GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Genetic-molecular characterization of backcross generations for sexual conversion in papaya (*Carica papaya* L.)

H.C.C. Ramos[1], M.G. Pereira[1], T.N.S. Pereira[1], G.B.A. Barros[1] and G.A. Ferreguetti[2]

[1]Laboratório de Melhoramento Genético Vegetal,
Centro de Ciências e Tecnologias Agropecuárias,
Universidade Estadual do Norte Fluminense Darcy Ribeiro,
Campos dos Goytacazes, RJ, Brasil
[2]Empresa Caliman Agrícola, Linhares, ES, Brasil

Corresponding author: H.C.C. Ramos
E-mail: helainecr@uenf.br

Genet. Mol. Res. 13 (4): 10367-10381 (2014)
Received April 2, 2014
Accepted August 8, 2014
Published December 4, 2014
DOI http://dx.doi.org/10.4238/2014.December.4.32

ABSTRACT. The low number of improved cultivars limits the expansion of the papaya crop, particularly because of the time required for the development of new varieties using classical procedures. Molecular techniques associated with conventional procedures accelerate this process and allow targeted improvements. Thus, we used microsatellite markers to perform genetic-molecular characterization of papaya genotypes obtained from 3 backcross generations to monitor the inbreeding level and parental genome proportion in the evaluated genotypes. Based on the analysis of 20 microsatellite loci, 77 genotypes were evaluated, 25 of each generation of the backcross program as well as the parental genotypes. The markers analyzed were identified in 11 of the 12 linkage groups established for papaya, ranging from 1 to 4 per linkage group. The average values for the inbreeding coefficient were 0.88 (BC_1S_4), 0.47 (BC_2S_3), and 0.63 (BC_3S_2). Genomic analysis revealed average values of the recurrent parent genome of 82.7% in

BC_3S_2, 64.4% in BC_1S_4, and 63.9% in BC_2S_3. Neither the inbreeding level nor the genomic proportions completely followed the expected average values. This demonstrates the significance of molecular analysis when examining different genotype values, given the importance of such information for selection processes in breeding programs.

Key words: Backcrossing; *Carica papaya* L.; Genetic diversity; Molecular markers; Parental proportion

INTRODUCTION

Papaya (*Carica papaya* L., Caricaceae) is the most economically important and famous species in its family. Despite significant production of this fruit tree, the narrow genetic basis of commercial types of papaya is well documented (Kim et al., 2002; Ma et al., 2004; Silva et al., 2008). Ming et al. (2008) found that a plausible explanation for the consolidation of genotypes of papaya might be a cultural preference, in addition to geographical isolation, forcing the selection of cultivars with a relatively narrow genetic basis and resulting in low diversity. Expansion of this genetic basis implies the introgression of an exotic germplasm in breeding programs as a source of new genes, in addition to the implementation of breeding programs to promote the hybridization of divergent genetic materials; this will favor the establishment of new gene combinations.

In Brazil, commercial varieties are primarily inbred lines and hybrid cultivars. The latter appear to have become a global trend (Oliveira et al., 2010) because of their high yield and fast return on investment (Chan, 2009). In the development and availability of new genetic materials, superior lines can be directly provided to producers and used in hybridization programs to develop stable hybrids. The use of inbred lines as commercial cultivars in the papaya crop may be based on floral biology, in which the reproductive structure of hermaphrodite plants allows self-fertilization, preventing inbreeding depression (Chan et al., 2009; Oliveira et al., 2010). Thus, hermaphroditism is relevant to papaya breeding programs aimed at developing new cultivars, in addition to favoring the development of fruits with commercially acceptable standards.

Hermaphroditism is one of the 3 sexual forms presented by papaya plants. Several hypotheses have been proposed to explain the genetic determination of these sexual forms (Storey, 1953; Hofmeyr, 1967; Ming et al., 2007). Advances in genomic research demonstrated, through the construction of a genetic map (Ma et al., 2004), physical map (Liu et al., 2004), *in situ* map (Yu et al., 2007), and sequencing results (Liu et al., 2004; Yu et al., 2008), that sex determination of the papaya trees is controlled by a pair of sexual chromosomes that evolved recently and show differences only at the molecular level. The chromosomes Y and Y^h present a small male-specific region responsible for the expression of hermaphroditism and masculinity; this sequence is approximately 4-5 Mb in length (Liu et al., 2004).

This knowledge allows for the transference of this genomic region to dioecious materials of great genetic and agronomic potential for breeding by backcrossing programs conducted by the Universidade Estadual do Norte Fluminense (UENF) research group (Silva et al., 2008; Ramos et al., 2011a). The efficiency of backcrossing can be significantly increased through the association with molecular markers, which is useful for estimating the genomic proportion of individuals and accelerating the recovery of the recurrent parent genome (Young and Tanksley, 1989; Servin and Hospital, 2002).

Despite recent trends of using emergent molecular markers to achieve high accuracy in genetic studies, microsatellites continue to be the most widely used class of markers for several reasons. The highly mutable nature of microsatellites makes them a powerful marker for distinguishing DNA polymorphisms among genotypes that are intimately related (Eustice et al., 2008). Other attributes of this marker include its multiallelic nature, reproducibility, high information content, codominant heritage, abundance, and extensive genome coverage. Additionally, their genomic distribution, evolutionary dynamics, biological function, and practical utility have been examined in several studies (Guichoux et al., 2011). In the papaya tree, hundreds of microsatellites have already been identified and characterized (Eustice et al., 2008; Oliveira et al., 2010), and the genomic location of many microsatellites has been determined by genetic mapping (Chen et al., 2007).

Therefore, in this study, we aimed to i) use microsatellite markers to detect polymorphisms in papaya genotypes; ii) conduct genetic characterization, seeking to estimate genotypic indices to quantify and organize genetic variability; iii) monitor the inbreeding level in the genotypes; and iv) estimate the genomic proportion of the recurrent parent in the population evaluated.

MATERIAL AND METHODS

Plant material

Seventy-five genotypes from the papaya genetic breeding program at the UENF were evaluated, as well as the parental genotypes (Cariflora and 'SS783'). The families evaluated were from 3 backcrossed generations: 25 from the first (BC_1), 25 from the second (BC_2), and 25 from the third generation (BC_3) of recurrent crossing with the Cariflora parent, as shown in Figure 1. These progenies were obtained using a procedure similar to the genealogic method, which is used to develop superior lines.

