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

Multivariate analysis to determine the genetic distance among backcross papaya (*Carica papaya*) progenies

H.C.C. Ramos[1], M.G. Pereira[1], L.S.A. Gonçalves[1], A.P.C.G. Berilli[2], F.O. Pinto[1] and E.H. Ribeiro[1]

[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]Instituto Federal de Educação, Ciência e Tecnologia de Colatina,
Colatina, ES, Brasil

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

Genet. Mol. Res. 11 (2): 1280-1295 (2012)
Received July 15, 2011
Accepted January 24, 2012
Published May 14, 2012
DOI http://dx.doi.org/10.4238/2012.May.14.2

ABSTRACT. Morpho-agronomic and molecular (RAPD and ISSR markers) data were used to evaluate genetic distances between papaya backcross progenies in order to help identify agronomically superior genotypes. Thirty-two papaya progenies were evaluated based on 15 morpho-agronomic characteristics, 20 ISSR and 19 RAPD primers. Manhattan, Jaccard and Gower distances were used to estimate differences based on continuous and binary data and combined analyses, respectively. Except for production, there were significant differences in the continuous variables among the genotypes. The molecular analysis revealed 193 dominant markers (ISSR and RAPD), being 53 polymorphic loci. Among the various clusters that were generated, the one based on a combined analysis of morpho-agronomic and molecular data gave the highest cophenetic correlation (0.72) compared to individual analysis, consistently allocating the progenies into six groups. We found that the Gower algorithm was more coherent in the discrimination of the genotypes, demonstrating that a combination of molecular and agronomic data is valuable for studies of genetic dissimilarity in papaya.

Key words: *Carica papaya*; Genetic diversity; Molecular markers; Gower algorithm

INTRODUCTION

The papaya (*Carica papaya* L.) belongs to the small family Caricaceae, which includes 35 species placed in six genera. Among all species, 32 are dioecious, two trioecious and one monoecious (Ming et al., 2007). The papaya is the only species of the genus *Carica*, also being the best known and most economically important within the family (Van Droogenbroeck et al., 2002), showing widespread cultivation in tropical and subtropical regions around the world. Its germplasm has considerable phenotypic variation for many characteristics of horticultural importance, including size and fruit shape, flesh color, flavor and soluble solids content, length of juvenile period, plant height, etc. (Kim et al., 2002). However, when referring to commercial cultivars, there is limited genetic variability, with the commercial plantations being established in Brazil basically of three main varieties (Sunrise Solo, Golden and the hybrid Tainung No. 1), thus evidencing the need for development of new genotypes.

Quantifying the level of genetic dissimilarity between genera, species, subspecies, populations and improved elite materials is essential for the genetics of populations (Reif et al., 2005), as for the success of breeding programs aimed at developing new cultivars (Marić et al., 2004). Besides providing a better understanding of the organization of germplasm and improving the efficiency in the sampling of genotypes, knowledge of the genetic distance also allows the selection of biologically oriented crossings (Vieira et al., 2007; Bertan et al., 2009), resulting in obtaining the segregating progenies with high genetic variability for selection (Marić et al., 2004). This genetic variability available in segregating populations is essential for breeding programs, because besides increasing the chance of identifying superior genotypes it is directly related to the genetic gain obtained by artificial selection. The results of diversity analysis can also be used to recommend new cultivars when the goal is to increase the genetic base of commercial cultivars (Vieira et al., 2007).

The study of genetic diversity has been defined as the process by which the variation between individuals, groups of individuals or populations is performed by a particular method or a combination of methods, from different data groups (Mohammadi and Prasanna, 2003). In this context, morphological and molecular analyses are among the tools and information most used to estimate the diversity (Marić et al., 2004), contributing substantially in the different stages of the breeding programs, allowing the determination of singularities and differences in the genetic and phenotypic constitution of genotypes (Franco et al., 2001).

Due to the fact that phenotypic expression is influenced by external factors such as environmental conditions, plant age, etc., genetic studies from morpho-agronomic characteristics have been considered of low accuracy (Vieira et al., 2007). The studies based on molecular methods have already become increasingly common, testifying to the great advances in breeding programs. These have been used as an additional tool in genetic studies because they have the advantage of having lower environmental effects and give direct information from the genome of each individual (Lefebvre et al., 2001; Marić et al., 2004). However, the analysis of these two categories of data separately can result in fragmented and often inaccurate inferences, making it difficult to understand the genetic relationships among the studied germplasm.

Some studies on papaya have reported intra- and inter-generic analysis using morphological data, isozyme, RFLP (restriction fragment length polymorphism), PCR-based molecular markers like AFLP (amplified fragment length polymorphism), RAPD (random amplified polymorphic DNA), ISSR (inter-simple sequence repeat), DAMD (directed amplification of minisatellite DNA), and microsatellites or SSR (simple sequence repeats) (Jobin-Decor et al., 1997; Parasnis et al., 1997; Van Droogenbroeck et al., 2002; Vitória et al., 2004; Kyndt et al., 2005; Saxena et al., 2005; Ocampo et al., 2006; Silva et al., 2007; Eustice et al., 2008; Oliveira et al., 2010). However, studies that use morpho-agronomic and molecular data together to assess the genetic material used in breeding programs are scarce.

In order to allow for inferences on the genetic variability of populations several methodologies for genetic data analysis have been proposed over the years. One methodology suggested by Gower in 1971 has been widely used recently. It enables the calculation of the distance between two observations, considering simultaneously the measures of categorical and continuous variables (Crossa and Franco, 2004). The Gower algorithm provides a semi-positive definite matrix with values between 0 and 1, making it necessary to standardize the variables used.

Thus, the purpose of this study was: i) to estimate the genetic distance between genotypes derived from backcrossing, using morpho-agronomic and molecular data and combined analysis, ii) to analyze the efficiency of such methodologies to access the genetic diversity and distinguish coherently the progenies evaluated in this study, and iii) to select superior genotypes for the progress of generation by self-fecundation.

MATERIAL AND METHODS

Plant material

Hermaphrodite plants were evaluated for 26 families derived from backcrossing (BC). Of these, sixteen are progenies of the first backcross generation (BC_1), one from the second (BC_2) and nine from the third (BC_3), as shown in Table 1. The different number of progenies per generation is due to the availability of families in each generation, which is the result of the selection pressure made in previous cycles. Four controls were also included in this study [Golden, SS783 (donor parent), SS72/12, UENF/Caliman01] and two progenies from the test cross (BC_3(2)XSS72/12 and BC_3(3)XSS72/12).

The segregating evaluated progenies were derived from the cross between the genotype Cariflora dioecious (recurrent parent) and the cultivar Sunrise Solo 783 (donor parent) in the backcross program. This program aims to transfer to the Cariflora genotype the genomic region that determines the expression of hermaphroditism in papaya, considering that this genotype has good combining ability (general and specific) when crossed with genotypes of the “Solo” group (Marin et al., 2006). The achievement of the Cariflora genotype converted to sex, that is, hermaphrodite, will allow promising lines to be obtained and hence the development of stable hybrids.

Table 1. Genetic materials evaluated in this study.

