https://doi.org/10.1590/1678-4499.20210267

PLANT BREEDING
Article

Genetic-molecular characterization in guava full-sib progeny

Julie Anne Vieira Salgado de Oliveira[1], Eileen Azevedo Santos[1,*], Alexandre Pio Viana[1], Fernando Henrique de Barros Walter[1], Rodrigo Moreira Ribeiro[1]

1. Universidade Estadual do Norte Fluminense Darcy Ribeiro – Centro de Ciências e Tecnologias Agropecuárias – Campos dos Goytacazes (RJ), Brazil.

Received: Sep.7, 2021 | **Accepted:** May 4, 2022

Section Editor: Gabriel Constantino Blain

***Corresponding author:** eileenazevedo@yahoo.com.br

How to cite: Oliveira, J. A. V. S., Santos, E. A., Viana, A. P., Walter, F. H. B. and Ribeiro, R. M. (2022). Genetic-molecular characterization in guava full-sib progeny. Bragantia, 81, e3322. https://doi.org/10.1590/1678-4499.20210267

ABSTRACT: Brazil is one of the world's largest producers of guava (*Psidium guajava* L.), a very promising fruit in the northern region of the state of Rio de Janeiro. Despite this, no guava cultivar has been developed for the region. Thus, this study proposed to examine a population of guava full sibs using microsatellite markers and to identify which genotypes are the most divergent for future crosses, to select cultivars better adapted to the soil and climatic conditions of northern Rio Janeiro. Ninety-six superior genotypes were selected according to their agronomic traits, which were characterized using 45 microsatellite markers. The genetic distance between the analyzed genotypes, their clustering pattern and the genetic structure of the population were estimated. Hierarchical cluster analysis by the neighbor joining method indicated the formation of three distinct groups. The use of molecular information revealed the existence of moderate genetic variability between the genotypes of the full-sib families. Bayesian analysis separated the genotypes into only two groups, as the individuals shared most of the analyzed genomic regions. The most genetically divergent guava genotypes, that is, those allocated to different groups, such as genotypes 5 and 85, should be recommended for future crosses to obtain segregating populations, thus giving continuity to the guava breeding program.

Key words: *Psidium guajava*, microsatellite markers, genetic diversity, plant breeding.

INTRODUCTION

Guava (*Psidium guajava* L.) is a fruit tree native to South America. Brazil is among the world's largest producers of guava, having large areas where soil-climatic conditions are favorable to the production of the fruit (Almeida et al. 2009). Guava growing is a promising activity in the northern region of the state of Rio de Janeiro, which, in addition to being close to port facilities, holds the potential to boost the local economy through fruit farming (Gomes Filho et al. 2010).

Despite this, Brazilian producers currently face a problem: the low number of available cultivars adapted to producing regions. Only 18 cultivars are registered in the National Cultivar Registry (Registro Nacional de Cultivares–RNC), and no guava cultivar has been developed or recommended for the state of Rio de Janeiro so far. One of the goals of the breeder is to obtain productive cultivars adapted to their producing region (Ramalho et al. 2010), besides acceptance by the consumer market.

In this respect, guava has high genetic diversity, which is favored by cross-pollination (Silva et al. 2017). Knowing the genetic variability of cultivars is an essential factor for breeding programs, as it allows the optimization of the breeding strategy to be applied (Silva et al. 2020).

A very important aid tool for understanding genetic variability used in breeding programs are molecular markers. Microsatellite markers (SSR) are highly polymorphic, providing a large amount of genetic information per locus. They are

abundant in the genome, multiallelic and easy to automate, in addition to being affordable (Turchetto-Zolete et al. 2017). SSR are also widely used in different countries as an efficient tool for the characterization of germplasm and in the study of genetic diversity in different *Psidium* species (Kareem et al. 2018).

