, Q. Wang, H. Wu, G. Mei, Y. Zang, Z. Han, C. Xu, W. Shen, D.
360   Yang, Z. Si, F. Dai, L. Zou, F. Huang, Y. Bai, Y. Zhang, A. Brodt, H. Ben-Hamo, X. Zhu, B. Zhou, X. Guan,
361   S. Zhu, X. Chen and T. Zhang (2019). "Gossypium barbadense and Gossypium hirsutum genomes provide
362   insights into the origin and evolution of allotetraploid cotton." Nat Genet 51(4): 739-748. DOI:
363   10.1038/s41588-019-0371-5.

364   Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang TH, Lan T, Welch
365   AJ, Juárez MJA, Simpson J, Fernández-Cortés A, Arteaga-Vázquez M, Góngora-Castillo E, Acevedo-
366   Hernández G, Schuster SC, Himmelbauer H, Minoche AE, Xu S, Lynch M, Oropeza-Aburto A, Cervantes-
367   Pérez SA, De Jesús Ortega-Estrada M, Cervantes-Luevano JI, Michael TP, Mockler T, Bryant D, Herrera-
368   Estrella A, Albert VA, Herrera-Estrella L. 2013. Architecture and evolution of a minute plant genome. *Nature*
369   498:94–98. DOI: 10.1038/nature12132.

370   Jeffreys, A. J., Neumann, R., & Wilson, V. 1990. Repeat unit sequence variation in minisatellites: a novel source of
371   DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* 60(3): 473-485.
372   DOI: 10.1016/0092-8674(90)90598-9.

373   Jurka J, Kapitonov V V., Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database
374   of eukaryotic repetitive elements. *Cytogenetic and Genome Research* 110:462–467. DOI: 10.1159/000084979.

375   Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative
376   studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111–120. DOI: 10.1007/BF01731581.

377   Koike A, Nakai K, Takagi T. 2002. The Origin and Evolution of Eukaryotic Protein Kinases. *Genome Letters* 1:83–
378   104. DOI: 10.1166/gl.2002.010.

379   Koonin E V., Csuros M, Rogozin IB. 2013. Whence genes in pieces: Reconstruction of the exon-intron gene
380   structures of the last eukaryotic common ancestor and other ancestral eukaryotes. *Wiley Interdisciplinary*
381   *Reviews: RNA* 4:93–105. DOI: 10.1002/wrna.1143.

382   Kurtz S. 2003. The Vmatch large scale sequence analysis software. *Ref Type: Computer Program*.

383   Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, Liang X, Huang G, Percy RG, Liu K,
384   Yang W, Chen W, Du X, Shi C, Yuan Y, Ye W, Liu X, Zhang X, Liu W, Wei H, Wei S, Huang G, Zhang X,
385   Zhu S, Zhang H, Sun F, Wang X, Liang J, Wang J, He Q, Huang L, Wang J, Cui J, Song G, Wang K, Xu X,
386   Yu JZ, Zhu Y, Yu S. 2015. Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1)

387   provides insights into genome evolution. *Nature Biotechnology* 33:524–530. DOI: 10.1038/nbt.3208.

388   Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, Chen W, Liang X, Shang H, Liu W, Shi
389   C, Xiao G, Gou C, Ye W, Xu X, Zhang X, Wei H, Li Z, Zhang G, Wang J, Liu K, Kohel RJ, Percy RG, Yu
390   JZ, Zhu YX, Wang J, Yu S. 2014. Genome sequence of the cultivated cotton gossypium arboreum. *Nature*
391   *Genetics* 46:567–572. DOI: 10.1038/ng.2987.

392   Liu Z, Liu Y, Liu F, Zhang S, Wang X, Lu Q, Wang K, Zhang B, Peng R. 2018. Genome-wide survey and
393   comparative analysis of long terminal repeat (LTR) retrotransposon families in four gossypium species.
394   *Scientific Reports* 8. DOI: 10.1038/s41598-018-27589-6.

395   Liu Y, Peng R, Liu F, Wang X, Cui X, Zhou Z, Wang C, Cai X, Wang Y, Lin Z, Wang K. 2016. A Gossypium BAC
396   clone contains key repeat components distinguishing sub-genome of allotetraploidy cottons. *Molecular*
397   *Cytogenetics* 9. DOI: 10.1186/s13039-016-0235-y.

398   Liu X, Zhao B, Zheng HJ, Hu Y, Lu G, Yang CQ, Chen JD, Chen JJ, Chen DY, Zhang L, Zhou Y, Wang LJ, Guo
399   WZ, Bai YL, Ruan JX, Shangguan XX, Mao YB, Shan CM, Jiang JP, Zhu YQ, Jin L, Kang H, Chen ST, He
400   XL, Wang R, Wang YZ, Chen J, Wang LJ, Yu ST, Wang BY, Wei J, Song SC, Lu XY, Gao ZC, Gu WY,
401   Deng X, Ma D, Wang S, Liang WH, Fang L, Cai CP, Zhu XF, Zhou BL, Chen ZJ, Xu SH, Zhang YG, Wang
402   SY, Zhang TZ, Zhao GP, Chen XY. 2015. Gossypium barbadense genome sequence provides insight into the
403   evolution of extra-long staple fiber and specialized metabolites. *Scientific Reports* 5. DOI: 10.1038/srep14139.

