Development and Characterization of SSR Markers in the Gossypium Barbadense Genome

Wenjuan ZHONG
Sichuan Academy of Agricultural Sciences, Industrial Crop Research Institute https://orcid.org/0000-0002-9804-8197

Can YUAN
Sichuan Academy of Agricultural Sciences, Industrial Crop Research Institute

Zhengjie CHEN
Sichuan Academy of Agricultural Sciences, Industrial Crop Research Institute

Yonghang ZHOU
Sichuan Academy of Agricultural Sciences, Industrial Crop Research Institute

Siwei Chen
Sichuan Academy of Agricultural Sciences, Industrial Crop Research Institute

Qingxia TANG
Sichuan Academy of Agricultural Sciences,Industrial Crop Research Institute

Chao ZHANG
Sichuan Academy of Agricultural Sciences, Industrial Crop Research Institute

Yiyun GONG
Sichuan Academy of Agricultural Sciences, Industrial Crop Research Institute

Zehu YANG
Sichuan Academy of Agricultural Sciences, Industrial Crop Research Institute

Zhengxuan MAO
Sichuan Academy of Agricultural Sciences, Indutrial Crop Research Institute

Fangsheng MU (fshmu@163.com)
Sichuan Academy of Agricultural Sciences, Institute Crop Research Institute https://orcid.org/0000-0001-8671-7974

Peicheng JI
Sichuan Academy of Agricultural Sciences, Institute Crop Research Institute

Research

Keywords: G. barbadense, Simple sequence repeats (SSRs), Microsatellite markers, Genetic diversity analysis

Posted Date: September 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-870780/v1

Abstract

Background

The fiber quality and resistance traits of *Gossypium barbadense* are considerably better than that of other *Gossypium* species. Simple sequence repeats (SSRs) are user friendly, low cost markers widely used in genetic studies. However, most SSRs have been developed from *G. hirsutum*, *G. arboreum*, and *G. raimondii*; no genome-wide SSRs have been developed from *G. barbadense*.

The *de novo* sequences of *G. barbadense* cv. Xinhai21 were utilized to develop SSR markers and scanned to detect SSRs using the MIcroSAtellite (*http://pgrc.ipk-gatersleben.de/misa/*) identification tool. And then *in silico* PCR analysis was conducted to evaluate these primers polymorphism in five *Gossypium* species.

Results

In total, 85,582 SSRs were identified with different motifs. 153,560 primer pairs were successfully designed for 73,419 SSRs. In *silico* analysis, we found that 8,466 primer pairs of 3,288 SSRs yielded one product (monomorphic) simultaneously in five *Gossypium* species. two *Gossypium* species (30 *G. hirsutum* and 27 *G. barbadense* accessions) were successfully separated by 300 primer pairs with the polymorphism information content (PIC) ranging from 0.00 to 0.93.

Conclusion

These newly developed SSR markers will be helpful for the construction of genetic linkage maps, genetic diversity analyses, QTL mapping, and molecular breeding of *Gossypium* species.

Background

Cotton (*Gossypium* spp.) is one of the most popular sources of natural fiber and oil. The *Gossypium* genus contains 45 diploid (2n = 26) and 6 tetraploid (2n = 52) species (Hawkins et al. 2006; Grover et al. 2015). Only four *Gossypium* species, including two tetraploids and two diploids, produce spinnable fiber. *Gossypium barbadense* (also known as sea-island cotton, extra-long staple cotton, American Pima, and Egyptian cotton) is one of two allotetraploid cultivated cotton species that produces extra-long fibers used in the manufacturing of superior textiles and contributes to 8% of the world's cotton output (Zhang et al. 2008). Due to its excellent fiber quality and high resistance to verticillium wilt, *G. barbadense* is considered an ideal genetic donor for improving the characteristics of *G. hirsutum* such as fiber quality and disease resistance. However, lack of genome-wide molecular markers in *G. barbadense* for genetic research and marker-assisted selection (MAS).

Simple sequence repeats (SSRs), or microsatellites, are repeats of 1–6 bp nucleotides representing high frequency, distribution, co-dominance, reproducibility, and high polymorphism (Powell et al. 1996; Gupta and Varshney 2000). Because of its aforementioned advantages, SSR markers have been widely used in

molecular finger-printing, genetic diversity analysis, mapping genetic linkage, and marker-assisted selection in many species over the past decades (Reddy et al. 2001; Feng et al. 2016). In addition, SSRs developed from expressed sequence tags (EST) possess a higher conserved sequence; hence the transferability of EST-SSRs is higher than that of genomic-SSRs across related species (Zhou et al. 2016). Therefore, EST-SSRs are appropriate for integrating different genetic linkage maps, QTL mapping, and evolutionary correlation (Tani et al. 2003).

Numbers of genomic SSRs have been developed in *Gossypium* species such as *G. hirsutum*, *G. arboreum*, and *G. raimondii*, and are widely used for constructing genetic linkage maps. In the absence of a reference genome for cotton, Reddy et al obtained more than 10,000 microsatellite-containing fragments in *G. hirsutum*, as well as designed primers for 307 out of 588 SSR markers, of which, 49% showed length polymorphism by polymerase chain reaction (PCR) amplification (Reddy et al. 2001). In the *G. hirsutum* cv. Guazuncho2 cDNA library, 846 clones were sequenced, but only 392 sequences containing SSRs had previously designed primers (Nguyen et al. 2004). Similarly, 966 sequences containing SSRs were identified from fiber/ovule cDNA libraries constructed for *G. hirsutum*, from which, 489 SSR primer pairs were developed (Han et al. 2006). Based on the ESTs derived from *G. arboreum* fibers 7–10 days postanthesis, 931 ESTs contained SSRs, of which, 544 had previously designed primers; only 99 were polymorphic between *G. hirsutum* cv. TM-1 and *G. barbadense* cv. Hai7124 (Han et al. 2004). Subsequently, genome of *Gossypium* species was sequenced and 136,345 SSRs were identified from 775.2 Mb sequences in *G. raimondii*; 112,177 primer pairs were designed for these SSRs (Zou et al. 2012). Moreover, 100,290, 83,160, and 56,937 SSRs were developed from the sequences of *G. hirsutum*, *G. arboreum*, and *G. raimondii*, respectively; thousands of primer pairs were designed (Wang et al. 2015). A number of SSR markers can be available from the CottonGen database (www.cottongen.org*)*, however, almost all were derived from *G. hirsutum* (Zhang et al. 2005; Qayyum et al. 2009; Wang et al. 2015), *G. abroreum* (Guo et al. 2006; Liu et al. 2006; Kantartzi et al. 2009), or *G. raimondii* (Zou et al. 2012), while only a few SSRs were designed based on *G. barbadense* sequences, including 214 sequences obtained from *G. barbadense* cv. acc3-79 (Zhang et al. 2009).

