

A peer-reviewed version of this preprint was published in PeerJ on 3 January 2020.

[View the peer-reviewed version](https://peerj.com/articles/8344) (peerj.com/articles/8344), which is the preferred citable publication unless you specifically need to cite this preprint.

Lu H, Cui X, Zhao Y, Magwanga RO, Li P, Cai X, Zhou Z, Wang X, Liu Y, Xu Y, Hou Y, Peng R, Wang K, Liu F. 2020. Identification of a genome-specific repetitive element in the *Gossypium* D genome. PeerJ 8:e8344 <https://doi.org/10.7717/peerj.8344>

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

Hejun Lu^{Equal first author, 1, 2}, Xinglei Cui^{Equal first author, 1}, Yanyan Zhao¹, Richard Odongo Magwang^{1, 3}, Pengcheng Li¹, Xiaoyan Cai¹, Zhongli Zhou¹, Xingxing Wang¹, Yuling Liu⁴, Yanchao Xu¹, Yuqing Hou¹, Renhai Peng⁴, Kunbo Wang^{Corresp., 1, 5}, Fang Liu^{Corresp. 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, China

² Gembloux Agro-bio Tech, University of Liège, Gembloux, Namur, Belgium

³ School of biological and physical sciences,, Jaramogi Oginga Odinga University of Science and Technology (JOUST), P.O Box, Bondo-Kenya

⁴ Anyang Istitute of Technology, Anyang, Henan, China

⁵ Tarium University, Alar, Xinjiang, China

Corresponding Authors: Kunbo Wang, Fang Liu
Email address: wkbcri@163.com, liufcri@163.com

The activity of genome-specific repetitive sequence is the main cause of the genome variation between *Gossypium A* and *D* genomes. Through the comparative analysis of the two genomes, we got a repetitive element (*ICRd* motif), which repeats massively in the diploid *Gossypium raimondii* (D_5) genome while almost absent in the diploid *Gossypium arboreum* (A_2) genome. We further explored the existence of *ICRd* motif in *G. raimondii*, *G. arboreum*, and two tetraploids (AADD) cotton *G. hirsutum* and *G. barbadense* by fluorescence *in situ* hybridization (FISH), and observed the *ICRd* motif exists in D_5 and *D*-subgenomes but not in A_2 and *A*-subgenome. The *ICRd* motif was investigated through its two constituents, a length variable tandem repeat region (TR) and a conservative sequence (CS), which highly repeat and evenly distribute in chromosomes of D_5 genome. The *ICRd* motif was revealed as the common conservative region of ancient LTR-TEs. The identifications and investigation of the *ICRd* motif promote the study on the *A* and *D* genome differences, facilitate the research on the *Gossypium* genome evolution, and provide assistance to subgenome identification and genome assembling.

1 Identification of a genome-specific repetitive element in 2 the *Gossypium* D genome

3 Hejun Lu^{1,2*}, Xinglei Cui^{1*}, Yanyan Zhao¹, Richard Odongo Magwanga^{1,3}, Pengcheng Li¹, Xiaoyan Cai¹,
4 Zhongli Zhou¹, Xingxing Wang¹, Yuling Liu⁴, Yanchao Xu¹, Yuqing Hou¹, Renhai Peng⁴, Kunbo Wang^{1,5†},
5 Fang Liu^{1†}

6 ¹ Research Base of Tarium University, State Key Laboratory of Cotton Biology, Institute of Cotton Research of Chinese
7 Academy of Agricultural Science, Anyang, Henan, 455000, China

8 ² Gembloux Agro-Bio Tech, University of Liège, Gembloux, 5030, Belgium

9 ³ School of biological and physical sciences (SBPS), Jaramogi Oginga Odinga University of Science and Technology
10 (JOUST), P.O Box 210-40601, Bondo-Kenya

11 ⁴ Anyang Institute of Technology, Anyang, Henan, 455000, China

12 ⁵ Tarium University, Alar, Xinjiang. 843300, China

13 *These authors contributed equally to this work.

14 †Correspondence: Kunbo Wang, E-mail: wkbcri@163.com; Fang Liu, liufcri@163.com

15 **Abstract:** The Activity of genome-specific repetitive sequence is the main cause of the genome variation between
16 *Gossypium* A and D genomes. Through the comparative analysis of the two genomes, we got a repetitive element
17 (*ICRd* motif), which repeats massively in the diploid *Gossypium raimondii* (D₅) genome while almost absent in the
18 diploid *Gossypium arboreum* (A₂) genome. We further explored the existence of *ICRd* motif in *G. raimondii*, *G.*
19 *arboreum*, and two tetraploids (AADD) cotton *G. hirsutum* and *G. barbadense* by fluorescence *in situ*
20 hybridization (FISH), and observed the *ICRd* motif exists in D₅ and D-subgenomes but not in A₂ and A-
21 subgenome. The *ICRd* motif was investigated through its two constituents, a length variable tandem repeat region
22 (TR) and a conservative sequence (CS), which highly repeat and evenly distribute in chromosomes of D₅ genome.
23 The *ICRd* motif was revealed as the common conservative region of ancient LTR-TEs. The identifications and
24 investigation of the *ICRd* motif promote the study on the A and D genome differences, facilitate the research on the
25 *Gossypium* genome evolution, and provide assistance to subgenome identification and genome assembling.

26 **Keywords:** *Gossypium*; D genome; Repetitive element; Genome-specific; Fluorescence *in situ* hybridization
27 (FISH); Evolution

28 1. Introduction

29 Repetitive DNA sequences are common in eukaryotic genomes, account for a huge fraction of the host genome
30 (Ibarra-Laclette et al., 2013) and are highly correlated with the size of the host genome (Feschotte, 2008). Repetitive
31 DNA can be divided into two major groups by their structures: tandem repeats and interspersed repeats (Jurka et al.,
32 2005). The tandem repeats are known as simple repeat sequences (SSR), including micro-satellites, mini-satellites,
33 and satellites (M.Lesk, 2002; Singh, 2015). The Interspersed repeats also were called as transposable elements (TEs)
34 or transposons.

35 After the first report of the TEs in maize (McCLINTOCK, 1950; Brink & Williams, 1973; Goldschmidt, 2002),
36 TEs were identified in many eukaryotic species (Munoz-Lopez & Garcia-Perez, 2010). There are thousands of
37 different TE families in plants, which display the extreme diversity (Sanmiguel & Bennetzen, 1998; Bennetzen,
38 2005; Morgante, 2006). Finnegan first proposed a TE classification system, which includes two classes based on
39 their transposition mechanisms: media by RNA (Retrotransposons) or DNA (DNA transposons) (Bowen & Jordan,
40 2002; Wessler, 2006; Arkhipova, 2018). Wicker unified the TEs nomenclatures and classification system applying
41 mechanistic and enzymatic criteria (Wicker et al., 2007, 2008, 2009; Seberg et al., 2009). TEs play important roles
42 in such as variations in intron size (Deutsch & Long, 1999; Zhang et al., 2011; Koonin, Csuros & Rogozin, 2013),
43 segmental duplication (Del Pozo & Ramirez-Parra, 2015), transfer of organelle DNA to the nucleus (Adams &
44 Palmer, 2003), expansion/contraction of tandem repeats and illegitimate recombination (Finnegan, 1989; Koike,
45 Nakai & Takagi, 2002). Long Terminal Repeat Retrotransposons (LTR-TEs), which are usually scattered throughout

46 genomes, is the most abundant TE type and can cause genome expansion over a short evolutionary period
 47 particularly in plant genomes (Piegu et al., 2006). The genome-specific TE is an efficient approach to study species
 48 formation and genome evolution in genome comparative research (Dong et al., 2018).

49 *Gossypium* diverged from the common ancestor with *Theobroma cacao* approximately 33.7 MYA (Wang et
 50 al., 2012). *Gossypium* comprises eight diploids ($2n=2x=26$) genomic groups: A, B, C, D, E, F, G, K, and one
 51 allotetraploid ($2n=4x=52$) genomic group: AD (WANG, WENDEL & HUA, 2018). They are good materials for
 52 polyploidization, genomic organization and genome-size variation researches due to its dramatic genome diversity:
 53 from the smallest New World D genome of an average 885 Mb to the Australian K-genome of an average of 2576
 54 Mb (Hendrix & Stewart, 2005). The accumulation of different lineage-specific TEs was thought to be responsible
 55 for the variation of genome size in *Gossypium* genomes (Hawkins et al., 2006; Lu et al., 2018). Of the eight genomic
 56 groups of *Gossypium*, the A and D groups are the main subjects investigated (Du et al., 2018) in cotton genomics
 57 research, because the major cultivated cotton *G. hirsutum* was known as formed from the reuniting of the
 58 progenitors of *G. arboreum* (A_2) and *G. raimondii* (D_5) (Paterson et al., 2012). The key trait difference between *G.*
 59 *arboreum* and *G. raimondii* is the former producing spinnable fibers but the latter not, meanwhile in genomics, the
 60 former has a genome size (1,746 Mb/1C) that is around two times of the latter (885 Mb/1C) (Hendrix & Stewart,
 61 2005). Genome sequencing showed that the numbers of protein-coding genes between A (41,330) and D (37,505)
 62 genomes are not obviously different, while the lineage-specific TE content is the main reason for the size gap of A
 63 and D genome (Li et al., 2015; Du et al., 2018). Moreover, the transposable elements were suggested to play an
 64 important role during cotton genome evolution and fiber cell development (Wang, Huang & Zhu, 2016). Thus the
 65 research on the lineage-specific repetitive sequences between A and D genome is a meaningful path to investigating
 66 the specification dynamic.