404   Lu H, Cui X, Liu Z, Liu Y, Wang X, Zhou Z, Cai X, Zhang Z, Guo X, Hua J, Ma Z, Wang X, Zhang J, Zhang H,
405   Liu F, Wang K. 2018. Discovery and annotation of a novel transposable element family in Gossypium. *BMC*
406   *Plant Biology* 18. DOI: 10.1186/s12870-018-1519-7.

407   Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the*
408   *National Academy of Sciences* 101:12404–12410. DOI: 10.1073/pnas.0403715101.

409   Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR,
410   Gwadz M, Hurwitz DI, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C,
411   Geer LY, Bryant SH. 2017. CDD/SPARCLE: Functional classification of proteins via subfamily domain
412   architectures. *Nucleic Acids Research* 45:D200–D203. DOI: 10.1093/nar/gkw1129.

413   McCLINTOCK B. 1950. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of*
414   *Sciences of the United States of America* 36:344–55.

415   Morgante M. 2006. Plant genome organisation and diversity: the year of the junk! *Current Opinion in Biotechnology*
416   17:168–173. DOI: 10.1016/j.copbio.2006.03.001.

417   Munoz-Lopez M, Garcia-Perez J. 2010. DNA Transposons: Nature and Applications in Genomics. *Current*
418   *Genomics* 11:115–128. DOI: 10.2174/138920210790886871.

419   Okonechnikov, K., O. Golosova, M. Fursov and U. Team (2012). "Unipro UGENE: a unified bioinformatics
420   toolkit." Bioinformatics 28(8): 1166-1167. DOI: 10.1093/bioinformatics/bts091.

421   Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, Yoo
422   MJ, Byers R, Chen W, Doron-Faigenboim A, Duke M V., Gong L, Grimwood J, Grover C, Grupp K, Hu G,
423   Lee TH, Li J, Lin L, Liu T, Marler BS, Page JT, Roberts AW, Romanel E, Sanders WS, Szadkowski E, Tan X,
424   Tang H, Xu C, Wang J, Wang Z, Zhang D, Zhang L, Ashrafi H, Bedon F, Bowers JE, Brubaker CL, Chee PW,

425 Das S, Gingle AR, Haigler CH, Harker D, Hoffmann L V., Hovav R, Jones DC, Lemke C, Mansoor S,
426 Rahman MU, Rainville LN, Rambani A, Reddy UK, Rong JK, Saranga Y, Scheffler BE, Scheffler JA, Stelly
427 DM, Triplett BA, Van Deynze A, Vaslin MFS, Waghmare VN, Walford SA, Wright RJ, Zaki EA, Zhang T,
428 Dennis ES, Mayer KFX, Peterson DG, Rokhsar DS, Wang X, Schmutz J. 2012. Repeated polyploidization of
429 Gossypium genomes and the evolution of spinnable cotton fibres. *Nature* 492:423–427. DOI:
430 10.1038/nature11798.

431 Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O.
432 2006. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic
433 expansions in Oryza australiensis, a wild relative of rice. *Genome Research* 16:1262–1269. DOI:
434 10.1101/gr.5290206.

435 R Core Team. 2014. R Language Definition V. 3.1.1. *https://www.r-project.org/*:Accessed Nov 2015.

436 Sanmiguel P, Bennetzen JL. 1998. Evidence that a recent increase in maize genome size was caused by the massive
437 amplification of intergene retrotransposons. *Annals of Botany* 82:37–44. DOI: 10.1006/anbo.1998.0746.

438 Saranga Y. 2007. Special Issue: A century of wheat research - from wild emmer discovery to genome analysis.
439 *Israel Journal of Plant Sciences* 55:207–313.

440 Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF.
441 2003. Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Molecular Biology and
442 Evolution* 20:633–643. DOI: 10.1093/molbev/msg065.

443 Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: Computational challenges and
444 solutions. *Nature Reviews Genetics* 13:36–46. DOI: 10.1038/nrg3117.

445 Wang K, Guo W, Zhang T. 2007. Detection and mapping of homologous and homoeologous segments in
446 homoeologous groups of allotetraploid cotton by BAC-FISH. *BMC Genomics* 8. DOI: 10.1186/1471-2164-8-
447 178.

448 Wang K, Huang G, Zhu Y. 2016. Transposable elements play an important role during cotton genome evolution and
449 fiber cell development. *Science China Life Sciences* 59:112–121. DOI: 10.1007/s11427-015-4928-y.

450 Wang M, Tu L, Yuan D, Zhu D, Shen C, Li J, Liu F, Pei L, Wang P, Zhao G, Ye Z, Huang H, Yan F, Ma Y, Zhang
451 L, Liu M, You J, Yang Y, Liu Z, Huang F, Li B, Qiu P, Zhang Q, Zhu L, Jin S, Yang X, Min L, Li G, Chen
452 LL, Zheng H, Lindsey K, Lin Z, Udall JA, Zhang X. 2019. Reference genome sequences of two cultivated
453 allotetraploid cottons, Gossypium hirsutum and Gossypium barbadense. *Nature Genetics* 51:224–229. DOI:
454 10.1038/s41588-018-0282-x.