Progeny	Identification	Progeny origin
10	$BC_1XSS72/12\text{-}6IS_2$	1st generation of backcross, 2nd self-fecundation
2	$52BC_1\text{-}2\text{-}2S_3$	1st generation of backcross, 3rd self-fecundation
3	$52BC_1\text{-}36\text{-}16S_3$	1st generation of backcross, 3rd self-fecundation
4	$52BC_1\text{-}34\text{-}5S_3$	1st generation of backcross, 3rd self-fecundation
5	$52BC_1\text{-}34\text{-}10S_3$	1st generation of backcross, 3rd self-fecundation
6	$52BC_1\text{-}36\text{-}9S_3$	1st generation of backcross, 3rd self-fecundation
7	$52BC_1\text{-}34\text{-}9S_3$	1st generation of backcross, 3rd self-fecundation
8	$52BC_1\text{-}36\text{-}4S_3$	1st generation of backcross, 3rd self-fecundation
12	$BC_1XSS72/12\text{-}7IS_2$	1st generation of backcross, 2nd self-fecundation
9	$52BC_1\text{-}27\text{-}5S_3$	1st generation of backcross, 3rd self-fecundation
11	$BC_1XSS72/12\text{-}6IIS_2$	1st generation of backcross, 2nd self-fecundation
13	$BC_1XSS72/12\text{-}7IIS_2$	1st generation of backcross, 2nd self-fecundation
15	$BC_1XSS72/12\text{-}10S_2$	1st generation of backcross, 2nd self-fecundation
14	$BC_1XSS72/12\text{-}8S_2$	1st generation of backcross, 2nd self-fecundation
16	$17BC_2\text{-}7S_2$	1st generation of backcross, 2nd self-fecundation
1	$16BC_1\text{-}37\text{-}6S_3$	1st generation of backcross, 3rd self-fecundation
17	Segregating-S_1	Segregating progeny of unknown origin
18	$20BC_3\text{-}S_1$	3rd generation of backcross, 1st self-fecundation
19	$21BC_3\text{-}S_1$	3rd generation of backcross, 1st self-fecundation
20	$22BC_3\text{-}S_1$	3rd generation of backcross, 1st self-fecundation
21	$19BC_3\text{-}S_1$	3rd generation of backcross, 1st self-fecundation
22	$6BC_3\text{-}S_1$	3rd generation of backcross, 1st self-fecundation
23	$16BC_3\text{-}S_1$	3rd generation of backcross, 1st self-fecundation
24	$5IBC_3\text{-}S_1$	3rd generation of backcross, 1st self-fecundation
25	$5IIBC_3\text{-}S_1$	3rd generation of backcross, 1st self-fecundation
26	$4BC_3\text{-}S_1$	3rd generation of backcross, 1st self-fecundation

I = plant selected in replication 1 of the previous generation; II = plant selected in replication 2 of the previous generation.

The genotypes were evaluated based on morpho-agronomic characteristics (continuous variables) and DNA markers (binary variables).

Location and experimental design

Evaluations of morpho-agronomic variables were performed in a trial started in February 2008 in the commercial company Caliman Agrícola S/A (Fazenda Romana) located in Linhares (19°23'28" south latitude and 40°04'20" west longitude, and 33 m asl) in the State of Espírito Santo. The experimental design consisted of randomized complete blocks with 32 treatments, two replications and 15 plants per plot, spaced 3.60 m between row and 1.80 m between plants in the row. From the total of 960 plants of the experiment, only 508 genotypes were evaluated because of the rate of segregation between female and hermaphrodite plants of 2:1 generated by self-fecundation, as well as the loss of plants to disease.

The fertilization, the management, the control of pests and diseases, and the cultivation practices followed the same ones adopted in the company commercial plantations.

Evaluated characters

A total of fifteen morpho-agronomic characteristics were measured and used to analyze the progenies in this study. The evaluations were made according to Silva et al. (2008), with three additional characters, as: i) the number of deformed fruits - NDFr: determined by counting carpelloid and pentandric fruits in hermaphrodite plants, ii) number of nodes without fruit - NNWFr: determined by counting the nodes with no fruit growth (due to flower abortion or sex reversal) and iii) thickness of the pulp - TP (cm): determined by the average of two measures (thicker and smaller) after the cross section cut of the fruit.

Genomic DNA isolation

Molecular analysis using ISSR and RAPD markers was conducted at the Laboratory of Plant Breeding (LMGV) of North Fluminense State University (UENF). The genetic materials submitted to molecular analysis were collected in bulk in order to achieve greater representation of the allelic families evaluated. The bulk was composed of plant tissue samples of 10 plants per progeny. The isolation of the genomic DNA from young leaves was performed following the CTAB method (Doyle and Doyle, 1990) with some modifications suggested by Daher et al. (2002). After the isolation, the DNA was quantified by analysis on 0.8% agarose gel, and diluted to a working concentration of 10 ng/μL using the High DNA Mass Ladder marker (Invitrogen, USA). The gel was stained with a mixture of GelRed^{TM} and Blue Juice (1:1) and the image captured by photo-documentation MiniBis Pro (Bio-Imaging Systems).

ISSR and RAPD molecular markers

For the molecular characterization, 19 RAPD primers and 20 ISSR primers were used (Table 2). The analyses with the RAPD marker were performed as described by Williams et al. (1990). The PCR conditions for the ISSR marker analysis were as follows: 2 μL 10X buffer (500 mM KCl, 100 mM Tris-HCl, pH 8.4, 1% Triton X-100), 0.5 μM primer (100 μM), 2 mM $MgCl_2$, 100 μM of each dNTP, 0.6 U Taq DNA polymerase; 1 μL (5%) DMSO and 2 μL genomic DNA (5 ng/μL), completing with water to a final volume of 20 μL. The amplification reactions were performed in a Mastercycler Eppendorf 5331 gradient thermal cycler, according to the following program: 94°C for 4 min followed by 40 cycles at 94°C for 1 min, 46°-50°C (temperature ranged according to the primer) for 2 min, 72°C for 2 min, and a final extension at 72°C for 7 min. The amplification products (from the RAPD and the ISSR) were separated on 2% agarose gel, stained with a mixture of GelRed^{TM} and Blue Juice (1:1), and revealed through a system of photo-documentation MiniBis Pro (Bio-Imaging Systems). The DNA Ladder marker (Invitrogen, USA) of 250 bp was used during the electrophoresis runs to determine the size of the amplified fragments.

Data analysis

Initially, the normality condition of the 15 characteristics was checked by the Shapiro-Wilk test for a more accurate interpretation of the data. The variables, number of deformed fruits and commercial fruit, were subjected to data transformation, following the expression $\sqrt{X} + 1/2$. Afterward, the morpho-agronomic characteristics were subjected to analysis of variance (ANOVA) to verify the existence of statistically significant genetic variability among genotypes. For that, the source of variation ‘genotype’ was considered as a fixed effect, in other words, following the experimental design of type.

Next, the averages were compared using the least significant difference (LSD) test at 5% probability. The quantitative data were also subjected to analysis of genetic diversity, using as the method of dissimilarity the Manhattan distance, which was obtained by the distance matrix.