G. barbadense cv. Xinhai21 produces extra-long fibers and is used in the production of superior textiles, its genome sequence had been sequenced and gene annotation had been studied in 2015 (Liu et al. 2015). In this study, the *de novo* sequences of *G. barbadense* cv. Xinhai21 were utilized to develop genomic SSR markers and primer pairs were designed for the SSR markers. Further, *in silico* analysis were performed to evaluate the polymorphism of the SSR primer pairs in five *Gossypium* species. To evaluate the potential usefulness of the SSR markers, genetic diversity analysis was conducted in different *G. hirsutum* and *G. barbadense* species. Our results will help improve the application of SSR markers and uncover the genetic basis of dominant *G. barbadense* traits.

Results

Characteristics of genome-wide SSRs in G. barbadense cv. Xinhai21

The published ~ 2264 Mb of *G. barbadense* cv. Xinhai21 genome sequence was utilized to develop SSRs with different types of desirable repeat motifs ranging from mono- to hexa-nucleotides. As a result, a total of 85,582 SSRs were identified genome-wide. Among them, 78,109 SSRs were physically mapped to the 26 chromosomes of *G. barbadense* with an average density of 38.20 per Mb. However, the density of SSRs in the A_t sub-genome was 33.50 per Mb, which was smaller than the D_t sub-genome (46.17 per Mb) (Tables 1; Table 2; Fig. 1). Chromosome D07 had the highest density (52.36 per Mb), while chromosome A04 had the lowest density (28.27 per Mb) (Table 2). In total, 476 types of repeat motifs were detected for mono- (2), di- (4), tri- (10), tetra- (32), penta- (95), and hexa-nucleotides (333) (Table S1). The distributions of mono- to hexa-nucleotide SSRs in the A_t and D_t sub-genomes were different. Overall, the numbers of SSRs in A_t was greater than in D_t, even more than two times greater for mono-nucleotide SSRs (Table 3). From the occurrence frequency of different repeat motifs, hexa-nucleotide repeats were the most abundant (31,843, 37.21%), followed by tri- (18,200, 21.27%), penta- (12,827, 14.99%), di- (12,375, 14.46%), tetra- (7,089, 8.28%), and mono-nucleotides (3,248, 3.80%) (Table 3). Having ranked the different motifs, the top 10 were rich in AAT/ATT (12,870, 13.14%), AAAAAT/ATTTTT (9,616, 9.82%), AT/AT (8,218, 8.39%), AAAAT/ATTTT (5,188, 5.30%), A/T (4,690, 4.79%), AAAT/ATTT (4,074, 4.16%), AAG/CTT (4,015, 4.10%), AAAAAG/CTTTTT (3,609, 3.69%), AG/CT (3,272, 3.34%), and AATCAG/ATTCTG (2,366, 2.42%) (Table 4). Similarly, the numbers of the top 10 motifs in A_t were greater than in D_t. Additionally, some motifs were more frequent in A_t than in D_t. For instance, the number of the AATCAG motif was 2,324 in A_t, but 42 in D_t, and the number of the ACAGG motif was 211 in A_t, but 1 in D_t (Table S2).

Table 1
Overall frequency of SSRs in the *G. barbadense* cv. Xinhai21

Genome	number of SSRs	Genome Size (Mb)	SSR/Mb
Sub genome A_t	47,663	1442.7	33.04
Sub genome D_t	37,919	820.8	46.20
Whole genome	85,582	2263.5	37.81

Table 2
Distribution and average density of SSRs and SSR marker primer sets mapped on *G. barbadense* cv. Xinhai21 chromosomes

Chromosome	Analysis size (Mb)	Total_SSRs	SSR/Mb	SSRs_with_primers	SSR/Mb
A_t01	63.84	2,367	37.08	1,900	29.76
A_t02	98.77	2,837	28.72	2,352	23.81
A_t03	104.2	3,202	30.73	2,608	25.03
A_t04	80.37	2,272	28.27	1,794	22.32
A_t05	100.73	4,272	42.41	3,414	33.89
A_t06	115.58	3,331	28.82	2,731	23.63
A_t07	93.49	3,354	35.88	2,686	28.73
A_t08	116.27	3,650	31.39	2,970	25.54
A_t09	76.81	2,979	38.78	2,353	30.63
A_t10	110.16	3,758	34.11	2,956	26.83
A_t11	115.13	4,125	35.83	3,331	28.93
A_t12	103.35	3,636	35.18	2,940	28.45
A_t13	108.51	3,343	30.81	2,707	24.95
A_t	1287.21	43,126	33.5	34,742	26.99
D_t01	60.1	2,662	44.29	2,217	36.89
D_t02	66.13	2,765	41.81	1,787	27.02
D_t03	50.69	2,167	42.75	1,787	35.26
D_t04	49.32	2,151	43.61	1,788	36.26
Dt05	59.32	3,076	51.85	2,559	43.14
D_t06	60.1	2,650	44.09	2,193	36.49
D_t07	55.25	2,893	52.36	2,363	42.77
D_t08	63.71	2,977	46.73	2,433	38.19
D_t09	47.68	2,287	47.97	1,887	39.58

Chromosome	Analysis size (Mb)	Total_SSRs	SSR/Mb	SSRs_with_primers	SSR/Mb
D_t10	62.34	2,800	44.91	2,279	36.56
D_t11	65.08	3,055	46.94	2,512	38.6
D_t12	58.26	2,782	47.75	2,325	39.91
D_t13	59.7	2,718	45.53	2,227	37.3
D_t	757.67	34,983	46.17	28,357	37.43
Total	2044.88	78,109	38.2	63,099	30.86

Table 3
Distribution of six types of SSR in the *G. barbadense* cv. Xinhai21 genome

Repeat Motifs	Sub genome A_t	Sub genome D_t	Whole genome	Ration (%)
Mono-nucleotide	2,259	989	3,248	3.80
Di-nucleotide	6,767	5,608	12,375	14.46
Tri-nucleotide	10,133	8,067	18,200	21.27
Tetra-nucleotide	3,554	3,535	7,089	8.28
Penta-nucleotide	7,375	5,452	12,827	14.99
Hexa-nucleotide	17,575	14,268	31,843	37.21

Table 4
Top 10 repeat motifs in sub genome At, Dt and whole genome of *G. barbadense* cv. Xinhai21

