GMR

Genetic diversity of arboreal cotton populations of the Brazilian semiarid: a remnant primary gene pool for cotton cultivars

I.P.P. de Menezes[1], L.V. Hoffmann[2], T.H. de Lima[1], A.R. da Silva[1], V.S. Lucena[3] and P.A.V. Barroso[4]

[1]Departamento de Biologia, Instituto Federal Goiano, Urutaí, GO, Brasil
[2]Embrapa Algodão, Santo Antônio de Goiás, GO, Brasil
[3]Universidade Federal Rural de Pernambuco, Recife, PE, Brasil
[4]Embrapa Monitoramento por Satélite

Corresponding author: L.V. Hoffmann
E-mail: lucia.hoffmann@embrapa.br

Genet. Mol. Res. 16 (3): gmr16039659
Received March 7, 2017
Accepted August 30, 2017
Published September 27, 2017
DOI http://dx.doi.org/10.4238/gmr16039659

ABSTRACT. Mocó cotton belong to the same species as the cultivated species, *Gossypium hirsutum*, and cultivated forms were mainly landraces but also developed as cultivars, bearing good fiber quality and drought tolerant when cropped as a perennial species. The northeast Brazil crop system based on this cotton type is finished, with a few small area planted in the three main States, where it was previously cultivated (Ceará, Paraíba, and Rio Grande do Norte), but in others, maintenance is accomplished by single dooryard plants. Plants were found in all visited Northeast Brazil municipalities, sometimes in the North of the country, and were collected for *ex situ* preservation and evaluation. Most of seeds had no fuzz (62.2%) and 94.6% of the genotypes presented spot in flowers. Seventy-one alleles were revealed in 12 loci. The genetic structure of the population evaluated by microsatellite markers shows two main groups, one comprising the Seridó region where landraces were originated and other comprising

the state of Ceará, where a specific breeding program was developed. Genotypes collected in North Brazil States as well as those collected in Bahia, Alagoas, and Sergipe grouped with those collected in Ceará. The Mantel correlogram indicates a significant ($P < 0.05$) correlation between genetic and geographical distances up to 77 km. The *ex situ* maintenance and agronomical evaluation are the main concerns for mocó, as the use of the agricultural interesting traits, possibly introgressed to other genotypes, is predicted. The *in situ* preservation is still of interest since there is more diversity there than in the collected plants and some should be continued due to use as medicinal plant.

Key words: Mocó cotton; Microsatellite markers; Genetic diversity; Germplasm conservation; Genetic structure

INTRODUCTION

Cotton landraces may be an important genetic resource, especially the tetraploid ones, which are those of the primary gene pool of the cultivated cotton species, *Gossypium hirsutum* and *G. barbadense*.

Mocó is a landrace from Semiarid Northest Brazil cultivated as a perennial crop, usually from four to six years in the field. The economic importance was related to drought tolerance and the possibility to harvest the plants for four to six years, in average, as well as use plant leaves to feed cattle. It has been cultivated in an area of around 2.5 million hectares, composed 90% of Brazil fiber exports, but the reach of boll weevil associated to economic weakness lead caused from 1974 to 1993 a huge reduction of the cultivated area to around 16% of the previous one (Moreira et al., 1997). Since 2013, there was no more cultivated area detectable in official Brazilian statistics (IBGE, 2015). No other crop has shown sufficiently drought tolerance to substitute it in semiarid conditions of Northeast region of the country.

Mocó cotton has been classified as *Gossypium hirsutum* var. *marie galante* (Hutchinson, 1951), although there are differences between mocó cotton and the plants from Antilles (Boulanger and Pinheiro, 1972; Stephens, 1973). Commercial varieties usually have white lint, and spontaneous or backyard plants can be found with various brown shades. One colored cotton variety, BRS 200, was developed selecting natural hybrids of upland cotton and mocós with brown lint available at Embrapa germplasm bank. Seeds have no fuzz; long fibers are frequent (Giband et al., 2010).

Hybridization and introgressions with *G. barbadense*, *G. mustelinum*, and *G. hirsutum* var. *latifolium*, present in the region have occurred as evidenced by morphological (Freire et al., 1990; Moreira et al., 1995) and microsatellite (Lacape et al., 2007; Menezes et al., 2010) markers. Selection by local farmers also contributes significantly to present variability (Freire et al., 1990). Nowadays, plants are maintained in backyards and are the predominant cotton variety in Northeast Brazil (https://www.cnpa.embrapa.br/albrana).

An all-inclusive analysis for mocó cotton maintenance ought to include plants collected through in the complete region, where it has been cultivated and maintained, therefore composed by Northeast and North Brazil plants.

In a previous genetic study, 184 plants from Maranhão, Piauí, Ceará, Rio Grande do Norte, and Paraíba states have been analyzed (Menezes et al., 2010). The same molecular

markers were now used to genotype 102 newly collected plants from Pernambuco, Alagoas, Sergipe, Bahia, Amapá, and Roraima, so the results were combined. The *in situ* maintenance is described for the 187 previous collected in five of those states and 144 newly collected plants, in the states explored in the present study.

The data provide means to evaluate this cotton race as a genetic resource and perspectives for *in situ* preservation, showing that only twelve high polymorphic SSR can be enough for a consistently analyze the population structure.

MATERIAL AND METHODS

Survey of the *in situ* maintenance

The expeditions to collect mocó cotton plants were from 2004 to 2008 (CGEN - Genetic Heritage Management Council under the Ministry of Environment authorization No. 13973). The geographical position of each plant was recorded, as well as morphological traits and characteristics of the property where it had been maintained (urban or rural area, small or medium farm). The plant owner, when available for consultation, was interviewed about the plant use, care, and seed origin. A web facility for the data (www.cnpa.embrapa.br/albrana) is available in English and Portuguese.