67 Fluorescence *in situ* hybridization (FISH) is a versatile tool to visualize the distribution of sequences in
 68 chromosomes and plays a vital role in recent cytogenetic research. More and more repetitive sequences in the cotton
 69 genome were reported recently with FISH, and the identification and localization of these repetitive sequences
 70 would facilitate genome sequencing and understanding the mechanism of genome evolution (Lu et al., 2018). One
 71 lineage-specific repetitive element that repeats many times in A genome while absent in D genome was reported and
 72 suggested as an important contributor to the size gap between the A and D genome (Lu et al., 2018).

73 The D genomic group represents a diverse group of diploids that diverged from a branch of A, B, C, E, F, G,
 74 and K genomic groups about 5-10 million years ago (MYA) (Senchina et al., 2003). Although the D genome is the
 75 smallest one in genome size in *Gossypium*, a set of repeat elements with high proliferation in the D genome while
 76 absence in A genome was discovered in this work. The discovery and characterization of these novel repetitive
 77 elements provided new components for repetitive sequences database and insight into the evolution of *Gossypium*.

78 2. Materials and Methods

79 2.1 Plant Materials

80 The plant materials were obtained from National Wild Cotton Nursery in Hainan Island, China, sponsored by
 81 the Institute of Cotton Research of Chinese Academy of Agricultural Sciences (ICR-CAAS). They were also
 82 conserved in the greenhouse at ICR-CAAS' headquarter in Anyang City, Henan Province, China. The DNA and cell
 83 came from the plant materials of cotton species listed in Table 1, based on the newest nomenclatures of *Gossypium*
 84 species (WANG, WENDEL & HUA, 2018).

85 The genome sequences of *G. raimondii* (Paterson et al., 2012), *G. arboreum* (Li et al., 2014), *G. hirsutum*
 86 (AD)₁-BGI (Wang et al., 2017), (AD)₁-NBI (Zhang et al., 2015; Wang et al., 2019), (AD)₁-JGI (Li et al., 2015), *G.*
 87 *barbadense* (AD)₂-HAU (Yuan et al., 2015) were downloaded from the Cottongen (<https://www.cottongen.org/>).
 88 The other assemblies of *G. barbadense* (AD)₂-CAS (Liu et al., 2015) were obtained from the website
 89 (<http://database.chgc.sh.cn/cotton/index.html>).

90 Table 1. The plant materials involved in this work.

Species	Ploidy	Genome	Accession
---------	--------	--------	-----------

<i>G. arboreum</i>	2x	A ₂	Shixiya-1
<i>G. raimondii</i>	2x	D ₅	D5-07
<i>G. hirsutum</i>	4x	(AD) ₁	CCRI-12
<i>G. barbadense</i>	4x	(AD) ₂	Xinhai-7

91 2.2 Characterization of the Repetitive Element and Bioinformatics Analysis

92 Perl scripts were used in this work to do data management, such as parsing the software results, extracting
 93 sequences from genomes or databases, and whole genome insertion analysis. BLASTN was used to identify the
 94 element repeats in genomes or other databases, with a threshold of greater than or equal to 80% matching ratio
 95 meanwhile 80% similarity, with reference to the 80-80 rules suggested previously (Wicker et al., 2007). The TRs
 96 were identified with Tandem Repeats Finder (Benson, 1999). Alignments were performed using MUSCLE (Edgar,
 97 2004). The Unipro UGENE was used to present the alignments and train consensus sequences.(Edgar, 2004) The
 98 inner enzyme annotation was realized by online CD-search in NCBI (Marchler-Bauer et al., 2017). RepeatMasker
 99 was used to annotate the insertions and estimate the proportion of repetitive sequences in genomes.

100 The flanking LTRs of LTR-TEs were identified with the LTRharvest (Ellinghaus, Kurtz & Willhoeft, 2008).
 101 Subsequently, the Vmatch was used to cluster the LTRs (Kurtz, 2003). The divergence time of the LTR-TEs was
 102 estimated using the formula $T = d/2r$, where r represents a substitution rate of 1.3×10^{-8} per site per year (Ma &
 103 Bennetzen, 2004), and d means the distances of paired LTRs, which was calculated based on the Kimura two-
 104 parameter (Kimura, 1980). The insertions of repetitive sequences in genomes were illustrated by R language (R
 105 Core Team, 2014).

106 2.3 Fluorescence in situ hybridization (FISH)

107 The probe was designed with the PCR product of *ICRd* motif, which was obtained from the forward primer:
 108 TTCTATTTTATCCATCGTTATG, reverse: GGAGATAGGATTTGTTGCT; and followed the amplification
 109 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 2
 110 min. The final extension was done at 72°C for 6 min. The composition of the reaction mix using the following:
 111 gDNA (~5 µg/ml), primers (~0.8 µM), PCR Master Mix (Thermo), and H₂O. The gDNA extracted from leaves of
 112 cotton plants (Table 1). The probe was purified and labeled with digoxigenin-dUTP via nick translation, according
 113 to the instructions of the manufacturer (Roche Diagnostics, USA). Mitotic chromosome preparation and FISH
 114 procedures were conducted using a modified protocol (Wang et al., 2001).

115 3. Results

116 3.1. One Specific Repetitive Sequence in *Gossypium D₅* Genome

117 One segment in *G. raimondii* (D₅) genome (Chr05: 50639971-50641791) was filtered out as genome-specific
 118 in D₅, by comparative genome analysis of *G. raimondii* (Paterson et al., 2012) and *G. arboreum* (A₂) (Li et al., 2014)
 119 with BLAST. This sequence is highly repeated and spreading all over 13 chromosomes of the D₅ genome
 120 (Supplementary Table 1), while do not exist in A₂ genome. We queried it in Repbase (Chen et al., 2007a) and NCBI,
 121 but no related annotation was found. Then we performed LTRharvest (Ellinghaus, Kurtz & Willhoeft, 2008) and
 122 CD-search (Marchler-Bauer et al., 2017), which revealed it is neither LTR nor coding sequence.

123 Manual inspection revealed the structure of the genome-specific sequence having two constituents, a tandem
 124 repeats array (referred as TR hereafter) composed of 133 bp basic units, and an unknown conservative sequence
 125 (referred as CS hereafter) (Figure 1). Based on this feature, we totally identified 72 sequences from D₅ genome
 126 (Supplementary Table 2), for abbreviation, they were termed as the *ICRd* motif naming following our previous work
 127 (Lu et al., 2018). Among the 72 *ICRd* motifs, the TRs are length-variable having a variety of the basic unit content
 128 that 2-20 basic units (Figure 2a), while the CSs are stable and have an average size ~ 860 bp.

129 To verify the genome specificity and chromosome distribution of the *ICRd* motif, we used the PCR product of
 130 *ICRd* motif from *G. raimondii* to designed the probe for fluorescence *in situ* hybridization (FISH) on the mitotic

131 chromosomes of diploid A_2 and D_5 , and tetraploid *G. hirsutum* ($(AD)_1$), *G. barbadense* ($(AD)_2$). The probe generated
132 bright signals covering all the chromosomes of D_5 and D-subgenome, but none signal on the A_2 and A-subgenome
133 (Figure 3). These cytogenetic inspections were in accordance with the genomic comparative analysis and further
134 revealed that the *ICRd* motif is a genome-specific and highly repetitive element in the D_5 genome, as well as in the
135 D-subgenome of tetraploid cotton.

136 3.2 LTR-TEs Inserted with *ICRd* Motif

137 We compared the insertion loci of 72 *ICRd* motifs with the whole genome repeats annotation (gff file) of the
138 D_5 genome (Paterson et al., 2012) and found that each of the motifs is one to one harbored in 72 LTR-TEs
139 (Supplementary Table 3), which meant the former is the inner part of the latter.

