Optimal use of SSR markers for varietal identification of upland cotton

Guilherme da Silva Pereira[1], Ana Luíza Ramos Cazé[1], Michelle Garcia da Silva[1],
Vanessa Cavalcante Almeida[1], Fernanda Oliveira da Cunha Magalhães[2], João Luís da Silva Filho[2],
Paulo Augusto Vianna Barroso[3] and Lúcia Vieira Hoffmann[2]

[1]Universidade Estadual da Paraíba, Rua Aprigio Veloso, nº 882, Bairro Universitário, CEP 58429-900 Campina Grande, PB, Brazil. E-mail: guispb@gmail.com, analuizacaze@gmail.com, michellegarcia.silva@gmail.com, vancalmeida@hotmail.com [2]Embrapa Algodão, Caixa Postal 179, CEP 75375-000 Santo Antônio de Goiás, GO, Brazil. E-mail: fernanda.magalhaes@embrapa.br, joao.silva-filho@embrapa.br, lucia.hoffmann@embrapa.br [3]Embrapa Agroenergia, Caixa Postal 40.315, CEP 70770-901 Brasília, DF, Brazil. E-mail: paulo.barroso@embrapa.br

Abstract – The objective of this work was to identify polymorphic simple sequence repeat (SSR) markers for varietal identification of cotton and evaluation of the genetic distance among the varieties. Initially, 92 SSR markers were genotyped in 20 Brazilian cotton cultivars. Of this total, 38 loci were polymorphic, two of which were amplified by one primer pair; the mean number of alleles per locus was 2.2. The values of polymorphic information content (PIC) and discrimination power (DP) were, on average, 0.374 and 0.433, respectively. The mean genetic distance was 0.397 (minimum of 0.092 and maximum of 0.641). A panel of 96 varieties originating from different regions of the world was assessed by 21 polymorphic loci derived from 17 selected primer pairs. Among these varieties, the mean genetic distance was 0.387 (minimum of 0 and maximum of 0.786). The dendrograms generated by the unweighted pair group method with arithmetic average (UPGMA) did not reflect the regions of Brazil (20 genotypes) or around the world (96 genotypes), where the varieties or lines were selected. Bootstrap resampling shows that genotype identification is viable with 19 loci. The polymorphic markers evaluated are useful to perform varietal identification in a large panel of cotton varieties and may be applied in studies of the species diversity.

Index terms: *Gossypium hirsutum*, cultivar discrimination, intraspecific polymorphism, microsatellite.

Uso otimizado de marcadores SSR para identificação varietal de algodoeiro herbáceo

Resumo – O objetivo deste trabalho foi identificar marcadores polimórficos de sequências simples repetidas (SSR) para identificação varietal de algodão e avaliação da distância genética entre as variedades. Inicialmente, 92 marcadores SSR foram genotipados em 20 cultivares de algodão do Brasil. Desse total, 38 locos foram polimórficos, dos quais dois deles foram amplificados por um único par de iniciadores; o número médio de alelos por loco foi 2,2. Os conteúdos de informação polimórfica (PIC) e de poder de discriminação (DP) foram, em média, 0,374 e 0,433, respectivamente. A distância genética média foi de 0,397 (mínima de 0,092 e máxima de 0,641). Um painel de 96 variedades originárias de diversas regiões do mundo foi avaliado por 21 locos polimórficos derivados a partir de 17 pares de iniciadores selecionados. Entre essas variedades, a distância genética média foi de 0,387 (mínima de 0 e máxima de 0,786). Os dendrogramas elaborados a partir de média aritmética não ponderada (UPGMA) não refletiram as regiões do Brasil (20 genótipos) ou do mundo (96 genótipos), onde as variedades ou linhagens foram selecionadas. O procedimento de reamostragem bootstrap mostra que a identificação dos genótipos é viável com 19 locos. Os marcadores polimórficos avaliados são úteis na identificação varietal de um painel amplo de variedades de algodão e podem ser aplicados em estudos de diversidade da espécie.

Termos para indexação: *Gossypium hirsutum*, discriminação de cultivares, polimorfismo intraespecífico, microssatélite.

Introduction

The genetic diversity of the genus *Gossypium* is high among wild and domesticated cotton species. Wild cotton comprises 45 diploid and 5 tetraploid species (Campbell et al., 2010), of which four have been independently domesticated in four different regions worldwide. *Gossypium hirsutum* L., one of

the tetraploid species, is classified into seven different races: one wild and six domesticated (Lacape et al., 2007). The *G. hirsutum* race *latifolium* Hutch., or upland cotton, is economically the most important one and, besides being broadly adapted, is also the main fiber crop (Campbell et al., 2010). However, there is evidence that cotton genetic diversity has been declining in breeding programs (Paterson et al., 2004), which can also lead to a reduction in yield gain through breeding, since diversity is required for selection (Campbell et al., 2010).

Diversity can be increased by using wild genotypes in breeding. However, there are constraints related to crossing barriers (Pereira et al., 2012) and mainly to traits largely different from those required for commercial cotton production. For this reason, cotton breeding has usually been performed from a narrow genetic base. In this context, intraspecific polymorphic markers can assist breeders, by easily displaying relevant genetic diversity among cotton lineages from overlooked germplasm banks, for instance.

Moreover, the identification of cotton varieties is important during breeding and registration processes, and during seed production, trade, and inspection. The identification of intraspecific polymorphic markers for varietal and cultivar discrimination is necessary, considering the narrow cotton genetic base and consequent insufficiency of morphological descriptors (Zhu et al., 2014). The use of molecular markers serves as a modern and suitable approach to varietal and cultivar identification as it is more rapid and cost effective (Korir et al., 2013). In addition, breeding efforts are facilitated by information on the genetic diversity of available germplasm resources, including those from commercial seeds.

Currently, most of the available polymorphic molecular markers for cotton varieties are interspecific, and genetic diversity, as well as molecular mapping in cotton, has frequently been done with more than one species or race (Blenda et al., 2006; Menezes et al., 2008). Despite the high diversity of SSR markers in *Gossypium* genomes (Lacape et al., 2007), the documentation of markers that are intraspecifically polymorphic in *G. hirsutum* is still incipient.

The objective of this work was to identify polymorphic simple sequence repeat (SSR) markers for varietal identification of cotton and evaluation of the genetic distance among the species varieties.

Materials and Methods

Two collections of plant material were used. The first one was composed by 20 Brazilian genotypes. Of these, 5 were commercial cultivars, identified by the BRS prefix, and 15 were inbred lines, identified by the prefix of the state in which they were selected: two from Bahia (BA), five from Goiás (GO), five from Mato Grosso (MT), and three from Paraíba identified by (CNPA). This collection represents genotypes of the breeding programs of Embrapa Algodão, developed for the main producing areas in the country. Genomic DNA was isolated from the endosperm of one seed for each genotype, placed in microtubes with sodium dodecyl sulfate extraction buffer (McDonald et al., 1994) and grinded with beadbeater equipment (BioSpec Products, Inc., Bartlesville, OK, USA).