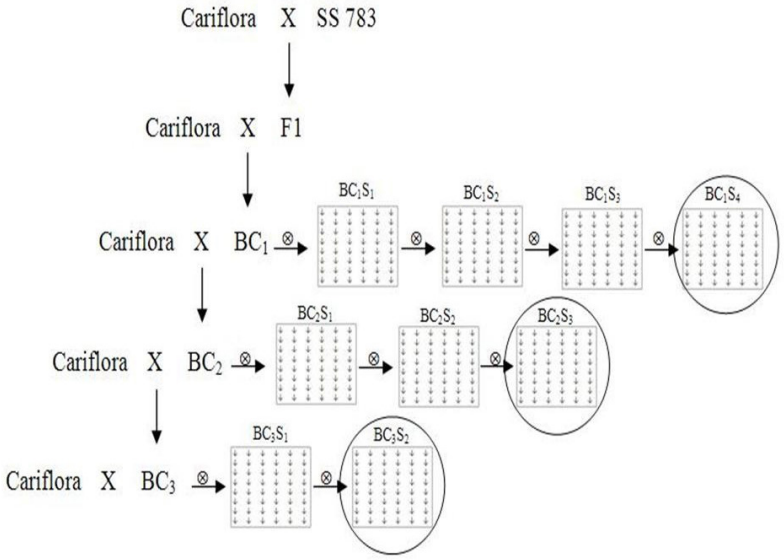

Figure 1. Breeding procedure used to achieve the genotypes evaluated in this study.

Segregant progenies were derived from the initial crossing between the dioecious genotype Cariflora (recurrent parent) and the cultivar Sunrise Solo 783 (SS783) in the backcross program. This program aims to transfer the hermaphroditism gene from 'SS783' to the dioecious genotype Cariflora because this genotype presents good combining ability (general and specific) when crossed with genotypes of the 'Solo' group (Marin et al., 2006). The elite genotype 'SS783', used as a donor parent (DP) of hermaphroditism, presents a high degree of homozygosity. Crosses between the hermaphrodite 'SS783' genotypes cause segregation of sex at the ratio of 2 hermaphrodite plants to 1 female plant, which is known as a gynoecious-andromonoic population.

Genomic DNA extraction

Total genomic DNA was extracted from young leaves using the Plant Genomics DNA Extraction Kit YGP 100 - RBC (BioAmerica, Linden, NJ, USA) according to manufacturer instructions. Samples of young leaves were individually collected for each genotype of the generations evaluated, while for the parental genotypes, the collection was conducted in bulk; this means that the analysis set was formed by 1 sample of 5 plants of each parent in order to sample the highest possible number of allelic forms present in each parent for each locus analyzed.

After extraction, the DNA was quantified by 0.8% agarose gel analysis using the High DNA Mass Ladder (Invitrogen, Carlsbad, CA, USA) marker. DNA samples were stained using the GelRed™ and Blue Juice mixture (1:1), and the image was captured using the MiniBis Pro photodocumentation system (DNR Bio-Imaging Systems, Israel). Later, the images were analyzed using the ImageJ software system and diluted to the working concentration of 10 ng/μL.

SSR analysis

The DNA of the parental genotypes was initially used to screen microsatellites markers to identify polymorphic SSR loci. Eighty-four genomic regions containing microsatellite sequences were analyzed, of which only 20 showed well-defined differentiation between the parents. These regions were accessed from SSR primers developed by Eustice et al. (2008), whose genomic locations were determined by genetic mapping performed by Chen et al. (2007).

Thus, amplification reactions of the segregant progeny and the parental genotypes were performed using 20 SSR primers in a final volume of 15 μL, as described by Ramos et al. (2011b); the annealing temperature was 60°-63.5°C according to each primer (Table 1). The amplification products were separated in a 4% MetaPhor agarose gel (Lonza, Basel, Switzerland) and stained using the GelRed™/Blue Juice mixture (1:1). The results were visualized using the MiniBis Pro photodocumentation system (DNR Bio-Imaging Systems, Israel).

Data analysis

Data obtained from amplification of the SSR primers were converted into numeric code for each allele per locus. This numeric matrix was developed by attributing values according to the number of alleles per locus as follows: 1 locus presenting 3 alleles is represented by 11, 22, and 33 for the homozygous forms (A_1A_1, A_2A_2, and A_3A_3, respectively) and 12, 13, and 23 for the heterozygous forms (A_1A_2, A_1A_3, and A_2A_3, respectively). Using this numeric base, we calculated the genetic distance between the genotypes studied using the GENES

(Cruz, 2008) software system with the weighted index as described by Ramos et al. (2011c). Cluster analysis of genotypes using a dendrogram was performed using the neighbor-joining hierarchical method (Saitou and Nei, 1987) with the Mega software system, version 5 (Tamura et al., 2011). A graphic dispersion of the genotypes was depicted using the principal coordinate analysis method using the Genalex 6.3 software system (Peakall and Smouse, 2009).

The software systems PowerMarker version 3.25 (Liu and Muse, 2005) and Popgene version 1.31 (Yeh et al., 1999) were used to estimate the values of polymorphism information content (PIC), Shannon Index, expected heterozygosity (H_E) estimated based on the expected proportion of heterozygotes upon random mating, observed heterozygosity (H_O) estimated based on the proportion of heterozygotes observed in a certain locus, and the inbreeding coefficient (F). The genotypes were also evaluated according to the genetic structure. For this, we used the Bayesian method with the Structure 2.3.1 software system (Pritchard et al., 2000). The admixture model and correlated allelic frequencies were employed using the "Burnin Period = 5000", followed by the extension of 50,000 replications during analysis.

The proportion of the recurrent parent genome transferred to the individuals of the 3 backcross generations was analyzed using the method proposed by Benchimol et al. (2005) with the formula PR = B + 0.5H/(B + H), where B is the number of plants with the recurrent allele and H is the number of plants with the heterozygote genotype. The data obtained were used to determine the proportion of the genomes of the recurrent and the donor parents in each individual of the populations.