455 Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, Zou C, Li Q, Yuan Y, Lu C, Wei
456 H, Gou C, Zheng Z, Yin Y, Zhang X, Liu K, Wang B, Song C, Shi N, Kohel RJ, Percy RG, Yu JZ, Zhu YX,
457 Cang J, Yu S. 2012. The draft genome of a diploid cotton Gossypium raimondii. *Nature Genetics* 44:1098–
458 1103. DOI: 10.1038/ng.2371.

459 Wang KB, Wang WK, Wang CY, Song GL, Cui RX, Li SH, Zhang XD. 2001. [Studies of FISH and karyotype of
460 Gossypium barbadense]. *Yi chuan xue bao = Acta genetica Sinica* 28:69–75.

461 WANG K, WENDEL JF, HUA J. 2018. Designations for individual genomes and chromosomes in Gossypium.
462 *Journal of Cotton Research* 1. DOI: 10.1186/s42397-018-0002-1.

463   Wendel JF, Cronn RC. 2001. Polyploidy and the evolutionary history of cotton. *Advances in Agronomy* 78:139–186.
464   DOI: 10.1016/S0065-2113(02)78004-8.

465   Wendel JF, Flagel LE, Adams KL. 2012. Jeans, genes, and genomes: Cotton as a model for studying polyploidy. In:
466   *Polyploidy and Genome Evolution*. 181–207. DOI: 10.1007/978-3-642-31442-1_10.

467   Wessler SR. 2006. Transposable elements and the evolution of eukaryotic genomes. *Proceedings of the National*
468   *Academy of Sciences* 103:17600–17601. DOI: 10.1073/pnas.0607612103.

469   Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O,
470   Paux E, SanMiguel P, Schulman AH. 2007. A unified classification system for eukaryotic transposable
471   elements. *Nature Reviews Genetics* 8:973–982. DOI: 10.1038/nrg2165.

472   Wickham, H. (2016). ggplot2: elegant graphics for data analysis, Springer.

473   Xie Y, Dong F, Hong D, Wan L, Liu P, Yang G. 2012. Exploiting comparative mapping among Brassica species to
474   accelerate the physical delimitation of a genic male-sterile locus (BnRf) in Brassica napus. *Theoretical and*
475   *Applied Genetics* 125:211–222. DOI: 10.1007/s00122-012-1826-6.

476   Yuan D, Tang Z, Wang M, Gao W, Tu L, Jin X, Chen L, He Y, Zhang L, Zhu L, Li Y, Liang Q, Lin Z, Yang X, Liu
477   N, Jin S, Lei Y, Ding Y, Li G, Ruan X, Ruan Y, Zhang X. 2015. The genome sequence of Sea-Island cotton
478   (Gossypium barbadense) provides insights into the allopolyploidization and development of superior
479   spinnable fibres. *Scientific Reports* 5. DOI: 10.1038/srep17662.

480   Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM, Hulse-Kemp AM,
481   Wan Q, Liu B, Liu C, Wang S, Pan M, Wang Y, Wang D, Ye W, Chang L, Zhang W, Song Q, Kirkbride RC,
482   Chen X, Dennis E, Llewellyn DJ, Peterson DG, Thaxton P, Jones DC, Wang Q, Xu X, Zhang H, Wu H, Zhou
483   L, Mei G, Chen S, Tian Y, Xiang D, Li X, Ding J, Zuo Q, Tao L, Liu Y, Li J, Lin Y, Hui Y, Cao Z, Cai C,
484   Zhu X, Jiang Z, Zhou B, Guo W, Li R, Chen ZJ. 2015. Sequencing of allotetraploid cotton (Gossypium
485   hirsutum L. acc. TM-1) provides a resource for fiber improvement. *Nature Biotechnology* 33:531–537. DOI:
486   10.1038/nbt.3207.

487   Zhang X, Tolzmann CA, Melcher M, Haas BJ, Gardner MJ, Smith JD, Feagin JE. 2011. Branch point identification
488   and sequence requirements for intron splicing in plasmodium falciparum. *Eukaryotic Cell* 10:1422–1428.
489   DOI: 10.1128/EC.05193-11.

490   Zhao, X., Y. Ji, X. Ding, D. M. Stelly and A. H. J. P. m. b. Paterson (1998). "Macromolecular organization and
491   genetic mapping of a rapidly evolving chromosome-specific tandem repeat family (B77) in cotton
492   (Gossypium)." 38(6): 1031-1041. DOI: 10.1023/A:1006073116627.

493   Zhao, X. P., Y. Si, R. E. Hanson, C. F. Crane, H. J. Price, D. M. Stelly, J. F. Wendel and A. H. Paterson (1998).
494   "Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton." Genome
495   Research 8(5): 479-492. DOI: 10.1101/gr.8.5.479/Genome Res. 1998. 8: 479-492.