DNA extraction and microsatellite markers

Leafs of seven plants were collected in the State of Amapá, 21 in Roraima, 52 in Pernambuco, 17 in Bahia, five in Alagoas, and five in Sergipe and transported in tubes with TE (TrisHCl and EDTA 0.5 M, pH 8.0). DNA was extracted from around 100 mg leaf tissue according to the procedure described by Doyle and Doyle (1990) with small modifications.

The twelve SSR loci have been previously used for mocó cotton diversity studies (Menezes et al., 2010) and were selected as the most polymorphic in mocó of 141 loci used: BNL 840, BNL 1421, BNL 1434, BNL 1551, BNL 2496, BNL 3103 (Liu et al., 2000), CIR 148, CIR 203, CIR 212, CIR 246, CIR 249, and CIR 311 (Nguyen et al., 2004). Some of the 184 genotypes previous analyzed from Maranhão, Piauí, Ceará, Rio Grande do Norte, and Paraíba (Menezes et al., 2010) were used as control of the allele sizes, assuring compatibility of the data for the analysis.

Data analysis

The allelic frequencies of polymorphic loci were calculated considering either the total number of collected plants or each state separately, as well as the number of alleles per locus, expected ($H_{E,}$ equivalent to genetic diversity) and observed (H_O) heterozygosities, and polymorphism information content (PIC) (Botstein et al., 1980) using Power Marker (Liu and Muse, 2005). The probability of identity (*PI*) was stipulated using Identity 1.0 (Wagner and Sefc, 1999). The number of private alleles, which are those found only in genotypes belonging to one of the analyzed states, was determined using GDA (Lewis and Zaykin, 2000). Wright's *F* statistics calculated at range interval of 95% with 10,000 bootstraping has been obtained using FSTAT (Goudet, 2001), including the fixation index (F_{IS}) and genetic differentiation ($F_{ST)}$. Neighbor joining tree at state level was constructed with base in genetic distance of Nei (1978) using Mega 5.05 (Tamura et al., 2007).

InStruct program, which produces reliable genetic structure with Bayesian analysis for often allogamous species (Gao et al., 2007), was run using the admixture model to values of *K* (number of groups) ranging from 1 to 10, with 10 independent replicates to each *K*. Each run was performed with a burn-in length of 50,000 iterations, followed by MCMC (Markov Chain Monte Carlo) algorithm with a simulation length of 100,000 iterations, and the maximum likelihood *K* was determined (Evanno et al., 2005). The optimal Q-matrix values for individual runs for each selected *K* was carried out using Clummp 1.1 program (Jakobsson and Rosenberg, 2007) and its output file served to construct graphic representations of Q-values for each genotype using Distruct program (Rosenberg, 2004). The individuals with membership coefficient (q_i) greater than 0.50 were assigned to the same group, and the genetic differentiation was tested by F_{ST} statistics. Neighbor-joining tree at individual level was construed based on the genetic distance (*GD*) obtained by proportion of shared alleles (*ps*, so $GD = 1 - ps$) (Bowcock et al., 1994) using Microsat (Minch et al., 1997).

The spatial relationship of sampling locations was evaluated by the multivariate Mantel test by comparing the distance matrices describing genetic and geographical relationships among the sampling locations. We computed the Mantel correlogram between the corresponding elements of the two distance matrices. The significance of correlation values at each lag of distance was tested by randomly permuting (1000 permutations) rows and columns of one matrix, while keeping the other constant, obtaining the sampling null distribution of Mantel correlation.

The individual-centred approach proposed by Manel et al. (2007), in which there is no need to pre assume the population size based on states boundaries, lead to further investigate the population concept. The geographical diversity distribution was studied using unbiased estimates of gene diversity (Nei, 1978) obtained within a moving window of 200-km radius (Rodriguez et al., 2015) over the sampling area. This was achieved using the R package biotools (Silva et al., 2017). The mean size of each neighborhood was 22.3 individuals. Around 53% of the neighbourhoods included at least 10 individuals. Afterwards, a heat map was designed using different R packages.

RESULTS

In situ maintenance

Mocó plants were localized in all the municipalities visited in the Northeastern states and some in North Brazil. The Caatinga Biome encompasses 88% of the collected samples. The others were collected in North and South Regions, in the states of Roraima, Amapá, Pará, and Paraná (Figure 1). A total of 331 accesses were found in 13 of the Brazilian states visited.

The cotton plants were predominantly (65.3%) in rural properties, a lesser proportion were in dooryards or backyards of urban residences (19%) and others at road sides (15.7%), possibly developed from seeds spilled out from trucks or remnant from plantations (Table 1). Part of the plants in rural properties are not as crops, but single plants to be used as a medicinal plant, mainly as anti-inflammatory, as a cotton wool or as an ornamental plant, and so classified as dooryard or backyards plants. Taking urban and rural areas together, 68.6% of the maintained mocó cotton were dooryards or backyards plants, therefore the main way of maintenance. The states of Ceará, Paraíba, and Rio Grande do Norte were exceptions, where mocó cotton was cultivated as commercial crop using mainly local (Paraíba) or bred varieties

(Ceará). A number of abandoned cotton fields were localized in Paraíba, where plants were able to reproduced, so the populations were considered feral.