For the analysis of the molecular data, the amplification products obtained by ISSR and RAPD primers were transformed into a matrix of binary data (assigning 0 for the absence and 1 for the presence of the band), and the dissimilarity matrix was obtained considering the arithmetic complement of the Jaccard index.

The analysis of the genetic dissimilarity was also performed considering simultaneously the molecular and morpho-agronomic data. For that, the estimation of the genetic distance matrix was obtained based on the similarity index proposed by Gower (1971). This index ranges from 0 to 1, calculated by:

$$S_{ij} = \frac{\sum_{k=1}^{p} W_{ijk} S_{ijk}}{\sum_{k=1}^{p} W_{ijk}}, \qquad \text{(Equation 1)}$$

where k is the number of variables (k = 1, 2, ..., p); i and j any two individuals; W_{ijk} is a weight given for ijk comparison, assigning 1 to valid comparisons and 0 for invalid comparisons (if the value of the variable is absent in one or both individuals); S_{ijk} is the contribution of variable k to the similarity between individuals i and j, with values between 0 and 1. This analysis allows the removal of differences between the ranges of variables, producing a value within the interval [0, 1] and equal weights.

Multivariate analyses were implemented using the hierarchical clustering techniques, based on the UPGMA (unweighted pair-group method using an arithmetic average) and the neighbor joining method (Saitou and Nei, 1987). The adjustment between the distance matrix and the cluster matrix was estimated by the cophenetic correlation coefficient (CCC) (Sokal and Rolf, 1962) and to estimate the correlation significance between the matrices the Mantel test (Mantel, 1967) was used with 1000 permutations.

The statistical programs used in the analysis were GENES (Cruz, 2008), ANOVA, the mean comparison test, and the analysis of correlation significance. The program R (R Development Core Team, 2006) was used for the genetic distance analysis (morpho-agronomic and molecular data) and to estimate the cophenetic correlation coefficient. The hierarchical clustering was performed with the aid of the MEGA version 5 software (Kumar et al., 2009) and graphic dispersion by the method of principal coordinates analysis (PCA) using the Genalex 6.3 program (Peakall and Smouse, 2009).

RESULTS AND DISCUSSION

Except for production, the other variables showed statistical significance at 5% probability by the F test. Although no significant difference was detected for the production, examining the maximum and minimum values (Table 2) it appears that some families had an average of up to four times the lower limit for this characteristic. Thus, the absence of significance may be related to high environmental influence on this trait, which results in high experimental error, making it difficult to differentiate between genotypes. This can be evidenced by analyzing the experimental coefficient of variation (CV_e) in which the variable production showed the highest value among all traits.

Table 2. Mean comparison test and the summary of the variance analysis for 15 characteristics in 32 progenies of papaya.

Genotypes	PH	HIFF	SD	NTFr	NDFr	NNWFr	NCoFr	AFrW	PROD	SS	FF	PF	DIAM	LENG	TP
$52BC_1$-27-$5S_3$	167.9	74.2	31.7	30.7	1.4	20.0	5.4	556.2	17442.7	10.8	94.6	72.0	9.1	14.9	2.5
$52BC_1$-2-$2S_3$	170.0	65.8	34.9	27.7	1.6	18.8	4.9	819.7	20796.4	10.5	111.2	62.6	10.7	16.5	2.6
$52BC_1$-36-$16S_3$	176.7	83.5	33.1	40.2	2.9	12.2	5.7	584.7	20146.6	10.4	112.2	82.8	9.8	13.3	2.4
$52BC_1$-34-$5S_3$	185.8	93.8	30.2	40.8	1.4	9.9	6.3	464.8	20742.9	11.1	107.3	83.8	9.1	14.1	2.0
$52BC_1$-34-$10S_3$	187.4	85.9	33.2	36.9	1.3	18.2	6.0	571.1	22679.4	11.2	114.5	79.2	9.2	16.0	2.1
$52BC_1$-36-$9S_3$	208.1	103.1	34.2	56.2	3.0	10.7	6.9	637.5	28454.4	10.3	106.9	77.6	9.9	14.9	2.4
$52BC_1$-34-$9S_3$	161.4	80.3	29.5	18.5	2.0	22.6	3.9	892.9	12648.3	10.7	113.5	80.1	10.7	16.9	2.7
$52BC_1$-36-$4S_3$	194.6	87.5	35.2	31.5	1.6	17.3	5.5	723.3	20034.3	10.3	105.6	75.6	9.6	15.9	2.3
$16BC_1$-37-$6S_3$	213.8	105.7	35.5	42.1	2.4	14.3	5.9	745.8	26020.6	10.5	94.1	71.9	10.4	15.6	2.3
BC_1XSS72/12-$7IS_2$	231.1	102.2	40.7	31.9	1.2	23.7	5.6	530.6	20622.4	12.4	125.4	72.4	8.7	16.7	2.1
BC_1XSS72/12-$6S_2$	239.7	116.2	43.8	47.0	0.9	17.4	6.9	413.9	18507.8	13.5	111.7	73.3	8.4	13.9	2.0
BC_1XSS72/12-$6IIS_2$	172.1	77.5	34.9	39.9	2.4	11.6	5.9	574.9	21455.1	10.5	91.7	74.3	9.5	14.4	2.3
BC_1XSS72/12-$7IIS_2$	213.0	135.3	36.7	27.7	2.7	11.8	4.5	476.0	10626.2	13.7	113.7	70.1	8.9	13.6	2.0
BC_1XSS72/12-$10S_2$	159.7	70.8	31.0	37.9	2.9	10.1	5.4	824.4	24187.1	9.3	99.6	74.7	10.0	17.2	2.7
BC_1XSS72/12-$8S_2$	188.7	100.6	35.8	28.5	2.2	11.2	4.9	646.5	15467.8	11.7	108.8	72.1	9.5	16.2	2.7
$17BC_2$-$7S_2$	207.5	96.6	38.7	22.9	2.2	23.1	4.4	782.1	15899.3	12.7	121.0	85.3	9.8	17.2	2.6
Segreganting-S_1	169.3	93.7	30.9	23.9	3.6	11.1	3.4	1180.3	21170.1	9.5	111.6	74.1	12.0	18.8	2.8
$20BC_3$-S_1	243.0	100.0	41.9	65.2	1.5	10.9	4.6	432.5	18860.0	13.1	121.8	81.3	9.2	13.2	1.8
$21BC_3$-S_1	216.7	98.3	38.5	25.5	1.2	20.2	4.9	1161.9	31564.6	10.9	111.8	77.7	10.8	19.9	2.9
$22BC_3$-S_1	214.2	97.3	33.2	32.3	1.5	15.2	5.5	732.4	27705.0	11.4	104.3	76.1	10.1	16.6	2.6
$19BC_3$-S_1	199.4	97.1	31.3	21.7	1.7	17.3	4.4	765.9	14486.6	11.2	93.6	71.9	10.8	18.2	2.4
$6BC_3$-S_1	188.6	93.5	34.4	20.3	2.6	19.4	3.7	1185.6	19323.7	11.1	112.2	80.0	11.3	20.5	2.8
$16BC_3$-S_1	195.1	90.2	31.5	19.5	1.7	22.9	4.1	747.1	15101.7	11.3	96.2	79.1	10.6	17.2	2.5
$5IBC_3$-S_1	212.0	108.5	38.7	25.1	2.6	19.0	4.4	719.9	13565.1	12.7	125.2	85.8	9.7	16.9	2.4
$5IIBC_3$-S_1	213.6	106.9	35.5	20.1	2.8	21.7	3.6	866.5	12228.3	11.9	106.4	87.9	10.1	17.3	2.5
$4BC_3$-S_1	190.2	92.1	34.2	18.4	1.5	28.3	4.1	839.4	31815.0	10.7	117.3	73.3	10.5	16.4	2.6
UC 01	222.7	123.9	32.9	15.6	1.6	24.6	3.7	1018.1	13832.3	13.1	111.3	80.7	10.5	20.5	2.5
SS783	234.9	123.2	37.9	26.2	1.2	19.9	5.1	644.6	16040.9	10.6	92.4	70.9	9.4	15.6	2.2
SS72/12	203.2	103.4	39.6	60.0	1.7	4.6	7.6	448.2	26798.1	11.5	108.1	65.7	8.4	13.8	1.9
Golden	225.1	125.4	35.5	37.0	0.9	20.5	6.1	458.4	16746.1	11.2	103.6	67.2	8.5	14.3	2.1
BC3(3)XSS72/12	240.0	125.0	41.0	30.8	1.2	18.1	5.5	669.2	19683.3	12.0	89.9	67.8	10.2	15.3	2.4
BC3(2)XSS72/12	219.6	121.1	34.9	24.8	1.8	15.2	4.7	663.4	15490.6	12.0	102.2	69.8	9.9	15.1	2.2
$LDS_{(5\%)}$	38.4	27.1	7.8	20.2	0.9	7.5	1.7	211.8	13793.0	1.2	18.6	12.7	0.9	2.0	0.2
GMS	1159.4**	558.35**	26.02**	299.60**	0.99**	57.94**	2.13**	90054.3**	60911521.7	2.37**	190.05*	75.25*	1.48**	7.74**	0.15**
RMS	355.06	175.87	14.67	98.11	0.21	13.42	0.72	10783.1	45736614.7	0.36	83.13	38.63	0.21	0.96	0.02
Average	202.04	99.35	35.36	32.11	1.93	16.93	5.11	711.81	19691.97	11.37	107.49	75.55	9.89	16.15	2.39
Maximum	270.0	143.46	48.0	87.0	3.8	31.57	8.56	1318.75	42746.25	13.98	139.02	97.91	12.34	20.95	3.11
Minimum	140.0	40.0	27.5	12.11	0.71	1.9	2.92	393.33	8300.0	9.25	83.13	53.4	8.14	13.1	1.83
CV_e (%)	9.33	13.35	10.01	30.84	23.92	21.63	16.58	14.59	34.34	5.32	8.48	8.23	4.63	6.08	5.17
CV_g (%)	9.92	14.37	6.74	31.26	32.41	27.85	16.42	27.97	13.99	8.81	6.80	5.66	8.08	11.40	10.97
I_v	1.06	1.08	0.62	1.01	1.35	1.29	0.99	1.92	0.41	1.65	0.80	0.69	1.74	1.85	2.12