140 We extracted the 72 LTR-TEs sequences from D_5 genome and parsed their structure, which showed all of 72
141 LTR-TEs are incomplete, lacking either enzyme or flanking LTRs, the required elements for an intact LTR-TE
142 (Wicker et al., 2007). We align all these LTR-TEs together and got their consensus accumulation histogram
143 (Supplementary Figure 1), which showed these TEs have a vast sequence variety among each other, however, an
144 only conservative region was observed, which is just the insertion region of the *ICRd* motif (Figure 4), revealing that
145 the *ICRd* motif is more stable than other elements along with the degradation and the evolution of the TEs.

146 Of the 72 LTR-TEs, 25 were identified having flanking LTRs, which were used to represent the classification
147 and evolution of these TEs. The LTR cluster results showed except two TEs have similarity in LTR region the other
148 23 TEs are absolutely different from each other, which further revealed they do not belong to the same family based
149 on the LTR similarity rules (Wicker et al., 2007). The estimated active date curve of these TEs almost all prior to 10
150 MYA and peak in ~30 MYA (Figure 5), in the close period with that *G. raimondii* and *T. cacao* diverged
151 approximately 33.7 MYA (Wang et al., 2012), but far earlier than the putative divergence time of *Gossypium* A and
152 D genomes (Wendel & Cronn, 2001). These revealed these LTR-TEs are ancient TEs and potentially contribute to
153 the speciation formation of *Gossypium*.

154 3.3 Abundant Constituents of *ICRd* Motif in D_5 genome

155 Toward the further analysis of the genomic feature of the *ICRd* motif, we separately investigated its two
156 constituents that TR and CS, on their content and distribution feature in the D_5 genome (Figure 6a). In total 350 TR
157 insertions were detected (Supplementary Table 2), which are different in length due to different times of the unit
158 repeating comprising among 2–21, and mainly 2–10 times of the basic unit (Figure 2b). The longest TR insertion in
159 D_5 (D_503 : 25689303–25697234) comprising 61 units up to 8 kb, which was extraordinary and unknown on how it
160 formed. On the other hand, in total 463 CSs were found (Supplementary Table 2). Combining analysis of the
161 insertion loci of the two constituents, we found 72 TRs are closely followed by 72 CSs, which just constitute the
162 *ICRd* motifs (Figure 1).

163 Further analysis proved the TR and CS are evenly distributed on the chromosomes based on χ^2 test (the
164 number of insertions is proportional to the size of the chromosome), where for the TR insertions, $\chi^2 = 5.56$ ($df = 12$,
165 $P > 0.9$), and for the CSs, $\chi^2 = 9.08$ ($df = 12$, $P > 0.5$). The even distributions meant the CS and TR are possible
166 ancient repetitive sequences that have evolved along with the chromosomes. The *G. raimondii* genome sequencing
167 work had reported that most TEs in *G. raimondii* are deletion derivatives lacking the domains that are typically
168 necessary for transposition that the only 3% of LTR base pairs derived from full-length LTR-TEs (Paterson et al.,
169 2012). Here the hundreds of constituents of *ICRd* motif in D_5 are potentially the fragments produced from the
170 ancient LTR-TEs.

171 3.4 The disappearance of *ICRd* Motif from *Gossypium*

172 With the aim to observe the disappearance of the *ICRd* motif in the newly formed *Gossypium* A genome, we
173 selected one pair segments from the highly collinear Chromosome 1 in the two cotton species to observe (Li et al.,
174 2014). The segment from Chromosome 1 of *G. raimondii* (D_501) harbor one *ICRd* motif and its homologous
175 segment from A_201 was got based on homologous SSR markers (Supplementary Table 4). The illustration of the

176 syntenic block of the two segments showed the *ICRd* motif together with its host LTR-TE were totally abandoned
 177 on the A₂01 segment, while their up- and downstream conservative regions remained (Figure 7). In the upstream, we
 178 observed two insertions rich in repeat sequences special on A₂01 segment (Supplementary Table 4), which was
 179 potentially due to the recent TE expanding event happened in A genome (Lu et al., 2018). Thus, we observed that
 180 the *ICRd* motifs and host LTR-TEs were directly abandoned from the genome with the recent formation of A
 181 genome (Wendel & Cronn, 2001; Wendel, Flagel & Adams, 2012), but remained in the D genome despite mass
 182 damage accumulation.

183 3.5 Distributions of *ICRd* Motifs in Tetraploid Cotton

184 Tetraploid cotton *G. hirsutum* and *G. barbadense* are the major cultivated fiber-producing cotton species.
 185 Research on the genome of these two species is an important approach to improve the cotton yield and quality.
 186 However, due to the huge amount of homologous segments between A and D-subgenomes, the tetraploid cotton
 187 genome assemblage has been a great challenge to molecular geneticists (Bowers et al., 2003; Chen et al., 2007b).
 188 Three versions of *G. hirsutum* genome assembly((AD)₁-BGI (Li et al., 2015), (AD)₁-NBI (Zhang et al., 2015),
 189 (AD)₁-JGI (Wang et al., 2017)), and two *G. barbadense* versions ((AD)₂-HAU (Yuan et al., 2015) and (AD)₂-CAS
 190 (Liu et al., 2015)) have been reported recently, however, the quality of the sequenced genomes require improvement
 191 in order to benefit cotton molecular breeders. Application of the lineage-specific repetitive element (LSR), the *ICRd*
 192 motifs are important tools in evaluating the quality of the genome assembly of the tetraploid cotton.

193 To observe the assembling quality of the *ICRd* motif in tetraploid genomes, we queried it with BLAST in the
 194 five tetraploid genome assemblies, including 3 versions of *G. hirsutum* ((AD)₁) and two versions of *G. barbadense*
 195 ((AD)₂) (Table 2). For the (AD)₁ assemblies, the blast result from the NBI version was in agreement to the FISH
 196 inspection that the *ICRd* motifs only generated the signals on the D-subgenome chromosomes (Figure 3). However,
 197 the BGI and JGI versions were inconsistent with the FISH inspection results, that *ICRd* motif was misassembled
 198 in the A-subgenome. For the (AD)₂ assemblies, the CAS showed a better assembling than HAU, because the *ICRd*
 199 motifs were located in all 13 D-sub chromosomes of CAS, but mainly matched with the unassembled scaffolds of
 200 HAU (Supplementary Table 5). This means the (AD)₂-CAS showed better scaffolds assembling than the (AD)₂-
 201 HAU. Thus, the *G. hirsutum* genome assembly (AD)₁-NBI, and the *G. barbadense* genome assembly (AD)₂-CAS
 202 are a conclusive version based on the comparison of the BLAST query and our cytogenetic experiment.

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

Assemblies	Reference	<i>ICRd</i> motif
(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
(AD) ₁ -NBI	(Zhang et al., 2015)	D _h 01-D _h 13; None in A-sub
(AD) ₁ -JGI	(Wang et al., 2017)	D _h 01-D _h 13; A _h 11
(AD) ₂ -CAS	(Liu et al., 2015)	D _b 01-D _b 13; None in A-sub
(AD) ₂ -HAU	(Yuan et al., 2015)	D _b 01, D _b 02, D _b 06-D _b 09, D _b 12; None in A-sub

204 4. Discussion

205 4.1 The Identification of *ICRd* Motif and *Gossypium* Evolution

206 TEs have played an important function in *Gossypium* speciation and the accumulation of different genomic-
 207 specific TEs were thought to be responsible for the variation of genome size in *Gossypium* genomes (Hawkins et al.,
 208 2006). Through FISH inspection, some A genome-specific repetitive elements have been well identified and
 209 characterized (Liu et al., 2016), but the similar work in the D genome has not yet been reported, this may be because
 210 the genome-specific repetitive sequences in A genome are much more than that in the D genome (Liu et al., 2018).
 211 However, in this work, starting with comparative genomic data, we screened out one kind of specific sequence in the
 212 D genome, and subsequently, we identified and characterized TEs.

213 The TEs harboring the *ICRd* motif showed an ancient active date approximately 10 MYA, while the time of
214 divergence of the A and D genomes from the common ancestor is estimated to have occurred 5-10 MYA (Grover et
215 al., 2004), thus the *ICRd* motifs existed in the ancestor and disappeared along with the formation of A genome.
216 Though the previous researches have stated the accumulation of lineage-specific TEs, which is thought to be
217 responsible for the variation of *Gossypium* genomes (Hawkins et al., 2006), and the LTR-TE activities after 5 MYA
218 mainly contributed to the two-fold size difference of the A and D genomes (Zhang et al., 2015). Based on our
219 analysis, we presumed that as same as the activity of new repetitive sequences, the extinction of ancient repetitive
220 sequences, such as the disappearance of *ICRd* motif in the A genome, also contributed significantly to the genome
221 evolution. Through FISH, we observed that the *ICRd* motifs were only distributed in D-subgenome chromosomes,
222 and the results were in agreement to the previous studies which reported that the TE have proliferated in the
223 progenitor genomes but were retained after allopolyploid formation in the D subgenome (Zhang et al., 2015).