Whole genome			**Sub genome A_t**			**Sub genome D_t**		
Motifs	Total	Ratio (%)	Motifs	Total	Ratio (%)	Motifs	Total	Ratio (%)
AAT/ATT	12870	13.14	AAT/ATT	7738	16.76	AAT/ATT	5132	13.93
AAAAAT/ ATTTTT	9616	9.82	AAAAAT/ ATTTTT	5560	12.04	AAAAAT/ ATTTTT	4056	11.01
AT/AT	8218	8.39	AT/AT	4584	9.93	AT/AT	3634	9.86
AAAAT/ ATTTT	5188	5.30	AAAAT/ ATTTT	3326	7.20	AAAT/ATTT	1992	5.41
A/T	4690	4.79	A/T	3252	7.04	AAG/CTT	1966	5.34
AAAT/ATTT	4074	4.16	AATCAG/ ATTCTG	2324	5.03	AAAAT/ ATTTT	1862	5.05
AAG/CTT	4015	4.10	AAAT/ATTT	2082	4.51	AG/CT	1600	4.34
AAAAAG/ CTTTTT	3609	3.69	AAG/CTT	2049	4.44	AAAAAG/ CTTTTT	1594	4.33
AG/CT	3272	3.34	AAAAAG/ CTTTTT	2015	4.36	A/T	1438	3.90

Development of Genome-wide SSR primers

A total of 69,750 (81.50%) out of 85,582 SSRs were successfully applied to design primer pairs from their unique flanking sequences, which obtained 209,250 primer pairs (Table S3). The majority of designed SSR primer sets were hexa-nucleotides (26,867, 38.52%), followed by tri- (13,194, 18.92%), penta- (11,072, 15.87%), di- (10,146, 14.55%), tetra- (5,774, 8.27%), and mono-nucleotides (2,697, 3.86%) (Table S3). Moreover, of these 69,750 designed SSR primers, 63,099 SSRs anchored to the 26 chromosomes of *G. barbadense* cv. Xinhai21 with an average marker density of 30.86 SSR markers per Mb (Table 2). Although the number of SSRs in A_t was greater than in D_t, the average marker density in D_t (37.43 per Mb) was greater than in A_t (26.99 per Mb) since the size of A_t (~ 1,442.7 Mb) was much larger than D_t (~ 820.8 Mb). Chromosome D05 had the highest density (43.14 per Mb), whereas chromosome A04 had the lowest density (22.32 per Mb) (Table 2).

In silico **PCR analysis**

The polymorphism of SSR marker primer pairs were evaluated by an *in silico* PCR analysis. It was revealed that the PCR products of 209,250 primer pairs varied with several PCR products including 0, 1, 2, 3, or > 3 products in *G. barbadense* cv. Xinhai21. Of these primers, 153,560 primer pairs were retained because they produced 1, 2, or 3 products. Specifically, 89,940 primer pairs produced 1 product, 49,462 produced 2 products, and 14,158 produced 3 products (Table 5). Thus, these three types of primer pairs were used to evaluate polymorphism in the other four *Gossypium* species. Among the four *Gossypium* species, the number of primer pairs that produced 0 products ranged from 23,378–73,107, primers that produced 1 product ranged from 42,145–81,590, primers that produced 2 products ranged from 4,897–44,602, primers that produced 3 products ranged from 1,405–18,827, and primers that produced > 3 products ranged from 724–24,608 (Table 5). Additionally, it is important to note that 8,466 primer pairs involving 3,288 SSR markers yielded 1 product in all five *Gossypium* species (Table S4).

Table 5
Number of products of primer sets designed from *G. barbadense* cv. Xinhai21 generated through in *silico* analysis in *G. barbadense* cv. acc3-79, *G. arboretum*, *G. hirsutum* and *G. raimondii* genomes

Genome	0 product	1 product	2 products	3 products	> 3 products
G. barbadense cv. Xinhai21	4,306 (2.1%)	89,940 (43.0%)	49,462 (23.6%)	14,158 (6.8%)	51,384 (24.5%)
G. barbadense cv.acc 3–79	23,378 (15.2%)	42,145 (27.4%)	44,602 (29.0%)	18,827 (12.3%)	24,608 (16.0%)
G. arboreum (BGI)	63,383 (41.3%)	81,590 (53.1%)	5,775 (3.8%)	1,568 (1.0%)	1,244 (0.8%)
G.hirsutum (TM-1)	28,044 (18.3%)	75,647 (49.3%)	41,701 (27.2%)	5,842 (3.8%)	2,326 (1.5%)
G. raimondii (JGI)	73,107 (47.6%)	73,427 (47.8%)	4,897 (3.2%)	1,405 (0.9%)	724 (0.5%)

PCR amplification and genetic diversity analysis

Three hundred SSR marker primer pairs (Table S7) were randomly selected from the 8,466 primer pairs, which were used for PCR amplification in the 30 *G. hirsutum* and 27 *G. barbadense* accessions (Table S8). Of these 300 primer pairs, 139 (46.33%) successfully amplified clear, single, expect fragments with noticeable differnces among the 57 *Gossypium* accessions, while 94 showed no difference, 57 showed no expect amplification, and 10 amplified nothing (Fig. 2). Therefore, the genotype bands of the 139 primer pairs were selected for PIC calculation and cluster analysis. Ultimately, PIC values were 0.00–0.56 with an average of 0.05 in the 30 *G. hirsutum* accessions and PIC values varied from 0.00 to 0.93with a mean 0.07 in the 27 *G. barbadense* accessions (Table S7). In the case of a similarity coefficient ≥ 0.61, the 57 *Gossypium* accessions were divided into two subgroups. One group contained the 30 *G. hirsutum* accessions and the other contained the 27 *G. barbadense* accessions (Fig. 3), confirming that these markers developed from *G. barbadense* cv. Xinhai21 could be used to determine genetic diversity in cotton and differentiate *G. barbadense* and *G. hirsutum*.

Analysis of GO terms and KEGG pathways for genes containing SSR markers

Genes of *G. barbadense* cv. Xinhai21 were blasted against the genome sequence of *G. barbadense* cv. acc3-79, which resulted in obtaining 74,371 genes. Of these genes, 50,713 had GO ID information, 20,539 had KEGG ontology, and 13,952 genes had a pathway ID (Table S5). There were 9,378 genes containing at least one SSR marker. Among them, 6,738 genes had GO ID information, 3,014 had KEGG ontology, and 2,002 had a pathway ID. The number of genes on each chromosome ranged from 204 (A_t04) to 533 (A_t05) (Table S6). Moreover, the GO enrichment of genes containing SSR markers were involved in many metabolic processes, such as zinc binding, ubiquitin-protein transferase activity, transcription factor activity, sucrose metabolism, polysaccharide catabolism, hydrolase activity, and microtubule-associated complex. As for the KEGG pathways, processes included ubiquitin-mediated proteolysis, plant hormone signal transduction, starch and sucrose metabolism, and vasopressin-mediated water reabsorption (Table S6).

