Theor Appl Genet (2012) 124:1173–1182
DOI 10.1007/s00122-011-1777-3

ORIGINAL PAPER

Localization of high level of sequence conservation and divergence regions in cotton

Kai Wang · Wenpan Zhang · Yujie Cao · Zhongxin Zhang · Dewei Zheng · Baoliang Zhou · Wangzhen Guo · Tianzhen Zhang

Received: 18 June 2011 / Accepted: 15 December 2011 / Published online: 3 January 2012

Abstract In a previous study, we observed that the variations in chromosome size are due to uneven expansion and contraction by comparing the structures and sizes of a pair of homoeologous high-resolution cytogenetic maps of chromosomes 12A and 12D in tetraploid cotton. To reveal the variation at the sequence level, in the present paper, we sequenced two pairs of homoeologous bacterial artificial chromosomes derived from high- to low-variable genomic regions. Comparisons of their sequence variations confirmed that the highly conserved and divergent sequences existed in the distal and pericentric regions, e.g., high- and low-variable genome size regions in these two pairs of cotton homoeologous chromosomes. Sequence analysis also confirmed that the differential accumulation of *Gossypium* retrotransposable *gypsy*-like element (*Gorge3*) accounted for the main contributions for the size difference between the pericentric regions. By fluorescence in situ hybridization analysis, we found that *Gorge3* has a bias distribution in the A_T/A proximal regions and is associated with the heterochromatin along the chromosomes in the entire *Gossypium* genome. These results indicate that, between A_T/A and D_T/D genomes, the distal and pericentric regions usually possess high level of sequence conservation and divergence, respectively, in cotton.

Introduction

The genus *Gossypium* consists of ~50 species, including 40–45 diploids ($2n = 2x = 26$) and 5 allotetraploids ($2n = 4x = 52$). Haploid genome size ranges threefold, from an average 885 Mb in the D-genome species to 2,572 Mb in K-genome species (Hendrix and Stewart 2005). Two groups of diploid species, designated A and D genome (designated A_T and D_T in tetraploid cotton), diverged from a common ancestor about 5–10 million years ago (MYA) and combined to form the allotetraploid species 1–2 MYA (Wendel 1989; Seelanan et al. 1997; Cronn et al. 2002). Despite the relatively young age, they have acquired a twofold difference in genome size (885 Mb in A vs. 1,697 Mb in D). Therefore, this wide range in genome sizes and a well-established phylogeny make *Gossypium* an excellent system for genome size evolution.

In cotton, large orthologous sequences have been compared to evaluate the relative effects of different mechanisms influencing genome size changes between the A_T/A and D_T/D genomes. By comparing 100+ kb of homoeologous sequence surrounding *CesA* gene in tetraploid cotton, high level of sequence conservation (95% sequence identity) between the A_T and D_T genomes was showed (Grover et al. 2004). In contrast, the homoeologous region surrounding the *AdhA* gene showed a nearly twofold difference in aligned sequence length, which is likely to mirror their relative difference in the overall genome size (885 vs. 1,697 Mbp) (Grover et al. 2007). Thus, the divergence of the A_T and D_T genomes did not occur uniformly in the genomes.

Communicated by P. Heslop-Harrison.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1777-3) contains supplementary material, which is available to authorized users.

K. Wang (✉) · W. Zhang · Y. Cao · Z. Zhang · D. Zheng · B. Zhou · W. Guo · T. Zhang (✉)
National Key Laboratory of Crop Genetics and Germplasm Enhancement, Cotton Research Institute, Nanjing Agricultural University, Nanjing 210095, Jiangsu, China
e-mail: kaiwang@njau.edu.cn

T. Zhang
e-mail: cotton@njau.edu.cn

We previously constructed a high-resolution map of chromosomes 12A and 12D, one pair of homoeologous chromosomes, using the genome-original BACs in tetraploid cotton (Wang et al. 2010). Comparisons of homoeologous regions demonstrated uneven genome size changes along these two chromosomes, in which the distal and pericentric regions showed the highest level of chromosomal size conservation and variation, respectively. These results suggest a higher level of sequence divergence in the pericentric regions than the distal regions. To test this hypothesis, we isolated and sequenced two pairs of BACs located in these regions. By sequence comparisons, we confirmed that the highly conserved and divergent sequences existed in the distal and pericentric regions in these two pairs of cotton homoeologous chromosomes. Sequence analysis also confirmed that the high level of sequence-size variation was mainly caused by the bias accumulation of TEs, especially the *Gorge3*-like retrotransposons. By fluorescence in situ hybridization (FISH) analysis, we found that, in the entire *Gossypium* genome, *Gorge3* has a bias massive accumulation in A or A_T genome. These results suggest that, between A_T/A and D_T/D genomes, the pericentric regions usually have a high level of sequence divergence in cotton; on the contrary, their distal regions may retain high level of sequence conservation.

Materials and methods

BAC selection, sequencing and analysis

G. hirsutum acc. TM-1 was used for cytological studies. BACs used for sequence analysis were identified by screening the *G. hirsutum* BAC libraries (Hu et al. 2009). The simple sequence repeat (SSR) markers used for BAC screening were selected from high-density genetic maps derived from populations of the tetraploid *Gossypium* species (Guo et al. 2008).

BAC selection

Previous cytogenetic analysis showed that the pericentric and distal regions displayed the highest and lowest levels of chromosomal size variations, respectively, between homoeologous chromosomes 12A and 12D (Wang et al. 2010). To make a comparison at the DNA sequence level, BACs from both regions were selected and then their sequences were analyzed.