Understanding the genetic structure of the population is essential for plant breeding. With a well-structured population, it is possible to select genotypes with desired and complementary characteristics. When the sharing of alleles between individuals is known, we are able to select more divergent genotypes for crosses, in order to generate a greater genetic variability and to look for a heterosis effect, which is useful in breeding programs (Bezerra et al. 2020).

This study aimed to characterize 96 pre-selected guava genotypes through 45 microsatellite loci, to estimate genetic variability in the population, and to identify and indicate the best crosses between genotypes with greater genetic distance.

MATERIAL AND METHODS

Evaluated population

The evaluated population was trained in the experimental area at the Antônio Sarlo School of Agriculture, Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF), located in Campos dos Goytacazes, northern region of Rio de Janeiro, Brazil (21º08'02" S and 41º40'47" W, 88 m above sea level). The climate in the area is the Aw type (tropical sub-humid and dry), with average annual temperature ranging from 22 to 25°C and average annual precipitation between 1,200 and 1,300 mm. The study was developed at the Plant Breeding Laboratory (Laboratório de Melhoramento Genético Vegetal–LMGV) at UENF.

Paiva et al. (2016), who selected the best genotypes in 17 full-sib families of guava using restricted maximum likelihood/best linear unbiased prediction method (Table 1), indicated the 96 individuals genotyped in this study.

Table 1. Origin and identification of 17 full-sib families of guava.

Description	Crossings	Total number of genotypes	Number of selected genotypes
Family 1	UENF 1834 × UENF 1833	24	12
Family 2	UENF 1831 × UENF 1830	24	12
Family 3	UENF 1831 × UENF 1832	24	1
Family 4	UENF 1831 × UENF 1837	24	0
Family 5	UENF 1831 × UENF 3839	24	0
Family 6	UENF 1833 × UENF 1832	24	11
Family 7	UENF 1834 × UENF 1839	24	1
Family 8	UENF 1835 × UENF 1834	24	16
Family 9	UENF 1834 × UENF 1836	24	0
Family 10	UENF 1836 × UENF 1835	24	15
Family 11	UENF 1833 × UENF 1836	24	2
Family 12	UENF 1831 × UENF 1835	24	10
Family 13	UENF 1833 × UENF 1835	24	5
Family 14	UENF 1832 × UENF 1833	24	0
Family 15	UENF 1834 × UENF 1837	24	5
Family 16	UENF 1834 × UENF 1831	24	0
Family 17	UENF 1832 × UENF 1835	24	6
Total			**96**

Genomic DNA extraction and polymerase chain reaction

Samples of young leaves of the selected genotypes were collected in the experimental area, forwarded to the LMGV at UENF. The samples were macerated in liquid nitrogen, and genomic DNA was extracted using the procedure proposed by Doyle and Doyle (1990), with adaptations (Supplementary Material).

After extraction, the DNA was quantified by analysis on 1% agarose gel with 1X TAE buffer (Tris, sodium acetate, EDTA, pH 8) using the 100-bp (100 ng) Lambda (λ) marker (100 ng·μL^{-1}) (Invitrogen, Carlsbad, CA, United States of America). The DNA samples were stained using a mixture of GelRedTM and Blue Juice (1:1), and the image was captured by the Loccus L-PIX EX gel documentation system. Based on the obtained images, the DNA concentration was estimated relative to the 100-bp marker, and the DNA samples were diluted to a working concentration of 10 ng·μL^{-1}.

To test the polymerase chain reaction (PCR) conditions, 192 primer pairs (GuavaMap 2008) designed to amplify SSR loci in *P. guajava* were tested in 10 individuals, with a temperature gradient ranging from 48 to 60 °C. After screening, a set of 45 primers was selected for the amplification reactions (Table 2).

Table 2. Sequence of 45 pairs of microsatellite primers used in the analysis of 96 genotypes of *Psidium guajava* L.