RESULTS AND DISCUSSION

Microsatellites markers

Eighty-four microsatellite markers were analyzed at the molecular level to distinguish the Cariflora and 'SS783' genotypes. Of these, 20 (23.8%) showed polymorphisms between the parental genotypes. Therefore, these polymorphic markers were used in the analysis of all genotypes evaluated in this study. These SSR loci generated 46 alleles, with the number of polymorphisms per locus from 2 to 4 and an average of 2.3 alleles/locus (Table 1). The number of alleles observed per locus was very similar to the minimum possible value, which is related to the bi-parental nature of the initial crossing that produced the population evaluated.

Although low, the number of alleles identified in the present study was higher than that found by Ramos et al. (2011b). This difference is likely because of the larger number of generations analyzed in this study and higher variability of the BC_3 families because these families were submitted to the recombination process during the breeding program. Additionally, the segregant nature of the recurrent parent itself favors higher heterozygosity in the progenies, mainly in the third backcross generation. Thus, our results were not unexpected.

The characteristics of the primers and their disposition among the linkage groups (LG) as described by Chen et al. (2007) are summarized in Table 1. Of the 20 regions accessed by the primers, 4 were genic regions located in LG3 (1), LG5 (1), and LG7 (2). Among the other primers, 2 were achieved from subclones of papaya BACs, and 14 were developed from information obtained by full genome sequencing. These markers are distributed in 11 of the 12 LGs established by Chen et al. (2007), where only LG11 was not sampled. The number of markers per group ranged from 1 (LG5, LG9, LG10, LG12) to 4 (LG4). The other linkage groups were represented by 2 markers each. The data showed that the use of markers from ge-

netic maps allows aggregation of relevant information to genetic studies, such as location and origin, leading to a better coverage of the genome and allowing for further precise inference regarding the study population.

Table 1. Sequence of the 20 pairs of microsatellite primers used to analyze the 75 genotypes derived from backcrossing and parental genotypes.

Locus	Sequence of the primer (5'→3')	LG	Ta (°C)	No. of alleles	H_E	H_O	PIC
P8K39CC	F: CGTCAAGTTGTTGGGTTGGTC R: TGACATCTCCGAAGAGCTGAGA	1	60	2	0.42	0.03	0.33
P3K1200CC	F: TGGTCCTCGAACGAATAGTGA R: TGATCATCATGCATCTACCAGC	1	63.5	2	0.39	0.20	0.32
P3K6912CC	F: TGAAGCCTCAGTGAATCCAAA R: CCCATGGGAACACATCTATTG	2	60	2	0.49	0.07	0.37
P3K1850CC	F: TTTCTCCCACATGACCCACA R: GGGGGTGCTTTGGAATCTTT	2	60	2	0.21	0.12	0.19
CPM1621CC	F: ATGGTAACCCAGCGTGAGGA R: ACGCCAAATATTCCCAACCC	3	60	2	0.48	0.16	0.37
ctg-14CC	F: GAAAAGGATATGGCGCAACCT R: AGTTCCAGGAAATTGCGGGT	3	60	3	0.48	0.08	0.43
P3K3968A5	F: TGCGATCGAAAGGTTCTTGAG R: TGGAAATGGCTGGTTTTGTCA	4	60	2	0.40	0.12	0.32
P3K1883CC	F: GGTTGAAACGTTAACGGCG R: GGGTAGAGAGTCAATGGATTTTGC	4	60	3	0.47	0.05	0.39
P6K268CC	F: ATGCTTGAGGGACAACCCTT R: AAAAGTATGCAGTCCCCAGTTG	4	63.5	2	0.40	0.21	0.32
P6K128CC	F: GCCGGCTCAGGAGGTTAAGA R: CAATGACCAAACGCCACACA	4	63.5	2	0.31	0.22	0.27
ctg-365A5	F: TTCTTTCACCCGCTCCTCTG R: AAACAACTCGGCCCAACTGA	5	60	3	0.46	0.05	0.38
P3K23CC	F: CGTAAAGGTCGGGTCAGCTA R: TGGTCTTCACATGAAATGAGCTT	6	60	2	0.39	0.22	0.31
P3K1382A5	F: ACAAATCCAGCAAATATCCCATT R: CAACATCTCAATTTGCAAAGCA	6	60	2	0.36	0.16	0.30
ctg-64CC	F: CATCCCGAACTACTCACATAAACA R: TGCTTGCTGCTCACTTATGG	7	60	2	0.47	0.40	0.36
ctg-41S5	F: TTCATCGTCTCGCTGAAATTGA R: CCAGTAGGCTCTCCAAATGGG	7	60	2	0.41	0.22	0.32
CPM766CC	F: TACCAAGTTCAGCAAGCGGT R: ATACTTTCTCCCCCTTCGGA	8	60	2	0.44	0.22	0.34
P3K170CC	F: CAATGGAGGGCAGTTTTGATG R: TGGGAGAAAAGGAAAGAACATGA	8	63.5	2	0.44	0.00	0.35
P3K1497CC	F: TGACGGTGAAAATTGCAACA R: AAAAGGGGAGTCCAAATTGGTT	9	60	3	0.65	0.57	0.58
P3K7484C0	F: CGGTAGCGACTCATCGGACT R: TTGACTCGCGAGGAAAGGAG	10	60	2	0.33	0.12	0.28
P3K3510C0	F: GTAGCCGAACGCACAACACA R: CGTGTAAAAGAAGCGGTAGATCG	12	60	4	0.69	0.88	0.64
Mean				2.3	0.44	0.21	0.36

LG = linkage group; Ta = annealing temperature; H_E = expected heterozygosity; H_O = observed heterozygosity; PIC = polymorphism information.

Previous studies investigated the number of markers necessary to achieve efficient control of genetic "background" in marker-assisted backcross programs (Hospital et al., 1992; Visscher et al., 1996), and found that analyzing 2-4 markers per chromosome provides good genetic control. According to Servin and Hospital (2002), the marker position in the chromosomes may be more relevant than their quantity and the optimum position of 2 markers in the chromosome is 20 cM from the telomere. They also found that better control of the genomic background could be obtained by analyzing a large number of markers per chromosome, which

are not in optimum positions and can be achieved by using few markers at ideal positions; thus, maximization of the expected proportion of the recurrent genome can be reached.