224 4.2 *ICRd* Motif Support Cytogenetic Markers for Tetraploid Cotton

225 The identification of *ICRd* motif provides new subgenome marker for the accurate assembling of tetraploid
226 cotton (Chen et al., 2007a). Chromosome identification is the foundation of plant genetics, evolution and genomics
227 researches (Saranga, 2007; Xie et al., 2012). Although many species have been sequenced, the rapid identification of
228 subgenome is still useful in applied researches. FISH has been used as a reliable cytological technique for
229 chromosome identification in many species (Wang, Guo & Zhang, 2007). The identification of cotton chromosomes
230 evolved recently with the FISH technique (Gan et al., 2012). In this study, the identified *ICRd* motifs can be used as
231 a new cytological marker in *Gossypium*, especially in tetraploid. And the repetitive sequence probes are easier and
232 more successful to be detected than other probes. Several similar makers have been reported (Liu et al., 2016). The
233 addition of these new cytological markers will enrich the marker database for chromosome identification and
234 facilitate cotton genomic studies.

235 Eukaryotic genomes have a high proportion of TEs and these TEs make the eukaryotic genomes assembly
236 much more difficult than simple genomes (Treangen & Salzberg, 2012). Many reported genome sequences have
237 gaps because of the high proportion of TEs (Adams et al., 2000). Allopolyploid genomes are especially difficult to
238 assemble homologous fragments from sub-genomes (Chen et al., 2007a). The incorrect assembling of the genomes
239 leads to ambiguity in research, which, in turn, produce biases and errors when interpreting results (Adams et al.,
240 2000). Though the three versions of genome assembly of *G. hirsutum*, two versions of *G. barbadense* have been
241 released, their accuracies are inconsistent as revealed by BLASTN of the *ICRd* motif. Here, our FISH results were in
242 agreement to two tetraploid cotton genome assemblies, (AD)₁-NBI of *G. hirsutum*, and (AD)₂-CAS of *G.*
243 *barbadense*. Moreover, the *ICRd* motifs also assist to assure the source of the unpackaged scaffolds in the genome
244 assemblies, for instance, in the scaffolds the presence of the *ICRd* motifs can aid in the assigning of the scaffolds to
245 the D-subgenome chromosomes

246 5. Conclusions

247 We identified a kind of repetitive sequence in *Gossypium* D genome but absent in A genome, the *ICRd* motifs,
248 were found to be retained in D-subgenome and not in A-subgenome. We analyzed their structure, genomic
249 distribution, affiliation, and evolution, which revealed a conserved region which harbored the ancient LTR-TEs, in
250 the D genome. The identification and characterization of *ICRd* motif provided new insight into understanding the TE
251 evolution along with the formation of the cotton genomes as well as providing a convenient and applicative tool to
252 distinguish the A and D genome subsets of the tetraploid cotton genome assembly.

253 **Supplementary Materials:** Figure S1: Supplementary Figure 1. The whole alignment of the 72 LTR-TEs, Table S1: Blast of the
254 1.8 kb sequences in *G. raimondii* genome, Table S2: The *ICRd* motifs and their constituents, Table S3: The structures of the
255 LTR-TEs harboring the *ICRd* motif, Table S4: The information of the two homologous segments, Table S5: Blast results of the
256 *ICRd* motif with tetraploid cotton.

257 **Author Contributions:** Conceptualization, Kunbo Wang and Fang Liu; Data curation, Hejun Lu and Xinglei Cui; Formal
258 analysis, Hejun Lu and Richard Odongo Magwanga; Funding acquisition, Fang Liu; Investigation, Hejun Lu, Yanyan Zhao,
259 Pengcheng Li and Yuling Liu; Methodology, Xinglei Cui, Yanyan Zhao and Yuqing Hou; Resources, Xiaoyan Cai, Zhongli

260 Zhou, Yanchao Xu and Renhai Peng; Software, Hejun Lu; Supervision, Kunbo Wang and Fang Liu; Validation, Hejun Lu and
261 Xingxing Wang; Writing original draft, Hejun Lu; Writing-review & editing, Xinglei Cui, Richard Odongo Magwanga, Yanyan
262 Zhao.

263 **Funding:** Please add: This research was supported by The National Key Research and Development Plan of China (grants
264 2016YFD0100306 and 2016YFD0100203) and The Natural Science Foundation of China (grants 31530053 and 31671745).

265 **Acknowledgments:** We are indebted to Dr. Syed Shan-e-Ali Zaidi of the University of Liege, Belgium, for his guidance in
266 analysis and interpretation of the data

267 **Conflicts of Interest:** The authors declare no conflict of interest.

268 References

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

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

298 Arkhipova IR. 2018. Neutral theory, transposable elements, and eukaryotic genome evolution.

- 299 *Molecular Biology and Evolution* 35:1332–1337. DOI: 10.1093/molbev/msy083.
- 300 Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in
301 flowering plants. *Current Opinion in Genetics and Development* 15:621–627. DOI:
302 10.1016/j.gde.2005.09.010.
- 303 Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids*
304 *Research* 27:573–580. DOI: 10.1093/nar/27.2.573.
- 305 Bowen NJ, Jordan IK. 2002. Transposable elements and the evolution of eukaryotic complexity.
306 *Current Issues in Molecular Biology* 4:65–76.
- 307 Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome
308 evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422:433–
309 438. DOI: 10.1038/nature01521.
- 310 Brink RA, Williams E. 1973. Mutable R-navajo alleles of cyclic origin in maize. *Genetics*
311 73:273–296.
- 312 Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Guo W, Chen X, Stelly DM,
313 Rabinowicz PD, Town CD, Arioli T, Brubaker C, Cantrell RG, Lacape J-M, Ulloa M, Chee
314 P, Gingle AR, Haigler CH, Percy R, Saha S, Wilkins T, Wright RJ, Van Deynze A, Zhu Y,
315 Yu S, Abdurakhmonov I, Katageri I, Kumar PA, Mehboob-ur-Rahman, Zafar Y, Yu JZ,
316 Kohel RJ, Wendel JF, Paterson AH. 2007a. Toward Sequencing Cotton (*Gossypium*)
317 Genomes. *Plant Physiology* 145:1303–1310. DOI: 10.1104/pp.107.107672.
- 318 Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Guo W, Chen X, Stelly DM,
319 Rabinowicz PD, Town CD, Arioli T, Brubaker C, Cantrell RG, Lacape J-M, Ulloa M, Chee
320 P, Gingle AR, Haigler CH, Percy R, Saha S, Wilkins T, Wright RJ, Van Deynze A, Zhu Y,
321 Yu S, Abdurakhmonov I, Katageri I, Kumar PA, Mehboob-ur-Rahman, Zafar Y, Yu JZ,
322 Kohel RJ, Wendel JF, Paterson AH. 2007b. Toward Sequencing Cotton (*Gossypium*)
323 Genomes. *PLANT PHYSIOLOGY* 145:1303–1310. DOI: 10.1104/pp.107.107672.
- 324 Deutsch M, Long M. 1999. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids*
325 *Research* 27:3219–3228. DOI: 10.1093/nar/27.15.3219.
- 326 Dong G, Shen J, Zhang Q, Wang J, Yu Q, Ming R, Wang K, Zhang J. 2018. Development and
327 Applications of Chromosome-Specific Cytogenetic BAC-FISH Probes in *S. spontaneum*.
328 *Frontiers in Plant Science* 9. DOI: 10.3389/fpls.2018.00218.
- 329 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
330 W, Liu Z, Zhu H, Ma L, Liu F, Yang D, Wang F, Fan W, Gong Q, Peng Z, Wang L, Wang