For the low-variable region, BACs identified by marker NAU3896 were selected as candidates for sequencing analysis. One of the D_T genome clone (BAC 067L14), evaluated containing the largest insert size, was sequenced first. PCR primers were then designed according to the different sequence positions along BAC 067L14. Candidate A_T BACs were evaluated for maximal overlap with the sequenced D_T BAC using these primers. Finally, A_T BAC 259M16, which shared the most PCR markers, was selected for sequencing.

For the high-variable region, BACs corresponding to the pericentromeric marker NAU1237 were selected for further sequencing analysis. BAC 215O23 was identified as the only D_T candidate for sequencing, because no other BACs were available after BAC library screening. Then, the A_T BAC 081K08 that was verified to share maximal overlap with clone 215O23 was selected for sequencing.

BAC sequencing

All BACs were shotgun-sequenced and assembled by the Beijing Genomics Institute (BGI, http://www.genomics.org.cn/bgi_new/english/index.htm). Briefly, the BAC DNA was sheared into 1.5- to 3-kb fragments by ultrasonication and then cloned into a pUC118 vector. The sequence reads, which had an average coverage of ~8×, were assembled and vector and low quality sequences were removed prior to sequence assembly. All these BAC sequences were deposited in GenBank (accessions HQ650105- HQ650108).

BAC sequence analysis

Sequence alignments of the homoeologous BAC pairs were performed using BLAST 2 and PAIRWISE-LAGAN software (Brudno et al. 2003). The outputs were compared and checked manually using BioEdit V7.0.9 (http://www.mbio.ncsu.edu/BioEdit/BioEdit.html). Potential genes among the aligned sequences were predicted using three software programs: GeneMark.hmm (Brudno et al. 2003), GENSCAN (Lomsadze et al. 2005) and FGENESH (http://linux1.softberry.com). The predicted genes were confirmed using BLASTN queries against the cotton EST database. To annotate genes, predicted proteins were searched against the non-redundant GenBank protein database using the program BLASTP. The gapped sequences were initially analyzed using BLASTN and CENSOR (Burge and Karlin 1997) to search for homologous elements in the NCBI nucleotide database and Repbase (Kohany et al. 2006). LTR retrotransposons were re-mined using an online program, LTR_FINDER(Jurka et al. 2005), and then subjected to BLASTX analysis for the confirmation of retrotransposon domains.

FISH and image analysis

To generate TE probes for FISH analysis, primers were designed according to their corresponding sequences (Table S2), and then the PCR products were labeled as probe using the FISH analysis. For the copy number evaluation, 50 ng of each labeled probe was used for the subsequent FISH analysis.

The chromosome spreads and FISH procedure were essentially the same as in previously published protocols (Wang et al. 2009). The repeated FISH procedure was similar to that previously described (Wang et al. 2007) with several modifications. After the first round of probing and image capture, the slides were soaked in a 1× PBS (phosphate-buffered saline) solution to remove the coverslips. The slides were dehydrated in an ethanol series (70, 90 and 100%, 5 min each) and used in a second round of FISH hybridization. Slides were examined under an Olympus BX51 fluorescence microscope. Images were captured and merged using Image-Pro Express software V5.0 with an Evolution VF CCD camera (Media Cybernetics, USA). All the probes were repeated in at least three individual FISH experiments, and more than 20 images were captured and analyzed for each probe. To provide identical conditions for semi-quantitative comparisons of different chromosomes or slides, all the images were captured under the same exposure time, 900 ms.

Results

High-level microcolinearity between BACs 259M16 and 067L14

Previous study showed that the distal regions had the highest level of chromosomal size conservation in chromosomes 12A and 12D (Wang et al. 2010). Here, two homoeologous BACs, 259M16 and 067L14, from the most distal regions (Fig. 1) were selected using in further sequence analysis. A_T BAC 259M16 contains 35.9 kb of chromosome 12A, and D_T BAC 067L14 contains 121.4 kb of chromosome 12D. The aligned length of two BACs is 39.7 kb, accounting for 32.3 kb in A_T and 38.6 in D_T (Fig. 2), indicating a 6.3-kb difference in length. The aligned region is relatively short due to a small insert in 259M16. However, as expected, both BACs showed a high similarity (95.3% excluding gaps; 77.8% including gaps) in sequences.

A total of 143 unequally distributed gaps were found between these two BACs. The small indels (<400 bp) represent only a 7-bp difference between the two BACs (Table 1). By contrast, one large indel (6,332 bp absent in A_T BAC 259M16) was determined to be nearly responsible for the size difference between the aligned regions of the A_T and D_T BACs. Sequence analysis by BLAST in NCBI showed that the large indel was an intact *copia* retrotransposon that contained well-defined domains of integrase, reverse transcriptase and RNase H. Their LTRs are 316-bp long and 91.5% identical to each other. However, FISH analysis using the intact *copia* retrotransposon as a probe showed that it was present at nearly equal frequencies in both A_T (gray value 9,670 ± 4,423) and D_T genomes (gray value 7,043 ± 3,378) (Fig. S1a, Table 2), despite this TE occurred only in the D_T genome in this ~40-kb distal region.