496

Plant materials used in this work, together with ploidy, studied genome, and specimen accession code.

1   Table 1. Plant materials used in this work, together with ploidy, studied genome, and specimen accession
2   code.

Species	Ploidy	Genome	Accession
G. arboreum	2x	A_2	Shixiya-1
G. raimondii	2x	D_5	D5-07
G. hirsutum	4x	$(AD)_1$	CCRI-12
G. barbadense	4x	$(AD)_2$	Xinhai-7

Table 2. The distribution of *ICRd* motifs on different genome assemblies of tetraploid cotton.

1   Table 2. The distribution of *ICRd* motifs on different genome assemblies of tetraploid cotton.

Tetraploid	Version	Reference	*ICRd* motif
G. hirsutum (AD)$_1$	BGI	(Li et al., 2015)	D$_h$01-D$_h$13; A$_h$02, A$_h$05, A$_h$07, A$_h$08
	NBI	(Zhang, Hu et al., 2015)	D$_h$01-D$_h$13; None in A-sub
	HAU	(Wang, Tu et al., 2019)	D$_h$01-D$_h$13; None in A-sub
	ZJU	(Hu, Chen et al., 2019)	D$_h$01-D$_h$13; None in A-sub
G. barbadense (AD)$_2$	CAS	(Liu et al., 2015)	D$_b$01-D$_b$13; None in A-sub
	HAUv1	(Yuan et al., 2015)	D$_b$01, D$_b$02, D$_b$06-D$_b$09, D$_b$12; None in A-sub
	HAUv2	(Wang, Tu et al., 2019)	D$_h$01-D$_h$13; None in A-sub
	ZJU	(Hu, Chen et al., 2019)	D$_h$01-D$_h$13; None in A-sub

Figure 1

The structure of *ICRd* motif

a: The self-blast of the *ICRd* motif showed the inner repeats; b: The structure of *ICRd* motif; c: The basic TR unit; d: The examples of the structure illustration of the LTR-TEs inserted with *ICRd* motif.

a

D$_5$05_50639971_50641791

b GTAGAATTACTGGCTTCAATGTACTCCACTGCAACTTCATGGAGGTAAAATCCGCCATCTTCGATCTGCTCCACTACTGCTTAGGGAGACAAGACCTGAAATCTTCAACCTGCTCCACTGTCTCGAGGGA
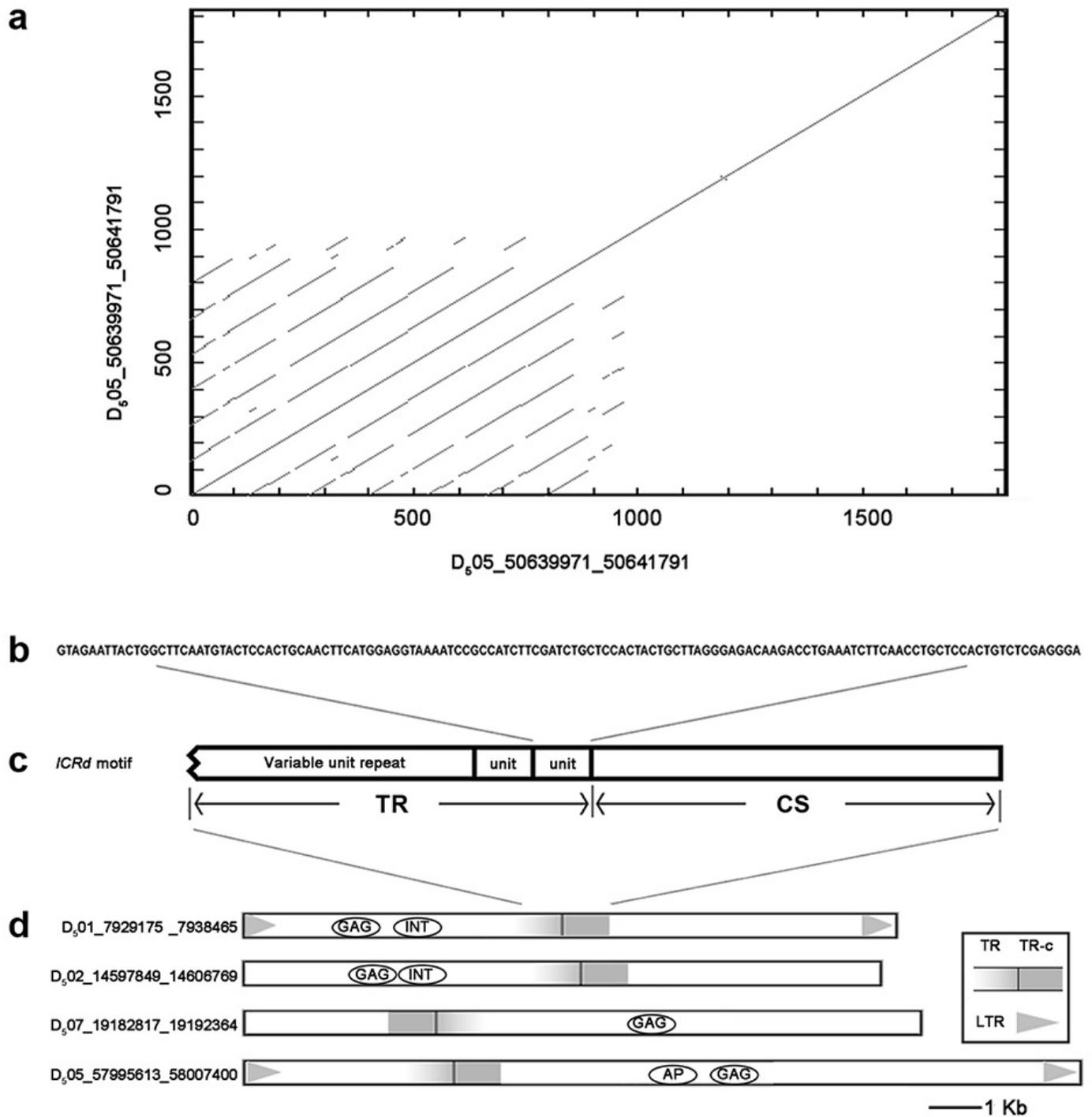
Figure 2