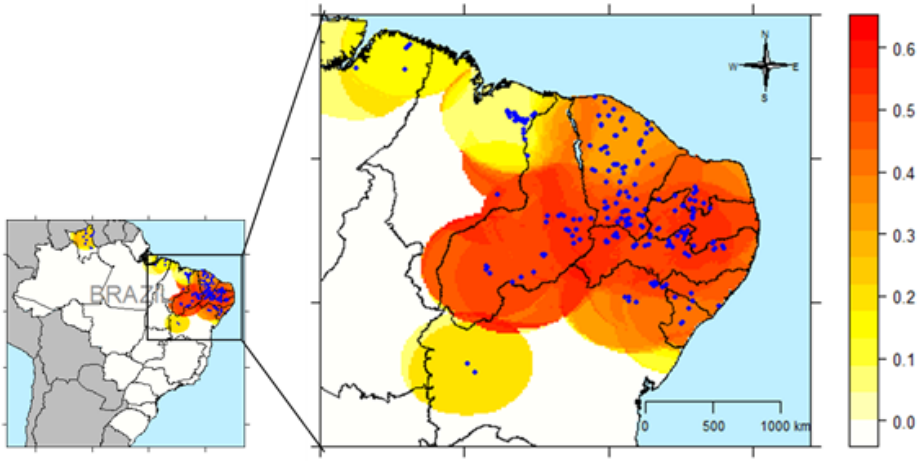

Figure 1. Gene diversity heat map of mocó cotton in Northern-Northeastern Brazil. The map was drawn based on the individual-centered approach considering a 200-km radius. Color key: from low (yellow) to high (red) diversity.

Table 1. Mocó cotton mantainance in 13 Brazilian states: Amapá (AP), Roraima (RR), Pará (PA), Maranhão (MA), Ceará (CE), Piauí (PI), Rio Grande do Norte (RN), Paraíba (PB), Pernambuco (PE), Bahia (BA), Alagoas (AL), Sergipe (SE), and Paraná (PR).

Kind of population	Number of collected plants according to the state													Total
	AP	RR	PA	MA	CE	PI	RN	PB	PE	BA	AL	SE	PR	
Dooryards or backyards	2	30	7	26	31	39	2	2	35	34	4	4	1	217
Spontaneous	-	1	-	1	21	-	3	8	21	2	2	1	-	60
Feral	-	-	-	-	-	-	10	-	-	1	-	-	-	11
Crop-bred variety	-	-	-	-	32	-	-	-	-	-	-	-	-	32
Crop-local variety	-	-	-	-	-	-	-	10	1	-	-	-	-	11
Total	2	31	7	27	84	39	15	20	57	37	6	5	1	331

Most (58.3%) of farmers, who were keeping the plants, declared that the seed was of unknown origin and 37.2% that it had been donated by familiars, friends of neighbors. Local varieties were present only in Paraíba and Pernambuco (Table 1). Several uses of the plants were reported, as medicinal, swab, lamp wick, ornamental, hand spinning, and commercial, comprising 49.8% of the sites, and slightly more than half of farmers declared that the plants were of non-use. People using the plants were usually the elderly, seemingly that young people in the region are not assuming responsibility on maintaining the traditional habits. Therefore, *in situ* conservation is expected to be discontinued.

All mocó plants presented green leaves and separated seeds characteristic of the botanical variety, while purple leaves and kidney seeds in this region would characterize *G. barbadense* plants. The absence of fuzz is also typical of mocó cotton, but only 62.2% of the plants presented seeds without fuzz. Only in the states of Paraíba and Pernambuco none of the seeds had fuzz. The presence of fuzz was predominant in the states of the North region (Amapá, Roraima, Pará), and the state in which there is a transition from North to Northeast region was Maranhão. There is possibly a sign of introgression from *G. barbadense* plants,

which is much more frequent than mocó in the North. Introgression of fuzz from herbaceous cotton is also possible.

The presence of purple spots in the petals also typifies mocó, and only 5.8% of the plants presented petals with no spots, considering only the 72.5% of plants that had flowers in the occasion of collection.

Genetic characterization

Mocó cotton exhibited 71 alleles, with a medium number of 5.92 per locus, varying from four to nine. Unequal distribution was observed for 8 of the 12 loci, where one of the alleles was present in more than 50% of the plants and may reveal fixation tendency. The medium frequency of the most common allele was 60% considering the twelve loci, and varied from 30 to 80%. The loci were high informative and medium PIC value was 0.505, varying from 0.209 for the locus revealed by the marker CIR 246 to 0.792 for the one revealed by BNL 2496. The medium genetic diversity was high, H_E = 0.535, while heterozigozity was low, H_O = 0.152. Hence, the fixation index f = 0.717 was high and significant (Table 2). The probability of identity coefficient varied from 0.089 (BNL 2496) to 0.635 (CIR 246) for the loci with the greatest and lowest discrimination power, respectively. The probability of two genotypes had the exactly same alleles profile was almost null, as indicated by the multilocus PI estimated for the twelve loci, equal to 4.6 x 10^{-7}.

Table 2. Descriptors of the genetic variability in the overall collection and according to the state of collection of mocó plants.

Overall collection	Size	Fr_{max}	*A*	PIC	H_E	H_O	F_{IS}	*PI*
Locus								
CIR148	141-151	0.753	5	0.388	0.411	0.059	0.858	0.417
CIR203	242-267	0.422	8	0.741	0.762	0.097	0.873	0.124
CIR212	130-144	0.868	3	0.221	0.236	0.072	0.697	0.628
CIR246	150-172	0.882	4	0.209	0.216	0.090	0.586	0.635
CIR249	193-199	0.768	4	0.363	0.388	0.156	0.600	0.451
CIR311	169-187	0.852	4	0.246	0.262	0.075	0.716	0.594
BNL840	149-169	0.531	5	0.570	0.622	0.007	0.988	0.305
BNL1421	195-247	0.584	8	0.550	0.594	0.370	0.380	0.299
BNL1434	236-278	0.502	7	0.568	0.628	0.228	0.640	0.324
BNL1551	157-192	0.425	7	0.653	0.700	0.153	0.783	0.237
BNL2496	108-166	0.365	9	0.792	0.808	0.363	0.553	0.089
BNL3103	187-209	0.304	7	0.761	0.788	0.155	0.805	0.134
Mean	-	0.605	5.917	0.505	0.535	0.152	0.717	4.6 x 10^{-7}