Locus	Sequence (5' 3')	Annealing temperature (°C)	Reference
mPgCIR027	F: AGCACTTAGGGACACAAATTCA R: CTCACTCTCCTCCATTCAAG	50	GuavaMap2008
mPgCIR029	F: CTCGCTTCAATCTCCATCTA R: AGCGACACAGACTCTTCATT	54	GuavaMap2008
mPgCIR030	F: CTCAAAGCACTATCATGTCG R: CCTTGTGGGTTCTCTTTTG	53	GuavaMap2008
mPgCIR035	F: TTGACCTGGCATTAACAGA R: CATTGGGAAAGGGAAGAA	56	GuavaMap2008
mPgCIR037	F: GCGACGTTGTTGACTGAT R: AGTGCGATAAAGACATTCAC	53	GuavaMap2008
mPgCIR038	F: AGCCTGTTTTACGCCTTC R:CGGCTGCTCTATTGTTATTT	53°C	GuavaMap2008
mPgCIR040	F: TGAATCTCCAGTGTCTTATCG R: TGATTTCAACTGCGTATGTC	55	GuavaMap2008
mPgCIR042	F: CTCACCCAAATCTACACAAG R: AAGGGACTGGACGATGTT	50	GuavaMap2008
mPgCIR094	F: CAACCTTCCCGTGATTATT R: CTAGCTTCTTCAGTGGGAAC	54	GuavaMap2008
mPgCIR096	F: ACGCTGCAAACGATACTAAT R: AACTCACACGAGCACAGAG	55	GuavaMap2008
mPgCIR099	F: TCAAAGTCCAAAACTCATGC R: GGGATGGAGTAAAGATGAAA	54	GuavaMap2008
mPgCIR102	F: AATTGGTGTAGCATCTGGA R: GCCTACCATGAACAGAGAAA	53	GuavaMap2008
mPgCIR106	F: GACTCGACAGGAAGGTCTC R: CAGCTTTGTATATCGCAAGA	54	GuavaMap2008
mPgCIR111	F: CAACCTCGTTTGAGTCTTCT R: AACATCATTGGGACCATTC	53	GuavaMap2008
mPgCIR131	F: GAGGTTGAGAGTTCAAGGT R: GGTTTGCTCTTGAAATCACTC	51	GuavaMap2008
mPgCIR135	F: CAGATAGCAAAACTGCCTCT R: ATATCCCTCTCGCCTTCTT	54	GuavaMap2008
mPgCIR138	F: GGTGAGACCACTGAGTTCC R: GAAAGACCAATGAAATCGAC	54	GuavaMap2008
mPgCIR140	F: GTGGTGAAGGAGTAAAGCTG R: GCAGTATAAAGCAACAGATGG	54	GuavaMap2008
mPgCIR143	F: TCTGGATTTTCTCCAATGC R: GTATGCACCACCATCTGC	54	GuavaMap2008

continue...

Table 2. Continuation...