Genetic diversity

Analysis of the SSR loci showed that the expected heterozygosity (H_E) in the population ranged from 0.21 to 0.69, with an average of 0.44, while the observed heterozygosity (H_O) ranged from 0.00 to 0.88, with an average of 0.21 (Table 1). The low heterozygosity of the population was expected and is related to the degree of selection. However, the difference between the expected and observed heterozygosity values may be related to allele failure (drop-out) during amplification in the PCR reaction as well as the structure of the population in inbreeding (Fukunaga et al., 2005).

Similar to heterozygosity, the PIC can be used to quantify genetic polymorphisms in each locus in the population. Markers were classified as informative when PIC $\geq$ 0.5. The highest PIC value was observed in locus P3K3510C0 (0.64); the lowest was in P3K1850CC (0.19), with an average of 0.36. PIC values indicate that the locus P3K3510C0, located in LG12, has the highest discriminatory power among the loci analyzed, which is supported by the higher number of alleles found in this genomic region. This result agrees with those of Mateescu et al. (2005), who described that the increase in the number of alleles per microsatellites locus translates into an increase in PIC, and PIC was closer to the expected heterozygosity. Pervaiz et al. (2010) examined genetic diversity in rice using microsatellite markers and found that the PIC values showed a significant and positive linear correlation with number of alleles at the SSR locus. Table 1 shows that the lowest difference between PIC and H_E was in locus P3K3510C0.

The Shannon index has been used in genetic studies to measure diversity within populations and is similar to an index of genotypic richness. For this index, the closer the estimated values are to unity, the higher the diversity. For all genotypes analyzed, index values ranged from 0.35 to 1.26. The average value of 0.67 revealed the existence of moderate variability in this population, which is sufficient for maintaining this breeding program. When each backcross generation was considered separately, the index average values were 0.64, 0.26, and 0.63 for the BC_1S_4, BC_2S_3, and BC_3S_2 generations, respectively, indicating lower genetic diversity for the BC_2S_3 generation.

Analysis of genetic diversity among the genotypes based on estimated heterozygosity revealed that the expected values were generally higher than were those observed. The exception was verified for the recurrent parent (RP), Cariflora, showing H_E and H_O values of 0.43 and 0.60, respectively. The expected heterozygosity ranged from 0.43 to 0.68, while the observed heterozygosity ranged from 0.05 to 0.60 (Table 2). Considering the three backcross generations individually, H_O values ranged from 0.05 to 0.40 for individuals of the BC_1S_4 generation, 0.05 to 0.11 in the BC_2S_3 generation, and 0.15 to 0.50 in the BC_3S_2 generation; averages were 0.23, 0.08, and 0.29, respectively. Reduced estimated heterozygosity may result from the presence of null alleles, a problem inherent to codominant markers. This can be solved using a large number of molecular markers, resulting from estimates closer to the expected values (Oliveira et al., 2010).

Analysis of these genotypic indices revealed higher diversity between individuals in the BC_1S_4 and BC_3S_2 populations. This higher variability found for the BC_1S_4 and BC_3S_2 generations can be explained by the existing variation in the pedigree of the families from which they were comprised. The BC_3S_2 generation, in addition to higher variation among its

progenies, showed a lower number of self-fertilization cycles, leading to a higher level of heterozygosity. Alternatively, the results observed in the BC_2S_3 can be explained by the poor sampling of this generation because the individuals evaluated were from a single plant of the previous generation (BC_2S_2). Because of the peculiarities of the backcross method, there is a trend for decreased values generally decreased for PIC, Shannon index (or genotypic population richness), and observed heterozygosity over the generations, which was also observed in generations advanced from self-fertilization. Thus, there was a progressive variability loss.

Table 2. Genotypes derived from 3 backcross generations and respective values of the expected heterozygosity (H_E), observed heterozygosity (H_O), and inbreeding coefficient (f).

Genotypes	H_E	H_O	f	Genotypes	H_E	H_O	f
Cariflora	0.43	0.60	-0.41	17BC2-20S3-19	0.67	0.05	0.93
SS783	0.65	0.05	0.93	17BC2-20S3-20	0.66	0.05	0.93
52BC1-7S4-1	0.68	0.10	0.86	17BC2-20S3-21	0.65	0.10	0.85
52BC1-8S4-1	0.64	0.05	0.92	17BC2-20S3-22	0.65	0.10	0.85
52BC1-9S4-1	0.53	0.18	0.68	17BC2-20S3-23	0.64	0.05	0.92
52BC1-11S4-1	0.64	0.16	0.76	17BC2-20S3-24	0.63	0.10	0.85
52BC1-12S4-1	0.59	0.30	0.50	17BC2-20S3-25	0.64	0.10	0.85
52BC1-13S4-1	0.57	0.20	0.66	17BC2-20S3-14	0.66	0.05	0.93
52BC1-15S4-1	0.65	0.10	0.85	17BC2-20S3-15	0.65	0.10	0.85
52BC1-15S4-2	0.66	0.05	0.92	17BC2-20S3-16	0.66	0.05	0.92
52BC1-15S4-3	0.65	0.11	0.84	17BC2-20S3-17	0.67	0.05	0.93
52BC1-15S4-4	0.65	0.10	0.85	17BC2-20S3-12	0.65	0.10	0.85
52BC1-22S4-1	0.63	0.15	0.77	17BC2-20S3-13	0.66	0.05	0.93
52BC1-24S4-1	0.66	0.10	0.85	21BC3-27S2-1	0.54	0.37	0.33
52BC1-21S4-1	0.64	0.17	0.74	21BC3-27S2-2	0.59	0.50	0.16
52BC1-21S4-2	0.63	0.40	0.37	22BC3-28S2-1	0.56	0.20	0.65
52BC1-21S4-3	0.64	0.40	0.38	22BC3-28S2-2	0.57	0.25	0.57
52BC1-21S4-4	0.61	0.35	0.43	22BC3-28S2-3	0.56	0.26	0.54
52BC1-21S4-5	0.59	0.30	0.50	22BC3-29S2-1	0.59	0.15	0.75
52BC1-21S4-6	0.64	0.35	0.46	22BC3-29S2-2	0.56	0.20	0.65
52BC1-21S4-7	0.64	0.37	0.43	19BC3-30S2-1	0.56	0.25	0.56
52BC1-21S4-8	0.61	0.25	0.59	19BC3-30S2-2	0.55	0.30	0.46
52BC1-21S4-9	0.64	0.32	0.52	19BC3-30S2-3	0.53	0.30	0.44
52BC1-21S4-10	0.66	0.25	0.63	6BC3-31S2-1	0.51	0.26	0.49
52BC1-21S4-11	0.64	0.40	0.38	6BC3-31S2-2	0.50	0.32	0.38
52BC1-21S4-12	0.64	0.40	0.38	6BC3-32S2-1	0.61	0.16	0.75
52BC1-21S4-13	0.63	0.30	0.53	6BC3-32S2-2	0.65	0.28	0.58
17BC2-20S3-1	0.65	0.10	0.85	6BC3-32S2-3	0.58	0.15	0.75
17BC2-20S3-2	0.64	0.05	0.92	16BC3-33S2-1	0.48	0.35	0.28
17BC2-20S3-3	0.66	0.10	0.85	16BC3-33S2-2	0.50	0.42	0.16
17BC2-20S3-4	0.67	0.11	0.85	5BC3-34S2-1	0.51	0.26	0.49
17BC2-20S3-5	0.66	0.05	0.93	5BC3-34S2-2	0.58	0.25	0.58
17BC2-20S3-6	0.65	0.10	0.85	5BC3-34S2-3	0.51	0.37	0.28
17BC2-20S3-7	0.66	0.10	0.85	5BC3-35S2-1	0.56	0.37	0.35
17BC2-20S3-8	0.65	0.10	0.85	5BC3-35S2-2	0.56	0.39	0.31
17BC2-20S3-9	0.65	0.10	0.85	4BC3-36S2-1	0.54	0.25	0.54
17BC2-20S3-10	0.66	0.05	0.93	4BC3-36S2-2	0.49	0.30	0.40
17BC2-20S3-11	0.66	0.10	0.85	4BC3-36S2-3	0.60	0.40	0.34
17BC2-20S3-18	0.66	0.05	0.93	Mean	0.61	0.20	0.65