Discussion

Comparison of SSR markers ddeveloped in the present study and previous studies

Although the development of genome-wide SSR markers in *G. hirsutum*, *G. arboreum*, and *G. raimondii* were reported previously (Wang et al. 2015; Zou et al. 2012), there are no reports for *G. barbadense*. In this study, using the *de novo* sequences of *G. barbadense* cv. Xinhai21 allowed for the identification of 85,582 SSR markers with an average density of 37.81 per Mb. The number of SSRs identified in this study is greater than those identified by Wang et al. for *G. arboreum* (83,160) and *G. raimondii* (56,937), but less those identified by Zou et al. for *G. raimondii* (136,345) and Wang et al. for *G. hirsutum* (100,290). The density of SSRs in *G. barbadense* cv. Xinhai21 was close to *G. hirsutum* (41.2 per Mb), but much lower than *G. arboreum* (49.1 per Mb) and *G. raimondii* (74.8 per Mb and 175.88 per Mb) (Wang et al. 2015; Zou et al. 2012). Interestingly, the distributions of microsatellite length in this study were similar to those identified by Wang et al. both ranking as hexa- (37.21% vs. 39.40%), tri- (21.27% vs. 22.40%), penta- (14.99% vs. 14.90%), di- (14.46% vs. 12.40%), tetra- (8.28% vs. 9.00%), and mono-nucleotides (3.80% vs. 1.80%) (Wang et al. 2015). Additionally, the AAT/ATT motif was the most abundant in *G. barbadense* cv. Xinhai21, accounting for 13.14%, which was consistent with the findings for *G. hirsutum* (14.04%). Although the AAT/ATT motif was the second degree abundant, it accounted for 11.15 and 12.69% in *G. raimondii* and *G. arboreum*, respectively (Wang et al. 2015). These results suggest that the pattern of microsatellite distribution is relatively conserved in *Gossypium* species and remains unchanged in the formation of allotetraploid cotton.

Genotyping advantage of monomorphic SSR markers

Only one PCR product yielded in PCR amplification using monomorphic SSR markers. Consequently, it is easy to distinguish and analyze the genotype. Using 17 monomorphic SSR markers, 14 were identified

that showed differences in alleles among 21 individuals, indicating that there was considerable variation among the flanking sequences of monomorphic SSR markers (Nazareno and Reis 2011). In this study, 8,466 primer pairs yielded one product in *in silico* PCR analysis, of which, 300 monomorphic SSR marker primer pairs were used for PCR amplification in the *G. hirsutum* and *G. barbadense* accessions; 139 (46.33%) out of 300 amplified clear, single fragments, which was agreed with the results of the *in silico* PCR analysis, and successfully classified these two *Gossypium* species. Additionally, the effective amplification of the primer sets (46.33%) was close to that (50.25%; 266 out of 511) identified by Wang et al. (2015), suggesting that monomorphic SSR markers can be widely used in genetic and evolutionary studies.

Application of SSRs developed from G. barbadense in cotton breeding

The quality of *G.bardadense* fiber is very good, thus one breeding goal is to introduce good fiber alleles into other high yield *Gossypium* species, in order to breed better quality fiber and higher yield varieties. According to the GO enrichment analysis, many genes containing SSRs were found to be possibly involved in fiber development (Table S6). For instance, 11 SSRs located on chromosomes A05 (1), A12 (2), D04 (1), D08 (3) and D12 (4) were identified with the term GO: 0009733, which is related to auxin biosynthesis and plant hormone signal transduction.

Expression of the IAA biosynthetic gene, *iaaM*, by transgenic methods increased the IAA levels during the fiber initiation stage, resulting in significantly improved lint yield and fiber fineness (Zhang et al. 2011). Some studies inferred that secondary metabolic compounds, such as phenylpropanoids, could affect cotton fiber (Hovav et al. 2008; Yang et al. 2006), and phenylpropanoid-related genes exhibited spatial distribution changes during the development of fiber (Yves et al. 2009). There were 25 SSRs associated with phenylpropanoid biosynthesis, which were located on chromosomes A03 (2), A04 (1), A05 (2), A09 (2), A11 (2), A12 (1), A13 (1), D02 (1), D03 (1), D04 (1), D05 (7), D09 (2), D12 (1), D13 (1), and UKA (1), and all are involved in a single pathway, ko00940, that encodes different enzymes, including O-methyltransferases, peroxidases, transferases, and metabolic enzymes.

As is previously reported, cellulose accounts for more than 94% of mature cotton fiber (Haigler et al. 2001). Five SSRs identified with the term GO: 0030244 were associated with cellulose biosynthesis and were distributed on chromosomes A02 (2), A06 (1), D05 (1), and D11 (1). The formation of cellulose needs sucrose synthase (*Sus*) to catalyze uridine diphosphate glucose (UDP-G), the precursor for cellulose (Amor et al. 1995; Ruan 2007). Eight SSRs identified with the term GO: 0016157 were associated with sucrose synthase, regulating starch, and sucrose metabolism (ko00500). Moreover, extensive arrays of microtubules are essential for the elongation of oriented cellulose microfibrils (He et al. 2008). Several β-tubulin (TUB) genes are highly expressed in elongating cotton fiber cells. Eight genes rescued the lethality phenotype when 9 TUB genes were transformed into the *tub2* mutant, which is deficient in b-tubulin (He et al. 2008). Two SSRs located on chromosomes A04 and D05 that were identified with the term GO: 0048487 are related to β-tubulin binding. Collectively, the SSR markers developed in this study can be used to uncover the genetic basis of fiber development and improve the

quality of fiber by MAS. Additionally, these SSRs will be helpful for studying other traits, including biotic and abiotic resistance.

Conclution

In summary, a genome-wide development of SSR markers in sea-island cotton (*G. barbadense* cv. Xinhai21) was performed. We characterized the genomic SSRs and developed the genome-wide SSR

primer from assembled genomic sequence, also evaluated polymorphism in different species. Furthermore, we conducted GO and KEGG enrichment analysis for genes containing SSRs. These results indicate that the newly development SSR markers can serve as a useful tool for genetic analysis and molecular breeding, especially on the construction of genetic linkage maps, genetic diversity analyses, QTL mapping of important traits, such as fiber quality and biotic resistance traits.

Methods

Acquisition of genome sequences

Genome sequence of *G. barbadense* cv. Xinhai21 was available from the NCBI database (https://www.ncbi.nlm.nih.gov/search/all/?term=PRJNA251673*)*, which was used to develop SSR markers. Genome sequences of *G. barbadense* (acc3-79), *G. arboreum* (BGI), *G. hirsutum* (TM-1), and *G. raimondii* (JGI) were downloaded from the CottonGen database (https://www.cottongen.org*)*, which was used for in *silico* analysis to evaluate the polymorphism of SSR primer pairs.

Simple sequence repeat (SSR) identify and primer design

Genome sequences of *G. barbadense* cv. Xinhai21 were scanned to detect SSRs using the MIcroSAtellite (http://pgrc.ipk-gatersleben.de/misa/*)* identification tool with basic motifs ranging from mono- to hexa-nucleotides (Thiel et al. 2003). The minimum length criteria were defined as 18, 9, 6, 5, 4, and 3 repeat units for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide motifs, respectively. Primers were designed based on the flanking sequences of SSR loci using Primer 3 v2.2.3 with the following parameters: 18–27 bp primer length, 57°C–63°C melting temperature, 30–70% GC content, and 100–280 bp product size (Rozen and Skaletsky 2004). For each SSR, three primer pairs were designed; if one sequence of the two primer pairs was identical, the two pairs were merged into one primer pair.