The second plant material, also from the germplasm bank of Embrapa Algodão, was composed by 96 worldwide genotypes: 22 current Brazilian varieties, 51 obsolete varieties (17 from Brazil, 17 from the USA, 7 from Mexico, 2 from Argentina, 1 from Venezuela, 2 from India, 2 from China, 1 from Uzbekistan, and 2 from Africa), and 23 lineages of unknown-origin selected or maintained for having special traits (mainly disease resistance and superior fiber traits). Genomic DNA was extracted from young leaves through the CTAB method (Plant..., 2014). After both extraction batches, DNA was quantified for comparison with well-known amounts of lambda phage DNA (Invitrogen, Carlsbad, CA, USA) in 0.8% (w/v) agarose gels stained by Sybr Gold (Invitrogen, Carlsbad, CA, USA).

For polymorphism screening in the collection of 20 genotypes, 92 SSR markers were included: 37 BNL (Liu et al., 2000; Lacape et al., 2007), 40 CIR (Nguyen et al., 2004), 4 JESPR (Reddy et al., 2001), and 11 NAU (Han et al., 2004). These SSR loci are summarized in Table 1 according to information available at the CottonGen database (Yu et al., 2014). Polymerase chain reactions (PCR) were performed on 20 μL solutions containing PCR buffer (10 mmol L^{-1} Tris-HCl pH 8.3, 50 mmol L^{-1} KCl, and 0.1% Triton X-100), 0.2 mmol L^{-1} of each dNTP, 1.0 U of Taq DNA polimerase, 0.15 μmol L^{-1} of each primer, and 20 ng (for BNL primers) or 25 ng (for CIR, JESPR, and NAU primers) of genomic DNA. Magnesium chloride was added to each primer according to the developer's specifications (Table 1).

Table 1. Information on repeat motifs, MgCl$_2$ concentration (mmol L^{-1}), annealing temperature (AT, ºC), and chromosome location for all simple sequence repeat (SSR) markers used in the intraspecific polymorphism screening of the collection of 20 upland cotton (*Gossypium hirsutum* race *latifolium* Hutch.) genotypes[1].

SSR marker	Repeat motif	MgCl$_2$	AT	Map position	References	SSR marker	Repeat motif	MgCl$_2$	AT	Map position	References
BNL409	(GA)12	2.5	55	13	Lacape et al. (2007)	CIR085	(AC)7	2.0	55	26	Nguyen et al. (2004)
BNL786	(AG)14	2.5	55	15	Liu et al. (2000)	CIR094	(TA)6(TG)9	2.0	55	20	Nguyen et al. (2004)
BNL1030	(GT)16, (CA)13	2.5	55	23	Lacape et al. (2007)	CIR097	(GT)7(GA)7	2.0	55	14	Nguyen et al. (2004)
BNL1053	(AC)16	3.0	55	21	Liu et al. (2000)	CIR099	(GT)8	2.0	55	18	Nguyen et al. (2004)
BNL1064	(CA)15, (GT)13	3.0	55	06	Liu et al. (2000)	CIR100	(GT)7	2.0	55	16	Nguyen et al. (2004)
BNL1066	(GT)10, (GA)9	2.5	55	11	Lacape et al. (2007)	CIR104	(AC)7	2.0	55	10	Nguyen et al. (2004)
BNL1421	(AG)29, (AG)14	2.5	55	13	Lacape et al. (2007)	CIR105	(AC)9	2.0	55	15	Nguyen et al. (2004)
BNL1434	(AG)13	2.5	55	02	Liu et al. (2000)	CIR107	(AC)8	2.0	55	16	Nguyen et al. (2004)
BNL1440	(AG)15	3.0	55	11, 25	Liu et al. (2000)	CIR148	(TG)8	2.0	55	12	Nguyen et al. (2004)
BNL1495	(AG)14	2.5	55	13	Lacape et al. (2007)	CIR165	(AC)13	2.0	55	19	Nguyen et al. (2004)
BNL1551	(AG)22	2.5	55	16, 21	Lacape et al. (2007)	CIR166	(AC)8(TC)8(AT)5	2.0	55	20	Nguyen et al. (2004)
BNL1681	(AG)11	2.5	55	11, 19, 21	Lacape et al. (2007)	CIR167	(TG)8	2.0	55	26	Nguyen et al. (2004)
BNL1705	(AG)27, (CT)28	2.5	55	21	Lacape et al. (2007)	CIR169	(AC)8	2.0	55	07, 16	Nguyen et al. (2004)
BNL1707	(AG)16	2.5	55	09	Lacape et al. (2007)	CIR170	(TG)7	2.0	55	26	Nguyen et al. (2004)
BNL1721	(AG)17	2.5	55	18	Liu et al. (2000)	CIR197	(AC)6	2.0	55	N/A	Nguyen et al. (2004)
BNL2496	(GA)15	2.5	55	17	Liu et al. (2000)	CIR203	(TG)11(N)1(TG)5	2.0	55	06, 11	Nguyen et al. (2004)
BNL2499	(GA)14, (CT)15	2.5	55	24	Lacape et al. (2007)	CIR212	(TG)20	2.0	55	19	Nguyen et al. (2004)
BNL2570	(GA)13	2.5	55	20	Lacape et al. (2007)	CIR228	(2)	2.0	55	03	Nguyen et al. (2004)
BNL2590	(AG)11	3.0	55	23	Liu et al. (2000)	CIR246	(TG)6	2.0	55	14	Nguyen et al. (2004)
BNL2847	(GA)17	2.5	55	09	Lacape et al. (2007)	CIR249	(CA)8(N)3(AT)7	2.0	55	04	Nguyen et al. (2004)
BNL2961	(GA)11	2.5	55	08, 24	Lacape et al. (2007)	CIR261	(TG)7	2.0	55	17	Nguyen et al. (2004)
BNL3065	(AG)21	3.0	55	16	Liu et al. (2000)	CIR280	(GT)7	2.0	55	06	Nguyen et al. (2004)
BNL3261	(GA)18	2.5	55	12	Lacape et al. (2007)	CIR292	(3)	2.0	55	14	Nguyen et al. (2004)
BNL3279	(AG)15	2.5	55	21	Liu et al. (2000)	CIR293	(CA)6	2.0	55	12	Nguyen et al. (2004)
BNL3398	(AC)18, (AT)5	2.5	55	03	Lacape et al. (2007)	CIR316	(AC)8(AT)5	2.0	51	21	Nguyen et al. (2004)
BNL3408	(GT)2AT(GT)12	2.5	55	03, 17	Liu et al. (2000)	CIR347	(AT)6(GT)7	2.0	51	17	Nguyen et al. (2004)
BNL3441	(4)	2.5	55	03	Liu et al. (2000)	CIR353	(AT)8(T)7(GT)7	2.0	51	09	Nguyen et al. (2004)
BNL3479	(5)	3.0	55	13, 18	Liu et al. (2000)	CIR372	(GT)12	2.0	51	10	Nguyen et al. (2004)
BNL3482	(AC)12	2.5	55	20, 26	Liu et al. (2000)	CIR373	(AC)9	2.0	51	05	Nguyen et al. (2004)
BNL3502	(AC)12, (AT)2	2.5	55	14	Lacape et al. (2007)	CIR399	(CA)9(TA)8(AG)11	2.0	50	11	Nguyen et al. (2004)
BNL3563	(CA)13(TA)4	3.0	55	10	Liu et al. (2000)	CIR413	(CA)33	2.0	51	24	Nguyen et al. (2004)
BNL3835	(TG)18	-	55	04, 12	Lacape et al. (2007)	JESPR153	(CTA)18	2.5	55	13, 18	Reddy et al. (2001)
BNL3902	(GT)18	2.5	55	15	Liu et al. (2000)	JESPR157	(6)	2.5	55	24	Reddy et al. (2001)
BNL3948	(AC)11	2.5	55	09, 10, 20	Lacape et al. (2007)	JESPR228	(GA)21	2.5	55	07, 16	Reddy et al. (2001)
BNL3994	(CT)25, (GA)25	2.5	55	22, 26	Lacape et al. (2007)	JESPR292	(CTT)7	2.5	55	16	Reddy et al. (2001)
BNL4053	(CA)10	2.5	55	09, 23	Lacape et al. (2007)	NAU1169	(TA)12	2.5	55	10, 20	Han et al. (2006)
BNL4060	(AC)13	2.5	55	02	Lacape et al. (2007)	NAU1215	(AAC)6	2.5	55	13	Han et al. (2006)
CIR017	(TG)6	2.0	55	06, 24	Nguyen et al. (2004)	NAU2469	TA(6), GAT(6)	2.5	55	N/A	Han et al. (2006)
CIR020	(TG)8	2.0	55	20	Nguyen et al. (2004)	NAU2498	AT(6)	2.5	55	N/A	Han et al. (2006)
CIR043	(T)16(TG)9	2.0	55	20	Nguyen et al. (2004)	NAU2665	(GAA)6	2.5	55	08, 24	Han et al. (2006)
CIR055	(AT)8(GT)15	2.0	55	01	Nguyen et al. (2004)	NAU3408	(AAAT)4AAT(AT)5	2.5	55	N/A	Han et al. (2006)
CIR062	(AC)7	2.0	51	05	Nguyen et al. (2004)	NAU3450	(CTGT)4	2.5	55	05	Han et al. (2006)
CIR063	(CA)6	2.0	55	20	Nguyen et al. (2004)	NAU4085	(GGT)6	2.5	55	N/A	Han et al. (2006)
CIR075	(TG)7	2.0	55	N/A	Nguyen et al. (2004)	NAU858	(TTC)6	2.5	55	09	Han et al. (2006)
CIR081	(AC)7	2.0	55	12	Nguyen et al. (2004)	NAU864	(TG)10	2.5	55	09, 23	Han et al. (2006)
CIR082	(GT)8	2.0	55	10	Nguyen et al. (2004)	NAU966	(AAGCTC)4	2.5	55	09, 16	Han et al. (2006)