H_E = expected heterozygosity; H_O = observed heterozygosity; f = inbreeding coefficient.

In breeding programs targeting the development of inbred lines, analysis of the inbreeding coefficient, or fixation index, is a very important parameter for measuring the level of homozygosity and heterozygosity in the population. Among the parental genotypes used in this study, the fixation index was -0.41 for Cariflora (recurrent parent) and 0.93 for 'SS783' (donor parent) (Table 2), which is similar to the expected values given their genetic nature.

However, negative values for the inbreeding coefficient are not common. This result verified that the recurrent parent could be associated with a higher value of observed heterozygosity compared to the expected heterozygosity, suggesting a possible excess of loci in heterozygosity in this individual.

Analysis of the inbreeding coefficient among individuals from different backcross generations showed values ranging from 0.37 to 0.92 in BC_1S_4, 0.85 to 0.93 in BC_2S_3, and 0.16 to 0.75 in BC_3S_2, with average values of 0.63, 0.88, and 0.47, respectively. Considering the backcross generation in which the genotypes are located, and concomitantly, the number of generations of self-fertilization to which they were submitted, the expected average inbreeding coefficients were 0.95, 0.92, and 0.86 for BC_1S_4, BC_2S_3, and BC_3S_2, respectively. However, values measured were significantly lower than expected values, mainly for the BC_1S_4 and BC_3S_2 generations, demonstrating that a high level of heterozygosity was maintained based on the microsatellite loci analyzed.

Optimum indices for the inbreeding coefficient were found by Oliveira et al. (2010) in their analysis of 83 pure lines and 3 segregant papaya populations. The authors used 20 polymorphic microsatellite loci in marker-assisted selection and found inbreeding coefficient values ranging from 0.63 to 1.00. Eleven lines were identified with 100% inbreeding, in addition to 18 lines with values very close to the maximum (0.95-0.96). This study demonstrated that the use of microsatellite markers in assisted selection is efficient for developing inbred lines of papaya.

Parent genomic proportion

Because of the codominant nature of microsatellite markers, it was possible to estimate both the level of fixation of loci and the genomic constitution of individuals. For the average proportion of recovery of the recurrent genome (RP) in progenies, it is expected that each *t* backcross generation will be similar to the recurrent parent, on average, in $1-(1/2)^{t+1}$ (Allard, 1971). Analysis of the parent genomic proportion in the progeny revealed an average of 70.3% of the RP genome for all evaluated genotypes (Figure 2). The analysis per generation revealed that a higher proportion of the RP genome was presented by the genotypes of the BC_3S_2 generation, with values of 66.7-97.5% and an average of 82.7%. The second highest proportion was observed in the BC_1S_4 generation, with an average of 64.4% and variation among the genotypes from 47.5 to 76.3%. The BC_2S_3 generation showed the lowest average proportion of the RP genome (63.9%), with individual values of 55-70% (Figure 2). Variation in the individual values in each generation demonstrates that some genotypes present RP genomic proportions within expected values, except for the second generation. After analyses, the average of parent genomic proportion per generation was below the theoretically expected value. This was because the genomic similarity of each segregant individual with the recurrent parent depends on the number of backcross generations and the level of recombination, although the latter is restricted by the advance of self-fertilized generations.

Frisch et al. (1999) explained that gene blocks linked to genes of interest (linkage drag) from the parental donor could be inserted into the recurrent genotype, thus helping to prevent the expected proportions from being found over generations. Additionally, these segments from the donor parent may carry unfavorable genes. In this case, phenotypic selection and molecular selection can maximize the possibility of excluding these traits from the breeding program.