The content of the basic unit in the TRs

a: The basic unit content in the TRs involved in the *ICRd* motifs, displayed from small to large; b: The number of *ICRd* TRs that harboring different unit content, the x-axis adopt the intervals of unit content for convenient exhibition.
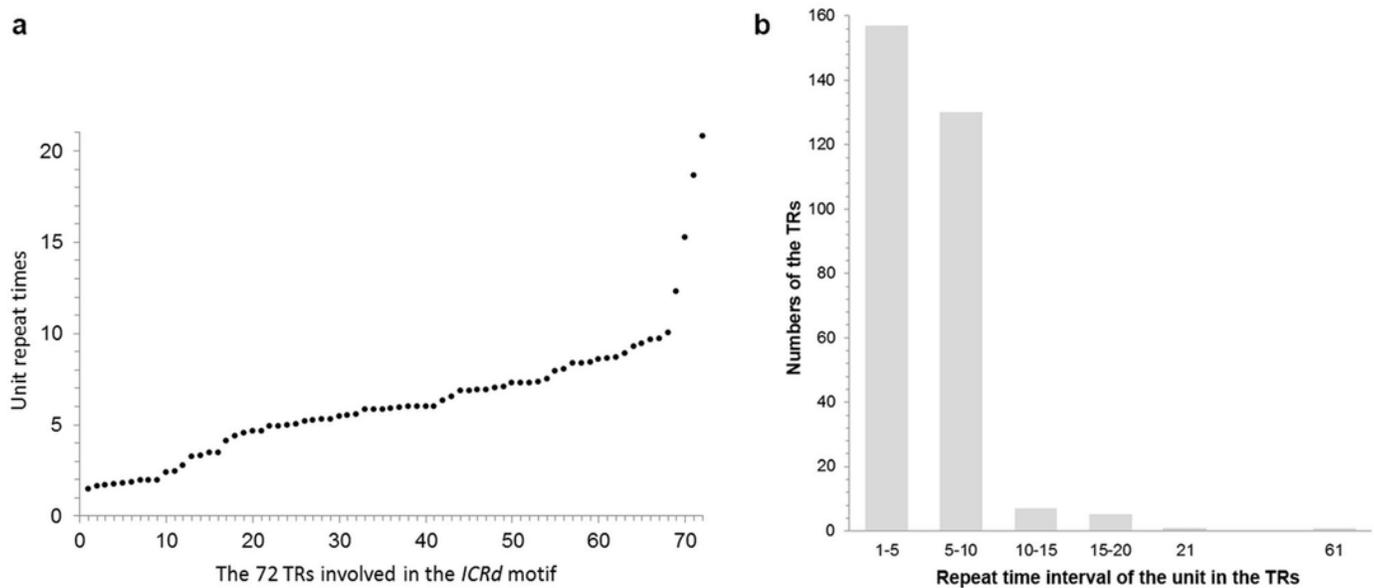
Figure 3

The FISH images of *ICRd* motif (red) hybridized to mitotic chromosomes of four species.

a: *G. arboreum* (AA); b: *G. hirsutum* (AADD); c: *G. barbadense* (AADD); d: *G. raimondii* (DD).