Locus	Sequence (5' 3')	Annealing temperature (ºC)	Reference
mPgCIR147	F: ACTGACATCTCTGACCATAGC R: GATTGCCATAGGAACTGAAA	54	GuavaMap2008
mPgCIR149	F: CTTCGTGGAAGAGGATGAC R: AATATAAACATCGCCACAGG	51	GuavaMap2008
mPgCIR153	F: GCCTCTGGTAAATCTGTTGA R: ACATACGGATCAAGTCCAAA	54	GuavaMap2008
mPgCIR160	F: TGGCTATAAGAATGGGAGAT R: GACGAGCTTAGCCTCTGAAT	56	GuavaMap2008
mPgCIR161	F: TCTCAAGGACCAACAAGAAG R: AGGACTTAGCTTGGGTTTTC	54	GuavaMap2008
mPgCIR163	F: TCTTTGCACATCAAACTCG R: CATGGTATCAATAGGTCAAGC	53	GuavaMap2008
mPgCIR164	F: TCCCGTAAGTTGTTCTGTTT R: CGTAAGAGATCGTGAAGGAG	54	GuavaMap2008
mPgCIR166	F: CTTTCCCATCAAAACGTAAG R: CCAATTCATGCACTTAGACA	54	GuavaMap2008
mPgCIR169	F: TTCAGGCAGATCGTGTTACT R: GTGCCTAACCTACACCCTAA	55	GuavaMap2008
mPgCIR180	F: CATGGATTCAACTCTTGTCG R: CTACATTGGAAGCAGAATGG	55	GuavaMap2008
mPgCIR181	F: AGACACTACCTCACCGCTAT R: ATAGAGAAGCCGAAGGAAAC	55	GuavaMap2008
mPgCIR184	F: AAGCTACAATCGACGAAAAC R: CACTATTAGCGAACCTGCAT	55	GuavaMap2008
mPgCIR187	F: AACGCCTAATAATGCGAAGT R: TCTTTCCCAGGATGGAGTA	52	GuavaMap2008
mPgCIR188	F: TACAGCGATTCTATCCCCTA R: TTTGTGGGGAAGAAACTACTG	53	GuavaMap2008
mPgCIR191	F: GACCCTCCACTTATATTTTG R: AAGCTGACATAACAGTCGAA	55	GuavaMap2008
mPgCIR193	F: GAACGTGGGTTACATACCAT R: ATCACCGTCCTCCTAAATCT	52	GuavaMap2008
mPgCIR200	F: CCTTGCTTTGGTGAGGTC R: GCTAATTCAGTCCTTCCAACT	51	GuavaMap2008
mPgCIR202	F: CCATTAGAGCCGACAAAA R: GACGAGAAACCCTAAACGTA	53	GuavaMap2008
mPgCIR204	F: GTCGGATCATGGAGAGATCA R: GCGGCTAAAGAAATCTGC	58	GuavaMap2008
mPgCIR207	F: CAAGATTTGCCTCAAAGAAAC R: AACTAAATAGCCTGCTGGTG	55	GuavaMap2008
mPgCIR209	F: CTAAAGCCACATCCAGCA R: CTAACATTTGCCTTCTACAGC	54	GuavaMap2008
mPgCIR210	F: CTACGAGGGTGTAACGAAAA R: CTACAACAGCCAACGTGAG	55	GuavaMap2008
mPgCIR221	F: CTAAGCCTGAAGTCCCAAAT R: CCTCTTCTAAAGGCAACGAC	55	GuavaMap2008
mPgCIR227	F: GGGATGCTCAAAACTGTAAG R: CCTGTTACATTGACGAATCA	51	GuavaMap2008
mPgCIR230	F: CACATTTGCTCCTGATTTTC R: GCTCTTCAACGACCATCTT	55	GuavaMap2008
mPgCIR249	F: TTTGTCTGGTCGTCCTAGTT R: CTTCAGTCCATCAGCAAAAT	53	GuavaMap2008

The PCR reactions were performed in thermocyclers from Applied Biosystems/Veriti 96 well, in a 38-cycle program. The final volume was 13 µL for each sample, which contained 2-µL DNA (10 ng·µL), 1.50-µL 10X Buffer (NH_4SO_4), 1.5-µL

$MgCl_2$ (25 mM), 1.5-µL dNTPs (2 mM), 1-µL primer (R+F) (5 µM) and 0.12-µL Taq-DNA polymerase (5 U/µL) (Invitrogen, Carlsbad, CA, United States of America).

Amplification products were separated on 4% MetaPhor agarose gel, immersed in TAE buffer [90 mM Tris-acetate (pH 8) + 10 mM EDTA], stained with Gel RedTM and Blue Juice (1:1), visualized by the Loccus L-PIX EX gel documentation system and compared with the 100-bp High DNA Mass Ladder (0.5 μL^{-1}) marker (Invitrogen, Carlsbad, CA, United States of America) during the runs to determine amplified fragments.