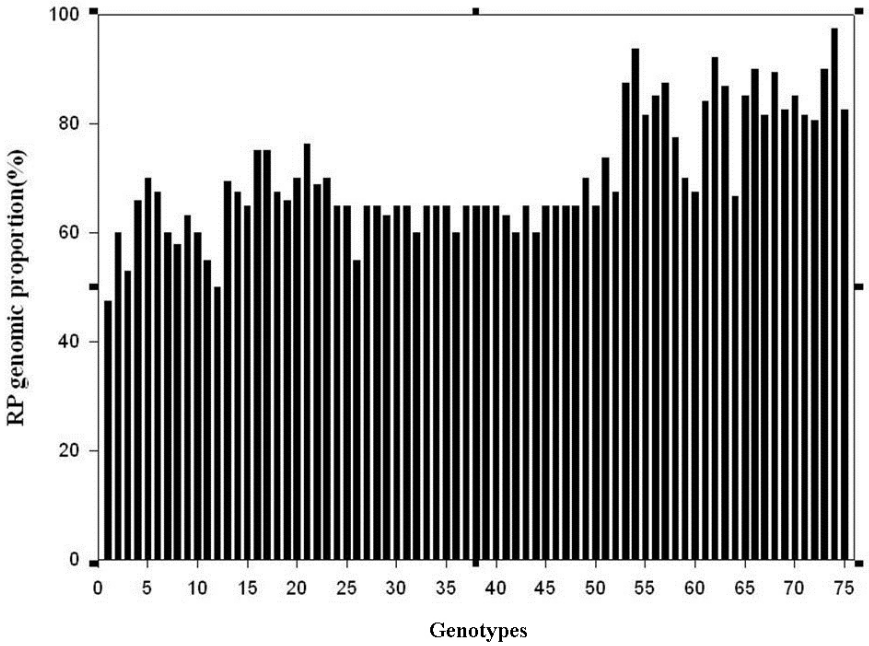

Figure 2. Genomic proportion of the recurrent parent (RP) of the genotypes belonging to the generations BC_1S_4 (1-25), BC_2S_3 (26-50), and BC_3S_2 (51-75).

Similar to the present study, Ramos et al. (2011b) analyzed 3 BC_1 progenies from the same breeding program using microsatellite markers and found that genomic proportions of the recurrent parent were lower than the expected values. In this specific case, the progenies evaluated were derived from a single BC_1 plant, giving a punctual result to this study, i.e., specific for only one backcross generation. Similarly, Silva et al. (2007) used the random amplified polymorphic DNA markers to select BC_1 genotypes to achieve the second backcross generation and found that the proportions were lower than expected. Because the main objective of this breeding program was to transfer the hermaphroditism gene and that the commercial varieties of papaya in Brazil are of the hermaphrodite type and agronomically similar to the donor parent, selecting for phenotypic attributes may have fostered deviation of the selection in favor of the donor parent, slowing recovery of the recurrent genome (Ramos et al., 2011b).

In the present study, it was possible to verify genotypes with satisfactory genomic proportions, indicating that even with the trend of directing selection in favor of the donor parent, the program showed acceptable results. Servin and Hospital (2002) found that even when selection using markers is not effectively applied, addition of the recurrent genome to the population would occur because of backcrossing. However, without effective selection, this recovery can be slow, requiring a larger number of generations to achieve the expected results.

Analysis of the genetic structure of the backcross generations was also conducted using the Structure software system (Figure 3). Based on this analysis, it was possible to clearly distinguish between the generations evaluated. The genotypes derived from BC_1 presented higher similarity, on average, for the donor parent (P2) and lower definition of hybrid zones in

the structuring of the individuals. We verified that the genetic structure of the descendants of BC_2 supported that they originated from a single plant of the previous generation and that this plant may be genetically more similar to the donor parent with a higher proportion of fixed alleles. In contrast, the genotypes from the BC_3 generation showed higher levels of hybrid zones in the structure of individuals, a large part of which, however, is shared with the recurrent parent. These results agree with the estimates of diversity and the record of the alleles of the recurrent parent in the progenies.

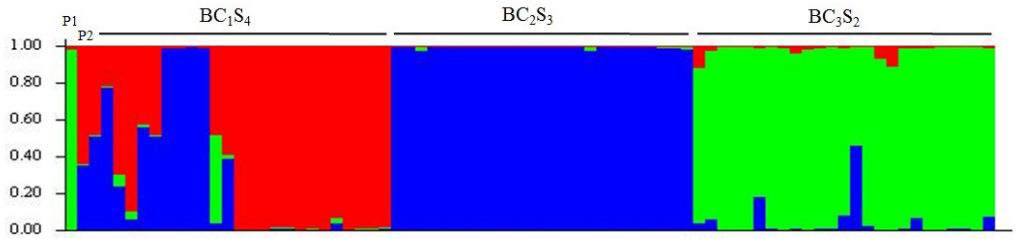

Figure 3. Analysis of the genetic structuring of the 75 genotypes of papaya from the BC_1S_4, BC_2S_3, and BC_3S_4 generations, plus the 2 parental genotypes [Cariflora (P1) and SS783 (P2)].

Genetic distance among backcross generation

Cluster analysis used the hierarchical neighbor-joining method (Figure 4), considering the genetic distance value of 0.5, and revealed the formation of 3 main groups. Group I contained the lowest number of individuals and was formed by individuals from BC_1 and the donor parent ('SS783'); group II contained the largest number of genotypes, clustering all progenies derived from BC_2 and some representatives of BC_1; group III included BC_3 individuals and the recurrent parent (Cariflora). Twenty-five individuals from the BC_1 generation were divided into 2 groups. One included genotypes from a single plant of the BC_1S_3 generation (previous generation), clustering together with BC_2 genotypes in group II; the second group included genotypes with higher variation in relation to pedigree. This subdivision among BC individuals supports the analysis of some parameters discussed above, showing similar levels of diversity in BC_1 and BC_3. Regarding the parents used in this breeding program, the clustering occurred coherently because each backcross generation was expected to increase the genomic proportion of the recurrent parent in the progeny.

Analysis of the genetic relations among the genotypes evaluated was also verified by graphic dispersion via principal coordinate analysis (Figure 5). The 2 first coordinates explained 72.1% of the total variation in the data; 44.45% of this variation was explained by coordinate 1, while 27.65% was explained by coordinate 2. These values are high, indicating good reliability of the dispersion presented. Clear separation was verified for individuals from each backcross generation, except some BC_1 genotypes that clustered close to the BC_2 genotypes. Regarding variation within each generation, the lowest variation was among BC_2 genotypes and the highest was among BC_3 genotypes. These results agree with the cluster analysis generated using the neighbor-joining method, thus validating the analyses conducted in this study.