- 331 X, Xu S, Shang H, Lu C, Zheng H, Huang S, Lin T, Zhu Y, Li F. 2018. Resequencing of
332 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of
333 key agronomic traits. *Nature Genetics* 50:796–802. DOI: 10.1038/s41588-018-0116-x.
- 334 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
335 throughput. *Nucleic acids research* 32:1792–7. DOI: 10.1093/nar/gkh340.
- 336 Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de
337 novo detection of LTR retrotransposons. *BMC Bioinformatics* 9. DOI: 10.1186/1471-2105-
338 9-18.
- 339 Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nature*
340 *reviews. Genetics* 9:397–405. DOI: 10.1038/nrg2337.
- 341 Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends in Genetics*
342 5:103–107. DOI: 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
344 chromosome identification, chromosomal collinearity and genetic-physical integrated map
345 in *Gossypium darwinii* and four D genome cotton species revealed by BAC-FISH. *Genes &*
346 *genetic systems* 87:233–41.
- 347 Goldschmidt RB. 2002. Marginalia to McClintock’s Work on Mutable Loci in Maize. *The*
348 *American Naturalist* 84:437–455. DOI: 10.1086/281640.
- 349 Grover CE, Kim HR, Wing RA, Paterson AH, Wendel JF. 2004. Incongruent patterns of local
350 and global genome size evolution in cotton. *Genome Research* 14:1474–1482. DOI:
351 10.1101/gr.2673204.
- 352 Hawkins JS, Kim HR, Nason JD, Wing RA, Wendel JF. 2006. Differential lineage-specific
353 amplification of transposable elements is responsible for genome size variation in
354 *Gossypium*. *Genome Research* 16:1252–1261. DOI: 10.1101/gr.5282906.
- 355 Hendrix B, Stewart JM. 2005. Estimation of the nuclear DNA content of *Gossypium* species.
356 *Annals of Botany* 95:789–797. DOI: 10.1093/aob/mci078.
- 357 Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang
358 TH, Lan T, Welch AJ, Juárez MJA, Simpson J, Fernández-Cortés A, Arteaga-Vázquez M,
359 Góngora-Castillo E, Acevedo-Hernández G, Schuster SC, Himmelbauer H, Minoche AE,
360 Xu S, Lynch M, Oropeza-Aburto A, Cervantes-Pérez SA, De Jesús Ortega-Estrada M,
361 Cervantes-Luevano JI, Michael TP, Mockler T, Bryant D, Herrera-Estrella A, Albert VA,
362 Herrera-Estrella L. 2013. Architecture and evolution of a minute plant genome. *Nature*

- 363 498:94–98. DOI: 10.1038/nature12132.
- 364 Jurka J, Kapitonov V V., Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase
365 Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*
366 110:462–467. DOI: 10.1159/000084979.
- 367 Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions
368 through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*
369 16:111–120. DOI: 10.1007/BF01731581.
- 370 Koike A, Nakai K, Takagi T. 2002. The Origin and Evolution of Eukaryotic Protein Kinases.
371 *Genome Letters* 1:83–104. DOI: 10.1166/gl.2002.010.
- 372 Koonin E V., Csuros M, Rogozin IB. 2013. Whence genes in pieces: Reconstruction of the exon-
373 intron gene structures of the last eukaryotic common ancestor and other ancestral
374 eukaryotes. *Wiley Interdisciplinary Reviews: RNA* 4:93–105. DOI: 10.1002/wrna.1143.
- 375 Kurtz S. 2003. The Vmatch large scale sequence analysis software. *Ref Type: Computer*
376 *Program*.
- 377 Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, Liang X, Huang G,
378 Percy RG, Liu K, Yang W, Chen W, Du X, Shi C, Yuan Y, Ye W, Liu X, Zhang X, Liu W,
379 Wei H, Wei S, Huang G, Zhang X, Zhu S, Zhang H, Sun F, Wang X, Liang J, Wang J, He
380 Q, Huang L, Wang J, Cui J, Song G, Wang K, Xu X, Yu JZ, Zhu Y, Yu S. 2015. Genome
381 sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into
382 genome evolution. *Nature Biotechnology* 33:524–530. DOI: 10.1038/nbt.3208.
- 383 Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, Chen W, Liang X,
384 Shang H, Liu W, Shi C, Xiao G, Gou C, Ye W, Xu X, Zhang X, Wei H, Li Z, Zhang G,
385 Wang J, Liu K, Kohel RJ, Percy RG, Yu JZ, Zhu YX, Wang J, Yu S. 2014. Genome
386 sequence of the cultivated cotton *Gossypium arboreum*. *Nature Genetics* 46:567–572. DOI:
387 10.1038/ng.2987.
- 388 Liu Z, Liu Y, Liu F, Zhang S, Wang X, Lu Q, Wang K, Zhang B, Peng R. 2018. Genome-wide
389 survey and comparative analysis of long terminal repeat (LTR) retrotransposon families in
390 four *Gossypium* species. *Scientific Reports* 8. DOI: 10.1038/s41598-018-27589-6.
- 391 Liu Y, Peng R, Liu F, Wang X, Cui X, Zhou Z, Wang C, Cai X, Wang Y, Lin Z, Wang K. 2016.
392 A *Gossypium* BAC clone contains key repeat components distinguishing sub-genome of
393 allotetraploidy cottons. *Molecular Cytogenetics* 9. DOI: 10.1186/s13039-016-0235-y.
- 394 Liu X, Zhao B, Zheng HJ, Hu Y, Lu G, Yang CQ, Chen JD, Chen JJ, Chen DY, Zhang L, Zhou

- 395 Y, Wang LJ, Guo WZ, Bai YL, Ruan JX, Shangguan XX, Mao YB, Shan CM, Jiang JP,
396 Zhu YQ, Jin L, Kang H, Chen ST, He XL, Wang R, Wang YZ, Chen J, Wang LJ, Yu ST,
397 Wang BY, Wei J, Song SC, Lu XY, Gao ZC, Gu WY, Deng X, Ma D, Wang S, Liang WH,
398 Fang L, Cai CP, Zhu XF, Zhou BL, Chen ZJ, Xu SH, Zhang YG, Wang SY, Zhang TZ,
399 Zhao GP, Chen XY. 2015. Gossypium barbadense genome sequence provides insight into
400 the evolution of extra-long staple fiber and specialized metabolites. *Scientific Reports* 5.
401 DOI: 10.1038/srep14139.
- 402 Lu H, Cui X, Liu Z, Liu Y, Wang X, Zhou Z, Cai X, Zhang Z, Guo X, Hua J, Ma Z, Wang X,
403 Zhang J, Zhang H, Liu F, Wang K. 2018. Discovery and annotation of a novel transposable
404 element family in Gossypium. *BMC Plant Biology* 18. DOI: 10.1186/s12870-018-1519-7.
- 405 M.Lesk A. 2002. Introduction to Bioinformatics Introduction to Bioinformatics VSN-S.
406 *Introduction to Bioinformatics* 2002:1–16.
- 407 Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes.
408 *Proceedings of the National Academy of Sciences* 101:12404–12410. DOI:
409 10.1073/pnas.0403715101.
- 410 Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer
411 RC, Gonzales NR, Gwadz M, Hurwitz DI, Lu F, Marchler GH, Song JS, Thanki N, Wang
412 Z, Yamashita RA, Zhang D, Zheng C, Geer LY, Bryant SH. 2017. CDD/SPARCLE:
413 Functional classification of proteins via subfamily domain architectures. *Nucleic Acids*
414 *Research* 45:D200–D203. DOI: 10.1093/nar/gkw1129.
- 415 McCLINTOCK B. 1950. The origin and behavior of mutable loci in maize. *Proceedings of the*
416 *National Academy of Sciences of the United States of America* 36:344–55.
- 417 Morgante M. 2006. Plant genome organisation and diversity: the year of the junk! *Current*
418 *Opinion in Biotechnology* 17:168–173. DOI: 10.1016/j.copbio.2006.03.001.
- 419 Munoz-Lopez M, Garcia-Perez J. 2010. DNA Transposons: Nature and Applications in
420 Genomics. *Current Genomics* 11:115–128. DOI: 10.2174/138920210790886871.
- 421 Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC,
422 Shu S, Udall J, Yoo MJ, Byers R, Chen W, Doron-Faigenboim A, Duke M V., Gong L,
423 Grimwood J, Grover C, Grupp K, Hu G, Lee TH, Li J, Lin L, Liu T, Marler BS, Page JT,
424 Roberts AW, Romanel E, Sanders WS, Szadkowski E, Tan X, Tang H, Xu C, Wang J,
425 Wang Z, Zhang D, Zhang L, Ashrafi H, Bedon F, Bowers JE, Brubaker CL, Chee PW, Das
426 S, Gingle AR, Haigler CH, Harker D, Hoffmann L V., Hovav R, Jones DC, Lemke C,
427 Mansoor S, Rahman MU, Rainville LN, Rambani A, Reddy UK, Rong JK, Saranga Y,