Bar = 5µm.
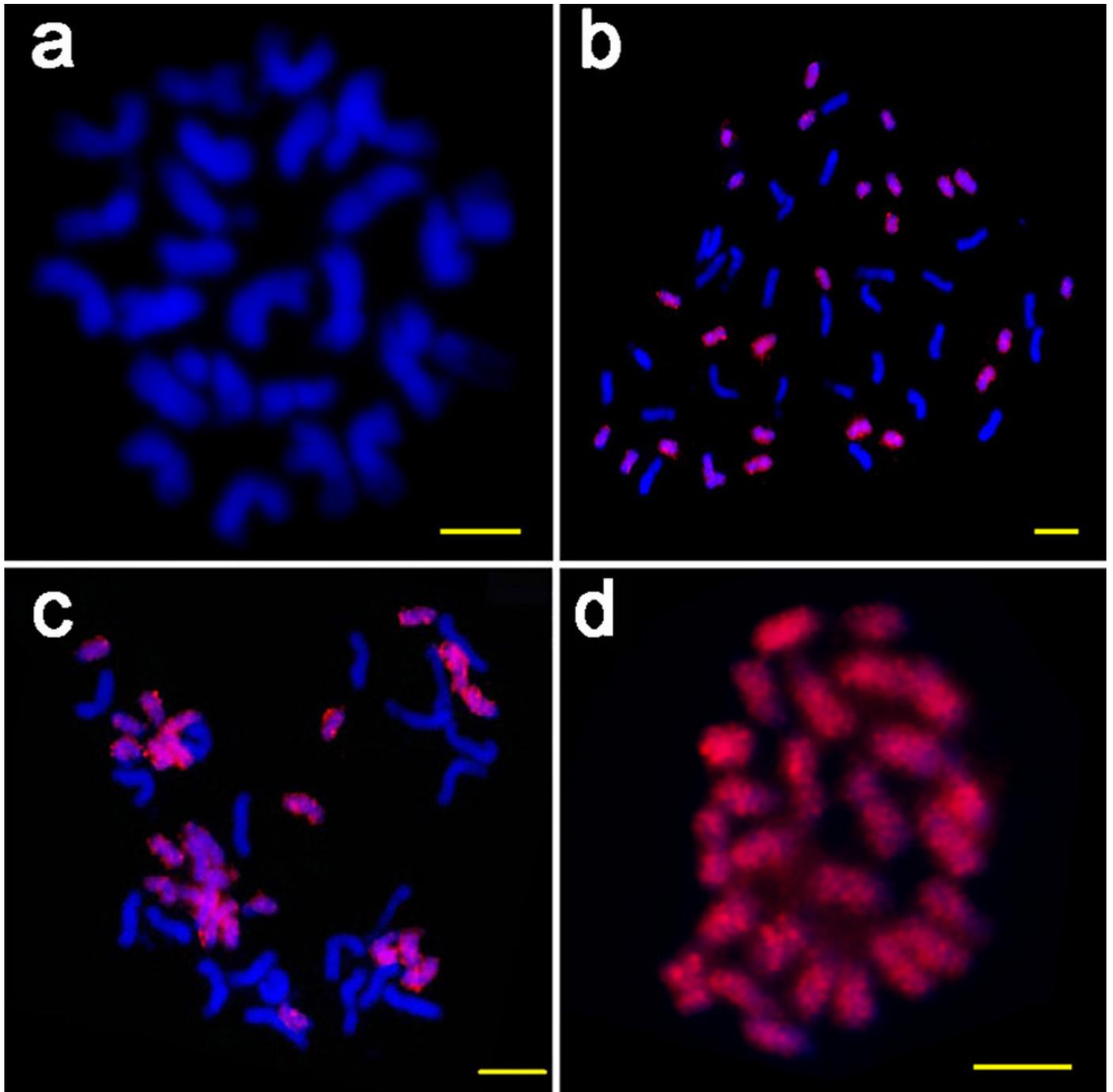
Figure 4

The consensus accumulation histogram from the whole alignment of the 72 LTR-TEs .

The region marked with the black line is the *ICRd* motif region.
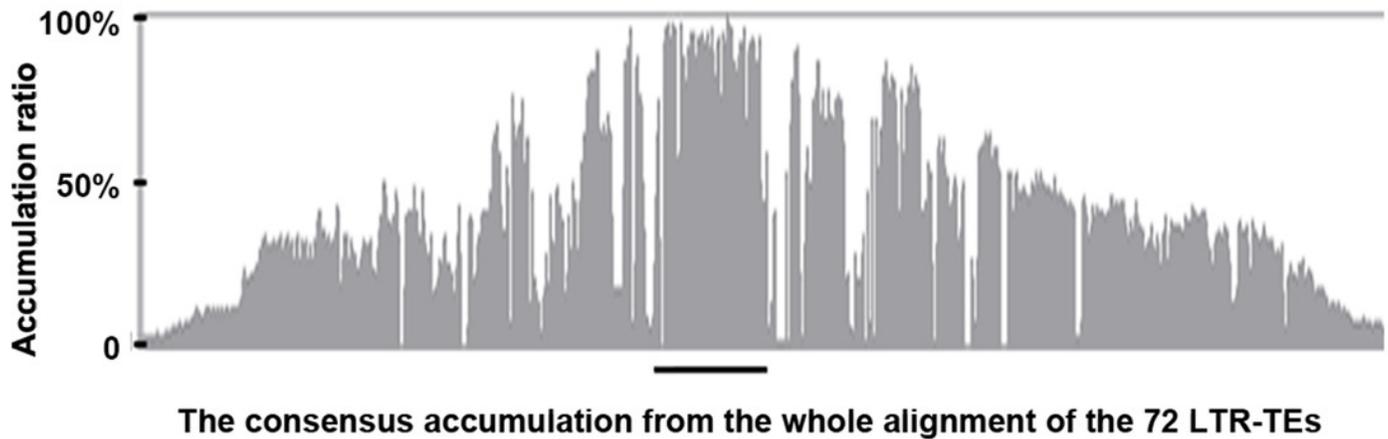
Figure 5