[1]Data were collected from the CottonGen (Yu et al., 2014) and CottonDB (Cotton…, 2011) online databases and from the variety developers. [2](TG)12(N)3(TGTA)11. [3](CA)6(GA)6(N)out(GA)7. [4](AC)18, (AT)2(AC)18(AT)4. [5](AC)15, (TC)6T(AC)15G(CA)2. [6](CTT)9(Y)10(CTT)4(Y)3(CTT)4.

The amplification reaction for the BNL primers was performed by an initial denaturation at 95°C for 12 min, followed by 30 cycles at 93°C for 1 min, 55°C for 2 min, and 72°C for 3 min, with a final extension at 72°C for 7 min. For the CIR, JESPR, and NAU primers, the initial denaturation was at 94°C for 5 min, followed by 35 cycles at 94°C for 30 s, annealing temperature recommended for each primer pair (Table 1) at 1 min, and 72°C for 1 min, with a final extension at 72°C for 8 min. PCR products were electrophoresed on 6% (w/v) polyacrylamide gels and stained with silver nitrate. SSR data were scored visually, and fragment size estimates were based on their mobility relative to a 50-bp ladder size standard (Invitrogen, Carlsbad, CA, USA).

Two measures of marker informativeness were obtained for each polymorphic SSR loci in the collection of 20 genotypes. Polymorphic information content (PIC) and discrimination power (DP) values were calculated as proposed by Botstein et al. (1980) and Tessier et al. (1999), respectively. Furthermore, individual observed heterozygozity (H_i), i.e. the percentage of heterozygous SSR loci, was calculated for each *G. hirsutum* cultivar.

For varietal identification, a number of markers was defined by performing 1,000 bootstrap resampling over an increasing number of polymorphic loci using the GenClone software, version 2.0 (Arnaud-Haond & Belkhir, 2007). Afterwards, genotyping was carried out in a panel of 96 genotypes for varietal identification. Amplification was performed according to the recommendations of the respective developers, as described before, multiplexing up to five primer pairs with similar annealing temperature and amplifying fragments of contrasting molecular weights. One primer of each pair was labeled with either 6-FAM, HEX, or NED (Applied Biosystems Inc., Foster City, CA, USA) and was combined to another in batches. Initial denaturation was at 95°C for 15 min, followed by 40 cycles at 94°C for 1 min, annealing temperature at 51 or 55°C for 90 s (depending on the primer combination), and 72°C for 90 s, with a final extension at 72°C for 8 min. The obtained PCR products were run in the ABI 3500xL sequencer and scored using the GeneMapper software (Applied Biosystems Inc., Foster City, CA, USA).

Posteriorly, the GenClone software, version 2.0 (Arnaud-Haond & Belkhir, 2007), was used to reevaluate the 1,000 bootstrap resampling over the number of markers scored in the collection of 96 genotypes. This was done in order to verify the reliability of this number of markers in discriminating the so-called multilocus genotypes (MLG).

For both collections of genotypes, genetic distances, defined as the proportion of shared alleles (Bowcock et al., 1994), were calculated over the means of 1,000 bootstrap resampling in the Microsat software, version 1.5 (Stanford University, Stanford, CA, USA). Cluster analysis was done using the unweighted pair group method with arithmetic average (UPGMA) in the Mega software, version 5.05 (Tamura et al., 2011).

Results and Discussion

Ninety-two primer pairs amplified 93 loci in the collection of 20 genotypes. From this total, 38 loci were polymorphic, totaling 40.8% (Table 2). The polymorphic loci presented a total of 84 alleles, with 2.2 alleles per locus in average, ranging from two to four alleles. The most polymorphic marker was BNL1551, which presented four different alleles. The highest PIC values (≥0.5) were those calculated for the CIR055, CIR165, and CIR249 loci, and the PIC average was 0.374. However, BNL1551, CIR249, and JESPR153b showed the highest DP values (>0.9), and the DP average was 0.433. The PIC values show the degree of marker informativeness within the *latifolium* race, which was smaller than the ones obtained using different cotton species (Liu et al., 2000; Lacape et al., 2007). However, DP seems to be more useful than PIC to select primer pairs for varietal identification, since it considers the number of evaluated individuals and shows the probability of randomly-selected genotypes being discriminated by each marker.

Exclusive alleles indicate a certain differentiation and can be used as a direct tool for varietal identification, as well as to check genetic purity (Schuster et al., 2006) and hybridization for that specific genotype (Selvakumar et al., 2010). Two loci, BNL786 and JESPR153b, revealed exclusive alleles for the CNPA 5052 inbred line, whereas two other loci, BNL1551 and CIR246, revealed exclusive alleles for the BRS Cedro commercial variety. The commercial variety BRS Seridó, the only one in the collection developed to be planted in Northeast Brazil, showed exclusive alleles revealed by BNL3994 and CIR105.

In addition, JESPR153 amplified two polymorphic loci, each one presenting sets of alleles with different sizes, both identified on previous cotton linkage maps (Ali et al., 2009), as a consequence of cotton allopolyploidization and the presence of homeologous loci from the A and D genomes (Lacape et al., 2009).