State	Sample	A (Ap)	Private alleles	H_E	H_O	F_{IS}
Amapá	7	23 (2.83)	0	0.165	0.060	0.722
Roraima	21	24 (2.71)	0	0.212	0.028	0.880
Maranhão	31	26 (2.27)	0	0.116	0.008	0.933
Piauí	47	53 (4.42)	2	0.556	0.161	0.726
Ceará	72	47 (4.18)	0	0.336	0.226	0.342
Paraíba	17	40 (3.33)	0	0.442	0.185	0.594
Rio G. do Norte	20	39 (3.26)	1	0.332	0.181	0.496
Pernambuco	52	53 (4.42)	6	0.469	0.189	0.610
Bahia	17	41 (3.42)	0	0.461	0.131	0.747
Alagoas	5	28 (2.78)	1	0.303	0.083	0.816
Sergipe	5	23 (2.38)	0	0.222	0.138	0.564
Total	294	71	10		-	-
Mean	-	-	-	0.327	0.126	0.675

A, number of alleles for locus; Ap, avarege number of polymorphic alleles for locus in each state; Fr_m, Major Allele Frequency; H_E, expected heterozygozity or genetic diversity; H_O, observed heterozygozity; F_{IS}, fixation index and *PI*, identity probability.

The topography of the distribution of mocó cotton using individual centred approach (Figure 1) shows that diversity is greater in the centre of Northeastern Brazilian semiarid region, not in the coastal region.

Ten of the alleles were exclusively present in one of the States (Table 2): 6 in Pernambuco, 2 from Piauí, 1 from Grande do Norte, and 1 from Alagoas. Only 2 of this private alleles presented high frequencies, and both from Pernambuco, BNL 840_157 pb (0.32) and CIR 203_264 pb (0.57), while for all others frequencies were smaller than 5%.

The number of alleles per locus per state was 3.27 in average. Pernambuco and Piauí were the states with the greater number of alleles, 53 (Table 2). Sergipe and Amapá had the lowest allele number, 23. The medium genetic diversity was high and expressive (H_E = 0.353), with the highest estimated values for Piauí (0.582) and Pernambuco (0.483). The medium observed heterozigosity was low, H_O = 0.128, with the greatest value for Ceará (0.228) and lowest for Maranhão (0.008). Endogamy was high (f = 0.664), Ceará presented the lowest fixation index value (0.333) and Maranhão the highest (0.933).

The medium genetic distances between states was high, 0.340. The most divergent were Maranhão and Sergipe (GD = 0.845), and the most similar those from Paraíba and Rio Grande do Norte (DG = 0.016), followed by Ceará and Bahia (0.053). The elevated genetic diversity can be noticed also by the medium F_{ST} value significantly different from zero (F_{ST} = 0.306, 95%CI = 0.194-0.404). The distinction between states was maintained despite the paired comparison shown in Table 3, with the exception of neighbors as Amapá/Roraima, Paraíba/Rio Grande do Norte, and Bahia/Alagoas/Sergipe, for which F_{ST} values were not significant. The distinctiveness was not marked between the neighbors Piauí/Paraíba, Bahia/Pernambuco and between Bahia and Ceará. Despite the genetic differentiation obtained organizing the genotypes organized according to the states of collection, the greatest part of the genetic variability was within states (H_S = 0.365), representing 68% of the total genetic diversity (H_T = 0.539).

Table 3. Pairwise genetic differentiation (F_{ST}) and unbiased Nei's genetic distance (GD) between the landraces by states and Structure's groups. F_{ST} in lower and GD in upper diagonal. AP - Amapá, RR-Roraima, MA - Maranhão, PI - Piauí, CE - Ceará, PB - Paraíba, RN - Rio Grande do Norte; PE - Pernambuco, BA - Bahia, AL - Alagoas e SE - Sergipe.

State	GD – Genetic distance										
	AP	RR	MA	PI	CE	PB	RN	PE	BA	AL	SE
AP	0	0.115	0.605	0.406	0.235	0.323	0.423	0.264	0.227	0.339	0.386
RR	0.248^{ns}	0	0.447	0.481	0.172	0.379	0.465	0.172	0.178	0.130	0.266
MA	0.730*	0.640*	0	0.178	0.602	0.234	0.210	0.526	0.600	0.502	0.845
PI	0.260*	0.329*	0.251*	0	0.484	0.139	0.170	0.407	0.445	0.568	0.672
CE	0.306*	0.262*	0.562*	0.327*	0	0.313	0.346	0.152	0.053	0.165	0.205
PB	0.306*	0.388*	0.408*	0.094*	0.310*	0	0.016	0.350	0.329	0.407	0.574
RN	0.417*	0.469*	0.412*	0.140*	0.353*	0.013^{ns}	0	0.438	0.392	0.433	0.667
PE	0.235*	0.203*	0.455*	0.224*	0.174*	0.241*	0.310*	0	0.138	0.247	0.314
BA	0.261*	0.241*	0.595*	0.244*	0.066*	0.235*	0.316*	0.100*	0	0.156	0.211
AL	0.370*	0.229*	0.648*	0.248*	0.210*	0.270*	0.353*	0.172*	0.099^{ns}	0	0.198
SE	0.468*	0.396*	0.755*	0.301*	0.257*	0.367*	0.468*	0.231*	0.193^{ns}	0.186^{ns}	0
	F_{ST} - Genetic differentiation										

^{ns}Not significant; *significantly different (95%CI: 10,000 bootstraps).