PH = plant height; HIFF = height of insertion of the first fruit; SD = stem diameter; NTFr = number of total fruits; NDFr = number of deformed fruits; NNWFr = number of nodes without fruit; NCoFr = number of commercial fruits; AFrW = average fruit weight; PROD = plant production; SS = soluble solids content; FF = firmness of the fruit; PF = pulp firmness; DIAM = diameter of the fruit; LENG = fruit length; TP = average thickness of the pulp. **, *Significant at 1 and 5% probability by the *t*-test, respectively.

The experimental CV_e for the evaluated characteristics was relatively low (4 to 23%), except for total and commercial fruits, which were higher than 30% (30.8 and 34.34%, respectively) (Table 2). On the other hand, the genotypic coefficient of variation (CV_g) ranged from 5 to 32%, where the highest values were observed for total and deformed fruits, nodes without fruit and mean fruit weight. This result is interesting for the breeding program because the existing genetic variation in traits, which contribute to the reduction of production (NDFr and NNWFr), might allow the selection of genotypes with low expression of these traits, increasing productivity. The relationship between these two parameters (CV_e and CV_g), that is, the index of variation (I_v) indicates that the situation is highly favorable for the selection of eleven variables (PH, HIFF, NTFr, NDFr, NNWFr, NCoFr, AFrW, SS, DIAM, LENG, and TP), according to Venkovsky (1987). However, for production and pulp and fruit firmness the response to the selection may not be as favorable, indicating limited genetic progress.

According to the mean comparison test determined by the method of the LSD (Table 2) of the 15 evaluated traits in this study, for four of them (SD, NTFr, PROD, and FF) there were no progeny with statistically lower means to the overall mean, and this result is not very interesting for node without fruit (NNWFr), since the goal is to reduce the expression of this trait in the population. On the other hand, only the characteristics of the fruit and pulp firmness progenies did not show statistically higher means to the overall mean, indicating little possibility of genetic progress.

Analyzing the backcross generations separately, it appears that among the progenies derived from the BC_1 generation, there was a significant difference for all continuous variables evaluated. As for the progenies derived from the BC_3 generation, there was no statistical difference only for the first fruit height. Thus, based on the analysis of morpho-agronomic characteristics it is possible to infer that the progenies derived from BC_1 present greater variability than those derived from BC_3, since the soluble solids and node without fruit characteristics were the most divergent within BC_1 and BC_3, respectively. This result is in agreement with the expectations since there is a greater variation in terms of pedigree among the progenies derived from BC_1. On the other hand, only for plant height, deformation of fruit and yield were the progeny derived from BC_2 statistically different from the control.

For the analysis of genetic divergence using agglomerative methods, neither the groups obtained by the hierarchical UPGMA method nor those obtained by the neighbor joining showed good agreement with the genealogy of the progenies based on molecular and quantitative data. However, the neighbor joining method allowed more coherent groups and thus they were considered in this study.

The analysis cluster analysis based on continuous characters, estimated by the Manhattan distance, showed the formation of seven groups, with a cut in the distance of 0.35, this being the average genetic distance (Figure 1A). Based on this analysis, the most distant progenies were the Segregating-S_1 and $BC_1XSS72/12\text{-}6IS_2$ while the closest were $BC_3(2)XSS72/12$ and $20BC_3\text{-}S_1$. The groups VI and VII are characterized by having fruits of medium size, differing mainly in relation to average fruit weight. On the other hand, representatives of the groups I, II, III, IV, and V show small fruit, differing mainly in relation to plant yield and soluble solids. Based on these group descriptions, it can be inferred that the characters related to fruit size (average weight, length and diameter) clearly contributed to the distinction of the progeny allocating them into two main groups. However, plant vigor and low fruit firmness may have contributed to distinguishing the progeny $52BC_1\text{-}2\text{-}2S_3$ from the other progenies of Group VII.