The individual observed heterozygosity of the collection of 20 genotypes was relatively high (Table 3), with a mean of 4.8%. The expected rate for advanced lines was $H_i = 0.0\%$, observed in five genotypes, and unexpectedly of $H_i = 16.2\%$ for two genotypes. The individual observed heterozigosity of 15 plants was relatively high, i.e. $H_i > 2\%$ (Table 3), which was not expected since the plants are lineages and, therefore, supposed to be derived from successive self-pollinations. The occurrence of heterozygotes among advanced lines may be explained by a limited number of self-pollination, by variety release of genotypes selected in the first steps of the breeding program (Lacape et al., 2007), or by pollen contamination on seed production fields. Heterozygotes should facilitate individual identification, although they were not supposed be found very often within lineages or cultivars.

The number of markers required for varietal discrimination was 19, estimated by resampling 38 polymorphic loci from the collection of 20 genotypes; when the number of loci tends towards 19, asymptotic behavior is observed (Figure 1A). The resampling strategy allowed testing random combinations of increasing numbers of markers. This generated minimum, average, and maximum number of

Table 2. Characterization of 38 polymorphic microsatellite loci amplified in the evaluated upland cotton (*Gossypium hirsutum* race *latifolium* Hutch.) genotypes[1].

SSR loci	Allele number	Allele size range (bp)	Allele frequencies	PIC	DP
BNL786	2	113–117	0.05, 0.95	0.100	0.105
BNL1053	2	177–179	0.39, 0.61	0.475	0.503
BNL1421	2	193–233	0.15, 0.85	0.255	0.268
BNL1495	2	113–157	0.15, 0.85	0.255	0.268
BNL1551	4	172–184	0.10, 0.05, 0.70, 0.15	0.475	0.995
BNL1705	2	172–191	0.10, 0.90	0.180	0.189
BNL1707	2	154–164	0.24, 0.76	0.361	0.382
BNL1721	2	177–189	0.25, 0.75	0.375	0.395
BNL2496	2	109–112	0.30, 0.70	0.420	0.442
BNL2499	2	250–252	0.25, 0.75	0.375	0.400
BNL2590	2	196–198	0.90, 0.10	0.180	0.189
BNL3261	2	199–203	0.25, 0.75	0.375	0.397
BNL3279	2	112–119	0.21, 0.79	0.327	0.347
BNL3408	2	130–132	0.26, 0.74	0.388	0.409
BNL3482	3	132–136	0.19, 0.69, 0.13	0.477	0.517
BNL3902	2	193–197	0.43, 0.58	0.489	0.514
BNL3948	2	97–99	0.28, 0.73	0.399	0.420
BNL3994	3	92–96	0.06, 0.78, 0.17	0.364	0.405
BNL4082	2	160–162	0.80, 0.20	0.320	0.337
CIR017	2	130–132	0.64, 0.36	0.459	0.495
CIR030	2	274–275	0.60, 0.40	0.480	0.505
CIR055	2	187–189	0.50, 0.50	0.500	0.529
CIR081	2	225–226	0.70, 0.30	0.420	0.442
CIR099	2	85–87	0.15, 0.85	0.255	0.268
CIR105	3	91–102	0.24, 0.71, 0.06	0.443	0.471
CIR148	2	148–150	0.73, 0.28	0.399	0.420
CIR165	2	219–220	0.50, 0.50	0.500	0.526
CIR169	2	190–192	0.45, 0.55	0.495	0.521
CIR170	2	156–158	0.45, 0.55	0.495	0.521
CIR212	2	141–143	0.85, 0.15	0.255	0.268
CIR228	2	198–200	0.25, 0.75	0.375	0.395
CIR246	3	148–166	0.80, 0.15, 0.05	0.335	0.353
CIR249	3	188–196	0.18, 0.18, 0.65	0.516	0.954
CIR372	2	155–156	0.84, 0.16	0.266	0.281
CIR373	2	166–170	0.55, 0.45	0.494	0.522
JESPR153a	2	98–104	0.88, 0.12	0.208	0.221
JESPR153b	3	126–146	0.21, 0.06, 0.74	0.413	0.968
JESPR292	2	175–181	0.19, 0.81	0.305	0.325

[1]PIC, polymorphic information content; and DP, discrimination power.

Table 3. Observed individual heterozygosity (H_i) of 20 upland cotton (*Gossypium hirsutum* race *latifolium* Hutch.) genotypes by 38 polymorphic microsatellite loci.

Genotypes	H_i (%)
BA 2059	0.0
BA 33	6.7
BRS Araçá	8.8
BRS Araripe	16.2
BRS Buriti	2.6
BRS Cedro	2.6
BRS Seridó	0.0
CNPA 2571	5.6
CNPA 5052	5.6
CNPA 6504	6.9
GO 1947	2.6
GO 2043	2.6
GO 4744	0.0
GO 4996	0.0
GO 8022	2.8
MT 1515	5.7
MT 1832	16.2
MT 4412	5.4
MT 5314	0.0
MT 6011	5.4

discriminated MLG for each class number of sampled loci, ensuring that the chosen set of loci allows for a good estimate of the real number of MLG in the analyzed sample. Because of the arbitrary property of resampling loci, in theory, any subset of the 38 polymorphic loci could be used for exclusive identification of the collection of 20 genotypes.

The number of loci needed for varietal discrimination was confirmed through 17 selected markers evaluated in the collection of 96 cotton genotypes (Figure 1B). Out of these 17 markers, the following 12 were selected based on the polymorphism screening performed in the present study, considering the highest PIC and DP values: BNL2499, BNL3482, CIR030, CIR055, CIR081, CIR099, CIR165, CIR170, CIR249, CIR373, JESPR153, and JESPR292. Moreover, five SSRs evaluated in previous studies were also included in