The neighbor-joining tree consists of two main groups (Figure 2), where neighbor states tended to group together. The individual genotype grouping performed by the Bayesian analysis repeated this pattern (Figures 3 and 4).

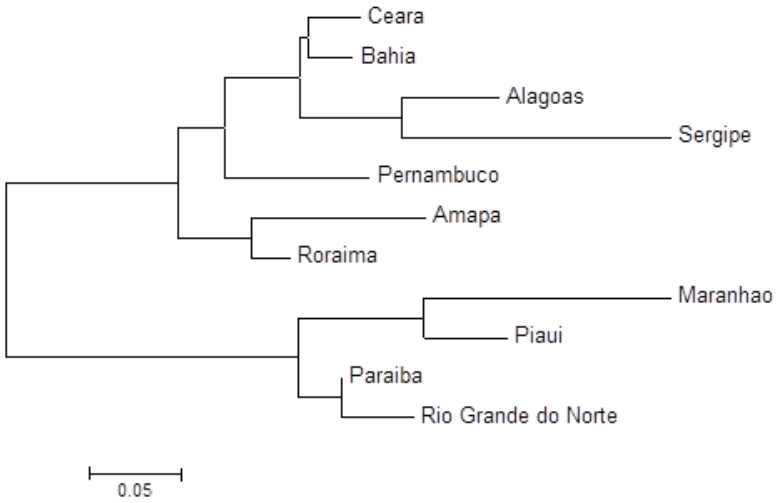

Figure 2. Dendrogram of the accession groups of mocó cotton collected by state. The clustering was based on the Nei (1978)'s genetic distance using the neighbor-joining method.

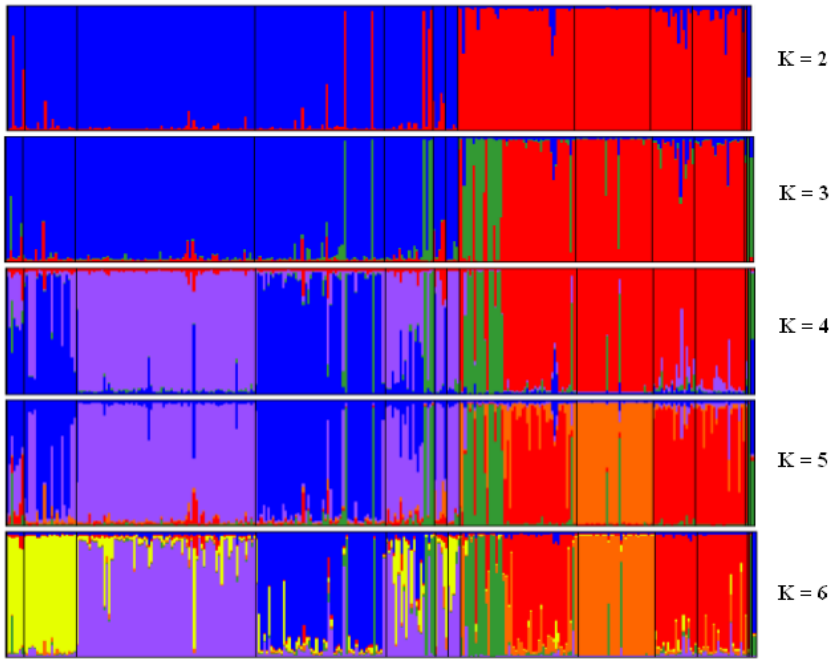

Figure 3. Inference of germoplasm structure of 294 cotton plants of landraces for ΔK raging of two to six using Instruct program. Each genetic group is represented by different colors, in which states and controls are separate by black vertical lines from left to right: Amapá, Roraima, Ceará, Pernambuco, Bahia, Alagoas, Sergipe, Piauí, Maranhão, Paraíba, Rio Grande do Norte, *Gossypium hirsutum*, *G. barbadense*, and *G. mustelinum*. Each column represents one of the collected cotton plants, and each plant has colors representing the genotype contribution from different K groups.

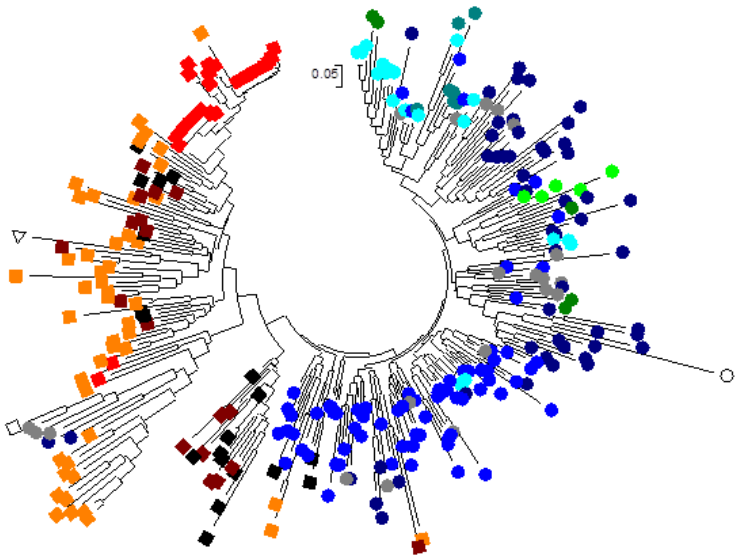

Figure 4. Neighbor-joining tree of 294 cotton carried out by Instruct. The colored circles and squares correspond to Instruct groups one and two. Interspecific controls *Gossypium hirsutum*, *G. barbadense*, *G. mustelinum* are indicated, respectively, by open triangle, open square, and circle.