- 428 Scheffler BE, Scheffler JA, Stelly DM, Triplett BA, Van Deynze A, Vaslin MFS,
429 Waghmare VN, Walford SA, Wright RJ, Zaki EA, Zhang T, Dennis ES, Mayer KFX,
430 Peterson DG, Rokhsar DS, Wang X, Schmutz J. 2012. Repeated polyploidization of
431 *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* 492:423–427.
432 DOI: 10.1038/nature11798.
- 433 Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing
434 RA, Panaud O. 2006. Doubling genome size without polyploidization: Dynamics of
435 retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice.
436 *Genome Research* 16:1262–1269. DOI: 10.1101/gr.5290206.
- 437 Del Pozo JC, Ramirez-Parra E. 2015. Whole genome duplications in plants: An overview from
438 *Arabidopsis*. *Journal of Experimental Botany* 66:6991–7003. DOI: 10.1093/jxb/erv432.
- 439 R Core Team. 2014. R Language Definition V. 3.1.1. <https://www.r-project.org/>:Accessed Nov
440 2015.
- 441 Sanmiguel P, Bennetzen JL. 1998. Evidence that a recent increase in maize genome size was
442 caused by the massive amplification of intergene retrotransposons. *Annals of Botany* 82:37–
443 44. DOI: 10.1006/anbo.1998.0746.
- 444 Saranga Y. 2007. Special Issue: A century of wheat research - from wild emmer discovery to
445 genome analysis. *Israel Journal of Plant Sciences* 55:207–313.
- 446 Seberg O, Petersen G, Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B,
447 Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2009.
448 Reply: A unified classification system for eukaryotic transposable elements should reflect
449 their phylogeny. *Nature Reviews Genetics* 10:276. DOI: 10.1038/nrg2165-c4.
- 450 Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, Noyes RD, Paterson AH, Wing RA, Wilkins
451 TA, Wendel JF. 2003. Rate variation among nuclear genes and the age of polyploidy in
452 *Gossypium*. *Molecular Biology and Evolution* 20:633–643. DOI: 10.1093/molbev/msg065.
- 453 Singh GB. 2015. Introduction to bioinformatics. In: *Modeling and Optimization in Science and*
454 *Technologies*. 3–10. DOI: 10.1007/978-3-319-11403-3-1.
- 455 Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing:
456 Computational challenges and solutions. *Nature Reviews Genetics* 13:36–46. DOI:
457 10.1038/nrg3117.
- 458 Wang K, Guo W, Zhang T. 2007. Detection and mapping of homologous and homoeologous
459 segments in homoeologous groups of allotetraploid cotton by BAC-FISH. *BMC Genomics*

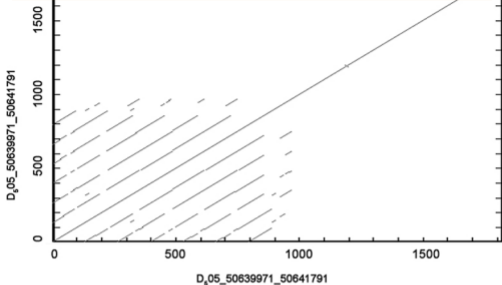
- 460 8. DOI: 10.1186/1471-2164-8-178.
- 461 Wang K, Huang G, Zhu Y. 2016. Transposable elements play an important role during cotton
462 genome evolution and fiber cell development. *Science China Life Sciences* 59:112–121.
463 DOI: 10.1007/s11427-015-4928-y.
- 464 Wang M, Tu L, Yuan D, Zhu D, Shen C, Li J, Liu F, Pei L, Wang P, Zhao G, Ye Z, Huang H,
465 Yan F, Ma Y, Zhang L, Liu M, You J, Yang Y, Liu Z, Huang F, Li B, Qiu P, Zhang Q, Zhu
466 L, Jin S, Yang X, Min L, Li G, Chen LL, Zheng H, Lindsey K, Lin Z, Udall JA, Zhang X.
467 2019. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium*
468 *hirsutum* and *Gossypium barbadense*. *Nature Genetics* 51:224–229. DOI: 10.1038/s41588-
469 018-0282-x.
- 470 Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, Zou C, Li Q,
471 Yuan Y, Lu C, Wei H, Gou C, Zheng Z, Yin Y, Zhang X, Liu K, Wang B, Song C, Shi N,
472 Kohel RJ, Percy RG, Yu JZ, Zhu YX, Cang J, Yu S. 2012. The draft genome of a diploid
473 cotton *Gossypium raimondii*. *Nature Genetics* 44:1098–1103. DOI: 10.1038/ng.2371.
- 474 Wang KB, Wang WK, Wang CY, Song GL, Cui RX, Li SH, Zhang XD. 2001. [Studies of FISH
475 and karyotype of *Gossypium barbadense*]. *Yi chuan xue bao = Acta genetica Sinica* 28:69–
476 75.
- 477 Wang Q, Wang S, Zhu X, Fang L, Hu Y, Chen X, Huang X, Du X, Chen S, Wan Q, Guo W,
478 Chen J, Liu C, Han B, Chen H, Li X, Pan M, Chen ZJ, Mei G, Chang L, Wu H, Huang T,
479 Xiang D, Wang Y, Cai C, Liu B, Zhou B, Zhang T, Gong H, Fang DD. 2017. Genomic
480 insights into divergence and dual domestication of cultivated allotetraploid cottons. *Genome*
481 *Biology* 18. DOI: 10.1186/s13059-017-1167-5.
- 482 WANG K, WENDEL JF, HUA J. 2018. Designations for individual genomes and chromosomes
483 in *Gossypium*. *Journal of Cotton Research* 1. DOI: 10.1186/s42397-018-0002-1.
- 484 Wendel JF, Cronn RC. 2001. Polyploidy and the evolutionary history of cotton. *Advances in*
485 *Agronomy* 78:139–186. DOI: 10.1016/S0065-2113(02)78004-8.
- 486 Wendel JF, Flagel LE, Adams KL. 2012. Jeans, genes, and genomes: Cotton as a model for
487 studying polyploidy. In: *Polyploidy and Genome Evolution*. 181–207. DOI: 10.1007/978-3-
488 642-31442-1_10.
- 489 Wessler SR. 2006. Transposable elements and the evolution of eukaryotic genomes. *Proceedings*
490 *of the National Academy of Sciences* 103:17600–17601. DOI: 10.1073/pnas.0607612103.
- 491 Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P,

- 492 Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2007. A unified
493 classification system for eukaryotic transposable elements. *Nature Reviews Genetics* 8:973–
494 982. DOI: 10.1038/nrg2165.
- 495 Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P,
496 Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2008. A universal
497 classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews*
498 *Genetics* 9:414–414. DOI: 10.1038/nrg2165-c2.
- 499 Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P,
500 Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2009. Reply: A unified
501 classification system for eukaryotic transposable elements should reflect their phylogeny.
502 *Nature Reviews Genetics* 10:276–276. DOI: 10.1038/nrg2165-c4.
- 503 Xie Y, Dong F, Hong D, Wan L, Liu P, Yang G. 2012. Exploiting comparative mapping among
504 Brassica species to accelerate the physical delimitation of a genic male-sterile locus (BnRf)
505 in Brassica napus. *Theoretical and Applied Genetics* 125:211–222. DOI: 10.1007/s00122-
506 012-1826-6.
- 507 Yuan D, Tang Z, Wang M, Gao W, Tu L, Jin X, Chen L, He Y, Zhang L, Zhu L, Li Y, Liang Q,
508 Lin Z, Yang X, Liu N, Jin S, Lei Y, Ding Y, Li G, Ruan X, Ruan Y, Zhang X. 2015. The
509 genome sequence of Sea-Island cotton (*Gossypium barbadense*) provides insights into the
510 allopolyploidization and development of superior spinnable fibres. *Scientific Reports* 5.
511 DOI: 10.1038/srep17662.
- 512 Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM,
513 Hulse-Kemp AM, Wan Q, Liu B, Liu C, Wang S, Pan M, Wang Y, Wang D, Ye W, Chang
514 L, Zhang W, Song Q, Kirkbride RC, Chen X, Dennis E, Llewellyn DJ, Peterson DG,
515 Thaxton P, Jones DC, Wang Q, Xu X, Zhang H, Wu H, Zhou L, Mei G, Chen S, Tian Y,
516 Xiang D, Li X, Ding J, Zuo Q, Tao L, Liu Y, Li J, Lin Y, Hui Y, Cao Z, Cai C, Zhu X,
517 Jiang Z, Zhou B, Guo W, Li R, Chen ZJ. 2015. Sequencing of allotetraploid cotton
518 (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nature*
519 *Biotechnology* 33:531–537. DOI: 10.1038/nbt.3207.
- 520 Zhang X, Tolzmann CA, Melcher M, Haas BJ, Gardner MJ, Smith JD, Feagin JE. 2011. Branch
521 point identification and sequence requirements for intron splicing in plasmodium
522 falciparum. *Eukaryotic Cell* 10:1422–1428. DOI: 10.1128/EC.05193-11.
- 523
- 524

Figure 1(on next page)

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**b**

GTGAAITACTGGCTTCAATGTACTCCACTGCCAACTTCATGGAGGTA AAAATCCGCCACTTCGATCTGCTCCACTACTGCTTAGGGAGACAAGACCTGAAATCTTCAACCTGCTCCACTGCTCGAGGGA

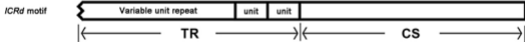
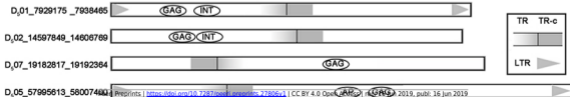
c**d**

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.