the present study due to shared specific interests: BNL3661, for resistance to the root-knot nematode, *Meloidogyne incognita* (Gutiérrez et al., 2010); CIR316, for resistance to *M. incognita* (Ulloa et al., 2010); JESPR101, related to four production or fiber traits (Zhang et al., 2013); JESPR110, related to fiber traits (Wang et al., 2011); and JESPR304, for resistance to *Fusarium oxysporum* wilt (Wang et al., 2009). Considering that four primer pairs (BNL3661, CIR165, CIR316, and JESPR101) amplified two distinct loci each in the collection of 96 genotypes, 21 loci were produced for varietal identification, i.e., two more than recommended for the collection of 20 genotypes. Bootstrap resampling over the 21 polymorphic loci from the collection of 96 genotypes indicated that these were sufficient to identify each variety (Figure 1B). For 19 loci, a minimum of 94 genotypes could be
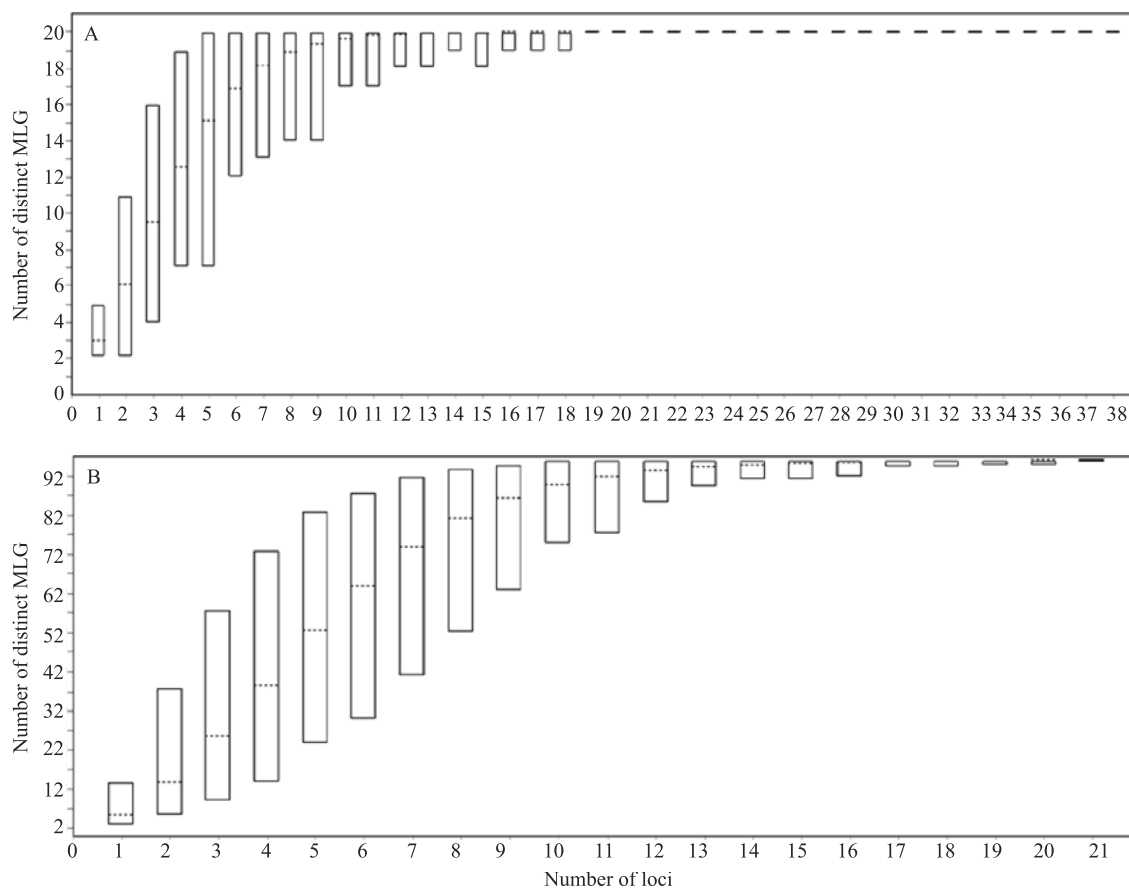

Figure 1. Minimum, average, and maximum numbers of distinct multilocus genotypes (MLG) identified by bootstrap resampling of 38 (A) and 21 (B) polymorphic microsatellite loci genotyped in the collections of 20 and 96 upland cotton (*Gossypium hirsutum* race *latifolium* Hutch.) genotypes, respectively. Nineteen markers can be successfully used for varietal identification in cotton.

discriminated properly. A maximum of 96 genotypes was achieved by a combination of at least ten loci.

For the collection of 20 genotypes, the genetic distances measured by the 38 polymorphic loci ranged from 0.092 to 0.641, with an average of 0.397 (Figure 2). The smallest genetic distance was observed between the GO 8022 and CNPA 2571 lines, and the highest between the BA33 line and the BRS Buriti commercial variety. For the collection of 96 genotypes, the average of genetic distance was 0.387, ranging from 0 to 0.786. The most divergent genotype pair was Hopiacala (USA) and Silvermine (unknown origin), whereas the less divergent ones were Roella (unknown origin) and SA 2628 (USA), and Coodetec 403 and Epamig 4 (both from Brazil). The analysis of the calculated genetic distance distributions of all genotype pairs showed a higher frequency of distances at the classes from 0.4 to 0.5 (Figure 2), representing 78 and 64% of the results for the collections of 20 and 96 genotypes, respectively.

Varieties selected in the same region were not placed in separate clusters in the grouping pattern shown by the UPGMA dendrograms (Figures 3 and 4). This may be explained by the genealogy of the lines, since some of them originated from the same crossings in the same breeding program and were then distributed to be evaluated in various regions. Another factor to be considered in this case are the similarities between selection parameters among the regions, as well as the lack of association between those selection parameters and the SSR markers. Since distances were small, all genotypes from both collections could be considered a single group; for the collection of 20 genotypes,
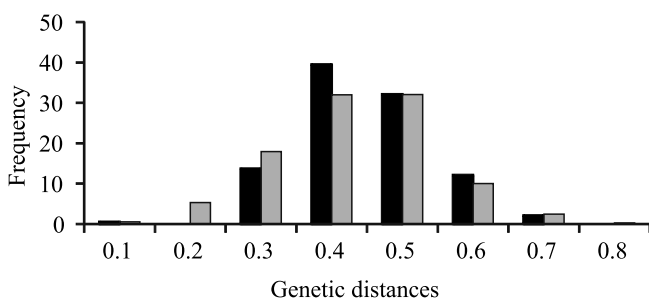
lineages could be subgrouped into two clusters based on distance 0.22 – one including 4 genotypes and another, the 16 remaining genotypes (Figure 3).

The recent discovery and advances on the analysis of polymorphism in cotton SSRs (Blenda et al., 2006; Chen & Du, 2006; Kebede et al., 2007; Lacape et al., 2007) led to the choice of high polymorphic markers. Microsatellites may be chosen instead of markers based on random PCR amplification because they are relatively easy to reproduce and their location in the genome can also be determined (Blenda et al., 2006). Furthermore, microsatellites are easy to perform and cost-effective, in comparison to SNP high-throughput technologies, because they are multiallelic (Gupta et al., 2005) and only a small number of markers are required for the analyses. Genomic SSRs are also more recommended than genic-derived SSRs, because gene regions tend to be less polymorphic (Kalia et al., 2011).

The usefulness of the markers is shown by the relatively high polymorphism level obtained (40.8%). Polymorphic primer pairs were distributed along 22 of the 26 amphidiploid cotton chromosomes; these markers were equally distributed on the chromosomes of both
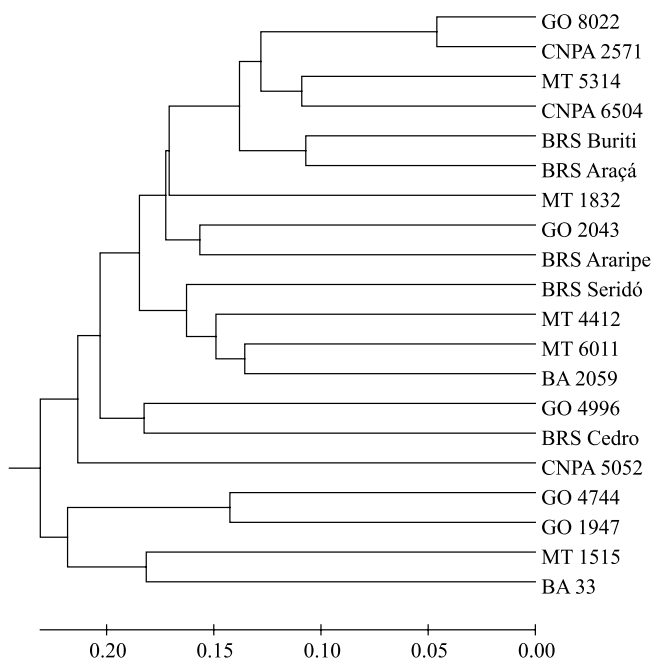
Figure 2. Relative frequency distribution of genetic distances obtained for 20 (black bars) and 96 (shaded bars) upland cotton (*Gossypium hirsutum* race *latifolium* Hutch.) genotypes by 38 and 21 polymorphic microsatellite loci, respectively.