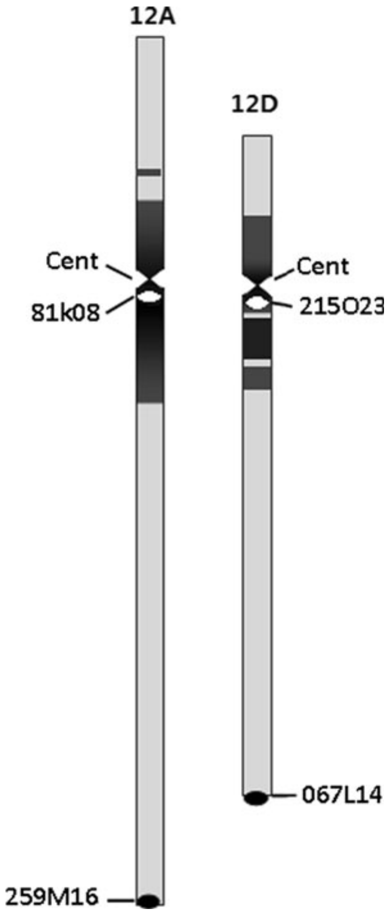

Fig. 1 BAC locations on the chromosomes 12A and 12D. Heterochromatic regions are indicated in *black*. The locations of BACs and heterochromatin are shown based on previous FISH mapping data (Wang et al. 2010)

Five genes were predicted in the co-linear segments present in both BACs (Table S1). The lengths of exons and introns between these five homoeologous genes were found to be nearly the same. The introns of these five genes contain a total of 10.5 and 10.7 kb for chromosomes 12A and 12D, respectively, with individual intron sizes ranging from 474 to 5,327 bp. The 218-bp difference between the total intron lengths in the two homoeologous regions was mainly caused by the presence of an intron of a gene encoding a putative sec10 (Table S1), which is 133-bp longer in 12D than in 12A. The remaining introns showed no obvious size differences.

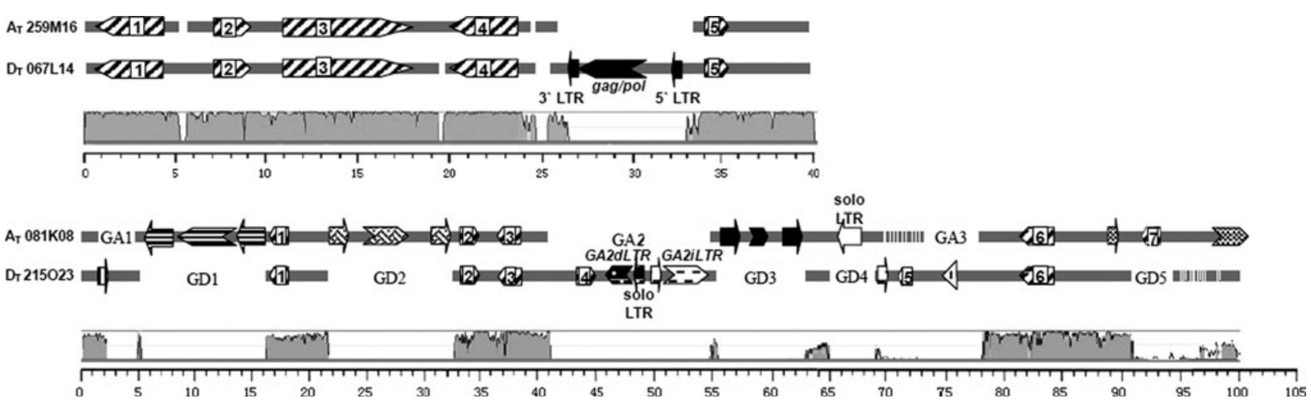

Fig. 2 Pairwise alignments of two pairs of homoeologous BACs: 259M16 and 067L14; and 081K08 and 215O23. Predicted genes are shown as numbered pentagons pointing in the direction of transcription. The numbers of genes in each alignment correspond to the list presented in Table S1. The LTR and *gag/pol* domains of the LTR retrotransposons are depicted by *arrows* and *hexagons*, respectively. Each retrotransposon is shown in a different background pattern. The *Mutator*-like transposon element in GA3 is indicated with a *triangle*. The large indels in the alignments of BACs 081K08 and 215O23 are named GA1–GA3 and GD1–GD5, respectively. Continuous windows of sequence identity as output of PAIRWISE-LAGAN (Brudno et al. 2003) are shown under the alignment between each pair of BACs. All are scaled from 50 to 100%, and the *scale bar* is in kilobases

Table 1 Comparison of indels between A_T and D_T genomes homoeologous BACs in *G. hirsutum* acc. TM-1

	259M16 and 067L14				081K08 and 215O23			
	A_T genome		D_T genome		A_T genome		D_T genome	
	No. indels	bp	No. indels	bp	No. indels	bp	No. indels	bp
Small indels	88	1,085	55	1,078	227	4,456	354	6,666
Large indels	1	6,332	0	0	3	21,998	5	33,903
Total	89	7,417	55	1,078	230	26,454	359	40,569

Table 2 FISH signal gray value estimates

	A_T genome	D_T genome	A genome	D genome
Gorge3-like TEs				
GA1LTR	–	–	–	–
GA2iLTR	39,976 ± 3,672	28,985 ± 5,399	33,123 ± 3,286	9,359 ± 2,585
GD1LTR	18,871 ± 2,445	–	8,056 ± 596	–
GD4LTR	21,179 ± 3,568	–	7,672 ± 535	–
GD5LTR	21,056 ± 2,775	–	24,879 ± 2,505	2,308 ± 732
Gorge3 total	101,082	28,985	73,730	11,667
Others	–	–	–	–
GA2dLTR	–	2,900 ± 304	–	5,717 ± 1,028
GA3Mu	–	3,094 ± 793	–	12,414 ± 2,197
GD2LTR	11,050 ± 1,924	8,600 ± 1,334	15,531 ± 2,589	10,584 ± 1,829
GD3LTR	8,121 ± 1,364	6,491 ± 1,244	6,827 ± 2,498	3,144 ± 1,329
Others total	19,171	21,085	22,358	31,859
Total	120,253	50,070	96,088	43,526
copia retrotransposon in BAC 067L14				
copia	9,670 ± 4,423	7,043 ± 3,378	31,325 ± 5,854	6,993 ± 2,257