The accumulation of putative active date of the LTR-TEs.
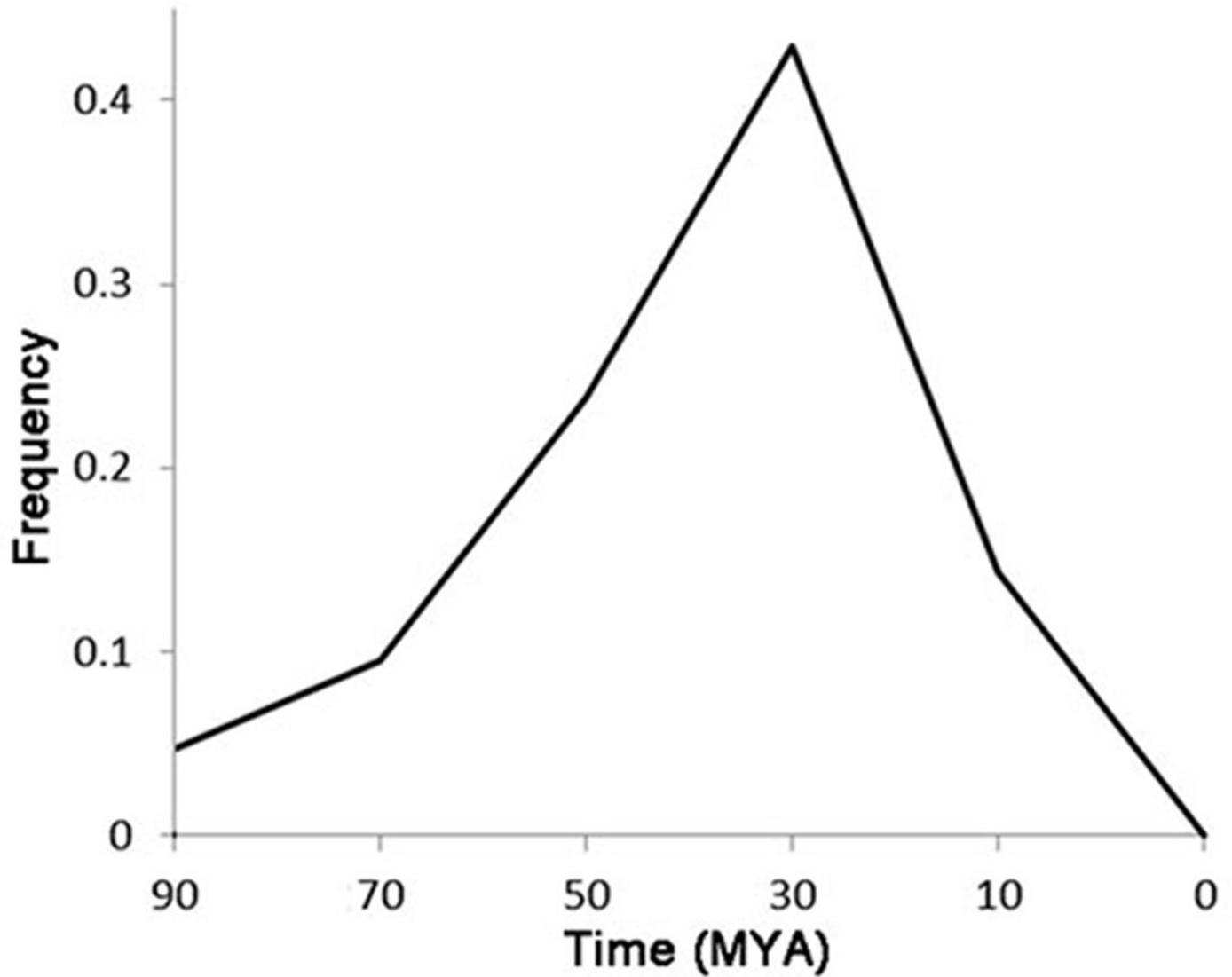
Figure 6