Figure 3. Clustering assessment obtained by the unweighted pair group method with arithmetic average (UPGMA), based on the genetic distances among 20 Brazilian cotton genotypes.

the A and D genomes, with 18 and 20 polymorphic loci, respectively. An equivalent distribution in diversity among the A and D genomes was also observed by Lacape et al. (2007), and seems to contradict a previous belief that the D genome would be more diverse than the A genome (Adams & Wendel, 2004).
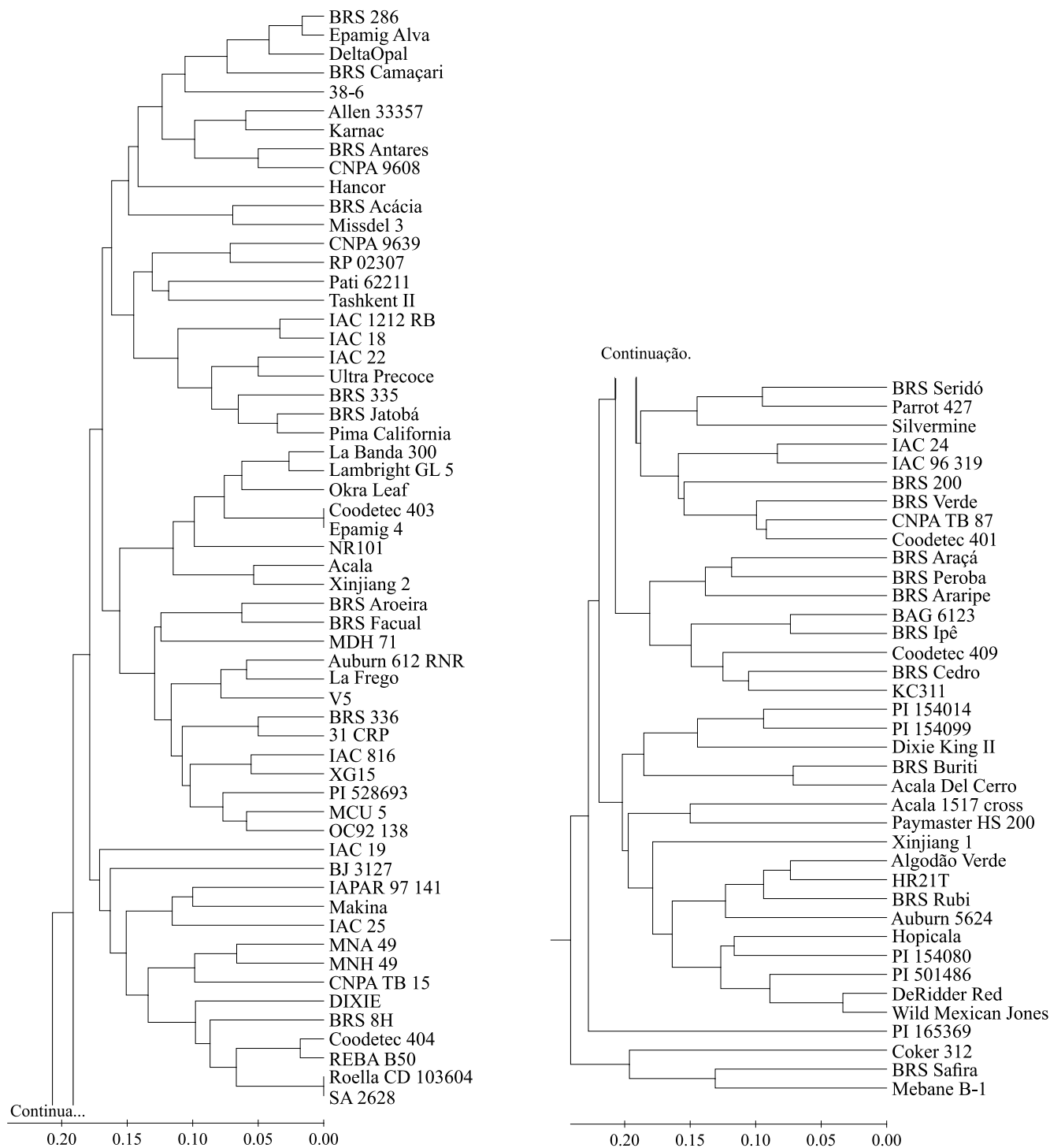

Figure 4. Clustering assessment obtained by the unweighted pair group method with arithmetic average (UPGMA), based on the genetic distances among 96 worldwide cotton genotypes.

The cultivated cotton race, *G. hirsutum* race *latifolium*, is less diverse than the other races (Bertini et al., 2006). Lacape et al. (2007) found a relatively low dissimilarity value of 0.20 within *latifolium* genotypes, but greater than when measured in other DNA marker studies, including those with RFLP (Brubaker & Wendel, 1994) or SSRs (Rungis et al., 2005; Tyagi et al., 2014). A decrease in the diversity of cultivated genotypes, when compared to the wild ones, is mostly related to the selection for crop domestication, which may be accompanied by a dispersal bottleneck (Van de Wouw et al., 2010). For cotton, an additional bottleneck occurred when just a few genotypes were transported from Mexico to the United States during the 19th century, from which the genotypes currently cultivated were derived (Paterson et al., 2004).

The intraspecific polymorphism of the markers may be significant for marker-assisted selection, since breeding programs might have to use some form of monitoring of allelic richness. The molecular basis of the cultivated cotton are reduced, but can be amplified by landraces or exotic germplasm introduction (Van de Wouw et al., 2010). The markers selected in the present study may be used to monitor genetic diversity among Brazilian or foreign genotypes and their crosses, as well as to select the most distant parental crosses that could foster genetic variance and, consequently, genetic gains, as shown for cotton by Gutiérrez et al. (2002).

The 19 SSRs chosen by permutation and resampling, combined with loci informativeness measures (Arnaud-Haond & Belkhir, 2007), can be used to discriminate upland cotton genotypes, with or without additional trait-linked markers. Genotypic discrimination should be used in germplasm banks and breeding programs to monitor the germplasm bank, to support breeders in variety protection or in monitoring the genetic variability of the genotypes used for crossings in the breeding program, and to understand reports of increased disease susceptibility in crops, as observed in a wheat variety (Simpfendorfer et al., 2013). Furthermore, the use of distinct multilocus genotypes should ensure variety protection in the world seed market and can be extended to experimental or commercial breeding when maximum genetic distances are required, as in the selection of parents for mapping or for other crossing purposes.

Conclusions

1. Genotype identification in upland cotton (*Gossypium hirsutum* race *latifolium* Hutch.) is viable with 19 SSR markers.

2. The geographical region where a commercial genotype is obtained by breeding does not influence clustering by the unweighted pair group method with arithmetic average (UPGMA) in cotton.