A high level of sequence divergence exists between pericentric BACs 081K08 and 215O23

BAC pair 081K08 (12A) and 215O23 (12D) were physically localized in the pericentric region where the highest level of size variation between these two homoeologous chromosomes presented (Fig. 1) (Wang et al. 2010). The insert length of A_T BAC 081K08 and D_T BAC 215O23 are 110.7 and 60.5 kb, respectively. The aligned length of two BACs is 100.6 kb, accounting for 74.1 kb A_T and 60.0 kb in

D_T, apparently indicating a 14.1-kb difference in length. Both BACs have very similar GC contents (34.1 in A_T vs. 34.5% in D_T), but their sequence identity was determined to be only 28.2% (84.6% in the only 33.6-kb gap excluded alignment).

Seven genes were predicted in the co-linear segments: four are present in both, one in A_T and two in D_T (Fig. 2 and Table S1). For the four shared genes, except guaiacol peroxidase, all the other three showed differences in total intron lengths (34–421 bp) between their homoeologous genes. However, the size difference is only 621 bp of the total intron length for the four shared genes and accounts for a very small part (~0.4%) of the total size difference between the two homoeologous BACs, suggesting that the intron length should not be a factor in their size variation.

The biased accumulation of transposition in the pericentric region

The 14.1-kb difference between the pericentric BACs is largely associated with 589 indels, including 230 indels with a total of 26.5 kb in the A_T BAC and 359 indels with 40.6 kb in the D_T BAC (Table 1). Among them, eight large indels, three in GA1–GA3 and five in GD1–GD5 (Fig. 2), account for a major fraction (11.9 kb, 84.3%) of the additional 14-kb sequence in the A_T BAC. The small indels, in contrast, account for a relatively small part of the 14-kb size difference (net gain of 2.2 kb in A_T, i.e., 15.7% of the 14 kb).

To figure out the main forces that caused the size difference, we examined the sequences of these eight large indels. Totally, nine TEs were confirmed to be involved in these eight large indels, because two LTR retrotransposons (designated as GA2iLTR and GA2dLTR here) were identified in the 13-kb large indel GA2. Among them, TE in GA3 (designated as GA3Mu) was identified as a conserved plant mobile domain (pfam10536) and was highly identical (88.3%) to a 700-bp region of *Mutator*-like transposon elements in *G. raimondii* (GB: EF457751). Except for it, all other TEs (except for GA2iLTR and GA2dLTR, all other were designated, respectively, as GA1LTR-GD5LTR) were identified as the LTR retrotransposons again by BLAST homoeology. Depending on LTR_FINDER predicting (Jurka et al. 2005) and well-defined domains of integrase, reverse transcriptase and RNase H by BLAST homoeology, four LTR retrotransposons, GD1LTR, GD2LTR, GD3LTR and GA2iLTR, were deduced as intact LTR retrotransposons.

To evaluate the relationship among these TEs, we further compared their sequence with each other. We found that five of the eight LTR retrotransposons, GA1LTR, GA2iLTR, GD1LTR, GD4LTR, and GD5LTR, showed relatively high degrees of homology with each other (Fig. 3). Among the remaining four, only GD2LTR and GA2dLTR showed a similarity (75.9%) in the 2.2-kb sequence. Further sequence comparison of the five LTR retrotransposons by BLASTN and BLASTX revealed that they were highly similar (71–100%, 364 hits in total) with one group of LTR retrotransposons, *Gorge3*, which was previously implicated to play a role in the genome size variation by lineage-specific amplification in *Gossypium* (Hawkins et al. 2006).

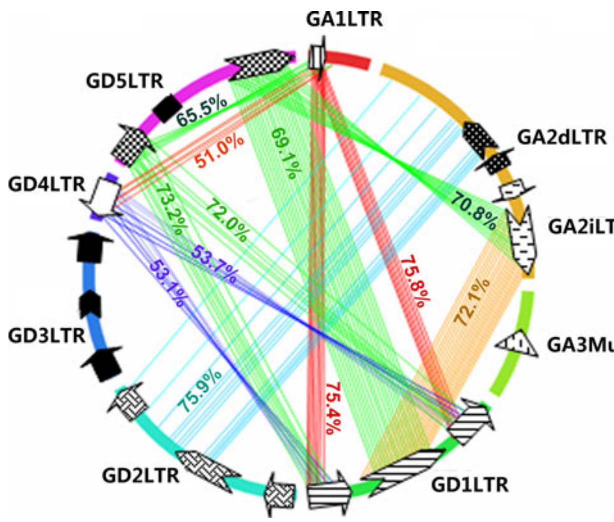

Fig. 3 Analysis of TE sequences. A *circle plot* summarizing the homoeology between these nine TEs found here. TEs shown as in Fig. 2 are drawn as arcs, and homoeologs are connected with lines. The average identities of homoeologs are indicated

Therefore, these observations confirmed that the 11.9-kb size difference in the pericentric region is mainly due to the biased accumulation of TEs, especially the *Gorge3*-like in cotton.