In silico analysis

In order to evaluate the polymorphism of the aforementioned SSR primer pairs, an *in silico* PCR analysis was conducted. First, sequences of SSR marker regions were aligned to the genome sequence of *G. barbadense* cv. Xinhai21 using e-PCR-2.3.11 (ftp://ftp.ncbi.nlm.nih.gov/pub/schuler/e-PCR/*)* with the following default parameters: 2 bp mismatch, 1 bp gap, 50 bp margin, and 50–1000 bp product size (Kirill et al. 2004; Shi et al. 2014; Wang et al. 2015). Only the SSR marker primer pairs that yielded one

product (monomorphic), two products, and three products were retained for further analyses. Then, these three types of SSR marker primer pairs were subjected to a BLAST analysis against the assembled contigs of *G. barbadense* (acc3-79), *G. arboreum* (BGI), *G. hirsutum* (TM-1), and *G. raimondii* (JGI) using e-PCR-2.3.11.

Evaluation of potential usefulness for SSR markers through genetic diversity analysis

To evaluate the potential usefulness of SSR marker primer pairs, a total of 300 primer pairs that yielded one product simultaneously in *G. barbadense* cv. Xinhai21, *G. barbadense* (acc3-79), *G. arboreum* (BGI), *G. hirsutum* (TM-1), and *G. raimondii* (JGI) were randomly chosen. Then, primer pairs were experimentally PCR amplified in 30 *G. hirsutum* and 27 *G. barbadense* accessions. Genomic DNA of the 30 *G. hirsutum* and 27 *G. barbadense* accessions were extracted from fresh young leaves using an improved cetyltrimethyl ammonium bromide (CTAB) method (Paterson et al. 1993). PCR amplification was performed on a 20.0 μL sample consisting of 10.0 μL 2×Tag Mastermix (Cat.CW0682L), 1.0 μL (10 μM) forward primer, 1.0 μL (10 μM) reverse primer, and 3.0 μL (50 ng/μL) gDNA template. The PCR cycling conditions were as follows: 2 min at 94°C, 35 cycles at 94°C for 40 s, 35 s at the annealing temperature for each specific primer, and 60 s at 72°C with a 7 min extension at 72°C in the final cycle. PCR products were separated on 8% denaturing polyacrylamide gel (PAGE) with 1×TBE buffer running at 180 V for 45–60 min and visualized by sliver-staining to check the amplification (Santos et al. 1993). Polymorphism information content (PIC) value was used to assess the allelic polymorphism of each SSR marker, which was calculated by using formula as previous proposed by Liu et al (2015). Then, the genetic relatedness of the 30 *G. hirsutum* and 27 *G. barbadense* accessions was analyzed. After SSR band data were standardized, cluster analysis was conducted based on the similarity coefficients and un-weighted pair group method with arithmetic averages (UPGMA) using the NTSYS-pc v2.10 package (Rohlf 1987; Xie et al. 2006).

GO and KEGG enrichment analyses of genes containing SSRs

In order to investigate the GO and KEGG enrichment of genes containing at least one SSR, GO terms of *G. barbadense* cv. Xinhai21 genes were annotated first, and then KEGG pathway was annotated since *G. barbadense* cv. Xinhai21 genes have no GO and KEGG enrichment information. In brief, sequences of *G. barbadense* cv. Xinhai21 genes were blasted against the genome sequence of *G. barbadense* cv. acc3-79 whic has GO term and KEGG pathway annotation information, such that obtain the GO and KEGG enrichment information of corresponding genes. Then, the aforementioned SSRs were anchored to the *G. barbadense* cv. Xinhai21 genome to determine the psychical position of SSRs and if gene contains SSR marker. Finally, genes containing SSR(s) were screened from the *G. barbadense* cv. Xinhai21 genes, thus obtaining the GO and KEGG enrichment information.

Abbreviations

SSR Simple sequence repeats

PCR Polymerase chain reaction

PIC Polymorphism information content

MAS Marker-assisted selection

EST Expressed sequence tags

QTL Quantitative trait loci

CTAB Cetyltrimethyl ammonium bromide

Declarations

Acknowledgements

We thank National Medium-term Gene Bank of Cotton in China and National Cotton Germplasm Resources Platform provided the 30 *G. hirsutum* and 27 *G. barbadense* accessions in Table S8. This work was funded by the National Key Technology R&D Program, the Ministry of Science and Technology (2016FYD0100203 and 2016YFD0101410).

Authors' contributions

Mu FS, Zhong WJ and Yuan C conceived and designed the experiments; Chen ZJ, Chen SW, Zhou YH, Ji PC, Gong YY and Yang ZH planted the materials and assisted in experiments, Zhong WJ , Tang QX , Yuan C performed most of the experiments and analyzed the data. Wen juan Zhong and Zhengjie Chen wrote the manuscript. Fangshen Mu and Ji PC revised the manuscript. Chao Zhang supervised the experiment. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflict of interest.

References

Amor Y, Haigler CH, Johnosn S, et al. membrane-associated form of sucrose synthase and its potential role in synthesis of cellulose and callose in plants. Proceedings of the National Academy of Sciences of

the United States of America.1995;92: 9353-7. https://doi.org/10.1073/pnas.92.20.9353.

Feng SG, He RF, Liu JJ, et al.Development of SSR markers and assessment of genetic diversity in medicinal *Chrysanthemum* morifolium cultivars. Frontiers in Genetics. 2016;7:113. https://doi.org/10.3389/fgene.2016.00113.

Grover CE, Zhu X, Grupp KK, et al. Molecular confirmation of species status for the allopolyploid cotton species, Gossypium ekmanianum Wittmack. Genet Resour Crop Evol.2015;62: 103–14. https://doi.org/10.1007/s10722-014-0138-x.

Guo WZ, Zhou BL, Yang LM, et al. Genetic diversity of landraces in *Gossypium arboreum* L. Race sinense assessed with simple sequence repeat markers. Journal of Intergrative Plant Biology. 2006;48: 1008-1017. https://doi.org/10.1111/j.1744-7909.2006.00316.x.

Gupta PK, Varsheny RK.The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. Euphytica.2000; 113: 163-85. https://doi.org/10.1023/A:1003910819967.

Haigler CH, Ivanova-Datcheva M, Hogan PS,et al. Carbon partitioning to cellulose synthesis. Plant Molecular Biology.2001; 47: 29-51.

Han ZG, Wang CB, Song XL, et al.Characteristics, development and mapping of Gossypium hirsutumderived EST-SSRs in allotetraploid cotton. Theoretical & Applied Genetics.2006; 112: 430-9. https://doi.org/10.1007/s00122-005-0142-9.