The Mantel correlogram indicates a significant ($P < 0.05$) correlation between genetic and geographical distances up to 77 km, approximately, with random fluctuations after that (Figure 5), meaning that closer populations in space tend to be genetically more similar than the expected by chance.

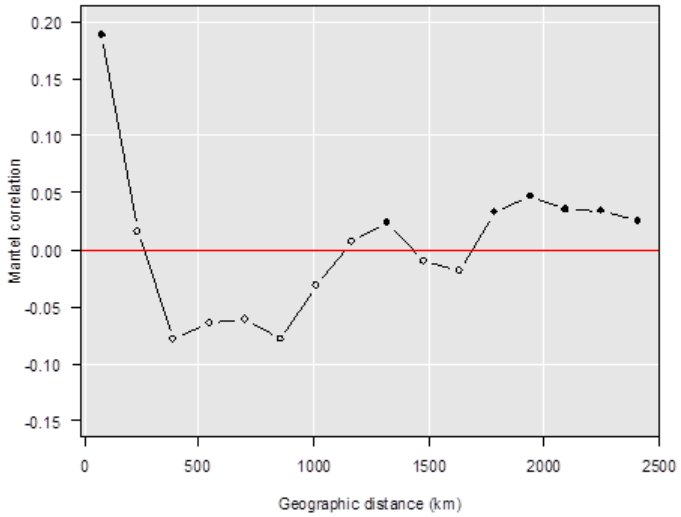

Figure 5. Mantel correlogram summarizing the spatial genetic variation among accessions of mocó cotton. Full points correspond to significant ($P < 0.05$) correlation after 1000 permutations.

The uppermost population structure of the 294 accesses was the stratification into two groups as shown by the plateau at ΔK = 2 (Evanno et al., 2005). Genotypes were allocated into group 1 with the average ancestry proportion of 96% and in group 2 it was 97%. The group 1 is composed by genotypes from Paraíba, Rio Grande do Norte, Maranhão, Piauí, three from Bahia, two from Pernambuco, and only one from Amapá; and the group 2 by the other accessions from North and Northeast Regions (Figure 3). Among the other species control plants, *G. hirsutum* and *G. barbadense* are in group 1, while *G. mustelinum* is in group 2, with the ancestry proportion, respectively, of 0.80, 0.98, and 0.57. The genetic pools presented elevated genetic differentiation (F_{ST} = 0.244, P < 0.05), slightly greater in group 1 (A = 62; H_E = 0.505) than in group 2 (A = 63; H_E = 0.435).

Furthermore, population sub-structures were indicated by secondary peaks at ΔK = 4 and ΔK = 6 with magnitude of five and ten times of that obtained at ΔK = 2, respectively. The average value of the LnP(D) of the K simulations represented progressive values from K = 2 to K = 6, then assumed a plateau distribution. The differences between the medium LnP(D) values for the successive Ks were significant as indicated by the Wilconxon test.

When K = 3, there is a subgroup composed by the *G. barbadense* plants and those collected in Piauí as well as of some of the collected in Pernambuco, Bahia, and Maranhão. It is shown in Figure 3 in green color and is stable from K = 3 to greater K values.

Plants collected in the State of Ceará grouped with most of plants collected in Bahia, Alagoas, and Sergipe when K = 4, and shown in purple in Figure 3. Although there is an admixture with Amapá and Roraima and in a lesser extent with Paraíba and Rio Grande Norte, the groups separate when K = 6. Considering the progressive K values from three to six, the admixture among groups grows, almost exclusively within the two groups primarily defined when ΔK=2 (Figure 3).

The neighbor-joining analysis divides the mocó accessions as the uppermost division shown when in the Bayesian groups for ΔK=2 (Figures 3 and 4).

DISCUSSION

Mocó *ex situ* preservation is necessary since plant *in situ* preservation is diminishing. Strategy must consider: i) the importance as genetic resource for disease resistance, fiber quality and color, drought resistance; ii) the possibility of any economical measure effective to control boll weevil making possible to be cultivated in the same region, and iii) the genetic diversity and population structure. This study includes most of the Brazilian mocó collection maintained at Embrapa, and may help to choose the seeds to be planted and evaluated.

The genetic diversity of the collection is greater than the shown by breeding lineages (Moiana et al., 2015; Pereira et al., 2015), *G. barbadense* (Almeida et al., 2009) or *G. mustelinum* (Barroso et al., 2010; Menezes et al., 2014). The PIC values were also more elevated than those found for cultivars (Yu et al., 2012; Hinze et al., 2015; Moiana et al., 2015) showing discriminative potential conferred by 12 loci. The large genetic basis of mocó cotton is due to multiple contributions during domestication and selection (Moreira et al., 1995), including introgressions (Freire and Moreira, 1991; Lacape et al., 2007) and selection by local farmers (Freire et al., 1990; Menezes et al., 2010). Consequently, multiple uses can be expected for the cotton type as a genetic resource.

Considering cotton mixed matting system, the homozygosis of all plants reveals reduced recombination taxes, and genetic diversity in autogamous species (Jarvis and Hodgkin, 1999) is threated by genetic drift (Hedrick, 2001). Gene flow is also low. The high

endogamy values (*f*) observed both in the overall collection as within states and multilocus F_{ST} value show a tendency to form families progressively homozygous. Similar structures of high diversity threatened by low genetic recombination are of the others allotetraploid cottons in Brazil: *G. barbadense* (Almeida et al., 2009), and *G. mustelinum* (Barroso et al., 2010; Alves et al., 2013; Menezes et al., 2014). The limited gene flow is explained by geographical distribution of the plants, since just one to a few individuals may be find together. Pollination agents cannot reach the distances.