Distribution analysis of TEs in cotton

To evaluate the genomic distributions of the above TEs, we isolated the TE sequences by PCR and then used them as probes in further FISH analysis. For the *Gorge3*-like TEs, all but GA1LTR hybridized strongly to the A_T genome of the tetraploid cotton (Fig. 4a, b, and Fig. S2b, data not shown for GD1LTR). Moreover, they also produced bright signals in the A genome (*G. arboreum*) (Fig. 4a1, b1, Fig. S1d and S1e). For GD5LTR, weak signals could be also found in *G. raimondii* (Fig. 4a2). In addition, GA2iLTR also hybridized strongly to the D_T genome (Fig. 4b) and D genome (Fig. 4b2). The FISH patterns for both GD5LTR and GA2iLTR are likely to reflect their element occurrence in both A_T and D_T (Fig. 2). For GA1LTR, we cannot detect any signal from tetraploid and diploid species. It may be due to its high level of degradation, which causes very short *Gorge3*-like element to remain (1,200 bp).

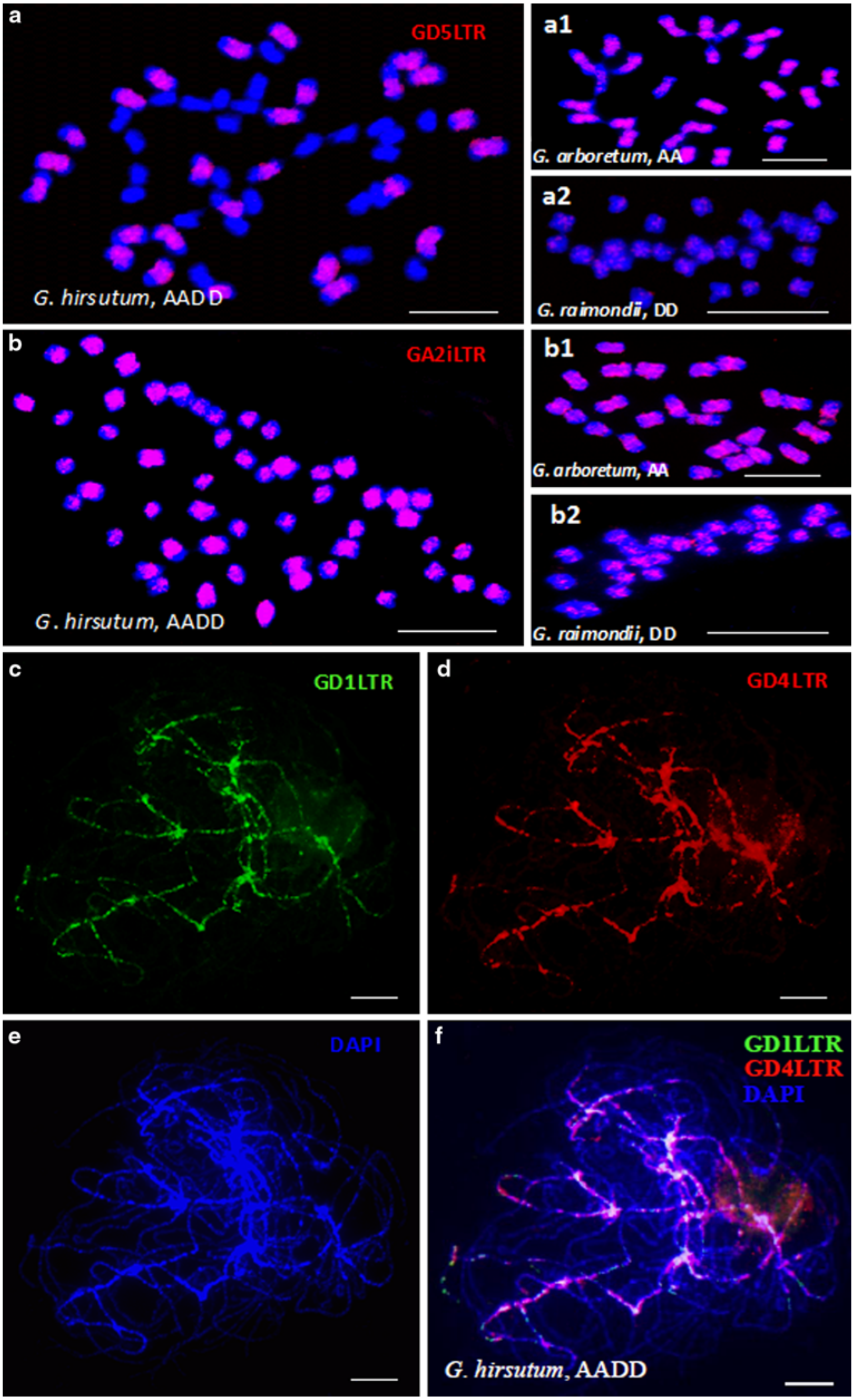

Fig. 4 Distribution analysis of *Gorge3*-like retroelements by FISH. (**a, b**) Probes of GD5LTR and GA2iLTR on *G. hirsutum*, respectively. (**a1** and **a2**) Probe GD5 on *G. arboreum* and *G. raimondii*, respectively. (**b1**, **b2**) Probe GA2iLTR on *G. arboreum* and *G. raimondii*, respectively. (**c–f**) Simultaneous detection by FISH of the GD1LTR (**c**) and GD4LTR (**d**) *Gorge3*-like retroelements on *G. hirsutum* pachytene chromosomes. Usually, the heterochromatin regions of pachytene chromosomes will be heavily stained by DAPI (Kapuscinski 1995). As shown in (**e**), the DAPI-stained heterochromatin regions are indicated as *bright blue* regions. Merged image (**f**) showed that GD1LTR (**c**) and GD4LTR (**d**) have a similar distributions regions, which co-localized with the heterochromatin in cotton. *Scale bar*, 10 µm (color figure online)

Interestingly, when *Gorge3*-like TEs were simultaneously tested in one FISH experiment, they produced nearly overlapping signal regions (data not shown). To reveal it, we further hybridized all of these *Gorge3*-like retrotransposons onto the pachytene chromosome. As shown in Fig. 4c–f, GD1LTR (Fig. 4c) and GD4LTR (Fig. 4d) hybridized to similar genomic regions, which overlapped with the A_T pericentric heterochromatin (Fig. 4e, f) in *G. hirsutum* pachytene chromosomes.