Han ZG, Guo WZ, Song XL, et al. Genetic mapping of EST-derived microsatellites from the diploid Gossypium arboreum in allotetraploid cotton. Mol Genet Genomics.2004; 272: 308-27. https://doi.org/10.1007/s00438-004-1059-8.

Hawkins JS, Kim HR, Nason JD, et al. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res. 2006;16:1252-61. https://doi.org/10.1101/gr.5282906.

He XC, Qin YM, Xu Y, et al. Molecular cloning, expression profiling, and yeast complementation of

19 beta-tubulin cDNAs from developing cotton ovules. Journal of Experimental Botany. 2008;59(10):2687-2695. https://doi.org/10.1093/jxb/ern127.

Hovav R, Udall JA, Chaudhary B, et al. The evolution of spinnable cotton fiber entailed prolonged development and a novel metabolism. PLoS Genet. 2008;4(2):e25. https://doi.org/10.1371/journal.pgen.0040025.

Kantartzi SK, Ulloa M, Sacks E, et al. Assessing genetic diversity in *Gossypium arboreum L.* cultivars using genomic and EST-derived microsatellites. Genetica.2009; 136: 141-7. https://doi.org/10.1007/s10709-008-9327-x.

Kirill R, Wonhee J, Schuler GD. A web server for performing electronic PCR. *Nucleic Acids Research*. 2004;32:108-12. https://doi.org/10.1093/nar/gkh450.

Liu DQ, Guo XP, Lin ZX, et al. Genetic diversity of Asian Cotton (*Gossypium arboreum L.*) in China evaluated by microsatellite analysis. Genetic Resources & Crop Evolution.2006; 53: 1145-52. https://doi.org/10.1007/s10722-005-1304-y.

Liu J, Qu JT, Yang C, et al. Development of genome-wide insertion and deletion markers for maize, based on next-generation sequencing data. BMC Genomics.2015; 16: 601. https://doi.org/10.1186/s12864-015-1797-5.

Liu X, Zhao B, Zheng HJ, et al. *Gossypium barbadense* genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. Scientific Reports.2015;5: 14139. https://doi.org/10.1038/srep14139.

Tani N, Takahashi T, Iwata H, et al. A consensus linkage map for sugi (Cryptomeria japonica) from two pedigrees, based on microsatellites and expressed sequence tags. Genetics. 2003;165:1551-68.

Nazareno A, Reis M. The same but different: Monomorphic microsatellite markers as a new tool for genetic analysis. American Journal of Botany.2011; 98: 265-7.https://doi.org/10.3732/ajb.1100163.

Nguyen TB, Giband M, Brottier P, et al. 2004. Wide coverage of the tetraploid cotton genome using newly developed microsatellite markers. Theoretical & Applied Genetics.2014; 109: 167-75. https://doi.org/10.1007/s00122-004-1612-1.

Paterson AH, Brubaker CL, Wendel JF. A rapid method for extraction of cotton (*Gossypium spp.*)

genomic DNA suitable for RFLP or PCR analysis. Plant molecular biology reporter.1993; 11: 122-7. https://doi.org/10.1007/BF02670470.

Powell W, Machray GC, Provan J. Polymorphism revealed by simple sequence repeats. Trends Plant

Sci.1996; 1: 215-22. https://doi.org/10.1016/1360-1385(96)86898-1.

Qayyum A, Murtaza N, Azhar F, et al. Biodiversity and nature of gene action for oil and protein contents in *Gossypium hirsutum L.* estimated by SSR markers. Journal of Food Agriculture and Environment. 2009; 7: 590-3.

Reddy, OUK, Pepper AE, Abdurakhmonov I, et al. New dinucleotide and trinucleotide microsatellite marker resources for cotton genome research. Journal of Cotton Science. 2001; 5: 103-13.

Rohlf FJ. NTSYS-pc: Microcomputer Programs for Numerical Taxonomy and Multivariate Analysis. The American Statistician. 1987; 41 (4): 330. DOI:10.2307/2684761.

Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods

in Molecular Biology. 2000;132: 365. https://doi.org/10.1385/1-59259-192-2:365.

Ruan YL. Rapid cell expansion and cellulose synthesis regulated by plasmodesmata and sugar: insights

from the single-celled cotton fibre. Functional Plant Biology. 2007; 34: 1-10.

DOI: 10.1071/FP06234.

Santos FR, Pena SDJ, Epplen JT. Genetic and population study of a Y-linked tetranucleotide repeat

DNA polymorphism with a simple non-isotopic technique. Hum Genet.1993; 90: 655-56. https://doi.org/10.1007/BF00202486.

Shi JQ, Huang SM, Zhan JP, et al. Genome-wide microsatellite characterization and marker development in the sequenced *Brassica* crop species. Dna Research. 2014; 21: 53-68. https://doi.org/10.1093/dnares/dst040.

Thiel T, Michalek W, Varshney R, et al. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare L.*). Theoretical & Applied Genetics. 2003; 106: 411-22. https://doi.org/10.1007/s00122-002-1031-0.

Wang Q, Fang L, Chen J, et al. Genome-wide mining, characterization, and development of microsatellite markers in *Gossypium* species. Scientific Reports. 2015; 5: 10638. https://doi.org/10.1038/srep10638.

Xie H, Sui Y, Chang FQ, et al. SSR allelic variation in almond (*Prunus dulcis Mill.*). Theoretical and

Applied Genetics. 2006; 112: 366–72. https://doi.org/10.1007/s00122-005-0138-5.

Yang SS, Cheung F, Lee JJ, et al. Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. The Plant Journal. 2006; 47: 761-75. https://doi.org/10.1111/j.1365-313X.2006.02829.x

Yves AG, Stephane B, Tony A, et al. Transcript profiling during fiber development identifies pathways in secondary metabolism and cell wall structure that may contribute to cotton fiber quality. Plant & Cell Physiology. 2009; 50: 1364-81. https://doi.org/10.1093/pcp/pcp084.

Zhang HB, Li YN, Wang BH, et al. Recent advances in cotton genomics. International Journal of Plant Genomics. 2008; 2008: 742304. https://doi.org/10.1155/2008/742304.

Zhang JF, Yang L, Cantrell RG, et al. Molecular marker diversity and field performance in commercial cotton cultivars evaluated in the Southwestern USA. Crop Science. 2005; 45: 1483-90. https://doi.org/10.2135/cropsci2004.0581.

Zhang M, Zheng XL, Song SQ, et al. Spatiotemporal manipulation of auxin biosynthesis in cotton ovule epidermal cells enhances fiber yield and quality. Nature Biotechnology.2011; 29: 453-8. https://doi.org/10.1038/nbt.1843.

Zhang PP, Wang XQ, Yang Y, et al. Isolation, characterization, and mapping of genomic microsatellite markers for the first time in Sea-Island Cotton (*Gossypium barbadense*). Acta Agronomic Sinica.2009; 35: 1013-20. DOI:10.3724/SP.J.1006.2009.01013.