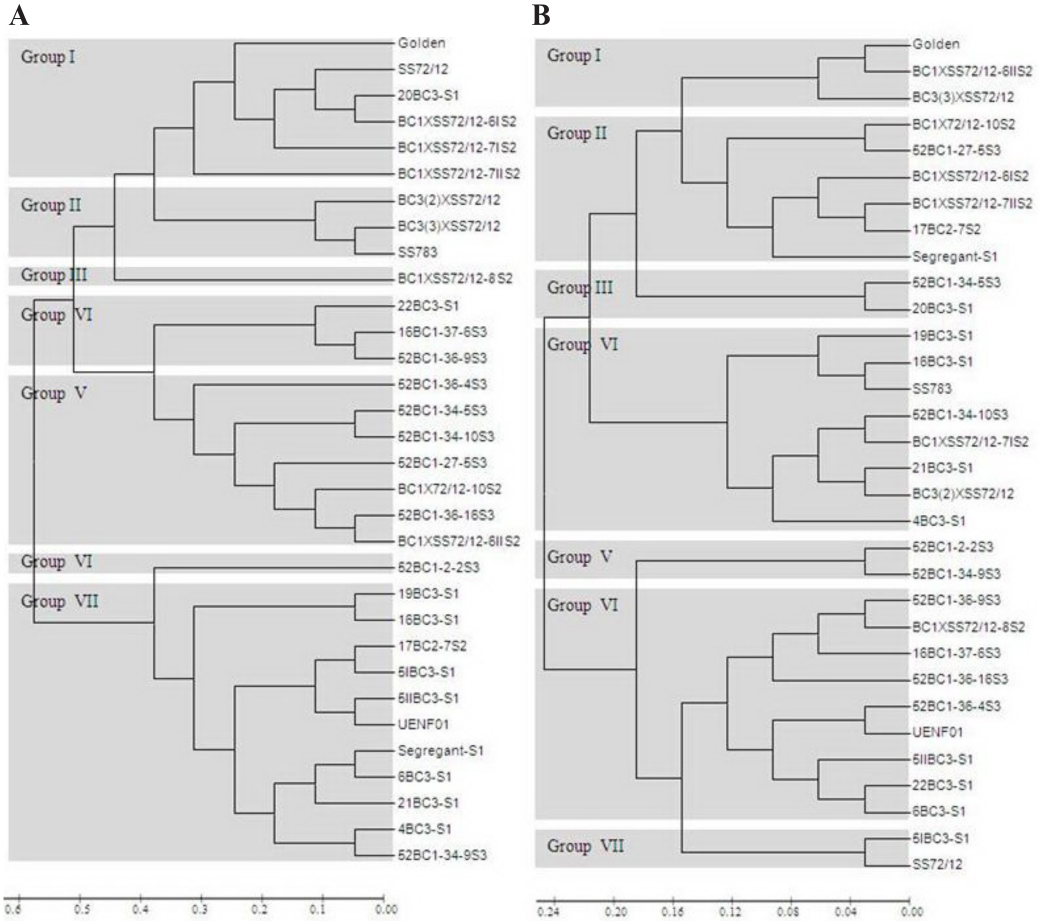

Figure 1. Dendogram obtained by the neighbor joining hierarchical method based on the analysis of 32 papaya families using: **A.** the Manhattan distance for the analysis of 15 quantitative characteristics (cophenetic correlation coefficient = 0.6) and **B.** arithmetic complement of the Jaccard index for the analysis of binary data (cophenetic correlation coefficient = 0.56).

The analysis of the binary variables consisted of the sum of the data obtained by ISSR and RAPD markers (Table 3). In total, 193 markers were analyzed of the dominant type, being 53 (27.5%) polymorphic, generating an average of 4.9 bands/primer, showing low genetic variability among the evaluated progenies. Considering that the progenies derived from BC_1S_3, BC_2S_2 and BC_3S_1 have an average of 90.62, 84.38 and 71.88% endogamy, respectively, a low degree of genetic heterogeneity among the progenies is an expected result. However, the genetic sampling done in bulk in an attempt to obtain a better allelic representation of the family may also have contributed to hiding possible differences between individuals of the same progeny.

According to Eustice et al. (2008), the high level of phenotypic diversity observed among papaya cultivars in the field is not correlated with the low level of polymorphism so far elucidated. Perhaps the fact that the improvement of papaya has been done over the years with a limited number of genotypes may have contributed to this situation. Another reason for the low variability may be related to reproductive barriers resulting from incompatibility between the species *C. papaya* L. and species from other genera of the family, creating a restricted gene pool.

Table 3. Sequences of ISSR and RAPD primers used in the analysis of 32 progenies and their annealing temperatures (Ta), total number of alleles and polymorphic alleles.

Primer No.	Sequence (5'-3')	Ta (°C)	Alleles		Amplified product
			Total	Polymorphic	
	ISSR primers				Min-max
1	AC AC AC AC AC AC AC AC AC	58	3	0	1200-1400
2	GTC GTC GTC GTC GTC GTC	58	5	3	350-700
3	AGC AGC AGC AGC AGCY	52	6	1	500-1100
4	AGC AGC AGC AGC AGCAY	65	7	0	400-2000
5	CA CA CA CA CA CA CA CARG	45	3	0	500-1500
6	AGC AGC AGC AGC AGCGR	65	4	0	600-2500
7	CAGA CAGA CAGA CAGA	56	6	0	400-1000
8	CT CT CT CT CT CT CT CTRC	45	5	1	500-1500
9	CT CT CT CT CT CT CT CTTG	42	2	0	1000-2000
10	AG AG AG AG AG AG AG AGYR	42	7	0	400-2000
11	CTC CTC CTC CTC CTC CTC	50	5	0	400-1000
12	GTC GTC GTC GTC GTC Y	58	4	1	400-800
13	GTG GTG GTG GTG GTGGR	50	6	0	500-1500
14	GA GA GA GA GA GA GA GA GAT	48	3	0	600-1500
15	GA GA GA GA GA GA GA GAYC	48	3	0	400-800
16	CA GA GA GA GA GA GA GA GA	52	7	1	300-1100
17	GC GA GA GA GA GA GA GA GA	56	7	4	300-1000
18	GGGTGGGGTGGGGTG	56	3	1	250-600
19	ATG ATG ATG ATG ATG ATGG	52	8	4	300-1500
20	AG AG AG AG AG AG AG AGYT	51	4	2	450-1000
	RAPD primers				
1	TCTGTGCTGG	36	7	1	450-1100
2	TGCGCCCTTC	36	4	1	500-2000
3	ACGGAAGCCC	36	4	0	550-1700
4	CCTGGGTCAG	36	4	0	450-1600
5	CTGTGTGCTC	36	7	3	400-1400
6	CACGAACCTC	36	2	1	400-550
7	TGAGCGGACA	36	5	2	250-1500
8	GTGTGCCCCA	36	6	1	250-1100
9	ACTGGGACTC	36	6	2	400-2500
10	ACCCGGTCAC	36	5	2	450-1500
11	GTGCAACGTG	36	4	3	400-750
12	TCTGGCGCAC	36	6	2	450-1300
13	GAGACGCACA	36	4	1	350-1400
14	GACCTACCAC	36	3	1	800-2000
15	GTAACCAGCC	36	7	0	350-2000
16	GGTGCACGTT	36	8	7	500-2000
17	CCCGTAGCAC	36	4	3	500-1000
18	CCCGTTGCCT	36	3	2	500-1000
19	GGGCCACTCA	36	7	2	400-1700
39			193	53	

Y = C or T; R = A or G.