The average fixation index and F_{ST} show that genetic distribution patterns are not by chance, reflecting the homozygosis and expressive differentiation of States. The H_S/H_T ratio shows that most of the variability is within states. Some historical and geographical aspects favor human migration and seed exchange within states, with some main cities in each state tending to influence vicinity municipalities. Most of plants within states keep a marked variability because most of them are true landraces (Freire et al., 1990), not derived from breeding programs, consistently with the declared origin of the seed (www.cnpa.embrapa.br/albrana), when available, as by the observed morphology in *ex situ* evaluation (Cardoso et al., 2015). Landraces ecotypes can be described according to the place of collection as mocozinho from Ceará and those with light brown fiber collected in the region of Seridó, which comprises parts of the States of Rio Grande do Norte and Paraíba (Freire and Moreira, 1991). There were two main breeding programs, one from the University of Ceará and other from Embrapa Cotton. The developed varieties tend to adapt in all region, but those from the University tended to be to Ceará, and those from Embrapa be distributed. The distribution of seeds has been organized by States based services (Moreira et al., 1982), what may explain partially the genetic organization.

Pernambuco and Piauí states have the greatest allele numbers and exclusive alleles, therefore may be chosen for programs for *in situ* or *ex situ* preservation.

One important structure of the genetic distribution, explained by considering Seridó Region the place of origin of mocó landraces (Freire and Moreira, 1991) and Ceará the only breeding program at a state level (Moreira et al., 1982), has been shown by the previous analysis with microsatellite markers, genetic distance and neighbor-joining methods. The states included in this study, such as Pernambuco, Alagoas, Sergipe, Bahia, Amapá, and Roraima, were all grouped with Ceará. Human migration and preservation of familiar seeds from Ceará may explain the similarity of the genotypes collected in North Region, mainly. In Bahia, Alagoas, and Sergipe, mocó cotton has been planted in the past, and Ceará cultivars may have been adapted. The Baeysian analysis introduced in the present studies confirms the genetic structure delineated by genetic distance and neighbor-joining methods.

The *in situ* preservation in backyards is not enough to preserve agricultural traits. The evaluation for such traits, as fiber quality and disease resistance may foster the use of these genotypes in breeding.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The support for this project was from Embrapa and a FINEP project coordinated by Dr Mauro Carneiro, Cenargen.

REFERENCES

Almeida VC, Hoffmann LV, Yokomizo GK, da Costa JN, et al. (2009). In situ and genetic characterization of *Gossypium barbadense* L. from the States of Pará and Amapá, Brazil. *Pesqui. Agropecu. Bras.* 44: 719-725. https://doi.org/10.1590/S0100-204X2009000700011

Alves MF, Barroso PAV, Yamaguishi Ciampi, A, Hoffmann LV, et al. (2013). Diversity and genetic structure among subpopulations of *Gossypium mustelinum* (Malvaceae). *Genet. Mol. Res*. 12: 597-609.

Barroso PAV, Hoffmann LV, Batista CE, Freitas RB, et al. (2010). In situ conservation and genetic diversity of three populations of *Gossypium mustelinum* Miers (ex Watt). *Genet. Resour. Crop Evol.* 57: 343-349. https://doi.org/10.1007/s10722-009-9472-9

Botstein D, White RL, Skolnick M and Davis RW (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32: 314-331.

Boulanger J and Pinheiro D (1972). Consequências genéticas da evolução da cultura algodoeira do nordeste do Brasil. *Pesqui. Agropecu. Nord.* 4: 45-52.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, et al. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368: 455-457. https://doi.org/10.1038/368455a0

Cardoso KCM, Abreu AG, Cardoso RM, Rocha ASNC, et al. (2015). Morfologia de algodoeiros de diferentes estados do brasil. In: 10º Simpósio de Recursos Genéticos para a América Latina e o Caribe, Bento Gonçalves.

Doyle JJ and Doyle JL (1990). Isolation of plant DNA from fresh tissue. *Focus* 12: 13-15.

Evanno G, Regnaut S and Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14: 2611-2620 https://doi.org/10.1111/j.1365-294X.2005.02553.x.

Freire EC and Moreira JA (1991). Relações genéticas entre o algodoeiro Mocó e diferentes espécies e raças de algodoeiro. *Rev. Bras. Genet.* 14: 393-411.

Freire EC, Santos MSS, Medeiros LC, Andrade FP, et al. (1990). Avaliação preliminar da coleção de germoplasmas de algodoeiro arbóreo no nordeste do Brasil. Available from Centro Nacional de Pesquisa do Algodão, Campina Grande, PB, Brazil. Publ. Documentos No. 38.

Gao H, Williamson S and Bustamante CD (2007). A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics* 176: 1635-1651. https://doi.org/10.1534/genetics.107.072371

Giband M, Dessauw D and Barroso PAV (2010). Cotton: Taxonomy, origin and domestication. In. Cotton: Technology for the 21st century (Wakelyn PJ, Chaudhry MR, ed.). International Cotton Advisory Committee, 3-15.

Goudet J (2001). FSTAT: program to estimate and test gene diversities and fixation indices (Software). Version 2.9.3. Available at [http://www2.unil.ch/popgen/softwares/fstat.htm].

Hedrick PH (2001). Conservation genetics: Where are we now? *Trends Ecol. Evol.* 16: 629-636. https://doi.org/10.1016/S0169-5347(01)02282-0

Hinze LL, Fang DD, Gore MA, Scheffler BE, et al. (2015). Molecular characterization of the *Gossypium* Diversity Reference Set of the US National Cotton Germplasm Collection. *Theor. Appl. Genet.* 128: 313-327. https://doi.org/10.1007/s00122-014-2431-7

Hutchinson JB (1951). Intra-specific differentiation in *Gossypium hirsutum*. *Heredity* 5: 161-193. https://doi.org/10.1038/hdy.1951.19

IBGE (2015). Instituto Brasileiro de Geografia e Estatística. Available at [www.sidra.ibge.gov.br].