In contrast, the other types of TEs generated much more weak signals and also showed some different distribution patterns in cotton. For example, both GA3Mu and GA2dLTR generated weak signals specific to the D_T (Fig. 5a and Fig S1f) and D genomes (Fig. 5b and Fig S1g). However, for GD2LTR, its signals clustered in the distal regions in both A_T (Fig. 5c) and A genomes (Fig. 5d), but dispersed in both D_T (Fig. 5c) and D genomes (Fig. 5e). Moreover, two pairs of bright spots were found in *G. raimondii* (Fig. 5e). For GD3LTR, it also produced similarly weak and dispersed signals in all chromosomes of both A_T and D_T subgenomes (Fig. 5f), and the A (Fig. S1h) and D (Fig. S1i). Additionally, the *copia* retrotransposon in distal BAC 067L14 was also analyzed by FISH. Its signals also clustered in the distal regions in both A_T (Fig. S1a) and A (Fig. S1b) genomes, but dispersed in both D_T (Fig. S1a) and D genomes (Fig. S1c), similar to those of GD2LTR (Fig. 5 c–e).

To account for the contributions of the above-mentioned TEs, we measured the gray value from FISH signal. As shown in Table 2, the total values of all the TEs (except for the *copia* retrotransposon in BAC 067L14) are 120,253 in A_T and 50,070 in D_T genomes, 96,088 in A and 43,526 in D genomes. The values in both A_T and A genomes have the same onefold higher values than those in D_T and D genomes. However, it is clear that *Gorge3*-like retrotransposons account for the main part of the gray values in both A_T and A genomes, 84.1% (101,082/120,253) in A_T and 76.7% (73,730/96,088) in A genomes. Interestingly, they also make a distinct contribution for the gray values in D_T and D genomes, 57.9% (28,985/50,070) in D_T and 26.8% (11,667/43,526) in D genomes.

Discussion

Non-homogeneous genome size evolution occurred along the genome in cotton

In plants, especially in the large genome species, the evolutionary events affecting genomic size may not have taken place uniformly in the whole genome (Vitte and Bennetzen 2006). Therefore, their genome size usually has an incongruent local and global evolution pattern (SanMiguel et al. 1998; Grover et al. 2004; Peterson-Burch et al. 2004; Choulet et al. 2010). By comparisons of two large homoeologous regions surrounding *CesA* (Grover et al. 2004) and *AdhA* (Grover et al. 2007) genes in cotton, uneven genome size evolution was also found. Our previous results based on cytogenetic mapping have shown that distal and pericentric regions have the highest level of conservation and variation, respectively, between A_T and D_T genomes in cotton (Wang et al. 2010). In this study, results based on sequence analysis indicated that it was mainly caused by the bias accumulation of *Gorge3*-like TEs. However, the preferred A_T/A proximal region distribution of *Gorge3*-like TEs indicates that the non-uniform pattern of chromosomal size variations should exist on the whole cotton genome.

To provide more evidence, we also try to evaluate the chromosomal localizations of *CesA* (Grover et al. 2004) and *AdhA* (Grover et al. 2007) genes based on their excellent sequence assay. Only one BAC was recovered from our BAC library by SSR marker NAU 2533, which was designed according to the *CesA* region sequence (GB: AY632360). Interestingly, it generated signals on the very distal regions of both chromosome 10A and 10D (figure not shown). Several SSR markers designed according to the *AdhA* region (GB: EF457752) have been localized onto the chromosome 1A (W.Z. Guo and T.Z. Zhang unpublished data). We did not identify any BACs by these *AdhA* markers. However, BACs identified by markers NAU2095 and NAU3690 which surround the *AdhA* region were all located in the proximal regions. Furthermore, they all showed A_T dispersed signal pattern, which indicated high-copy-number repeats. All these provide new evidence to support that the distal and proximal regions do exist with a high level of sequence conservation and variation, respectively, between A_T/A and D_T/D genomes in cotton. These results will be very helpful for us to reveal the structure of cotton genome and to complete the future whole genome sequencing.