The distribution of the *ICRd* motif and its constituent in the D_5 genome

a: Insertions of the *ICRd* motif and its constituents in the D_5 genomes; b, c: *ICRd* TR and TR-c chromosomal distribution, the expected (grey) and actual (white) distributions across all chromosomes are illustrated; in addition, the density per megabase is shown for each chromosome.
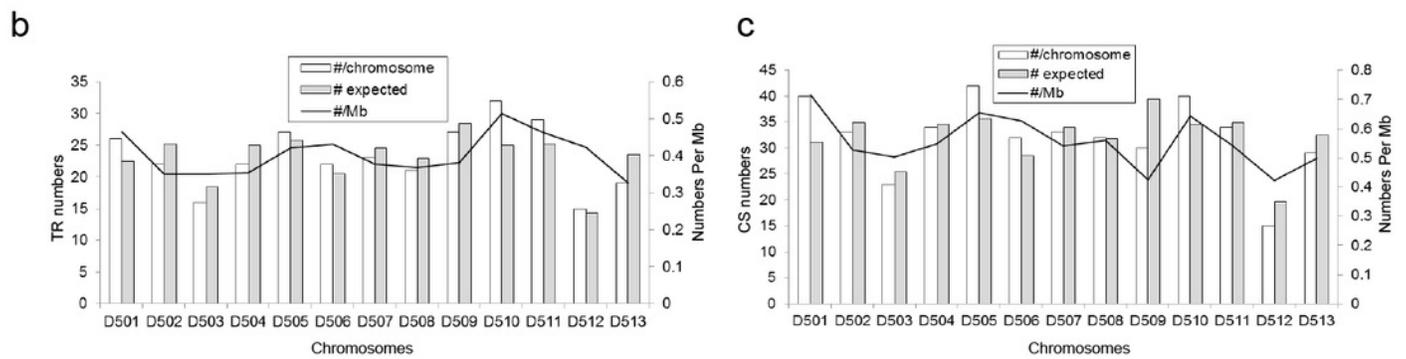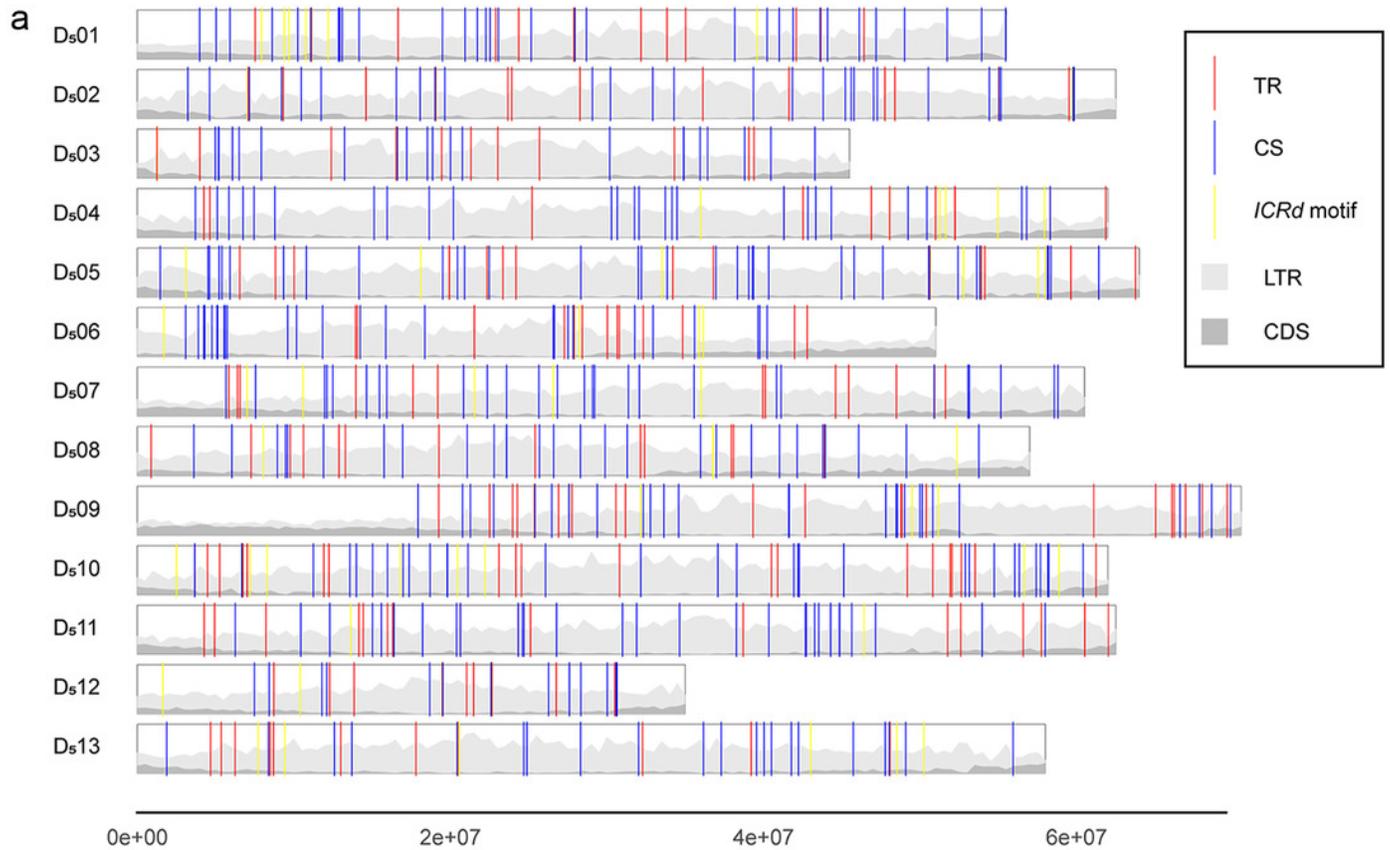
Figure 7

The colinearity of the two homologous segments.