The genetic dissimilarity estimated from the binary data revealed that the most dissimilar families were $52BC_1$-2-$2S_3$ and $5IBC_3$-S_1, while the most similar were $BC_3(2)XSS72/12$ and $20BC_3$-S_1. Cluster analysis allocates the 32 families evaluated into seven groups (Figure 1B), with progeny number per group ranging from two to nine. Similar to the analysis with quantitative data, the group generated by molecular data also showed low coherence with the genealogy of the material evaluated.

Despite the cluster analysis based on morpho-agronomic and molecular data having shown the same number of groups, their profiles were considerably different. Based on just the two main groups trained in both analyses, it appears that 59.4% of the progenies were placed in groups in a similar way. By increasing the number of considered groups, the cluster similarity drops significantly. The two analyses also differ as to the coherency among the formed groups and the genealogy of the progenies. It is noted that, although both groups are not 100% consistent with genealogy, that is, not allowing grouping of the progenies according to the generation of backcrossing, the analysis based on the quantitative data was the one closest to the expected.

This result shows that there was a high discrepancy between the clusters generated by the continuous and binary data, which can be confirmed by the low value found for the correlation between the two matrices (r = 0.04). This indicates that the genetic distance by molecular markers was not exactly representative of the genetic distance based on quantitative characters. This is a result that has been found in some studies of diversity involving the combined analysis of continuous and discrete variables (Marić et al., 2004; Roy et al., 2004).

According to Lefebvre et al. (2001), the relationship between molecular and phenotypic distance is closely related to the polygenic inheritance of the phenotypic characteristics used in the analysis, and the magnitude of the correlation coefficient between these two types of data is dependent on the association between the marker locus and the locus that controls quantitative trait loci. According to the authors, this correlation necessarily decreases with the increasing number of genomic regions involved in trait variation. Thus, given the low value of the correlation between the matrices of the continuous and binary data found in this study, it can be inferred that the association between the marker locus and the quantitative characteristics evaluated is almost non-existent or very weak, indicating that different genomic regions might be sampled by different variables. It also indicates the advantage of the combined analysis.

In the combined analysis of the continuous and binary data the most dissimilar genotypes found were BC_3(3)XSS72/12 and Segregat-S_1, while the most similar were 19BC_3-S_1 and 22BC_3-S_1. Groups generated by the neighbor joining method allowed the formation of six groups, considering the average distance (0.38) as the cutoff point (Figure 2). Regarding the morpho-agronomic characteristics, the genotypes of Group I present on average higher production and low firmness, whereas in Group III, the production is high, but presents greater uniformity than Group I. High yield is also found among members of Group VI, which also stand out for their greater weight and fruit size. In Groups II and V, the genotypes are characterized by having greater pulp and peel firmness and higher soluble solids content.

It is verified that the analysis of genetic similarity using the Gower algorithm showed greater coherence (94%) in the allocation of the progenies in the different groups in relation to the grouping considering the continuous and binary data separately, which can be confirmed by the graphical dispersion based on the PCA shown in Figure 3. The cophenetic correlation coefficient (r = 0.72) indicates a good agreement between the graphical layout of the genetic distance matrix and the original result confirmed by the coincidence in the indication of the more similar and dissimilar progenies.

The association between the matrices (Table 4) showed that the correlation between the combined matrix (continuous and binary data) and the matrix of continuous data was 0.51 while the correlation between the combined matrix and the molecular matrix was 0.04, indicating more agreement with the matrix of continuous data. Such agreement can be verified by greater similarity between the grouping based on the combined analysis and based on the

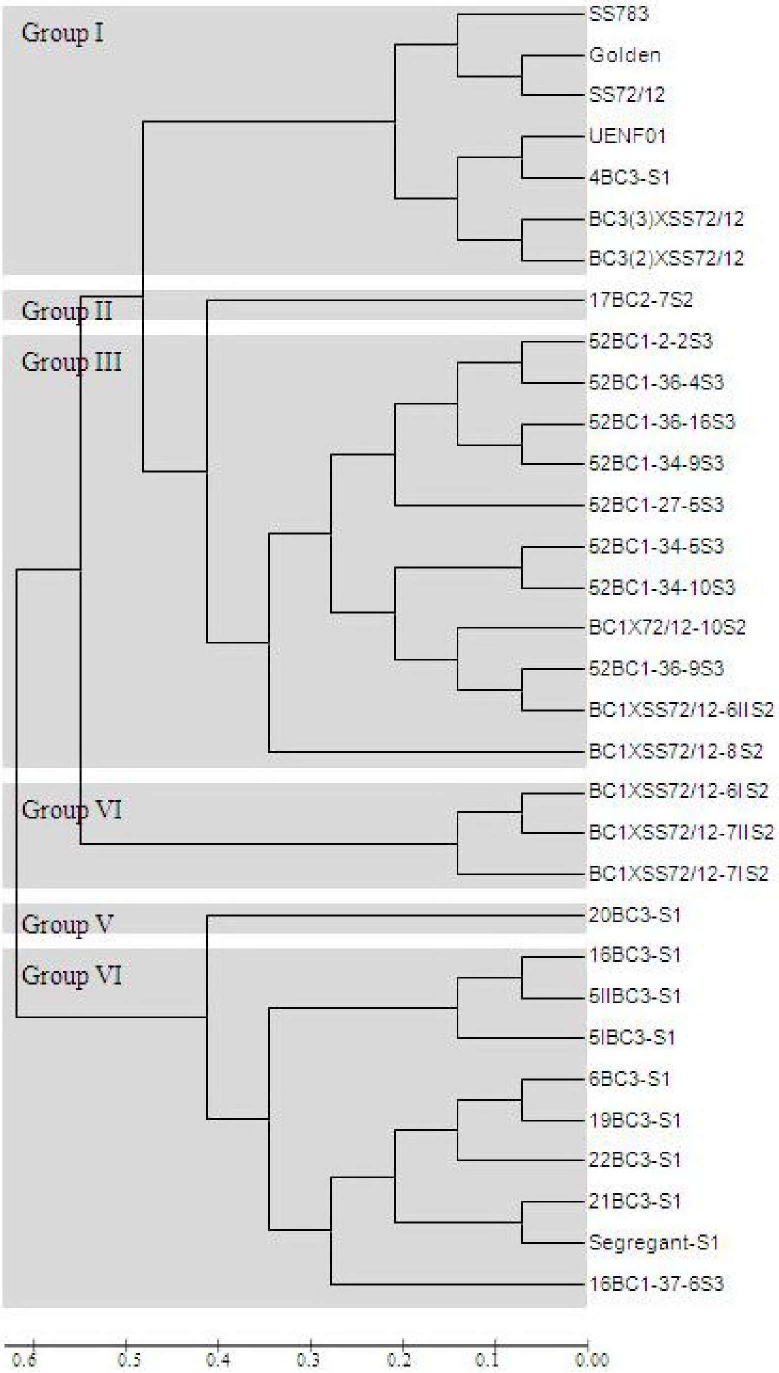

Figure 2. Dendogram obtained by the neighbor joining method based on the analysis of 32 papaya families using the Gower distance for the combined analysis of the continuous and binary data (cophenetic correlation coefficient = 0.72).