Gorge3-like retrotransposons played a prominent role in *Gossypium* genome size evolution

A previous study revealed that two types of retrotransposons, *copia*-like and *Gorge3*, accumulated differently in different diploid genomes in *Gossypium* (Hawkins et al. 2006). The *Gorge3* retrotransposon varied greatly in copy number among diploid genomes and accounted for the greatest portion of their difference in genome size (Hawkins et al. 2006; Grover et al. 2007). In this study, five LTR retrotransposons in pericentric BACs were identified as *Gorge3*-like retrotransposons based on high sequence similarities (71–100%) with previously found *Gorge3* retrotransposons (Grover et al. 2004; Hawkins et al. 2006; Grover et al. 2007; Hawkins et al. 2009). The much higher strengths of hybridization in the A_T/A than D_T/D genomes

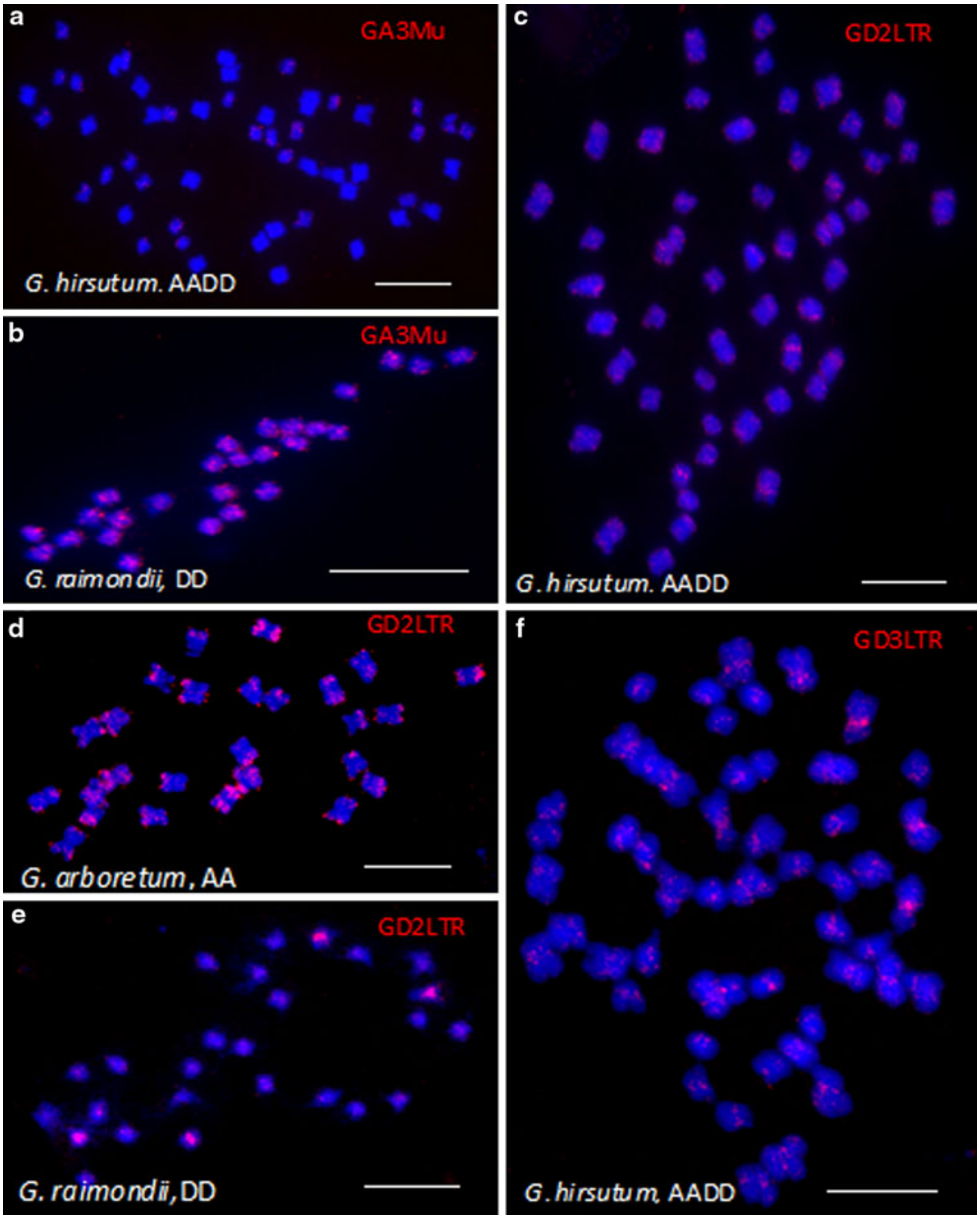

Fig. 5 FISH analysis of non-*Gorge3*-like TEs in cotton. (**a** and **b**) Probe of GA3Mu on *G. hirsutum* and *G. raimondii* genomes, respectively. (**c–e**) Probe of GD2LTR on *G. hirsutum*, *G. arboreum* and *G. raimondii*, respectively. (**f**) Probe of GD3LTR on *G. hirsutum*. *Scale bar*, 10 μm

(Table 2) probably indicated it had a bias massive accumulation in the A_T/A genomes.

The genomic location of *Gorge3*-like TE may provide us clues for revealing their acting on the genome. By FISH analysis, we could decide that the *Gorge3*-like TEs have a similar distribution tendency and prefer integrating into the proximal heterochromatic region (Fig. 4). To confirm it, we recovered the *Gorge3* sequences from *G. hirsutum*, *G. herbaceum* and *G. raimondii* by PCR (Hawkins et al. 2009). FISH tests showed that all of the genome-derived *Gorge3* had a similar preferred A_T/A proximal region distribution pattern (Fig. S2). In plant, retrotransposons usually constitute the bulk of the transposonome (Kumar and Bennetzen 1999; Dooner and Weil 2007) and tend to concentrate in centromere (Jiang et al. 2003), intergenic regions (Sanmiguel and Bennetzen 1998; Jiang et al. 2003) and terminal heterochromatic regions (Pearce et al. 1996). As we know, these regions usually contain no or very few active genes. Therefore, it may indicate that *Gorge3* retrotransposons have evolved to transpose primarily into relatively inactive regions to avoid mutating genes at a high frequency in cotton (SanMiguel et al. 1996). In this way, *Gorge3* retrotransposons could proliferate without being deleterious to the cotton host genome.