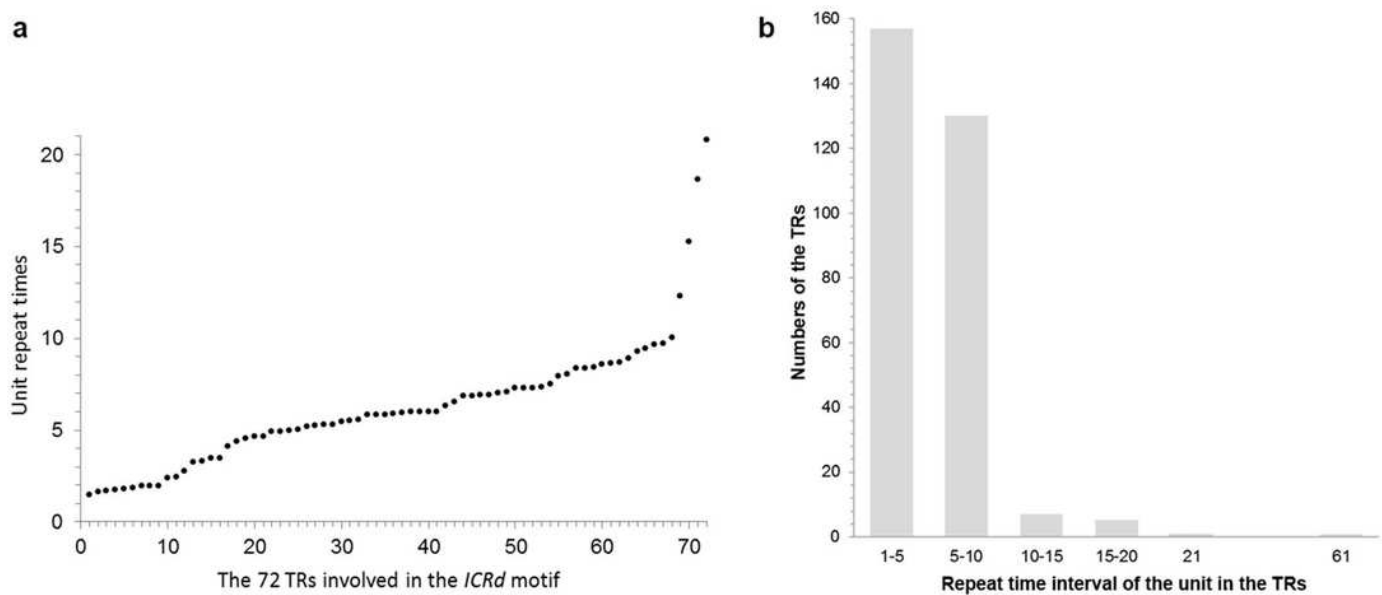


Figure 3

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 = 5 μ m.

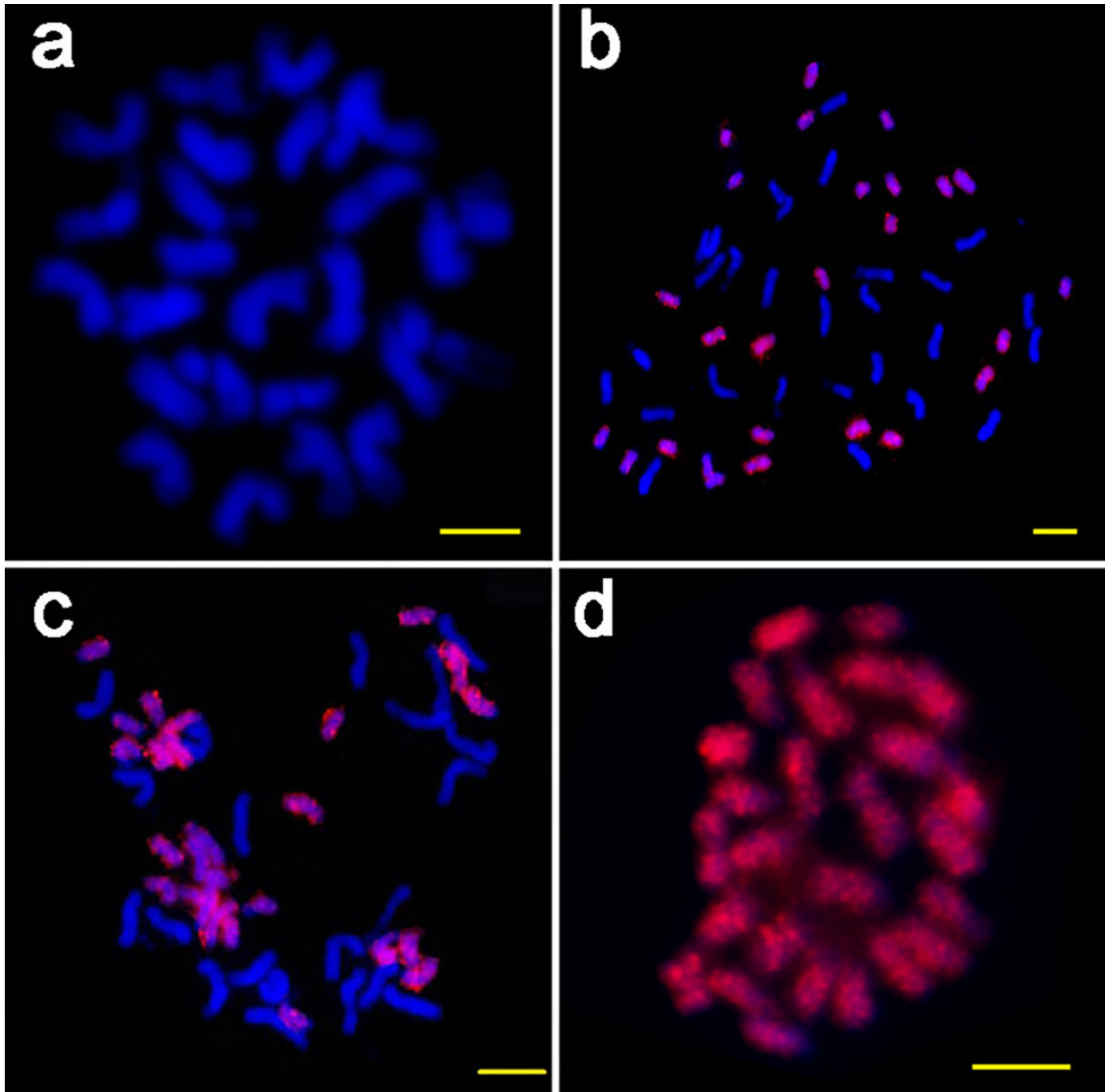


Figure 4 (on next page)

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.