Zhou Q, Lou D, Ma LC,et al. Development and cross-species transferability of EST-SSR markers in Siberian wildrye (*Elymus sibiricus L.*) using Illumina sequencing. Scientific Reports.2016; 6: 20549. https://doi.org/10.1038/srep20549.

Zou CS, Lu CR, Zhang YP, et al. Distribution and characterization of simple sequence repeats in *Gossypium raimondii* genome. Bioinformation.2012; 8: 801-6. https://doi.org/10.6026/97320630008801.

Figures

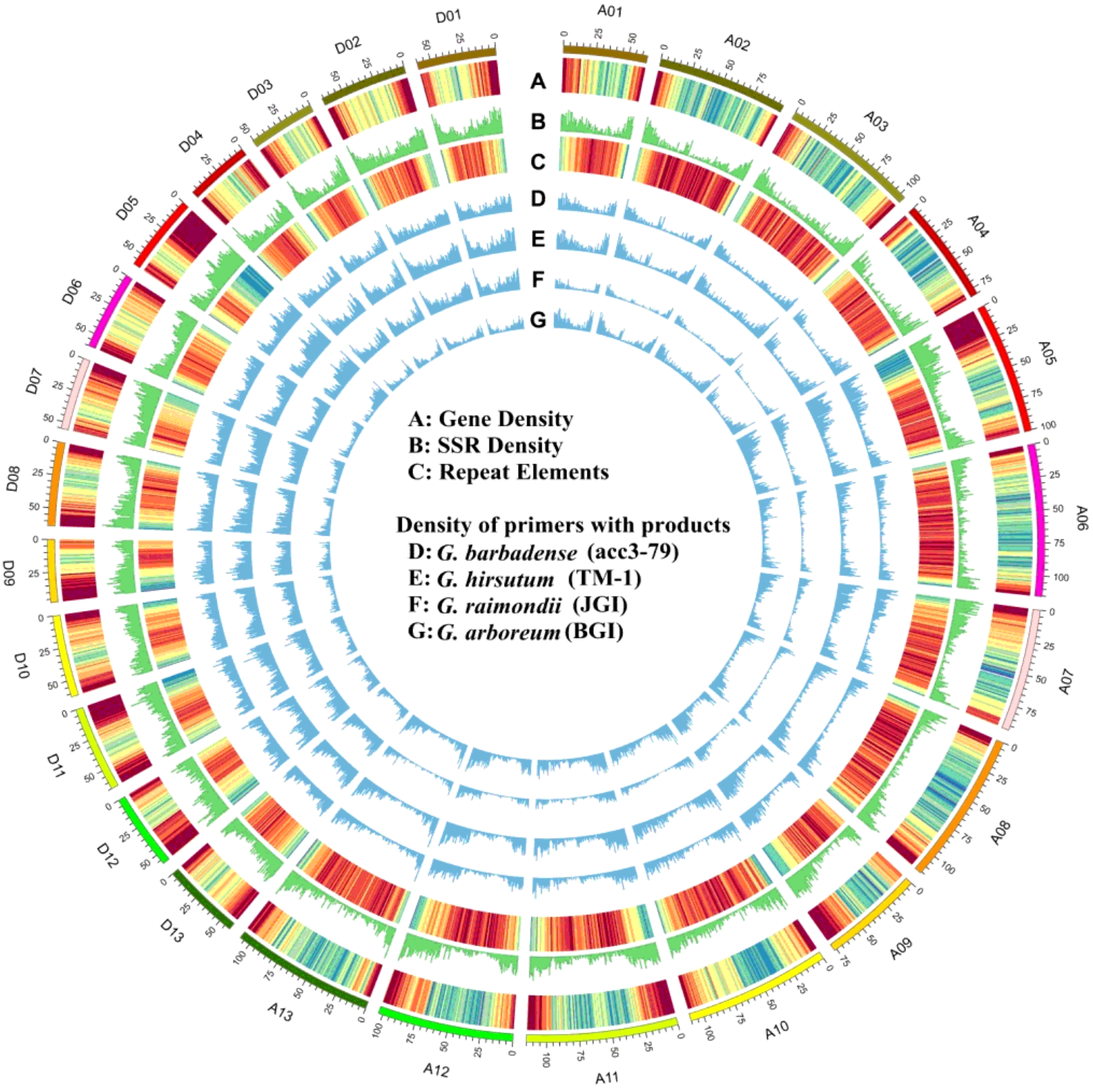

Figure 1

Heat map of gene density, SSR density and repeat elements of G. barbadense cv. Xinhai21 and density of primer sets with products. (A, C) Colors from red to yellow, to green indicate density from high to low (B, D-G). The higher the peak is, the higher the density is.