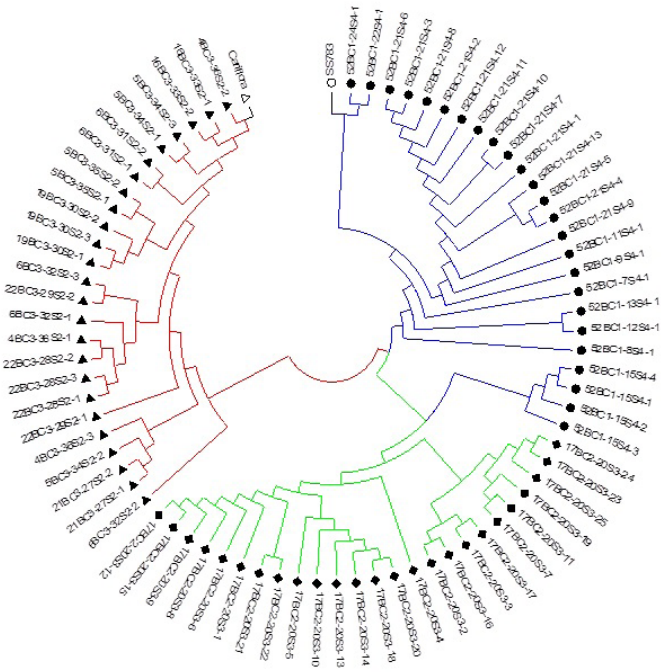

Figure 4. Dissimilarity dendrogram achieved by the neighbor-joining method illustrating the genetic relationship among 75 genotypes belonging to 3 backcross generations (BC_1, blue; BC_2, green; BC_3, red) and the parental genotypes Cariflora and SS783, (Cophenetic correlation coefficient = 0.91).

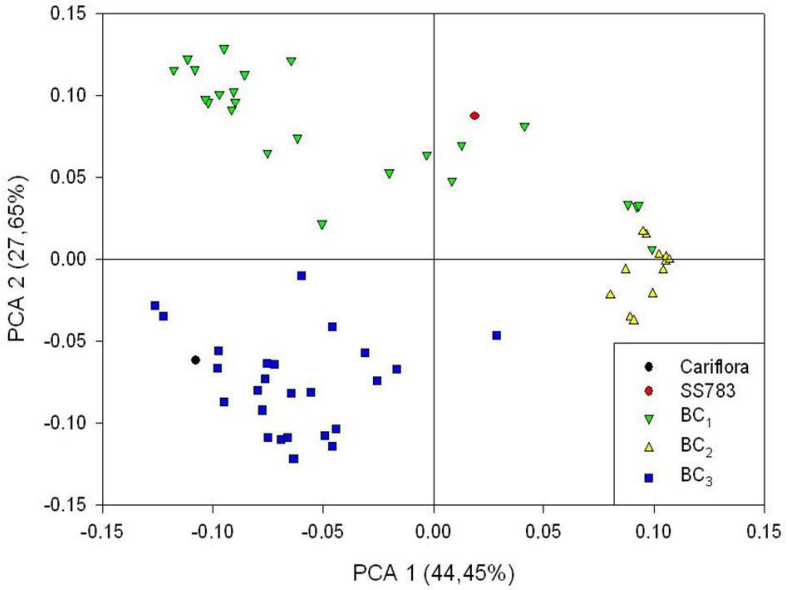

Figure 5. Principal coordinate analysis, considering 75 genotypes of papaya derived from backcrossing, the recurrent parent (Cariflora) and the donor parent, based on the distance matrix achieved by microsatellite marker analysis.

Previous studies comparing the strategies of marker-assisted backcrossing and conventional procedures have been conducted for some species (Davies et al., 2006; Oliveira et al., 2008). The results showed a statistically significant difference between the 2 strategies and an increase of 14.5% in the recovery efficiency of the recurrent parent genome when microsatellite marker-assisted backcrossing was used to monitor selection. This has been corroborated in genetic studies for papaya crop (Ramos et al., 2011b), confirming that using microsatellite markers as an auxiliary strategy for phenotypic analysis is advantageous and may be more efficient in the backcrossing process.

The results achieved in the present study revealed that neither the inbreeding coefficient nor the genomic proportions necessarily follow the expected average values, demonstrating the importance of molecular analysis in determining the respective values of different genotypes given the relevance of such information in selective processes in breeding programs. In contrast, analysis using clustering methods distinguished the progenies derived from the 3 backcross generations and their genetic proximity to the parental genotypes, corroborating the high discriminatory power of microsatellite markers.

Genomic constitution analysis of the individuals indicated that the objective of the sexual conversion for the dioecious genotype Cariflora has been reached, given the high values of the parent recurrent proportion found in some genotypes of the BC_3 generation. Therefore, we expect to achieve inbred lines of Cariflora that are converted for sex and are agronomically superior in few self-fertilization generations, which are assessed by analyzing their potential *per se* and the combining ability via a testcross to verify their genetic potential in crossing. These crosses may be used as varieties or as parents in controlled crossings.

Because of the intention to promote the development of new inbred lines, 1 or 2 future generations of self-fertilizing are needed to achieve the desirable inbreeding coefficient indexes in the progenies. In contrast, the moderate variability present in the BC_1 and BC_3 generations are satisfactory for the continuity of the breeding program, allowing the achievement of genetic gains with the selection of converted and agronomically superior genotypes. This supports the efficient monitoring of selection by molecular markers, mainly microsatellites, contributing to the reduction in time and financial resources necessary for developing new genetic materials.

ACKNOWLEDGMENTS

We thank the Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) for the PhD scholarship and the Caliman Agrícola Company (CALIMAN) for financial and logistic support.

REFERENCES

Allard RW (1971). Princípios do Melhoramento Genético de Plantas. 2nd edn. Edgar Blucher, São Paulo.

Benchimol LL, de Souza CL and de Souza AP (2005). Microsatellite-assisted backcross selection in maize. *Genet. Mol. Biol*. 28: 789-797.

Chan Y-K (2009). Breeding Papaya (*Carica papaya* L.). In: Breeding Plantation Tree Crops: Tropical Species (Jain SM and Priyadarshan PM, eds.). Malaysia.

Chen C, Yu Q, Hou S, Li Y, et al. (2007). Construction of a sequence-tagged high-density genetic map of papaya for comparative structural and evolutionary genomics in brassicales. *Genetics* 177: 2481-2491.