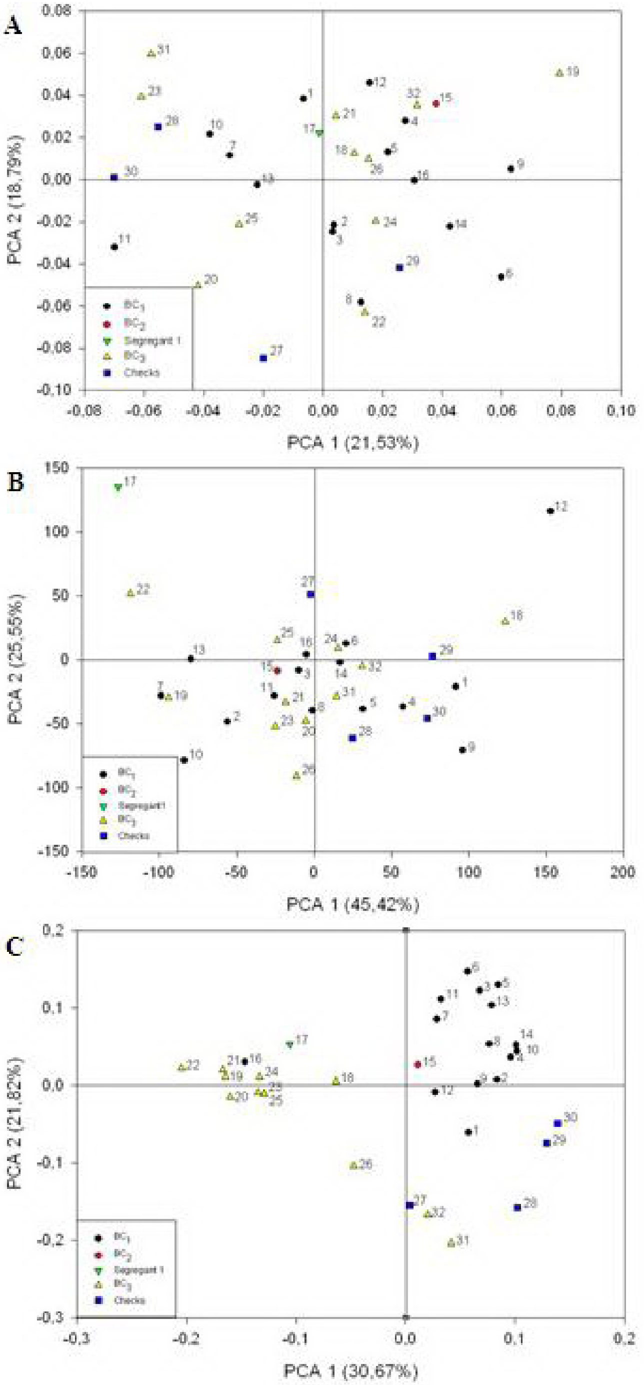

Figure 3. Analysis of principal coordinates, considering 28 papaya progenies and four controls, using morpho-agronomic data (**A**), molecular (**B**) and combined data analysis (**C**).

continuous data. Thus, it is observed from these results that in the set of data the reason the continuous tended to contribute with a high value to the combined analysis may be due to a set of quantitative attributes with extensive genomic representation.

Table 4. Correlation between genetic distance matrixes estimated from the analysis of morpho-agronomic characteristics, molecular markers (ISSR and RAPD) and by the combined analysis of quantitative and molecular data in 32 papaya progenies.

	Quantitative	Molecular	Combined
Quantitative	-	0.04*	0.51**
Molecular		-	0.04
Combined			-

**, * = Significant at 1 and 5% probability by the *t*-test, respectively.

However, the results obtained in this study are different from those found in several papers, disagreeing with the expected mainly due to the large number of molecular markers used in the analysis. Examples can be found in studies conducted with tomato (Gonçalves et al., 2008) and wheat (Marić et al., 2004; Vieira et al., 2007; Bertan et al., 2009), where the highest correlation coefficient was found between combined analysis and molecular data. However, when conducting a comparative analysis between the degree of polymorphism revealed by molecular markers in different studies, it is verified that the values ranged from 68 to 93.2% in the studies mentioned above, while in the present study this level of polymorphism was significantly lower (27.5%), resulting in a low correlation. Analyzed together, these data allow us to infer that the greater association between the matrix based on the combined analysis and the individual matrices of quantitative and molecular data depends on their discriminating power and not their number. In this context, it can be suggested that if a higher polymorphism degree of molecular data had been found in this study, a higher correlation between a binary and combined matrix could have been obtained.

The results achieved have indicated that several cycles of selection and endogamy caused by continuous generations of self-fecundation in this population may already be hindering the use of molecular markers to access genetic variability. However, we found that the combined use of the quantitative and molecular data allows aggregating different information, contributing to an adequate estimation of dissimilarity and a better understanding of the genetic relationships of the materials evaluated, leading to more consistent inferences regarding the genotype differentiation.

Faced with the greater agreement between the distance and cluster matrices, as well as the remarkable consistence in the formation of the groups, the results of the combined analysis were used as the basis for the indication of more divergent and agriculturally superior families to the advance of self-fecundation generation. Several studies have also used the combined analysis of continuous, discrete and binary variables to make inferences about the genetic structure of populations of several crops such as tomatoes (Gill et al., 2008), *Capsicum* (Sudre et al., 2010), cherry tomato (Rocha et al., 2010), banana (Mattos et al., 2010), pepper (Fonseca et al., 2008), etc.

Therefore, according to cluster analysis, the four most divergent progenies derived from BC_1 were $52BC_1$-34-$5S_3$, $16BC_1$-37-$6S_3$, $BC_1XSS72/12$-$10S_2$, and $BC_1XSS72/12$-$7IIS_2$, which differ mainly regarding plant height, height of first fruit, production, soluble solids and fruit size. However, despite being a good source of soluble solids, the $BC_1XSS72/12$-$7IIS_2$

shows the lowest production of all BC_1 progenies, besides being the least precocious, which may be replaced by the progeny BC_1XSS72/12-6S_2, which has good genetic distance in relation to the others, yield, but with high soluble solids content, high production of commercial fruit and low fruit deformation. Among the BC_3 progenies higher average dissimilarity was found in 20BC_3-S_1, 21BC_3-S_1, 6BC_3-S_1, and 4BC_3-S_1, diverging mainly for fruit size, soluble solids and firmness of the fruit, showing a high production, except the progeny 20BC_3-S_1 that shows a slightly below average production, but in the meantime it is a good source of soluble solids and firmness. Although it is the only representative of the BC_2 generation, the progeny 17BC_2-7S_2 emerges as promising by presenting a good average distance in relation to other progenies, as well as having good agronomic traits.