Jakobsson M and Rosenberg NA (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23: 1801-1806. https://doi.org/10.1093/bioinformatics/btm233

Jarvis DI and Hodgkin T (1999). Wild relatives and crop cultivars: detecting natural introgressões and farmer selection of new genetic combinations in agro-ecosystem. *Mol. Ecol.* 8: S159-S173. https://doi.org/10.1046/j.1365-294X.1999.00799.x

Lacape LM, Dessauw D, Rajab M, Noyer JL, et al. (2007). Microsatellite diversity in tetraploid *Gossypium* germplasm: assembling a highly informative genotyping set of cotton SSRs. *Mol. Breed.* 19: 45-58. https://doi.org/10.1007/s11032-006-9042-1

Lewis P and Zaykin D (2000). Genetic data analysis: computer program for the analyses of allelic data (software). Version 1.1 (win32). Available at [http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php].

Liu K and Muse SV (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21: 2128-2129. https://doi.org/10.1093/bioinformatics/bti282

Liu S, Saha S, Stelly D, Burr B, et al. (2000). Chromosomal assignment of microsatellite loci in cotton. *J. Hered.* 91: 326-332. https://doi.org/10.1093/jhered/91.4.326

Manel S, Berthoud F, Bellemain E, Gaudeul M, et al. (2007) A new individual-based spatial approach for identifying genetic discontinuities in natural populations. *Mol. Ecol.* 16: 2031-2043.

Menezes IPP, Barroso PAV, Hoffmann LV, Lucena VS, et al. (2010). Genetic diversity of mocó cotton (*Gossypium hirsutum* race *marie-galante*) from the northeast of Brazil: implications for conservation. *Botany* 88: 1-9.

Menezes IP, Gaiotto FA, Hoffmann LV, Ciampi AY, et al. (2014). Genetic diversity and structure of natural populations of Gossypium mustelinum, a wild relative of cotton, in the basin of the De Contas River in Bahia, Brazil. *Genetica* 142: 99-108. https://doi.org/10.1007/s10709-014-9757-6

Minch E, Ruiz-Linares A, Goldstein DB, Feldman MW, et al. (1997). Microsat (version 1.5): a computer program for calculating various statistics on microsatellite allele data. Stanford University Medical Center, Stanford, CA.

Moiana LD, Vidigal Filho PS, Gonçalves-Vidigal MC and Carvalho LP (2015). Genetic diversity and population structure of upland cotton Brazilian cultivars (*Gossypium hirsutum* L. race *latifolium* H.) using SSR markers. *Aust. J. Crop Sci.* 9: 143-152.

Moreira JAN, Beltrão NEM, Freitas EC, Novaes Filho MB, et al. (1997). Decadência do algodoeiro mocó e medidas para o seu soerguimento no nordeste brasileiro. Available at Centro Nacional de Pesquisa do Algodão, Campina Grande, PB, Brazil, Publ. Documento No. 43.

Moreira JAN, Freire EC, Santos RF, Barreiro Neto M, et al. (1982). Visão retrospectiva do melhoramento genético no algodoeiro mocó (*Gossypium hirsutum r. marie galante* Hutch) no nordeste do Brasil. Available at Centro Nacional de Pesquisa do Algodão, Campina Grande, PB, Brazil, Publ. Documento No. 12.

Moreira JAN, Freire EC, Santos RF and Vieira RM (1995). Use of numerical taxonomy to compare "Mocó" cotton with other cotton species and races. *Rev. Bras. Genet.* 18: 99-103.

Nei M (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89: 583-590.

Nguyen TB, Giband M, Brottier P, Risterucci AM, et al. (2004). Wide coverage of the tetraploid cotton genome using newly developed microsatellite markers. *Theor. Appl. Genet.* 109: 167-175. https://doi.org/10.1007/s00122-004-1612-1

Pereira GS, Cazé ALR, Silva MG, Almeida VC, et al. (2015). Optimal use of SSR markers for varietal identification of upland cotton. *Pesqui. Agropecu. Bras.* 50: 571-581. https://doi.org/10.1590/S0100-204X2015000700007

Rodriguez M, Rau D, Bitocchi E, Bellucci E, et al. (2015). Landscape genetics, adaptive diversity and population structure in *Phaseolus vulgaris*. *New Phytol.* 209: 1781-1794.

Rosenberg NA (2004). Distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes* 4: 137-138. https://doi.org/10.1046/j.1471-8286.2003.00566.x

Silva AR, Malafaia, G and Menezes IPP (2017) Biotools: an R function to predict spatial gene diversity via an individual-based approach *Gen. Mol. Res*. 16: gmr16029655.

Stephens SG (1973). Geographical distribution of cultivated cottons relative to probable centers of domestication in the new world. In: Genes, enzymes and populations (Adrian M, ed.). Plenum Press, New York, 239-254.

Tamura K, Dudley J, Nei M and Kumar S (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24: 1596-1599. https://doi.org/10.1093/molbev/msm092

Yu JZ, Fang DD, Kohel RJ, Ulloa M, et al. (2012). Development of a core set of SSR markers for the characterization of *Gossypium* germplasm. *Euphytica* 182: 203-213. https://doi.org/10.1007/s10681-012-0643-y

Wagner HW and Sefc KM (1999). Identity 1.0. Centre for applied genetics. University of Agricultural Sciences, Vienna.