Acknowledgments

References

ADAMS, K.L.; WENDEL, J.F. Exploring the genomic mysteries of polyploidy in cotton. **Biological Journal of the Linnean Society**, v.82, p.573-581, 2004. DOI: 10.1111/j.1095-8312.2004.00342.x.

ALI, M.A.; KHAN, I.A.; NAWAB, N.N. Estimation of genetic divergence and linkage for fibre quality traits in upland cotton. **Journal of Agricultural Research**, v.47, p.229-236, 2009.

ARNAUD-HAOND, S.; BELKHIR, K. GENCLONE: a computer program to analyse genotypic data, test for clonality and describe spatial clonal organization. **Molecular Ecology Notes**, v.7, p.15-17, 2007. DOI: 10.1111/j.1471-8286.2006.01522.x.

BERTINI, C.H.C. de M.; SHUSTER, I.; SEDIYAMA, T.; BARROS, E.G. de; MOREIRA, M.A. Characterization and genetic diversity analysis of cotton cultivars using microsatellites. **Genetics and Molecular Biology**, v.29, p.321-329, 2006. DOI: 10.1590/S1415-47572006000200021.

BLENDA, A.; SCHEFFLER, J.; SCHEFFLER, B.; PALMER, M.; LACAPE, J.-M.; YU, J.Z.; JESUDURAI, C.; JUNG, S.; MUTHUKUMAR, S.; YELLAMBALASE, P.; FICKLIN, S.; STATON, M.; ESHELMAN, R.; ULLOA, M.; SAHA, S.; BURR, B.; LIU, S.; ZHANG, T.; FANG, D.; PEPPER, A.; KUMPATLA, S.; JACOBS, J.; TOMKINS, J.; CANTRELL, R.; MAIN, D. CDM: a cotton microsatellite database resource for *Gossypium* genomics. **BMC Genomics**, v.7, article ID 132, 2006. DOI: 10.1186/1471-2164-7-132.

BOTSTEIN, D.; WHITE, R.L.; SKOLNICK, M.; DAVIS, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. **American Journal of Human Genetics**, v.32, p.314-331, 1980.

BOWCOCK, A.M.; RUIZ-LINARES, A.; TOMFOHRDE, J.; MINCH, E.; KIDD, J.R.; CAVALLI-SFORZA, L.L. High resolution of human evolutionary trees with polymorphic microsatellites. **Nature**, v.368, p.455-457, 1994. DOI: 10.1038/368455a0.

BRUBAKER, C.L.; WENDEL, J.F. Reevaluating the origin of domesticated cotton (*Gossypium hirsutum*; Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). **American Journal of Botany**, v.81, p.1309-1326, 1994. DOI: 10.2307/2445407.

CAMPBELL, B.T.; SAHA, S.; PERCY, R.; FRELICHOWSKI, J.; JENKINS, J.N.; PARK, W.; MAYEE, C.D.; GOTMARE, V.; DESSAUW, D.; GIBAND, M.; DU, X.; JIA, Y.; CONSTABLE, G.; DILLON, S.; ABDURAKHMONOV, I.Y.; ABDUKARIMOV, A.; RIZAEVA, S.M.; ABDULLAEV, A.; BARROSO, P.A.V.; PÁDUA, J.G.; HOFFMANN, L.V.; PODOLNAYA, L. Status of the global cotton germplasm resources. **Crop Science**, v.50, p.1161-1179, 2010. DOI: 10.2135/cropsci2009.09.0551.

CHEN, G.; DU, X.-M. Genetic diversity of source germplasm of upland cotton in China as determined by SSR marker analysis. **Acta Genetica Sinica**, v.33, p.733-745, 2006. DOI: 10.1016/S0379-4172(06)60106-6.

COTTON genome database. Washington: United States Department of Agriculture, 2011. Available at: <http://www.cottondb.org/wwwroot/cdbhome.php>. Accessed on: 10 Dec. 2014.

GUPTA, P.K.; RUSTGI, S.; KULWAL, P.L. Linkage disequilibrium and association studies in higher plants: present status and future prospects. **Plant Molecular Biology**, v.57, p.461-485, 2005. DOI: 10.1007/s11103-005-0257-z.

GUTIÉRREZ, O.A.; BASU, S.; SAHA, S.; JENKINS, J.N.; SHOEMAKER, D.B.; CHEATHAM, C.L.; MCCARTY, J.C. Genetic distance among selected cotton genotypes and its relationship with F_2 performance. **Crop Science**, v.42, p.1841-1847, 2002. DOI: 10.2135/cropsci2002.1841.

GUTIÉRREZ, O.A.; JENKINS, J.N.; MCCARTY, J.C.; WUBBEN, M.J.; HAYES, R.W.; CALLAHAN, F.E. SSR markers closely associated with genes for resistance to root-knot nematode on chromosomes 11 and 14 of upland cotton. **Theoretical and Applied Genetics**, v.121, p.1323-1337, 2010. DOI: 10.1007/s00122-010-1391-9.

HAN, Z.-G.; GUO, W.-Z.; SONG, X.-L.; ZHANG, T.-Z. Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in alloteraploid cotton. **Molecular Genetic Genomics**, v.272, p.308-327, 2004. DOI: 10.1007/s00438-004-1059-8.

HAN, Z.-G.; WANG, C.; SONG, X.; GUO, W.; GOU, J.; LI, C.; CHEN, X.; ZHANG, T. Characteristics, development and mapping of Gossypium hirsutum derived EST-SSRs in allotetraploid cotton. **Theoretical and Applied Genetics**, v.112, p.430-439, 2006. DOI: 10.1007/s00122-005-0142-9.

KALIA, R.K.; RAI, M.K.; KALIA, S.; SINGH, R.; DHAWAN, A.K. Microsatellite markers: an overview of the recent progress in plants. **Euphytica**, v.177, p.309-334, 2011. DOI: 10.1007/s10681-010-0286-9.

KEBEDE, H.; BUROW, G.; DANI, R.G.; ALLEN, R.D. A-genome cotton as a source of genetic variability for upland cotton (*Gossypium hirsutum*). **Genetic Resources and Crop Evolution**, v.54, p.885-895, 2007. DOI: 10.1007/s10722-006-9157-6.

KORIR, N.K.; HAN, J.; SHANGGUAN, L.F.; WANG, C.; KAYESH, E.; ZHANG, Y.Y.; FANG, J.G. Plant variety and cultivar identification: advances and prospects. **Critical Reviews in Biotechnology**, v.33, p.111-125, 2013. DOI: 10.3109/07388551.2012.675314.

LACAPE, J.-M.; DESSAUW, D.; RAJAB, M.; NOYER, J.-L.; HAU, B. Microsatellite diversity in tetraploid *Gossypium* germplasm: assembling a highly informative genotyping set of cotton SSRs. **Molecular Breeding**, v.19, p.45-58, 2007. DOI: 10.1007/s11032-006-9042-1.

LACAPE, J.-M.; JACOBS, J.; ARIOLI, T.; DERIJCKER, R.; FORESTIER-CHIRON, N.; LLEWELLYN, D.; JEAN, J.; THOMAS, E.; VIOT, C. A new interspecific, *Gossypium hirsutum* x *G. barbadense*, RIL population: towards a unified consensus linkage map of tetraploid cotton. **Theoretical and Applied Genetics**, v.119, p.281-292, 2009. DOI: 10.1007/s00122-009-1037-y.