Cruz CD (2008). Programa GENES: Diversidade Genética. Editora Universidade Federal de Viçosa, Viçosa.

Davies J, Berzonsky WA and Leach GD (2006). A comparison of marker-assisted and phenotypic selection for high grain protein content in spring wheat. *Euphytica* 152: 117-134.

Eustice M, Yu Q, Lai CW, Hou S, et al. (2008). Development and application of microsatellite markers for genomic analysis of papaya. *Tree Genet. Genomes* 4: 333-341.

Frisch M, Bohn M and Melchinger AE (1999). Comparison of selection strategies for marker-assisted backcross of a gene. *Crop Sci.* 39: 1295-1301.

Fukunaga K, Hill J, Vigouroux Y, Matsuoka Y, et al. (2005). Genetic diversity and population structure of teosinte. *Genetics* 169: 2241-2254.

Guichoux E, Lagache L, Wagner S, Chaumeil P, et al. (2011). Current trends in microsatellite genotyping. *Mol. Ecol. Resour.* 11: 591-611.

Hofmeyr JDJ (1967). Some genetic breeding aspects of *Carica papaya* L. *Agron. Trop.* 17: 345-351.

Hospital F, Chevalet C and Mulsant P (1992). Using markers in gene introgression breeding programs. *Genetics* 132: 1199-1210.

Kim MS, Moore PH, Zee F, Fitch MM, et al. (2002). Genetic diversity of *Carica papaya* as revealed by AFLP markers. *Genome* 45: 503-512.

Liu K and Muse SV (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21: 2128-2129.

Liu Z, Moore PH, Ma H, Ackerman CM, et al. (2004). A primitive Y chromosome in papaya marks incipient sex chromosome evolution. *Nature* 427: 348-352.

Ma H, Moore PH, Liu Z, Kim MS, et al. (2004). High-density linkage mapping revealed suppression of recombination at the sex determination locus in papaya. *Genetics* 166: 419-436.

Marin SLM, Pereira MG, Amaral Júnior AT, Martelleto LAP, et al. (2006). Heterosis in papaya hybrids from partial diallel of 'Solo' and 'Formosa' parents. *Crop Breed. Appl. Biotechnol.* 6: 24-29.

Mateescu RG, Zhang Z, Tsai K, Phavaphutanon J, et al. (2005). Analysis of allele fidelity, polymorphic information content, and density of microsatellites in a genome-wide screening for hip dysplasia in a crossbreed pedigree. *J. Hered.* 96: 847-853.

Ming R, Yu Q and Moore PH (2007). Sex determination in papaya. *Semin. Cell Dev. Biol.* 18: 401-408.

Ming R, Yu Q, Bias A, Chen C, et al. (2008). Genomics of Papaya, A Common Source of Vitamins in the Tropics. In: Genomics of Tropical Crop Plants (Moore PH and Nilno R, eds.). Springer, New York.

Oliveira EJ, Silva AS, Carvalho AM, Santos LF, et al. (2010). Polymorphic microsatellite marker set for *Carica papaya* L. and its use in molecular-assisted selection. *Euphytica* 173: 279-287.

Oliveira LK, Melo LC, Brondani C, Peloso MJ, et al. (2008). Backcross assisted by microsatellite markers in common bean. *Genet. Mol. Res.* 7: 1000-1010.

Peakall R and Smouse P (2009). GenAlEx Tutorials-Part 1: Introduction to Population Genetic Analysis. Australian National University, Australia.

Pervaiz ZH, Rabbani MA, Khaliq I, Pearce SR, et al. (2010). Genetic diversity associated with agronomic traits using microsatellite markers in Pakistani rice landraces. *Electron. J. Biotechnol.* 13: 1-12.

Pritchard JK, Stephens M and Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.

Ramos HCC, Pereira MG, Silva FF and Viana AP (2011a). Seasonal and genetic influences on sexual expression in segregating papaya population derived from backcross. *Crop Breed. Appl. Biotechnol.* 11: 97-105.

Ramos HC, Pereira MG, Silva FF, Goncalves LS, et al. (2011b). Genetic characterization of papaya plants (*Carica papaya* L.) derived from the first backcross generation. *Genet. Mol. Res.* 10: 393-403.

Ramos HCC, Pereira MG, Gonçalves LAG, Pinto FO, et al. (2011c). Comparison of multiallelic distances on the genetic diversity quantification in papaya. *Acta Sci.* 33: 59-66.

Saitou N and Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406-425.

Servin B and Hospital F (2002). Optimal positioning of markers to control genetic background in marker-assisted backcrossing. *J. Hered.* 93: 214-217.

Silva FF, Pereira MG, Campos WF, Damasceno Júnior PC, et al. (2007). DNA marker-assisted sex conversion in elite papaya genotype (*Carica papaya* L.). *Crop Breed. Appl. Biotechnol.* 7: 52-58.

Silva FF, Pereira MG, Ramos HCC, Damasceno Júnior PC, et al. (2008). Selection and estimation of the genetic gain in segregating generations of papaya (*Carica papaya* L.). *Crop Breed. Appl. Biotechnol.* 8: 1-8.

Storey WB (1953). Genetics of the papaya. *J. Hered.* 44: 70-78.

Tamura K, Peterson D, Peterson N, Stecher G, et al. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28: 2731-2739.

Visscher PM, Haley CS and Thompson R (1996). Marker-assisted introgression in backcross breeding programs. *Genetics* 144: 1923-1932.

Yeh FC, Boyle T, Rongcai Y, Ye Z, et al (1999). POPGENE. Microsoft Window-Based Freeware for Population Genetic Analysis. Version 1.31. Manual.

Young ND and Tanksley SD (1989). RFLP analysis of the size of chromosomal segments retained around the *tm-2* locus of tomato during backcross breeding. *Theor. Appl. Genet.* 77: 353-359.

Yu Q, Hou S, Hobza R, Feltus FA, et al. (2007). Chromosomal location and gene paucity of the male specific region on papaya Y chromosome. *Mol. Genet. Genomics* 278: 177-185.

Yu Q, Hou S, Feltus FA, Jones MR, et al. (2008). Low X/Y divergence in four pairs of papaya sex-linked genes. *Plant J.* 53: 124-132.