Some different views regarding the use of combined analysis to discriminate genotypes can be noted between the different genetic diversity studies that consider different data sources. However, based on the results obtained in this study, it is verified that the implementation of the Gower algorithm resulted in a coherent analysis for the discrimination of the evaluated progenies, creating new possibilities for studies of genetic dissimilarity. This indicates that the combination of molecular and continuous data stands out as a potential tool to be used in studies of dissimilarity, not only for characterization of germplasm, but also for analysis in advanced generation breeding, allowing for more accurate inferences and contributing to the maintenance of a proper genetic basis for successful improvement programs for several crops.

ACKNOWLEDGMENTS

Research supported by Fundação de Amparo à Pesquisa of the State of Rio de Janeiro (FAPERJ) and the Caliman Agrícola S/A Company (CALIMAN).

REFERENCES

Bertan I, Carvalho FIF, Oliveira AC, Benin G, et al. (2009). Morphological, pedigree, and molecular distances and their association with hybrid wheat performance. *Pesq. Agropec. Bras.* 44: 155-163.

Crossa J and Franco J (2004). Statistical methods for classifying genotypes. *Euphytica* 137: 19-37.

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

Daher RF, Pereira MG, Tupinambá EA, Amaral Júnior AT, et al. (2002). Assessment of coconut tree genetic divergence by compound sample RAPD marker analysis. *Crop Breed. Appl. Biotechnol.* 3: 431-438.

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

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.

Fonseca RM, Lopes R, Barros WS, Lopes MTG, et al. (2008). Morphologic Characterization and genetic diversity of Capsicum chinense Jacq. accessions along the upper Rio Negro - Amazonas. *Crop Breed. Appl. Biotechnol.* 8: 187-194.

Franco F, Crossa J, Ribaut JM, Betran J, et al. (2001). A method for combining molecular markers and phenotypic attributes for classifying plant genotypes. *Theor. Appl. Genet.* 103: 944-952.

Gonçalves LS, Rodrigues R, Amaral AT Jr, Karasawa M, et al. (2008). Comparison of multivariate statistical algorithms to cluster tomato heirloom accessions. *Genet. Mol. Res.* 7: 1289-1297.

Gower JC (1971). A General coefficient of similarity and some of its properties. *Biometrics* 27: 857-871.

Jobin-Decor MP, Graham GC, Henr RJ and Drew RA (1997). RAPD and isozyme analysis of genetic relationships between *Carica papaya* and wild relatives. *Genet. Resour. Crop Evol.* 44: 471-477.

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.

Kumar S, Nei M, Dudley J and Tamura K (2009). MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief. Bioinform.* 9: 299-306.

Kyndt T, Romeijn-Peeters E, Van Droogenbroeck B, Romero-Motochi JP, et al. (2005). Species relationships in the genus Vasconcellea (*Caricaceae*) based on molecular and morphological evidence. *Am. J. Bot.* 92: 1033-1044.

Lefebvre V, Goffinet B, Chauvet JC, Caromel B, et al. (2001). Evaluation of genetic distances between pepper inbred lines for cultivar protection purposes: comparison of AFLP, RAPD and phenotypic data. *Theor. Appl. Genet.* 103: 741-750.

Mantel N (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27: 209-220.

Marić S, Bolarić S, Martinčić J, Pejić I, et al. (2004). Genetic diversity of hexaploid wheat cultivars estimated by RAPD markers, morphological traits and coefficients of parentage. *Plant Breed.* 123: 366-369.

Mattos LA, Amorim EP, Amorim VBO, Cohen KO, et al. (2010). Agronomical and molecular characterization of banana germoplasm. *Pesq. Agropec. Bras.* 45: 146-154.

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

Mohammadi SA and Prasanna BM (2003). Analysis of genetic diversity in crop plants - Salient statistical tools and considerations. *Crop Sci.* 43: 1235-1248.

Ocampo J, D'Eeckenbrugge GC, Bruyère S, Bellaire LL, et al. (2006). Organization of morphological and genetic diversity of Caribbean and Venezuelan papaya germplasma. *Fruit* 61: 25-37.

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

Parasnis AS, Ramakrishna W, Chowdari KV, Gupta VS, et al. (1999). Microsatellite (GATA) n reveals sex-specific differences in papaya. *Theor. Appl. Genet.* 99: 1047-1052.

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

R Development Core Team (2006). A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Reif JC, Melchinger AE and Frisch M (2005). Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Sci.* 45: 1-7.

Rocha MC, Gonçalves LSA, Rodrigues R, Silva PRA, et al. (2010). Uso do algoritmo de Gower na determinação da divergência genética entre acessos de tomateiro do grupo cereja. *Acta Sci.* 32: 423-431.

Roy JK, Lakshmikumaran MS, Balyan HS and Gupta PK (2004). AFLP-based genetic diversity and its comparison with diversity based on SSR, SAMPL, and phenotypic traits in bread wheat. *Biochem. Genet.* 42: 43-59.

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

SAS Institute Inc. (1992). Statistical Analysis System. Release 6.12. SAS, Cary.

Saxena S, Chandra R, Srivastava AP, Mishra M, et al. (2005). Analysis of genetic diversity among papaya cultivars using Single Primer Amplification Reaction (SPAR) methods. *J. Hortic. Sci. Biotech.* 80: 291-296.

Silva FF, Pereira MG, Campos WF and Damasceno Júnior PC (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). Estimation of genetic parameters related to morphoagronomic and fruit quality traits of papaya. *Crop Breed. Appl. Biotechnol.* 8: 65-73.

Sokal RR and Rohlf FJ (1962). The comparison of dendrograms by objective methods. *Taxon* 11: 33-40.

Sudre CP, Goncalves LS, Rodrigues R, do Amaral Junior AT, et al. (2010). Genetic variability in domesticated Capsicum spp as assessed by morphological and agronomic data in mixed statistical analysis. *Genet. Mol. Res.* 9: 283-294.

Van Droogenbroeck B, Breyne P, Goetghebeur P, Romeijn-Peeters E, et al. (2002). AFLP analysis of genetic relationships among papaya and its wild relatives (*Caricaceae*) from Ecuador. *Theor. Appl. Genet.* 105: 289-297.

Venkovsky R (1987). Herança Quantitativa. In: Melhoramento e Produção do Milho. (Paterniani E and Viegas GP, eds.). Fundação Cargill, Campinas, 135-214.

Vieira EA, Carvalho FIF, Bertan I, Kopp MM, et al. (2007). Association between genetic distances in wheat (*Triticum aestivum* L.) as estimated by AFLP and morphological markers. *Genet. Mol. Biol.* 30: 392-399.

Vitória AP, Souza Filho GA, Bressan-Smith R, Pinto FO, et al. (2004). DNA fingerprint of *Carica papaya* L. genotypes by RAPD markers. *J. New Seeds* 6: 1-10.

Williams JG, Kubelik AR, Livak KJ, Rafalski JA, et al. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* 18: 6531-6535.