LIU, S.; SAHA, S.; STELLY, D.; BURR, B.; CANTRELL, R.G. Chromosomal assignment of microsatellite loci in cotton. **Journal of Heredity**, v.91, p.326-332, 2000. DOI: 10.1093/jhered/91.4.326.

MCDONALD, M.B.; ELLIOT, L.J.; SWEENEY, P.M. DNA extraction from dry seeds for RAPD analyses in varietal identification studies. **Seed Science and Technology**, v.22, p.171-176, 1994.

MENEZES, I.P.P. de; HOFFMANN, L.V.; ALVES, M.F.; MORELLO, C. de L.; BARROSO, P.A.V. Distância genética entre linhagens avançadas de germoplasma de algodão com uso de marcadores de RAPD e microssatélites. **Pesquisa Agropecuária Brasileira**, v.4, p.1339-1347, 2008. DOI: 10.1590/S0100-204X2008001000012.

NGUYEN, T.-B.; GIBAND, M.; BROTTIER, P.; RISTERUCCI, A.-M.; LACAPE, J.-M. Wide coverage of the tetraploid cotton genome using newly developed microsatellite markers. **Theoretical and Applied Genetics**, v.109, p.167-175, 2004. DOI: 10.1007/s00122-004-1612-1.

PATERSON, A.H.; BOMAN, R.K.; BROWN, S.M.; CHEE, P.W.; GANNAWAY, J.R.; GINGLE, A.R.; MAY, O.L.; SMITH, C.W. Reducing the genetic vulnerability of cotton. **Crop Science**, v.44, p.1900-1901, 2004. DOI: 10.2135/cropsci2004.1900.

PEREIRA, G.S.; SOUSA, R.L.; HOFFMANN, L.V.; SILVA, E.F.; BARROSO, P.A.V. Selective fertilization in interspecific crosses of allotetraploid species of *Gossypium*. **Botany**, v.90, p.159-166, 2012. DOI: 10.1139/b11-094.

PLANT DNA extraction protocol for DArT. Camberra: Diversity Arrays Technology, 2014. Available at: <https://www.diversityarrays.com/files/DArT_DNA_isolation.pdf>. Accessed on: 10 Dec. 2014.

REDDY, O.U.K.; PEPPER, A.E.; ABDURAKHMONOV, I.; SAHA, S.; JENKINS, J.N.; BROOKS, T.; BOLEK, Y.; EL-ZIK, K.M. New dinucleotide and trinucleotide microsatellite marker resources for cotton genome research. **Journal of Cotton Science**, v.5, p.103-113, 2001.

RUNGIS, D.; LLEWELLYN, D.; DENNIS, E.S.; LYON, B.R. Simple sequence repeat (SSR) markers reveal low levels of polymorphism between cotton (*Gossypium hirsutum* L.) cultivars. **Australian Journal of Agricultural Research**, v.56, p.301-307, 2005. DOI: 10.1071/AR04190.

SCHUSTER, I.; VIEIRA, E.S.N.; PADILHA, L. Marcadores moleculares no pós-melhoramento. In: BORÉM, A.; CAIXETA, E.T. (Ed.). **Marcadores moleculares**. Viçosa: Ed. da UFV, 2006. p.205-230.

SELVAKUMAR, P.; RAVIKESAVAN, R.; GOPIKRISHNAN, A.; THIYAGU, K.; PREETHA, S.; BOOPATHI, N.M. Genetic purity analysis of cotton (*Gossypium* spp.) hybrids using SSR markers. **Seed Science and Technology**, v.38, p.358-366, 2010. DOI: 10.15258/sst.2010.38.2.09.

SIMPFENDORFER, S.; MARTIN, A.; SUTHERLAND, M.W. Use of SSR markers to determine the genetic purity of a popular Australian wheat variety and consequences for stripe rust reactions. **Seed Science and Technology**, v.41, p.98-106, 2013. DOI: 10.15258/sst.2013.41.1.09.

TAMURA, K.; PETERSON, D.; PETERSON, N.; STECHER, G.; NEI, M.; KUMAR, S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. **Molecular Biology and Evolution**, v.28, p.2731-2739, 2011. DOI: 10.1093/molbev/msr121.

TESSIER, C.; DAVID, J.; THIS, P.; BOURSIQUOT, J.M.; CHARRIER, A. Optimization of the choice of molecular markers for varietal identification in *Vitis vinifera* L. **Theoretical and Applied Genetics**, v.98, p.171-177, 1999. DOI: 10.1007/s001220051054.

TYAGI, P.; GORE, M.A.; BOWMAN, D.T.; CAMPBELL, B.T.; UDALL, J.A.; KURAPARTHY, V. Genetic diversity and population structure in the US upland cotton (*Gossypium hirsutum* L.). **Theoretical and Applied Genetics**, v.127, p.283-295, 2014. DOI: 10.1007/s00122-013-2217-3.

ULLOA, M.; WANG, C.; ROBERTS, P.A. Gene action analysis by inheritance and quantitative trait loci mapping of resistance to root-knot nematodes in cotton. **Plant Breeding**, v.129, p.541-550, 2010. DOI: 10.1111/j.1439-0523.2009.01717.x.

VAN DE WOUW, M.; KIK, C.; VAN HINTUM, T.; VAN TREUREN, R.; VISSER, B. Genetic erosion in crops: concept, research, results and challenges. **Plant Genetic Resources**, v.8, p.1-15, 2010. DOI: 10.1017/S1479262109990062.

WANG, F.; GONG, Y.; ZHANG, C.; LIU, G.; WANG, L.; XU, Z.; ZHANG, J. Genetic effects of introgression genomic components from Sea Island cotton (*Gossypium barbadense* L.) on fiber related traits in upland cotton (*G. hirsutum* L.). **Euphytica**, v.181, p.41-53, 2011. DOI: 10.1007/s10681-011-0378-1.

WANG, P.; SU, L.; QIN, L.; HU, B.; GUO, W.; ZHANG, T. Identification and molecular mapping of a Fusarium wilt resistant gene in upland cotton. **Theoretical and Applied Genetics**, v.119, p.733-739, 2009. DOI: 10.1007/s00122-009-1084-4.

YU, J.; JUNG, S.; CHENG, C.H.; FICKLIN, S.P.; LEE, T.; ZHENG, P.; JONES, D.; PERCY, R.; MAIN, D. CottonGen: a genomics, genetics and breeding database for cotton research. **Nucleic Acids Research**, v.42, p.D1229-D1236, 2014. DOI: 10.1093/nar/gkt1064.

ZHANG, T.; QIAN, N.; ZHU, X.; CHEN, H.; WANG, S.; MEI, H.; ZHANG, Y. Variations and transmission of QTL alleles for yield and fiber qualities in upland cotton cultivars developed in China. **PLoS One**, v.8, article e57220, 2013. DOI: 10.1371/journal.pone.0057220.

ZHU, Q.-H.; SPRIGGS, A.; TAYLOR, J.M.; LLEWELLYN, D.; WILSON, I. Transcriptome and complexity-reduced, DNA-based identification of intraspecies single-nucleotide polymorphisms in the polyploid *Gossypium hirsutum* L. **G3-Genes Genomes Genetics**, v.4, p.1893-1905, 2014.