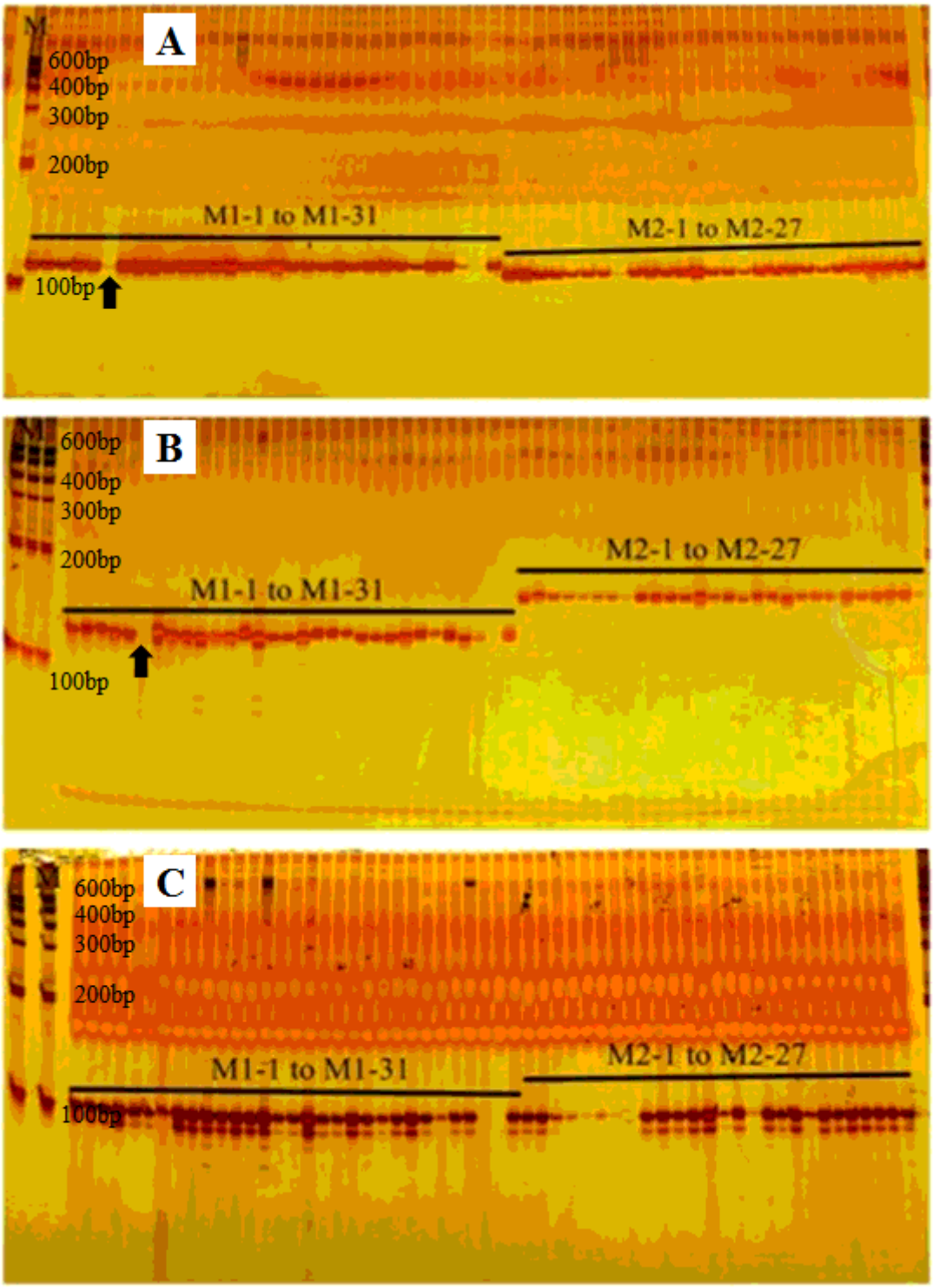

Figure 2

Gel electrophoresis profiles of SSR marker primer sets. (A) Primer XHSSR8, showing length polymorphism (B) Primer XHSSR32, showing length polymorphism (C) Primer XHSSR42, no polymorphism. DNA molecular standard with length (bp) is indicated on the left. Genotypes of 30 G. hirsutum and 27 G. barbadense are indicated in lanes M1-1 to M1-30 and M2-1 to M2-27, respectively. The lanes are pointed with black arrows indicate that no samples were loaded.

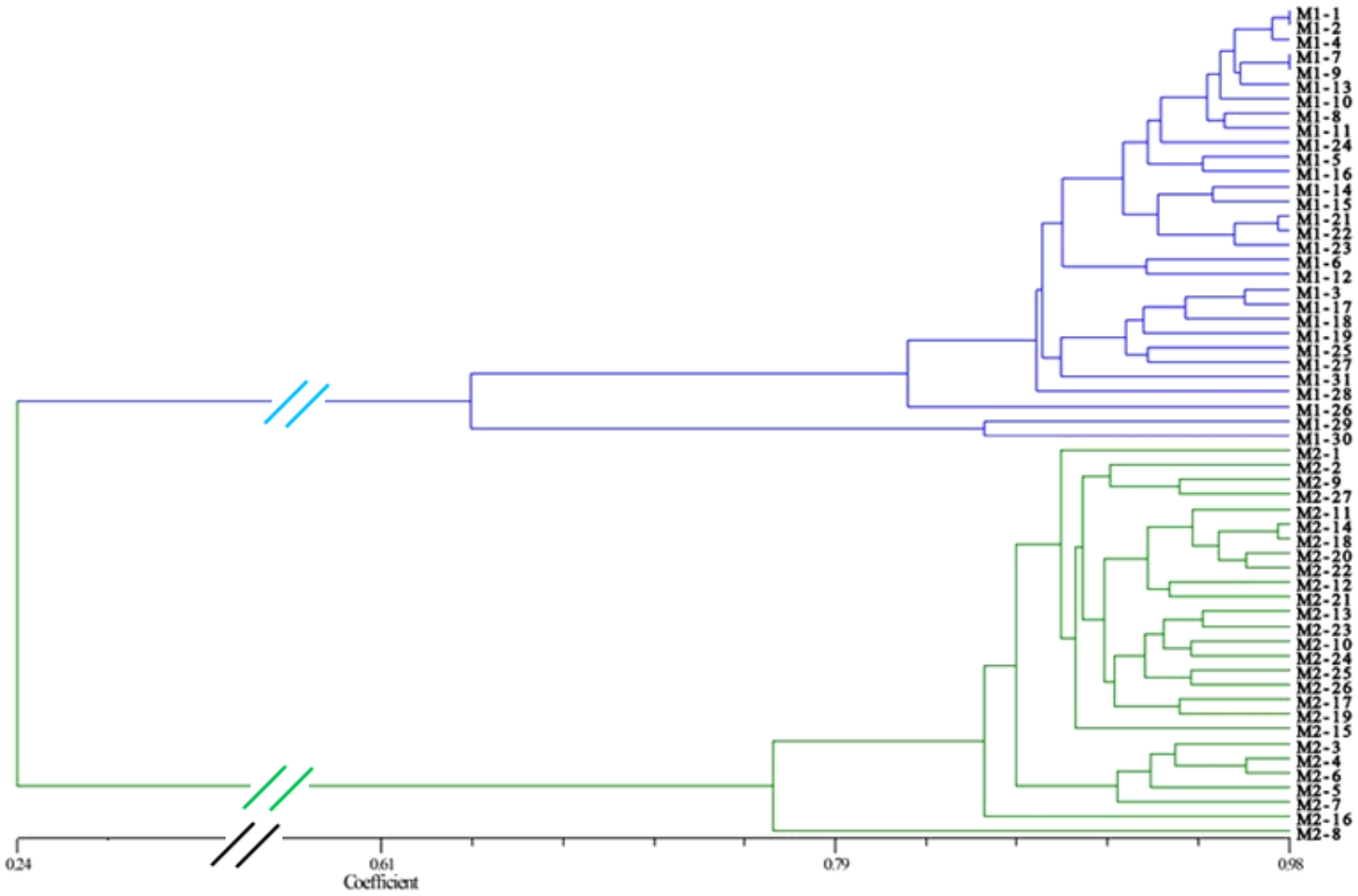

Figure 3

Phylogenetic relationship among 30 G. hirsutum and 27 G. barbadense accessions. At a similarity coefficient ≥ 0.61, G. hirsutum and G. barbadense accessions were divided into two subgroups.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.Repeatsequencesof476typesofmotifs.xlsx
- TableS2.Frequencyofdifferentmotifsinthegenome.xlsx
- TableS3.InformationofSSRmarkerprimersets.xlsx
- TableS4.Oneproductprimersetsinsilicoanalysis.xlsx
- TableS5.GOandKEGGenrichmentforXinhai21genes.xlsx
- TableS6.GenescontainingSSRmarkers.xlsx
- TableS7.300SSRprimersforgeneticdiversityanalysis.xlsx
- TableS8.Detailedinformationofthe57accessions.xlsx