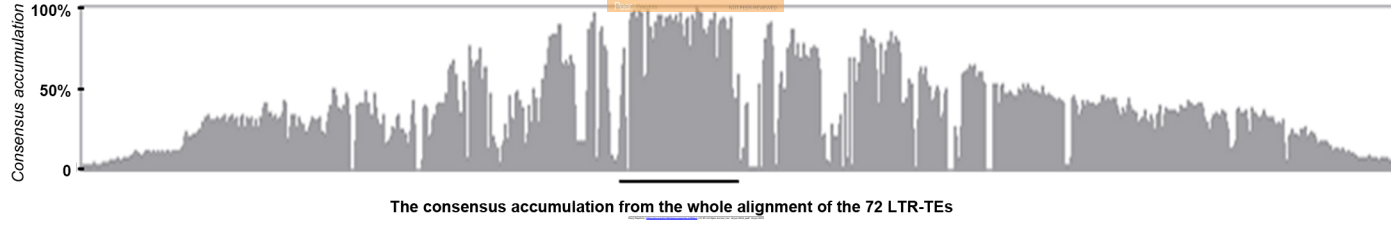


Figure 5

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

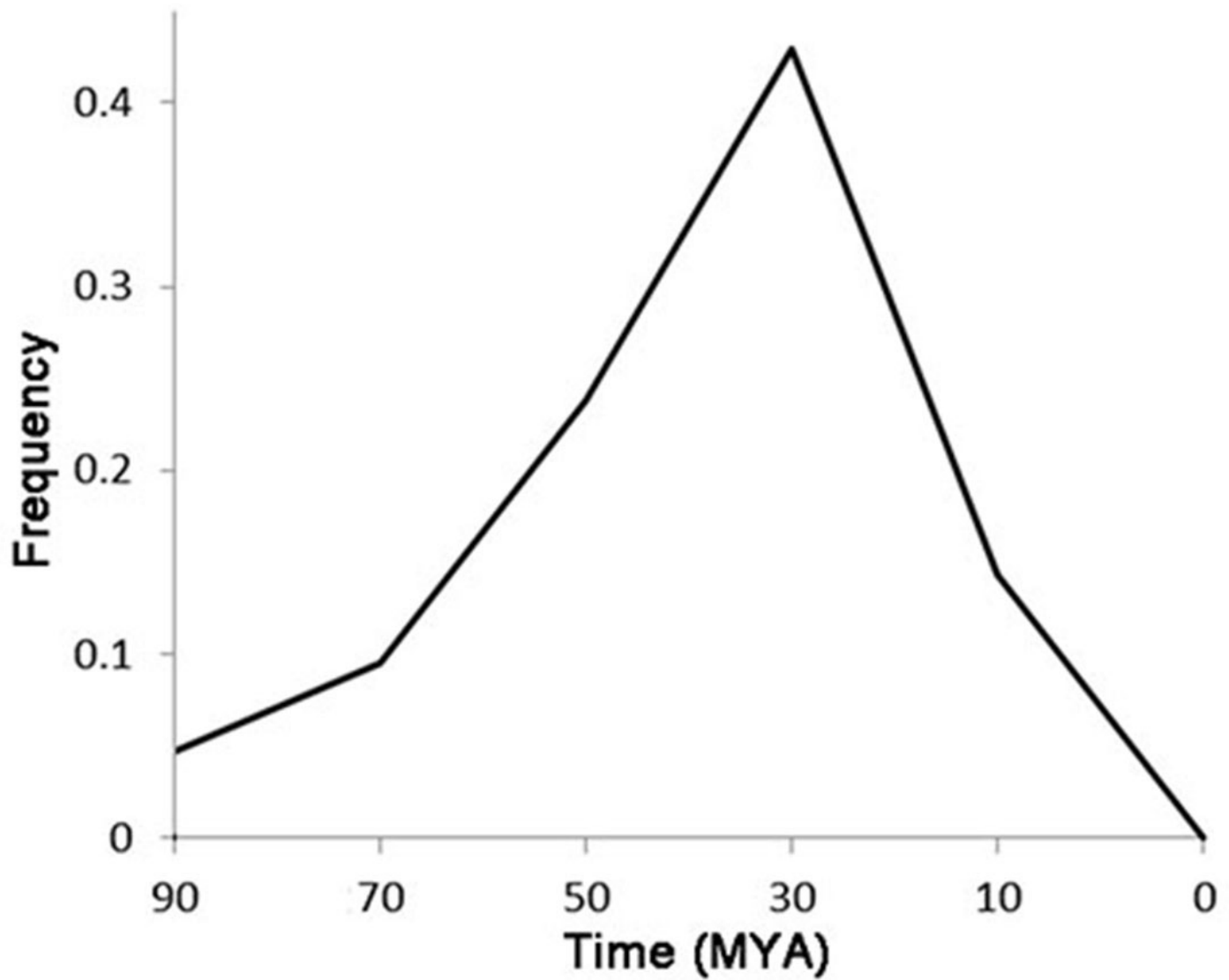


Figure 6(on next page)

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.

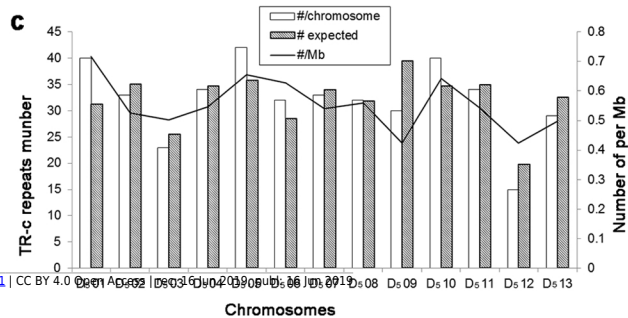
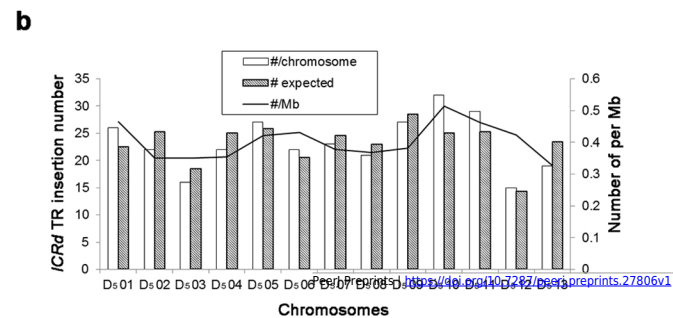
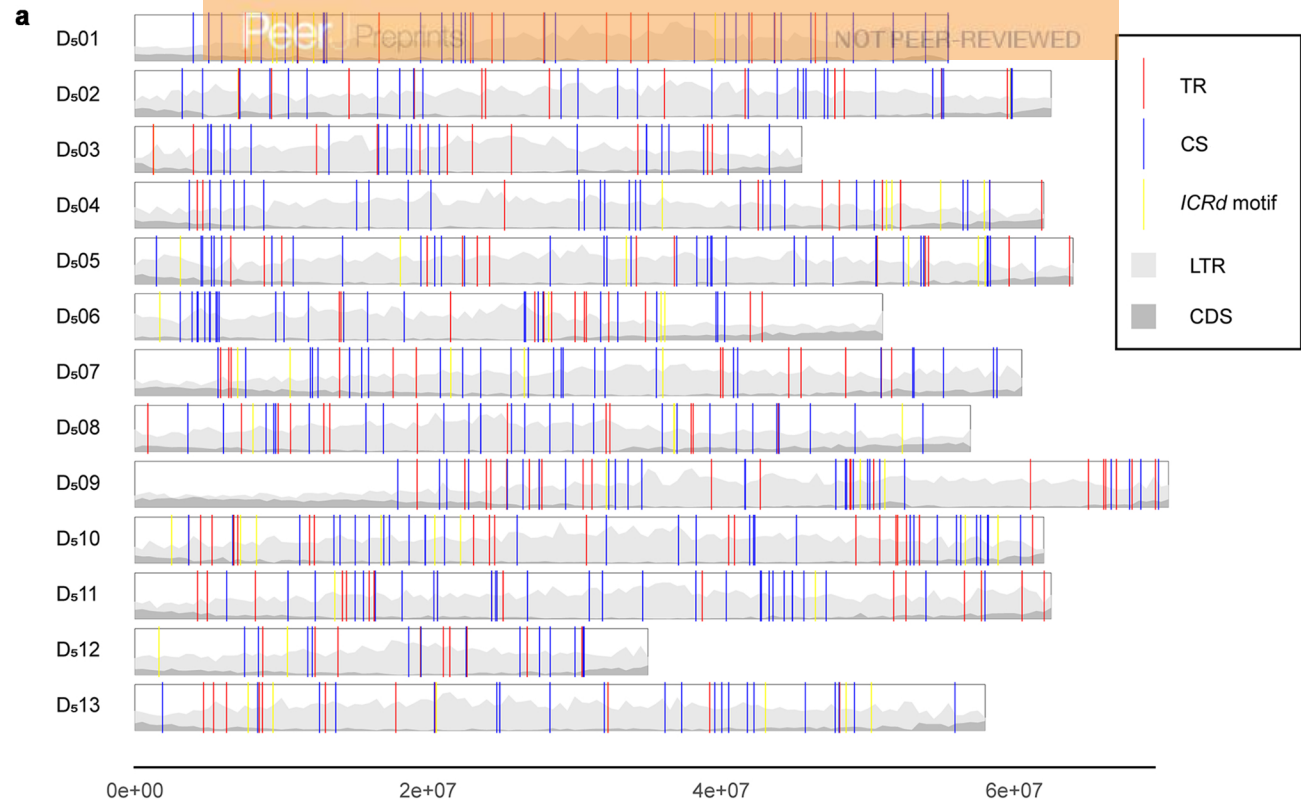
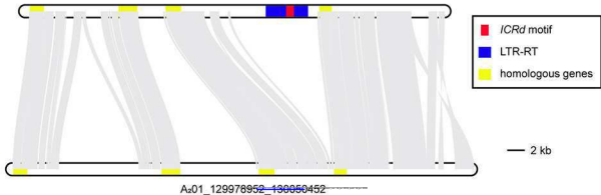


Figure 7 (on next page)

The colinearity of the two homologous segments.

D_s01_9633533_9699033

NOT PEER-REVIEWED



A₂01_129978952_130050452