Interestingly, our data also showed that some members of *Gorge3*-like retrotransposons, such as GA2iLTR or GD5LTR, could proliferate in both A_T and D_T genomes. But why did the *Gorge3*-like retrotransposons finally get a net lineage-specific amplification? It may be due to that all or some members of *Gorge3* retrotransposons undergo a bias DNA loss in the D_T/D genomes (Hawkins et al. 2009).

In addition to TE activity, the plant genome size may vary due to other mechanisms, such as variation in intron length (Deutsch and Long 1999) and small indel bias (Bensasson et al. 2001; Petrov 2002). Here, in accordance with previous studies (Wendel et al. 2002), no apparent correlation between the whole genome size and intron length was found. By contrast, a relatively substantial amount of sequence variation (2,210 bp, ~15.7% of overall difference) between the high divergence region BACs implicated a small indel bias in the investigation of the cotton genome size differences, even though the data revealed no evidence of an indel bias between the conserved region BACs in this study.

Acknowledgments We thank Dr. Jiming Jiang and Russell J. Kohel for revisingthe manuscript, Xiaowei Niu and Yanjie Jiang for the technical assistance, and Xiangdong Chen for the analysis of *Gorge3*-like retrotransposons. This work was supported by the National Natural Science Foundation of China (30700510, 31071460) and New Century Excellent Talents in University (NCET-10-0496). This work was also supported partly by The Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Bensasson D, Petrov DA, Zhang D-X, Hartl DL, Hewitt GM (2001) Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol Biol Evol 18:246–253

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721–731

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier M-C, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu S, Kong X, Jia J, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22:1686–1701

Cronn RC, Small RL, Haselkorn T, Wendel JF (2002) Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. Am J Bot 89:707–725

Deutsch M, Long M (1999) Intron–exon structures of eukaryotic model organisms. Nucleic Acids Res 27:3219–3228

Dooner HK, Weil CF (2007) Give-and-take: interactions between DNA transposons and their host plant genomes. Curr Opin Genet Dev 17:486–492

Grover CE, Kim H, Wing RA, Paterson AH, Wendel JF (2004) Incongruent patterns of local and global genome size evolution in cotton. Genome Res 14:1474–1482

Grover CE, Kim H, Wing RA, Paterson AH, Wendel JF (2007) Microcolinearity and genome evolution in the *AdhA* region of diploid and polyploid cotton (*Gossypium*). Plant J 50:995–1006

Guo W, Cai C, Wang C, Zhao L, Wang L, Zhang T (2008) A preliminary analysis of genome structure and composition in *Gossypium hirsutum*. BMC Genomics 9:314

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res 16:1252–1261

Hawkins JS, Proulx SR, Rapp RA, Wendel JF (2009) Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proc Natl Acad Sci USA 106:17811–17816

Hendrix B, Stewart JM (2005) Estimation of the nuclear DNA content of *Gossypium* species. Ann Bot 95:789–797

Hu Y, Guo WZ, Zhang TZ (2009) Construction of a bacterial artificial chromosome library of TM-1, a standard line for genetics and genomics in upland cotton. J Integr Plant Biol 51:107–112

Jiang J, Birchler JA, Parrott WA, Dawe RK (2003) A molecular view of plant centromeres. Trends Plant Sci 8:570–575

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Kapuscinski J (1995) DAPI: a DNA-specific fluorescent probe. Biotech Histochem 70:220–233

Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: Repbase-Submitter and Censor. BMC Bioinformatics 7:474

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33:479–532

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res 33:6494–6506

Pearce S, Pich U, Harrison G, Flavell A, Heslop-Harrison J, Schubert I, Kumar A (1996) The *Ty1-copia* group retrotransposons of *Allium cepa* are distributed throughout the chromosomes but are enriched in the terminal eterochromatin Chromosome Res 4:357–364

Peterson-Burch B, Nettleton D, Voytas D (2004) Genomic neighborhoods for Arabidopsis retrotransposons: a role for targeted integration in the distribution of the Metaviridae. Genome Biol 5:R78

Petrov DA (2002) Mutational equilibrium model of genome size evolution. Theor Popul Biol 61:531–544

Sanmiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot 82:37–44

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765–768

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45

Seelanan T, Schnabel A, Wendel JF (1997) Congruence and consensus in the cotton tribe (Malvaceae). Syst Bot 22:259–290

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103:17638–17643

Wang K, Zhang YJ, Guan B, Guo W, Zhang T (2007) Fluorescence in situ hybridization of bacterial artificial chromosome in cotton. Prog Biochem Biophys 34:1216–1222

Wang K, Yang Z, Shu C, Hu J, Lin Q, Zhang W, Guo W, Zhang T (2009) Higher axial-resolution and sensitivity pachytene fluorescence in situ hybridization protocol in tetraploid cotton. Chromosome Res 17:1041–1050

Wang K, Guo W, Yang Z, Hu Y, Zhang W, Zhou B, Stelly D, Chen Z, Zhang T (2010) Structure and size variations between 12A and 12D homoeologous chromosomes based on high-resolution cytogenetic map in allotetraploid cotton. Chromosoma 119: 255–266

Wendel JF (1989) New World tetraploid cottons contain Old World cytoplasm. Proc Natl Acad Sci USA 86:4132–4136

Wendel JF, Cronn RC, Alvarez I, Liu B, Small RL, Senchina DS (2002) Intron size and genome size in plants. Mol Biol Evol 19:2346–2352