Statistical analysis

The data obtained from the amplification of 45 SSR were converted into numerical code for each allele per locus. The numerical matrix was developed by assigning values from 1 to the maximum number of alleles per locus, as described next: for a locus with three alleles, the representations of 11, 22 and 33 were used for the homozygous forms (A1A1, A2A2 and A3A3); and 12, 13 and 23 for the heterozygotes (A1A2, A1A3 and A2A3). From this numerical matrix, the genetic distance between the studied genotypes was calculated using GENES software (Cruz 2013).

Three indices were tested to calculate the similarity between genotype pairs (Table 3), namely: the unweighted index, the weighted index and the Smouse and Peakall index (Cruz et al. 2011).

Table 3. Distance indices and clustering methods tested to estimate the similarity between pairs of guava genotypes.

Distance measurements	Grouping methods	Cophenetic correlation coefficient	Distortion (%)	Stress (%)
Weighted index	UPGMA	0.9302	1.6583	12.8785
	Neighbor joining	0.9229	46.2709	29.9213
	Ward	0.4404	—	—
Unweighted index	UPGMA	0.8745	3.754	19.3742
	Neighbor joining	0.8627	72.015	49.4014
	Ward	0.5551	—	—
Smouse and Peakall Index	UPGMA	0.8733	6.9976	26.4531
	Neighbor joining	0.8877	76.152	54.2757
	Ward	0.5417	—	—

UPGMA: unweighted pair-group method with arithmetic mean.

Three hierarchical clustering methods were tested (Table 3): unweighted pair-group method with arithmetic mean (UPGMA), which uses arithmetic means of dissimilarity measures, avoiding characterization through extreme values (Cruz et al. 2011); the neighbor joining method, proposed by Saitou and Nei (1987), which groups the closest individuals with data from the distance matrix; and the Ward method, in which the similarity measure used in the cluster is the sum of squares between two clusters (Hair et al. 2009).

The use of different clustering methods for the same goal, without indicating the choice criterion, can make it difficult to compare results, since they are influenced by the method selected for the construction of the clustering matrix (Cerqueira-Silva et al. 2009).

In this study, the selected method was the neighbor joining due to the high cophenetic correlation coefficient (CCC) and the similarity with the results obtained in Bayesian analysis.

The weighted index was chosen because it showed the highest cophenetic correlation, estimated by Eq. 1:

$$S_{ii=\frac{1}{2}}\sum_{j-1}^{L} p_j\, c_j \quad (1)$$

in which (Eq. 2) $p_j = \frac{a_j}{A}$ (2) = weight associated with locus j determined by a_j = total number of alleles at locus j; A = total number of alleles studied, in which (Eq. 3) $\sum_{j=1}^{L} p_j = 1$ (3); and c_j = the number of common alleles between the pairs of accessions i and i'.

The index deals with similarity measures, and in cluster analysis, it is recommended to use dissimilarity measures, defined by Eq. 4:

$$D = 1 - S \tag{4}$$

After the generation of the distance matrix, the cluster analysis of individuals was performed via dendrogram, by applying the neighbor joining method, using Mega software version 6 (Kumar et al. 2008).

The diversity indices of the 96 genotypes were estimated using Genalex 6.5 software (Peakall and Smouse 2012), based on the following parameters: number of alleles per polymorphic locus (Na), number of effective alleles (Ne), observed heterozygosity (Ho), expected heterozygosity (He), information index or Shannon index (I), and fixation index or inbreeding coefficient (F).

The Ne, that is, those actually found in the population, can be calculated by Eq. 5:

$$Ne = \frac{1}{\Sigma_P^2} \tag{5}$$

in which: P^2 = the sum of the frequency of homozygous and heterozygous alleles.

The information index, known as the Shannon index (I), is used to indicate diversity, which can be calculated by Eq. 6:

$$I = -1.\sum[Pi.Ln(Pi)] \tag{6}$$

in which: Pi = allelic frequency for each of the alleles in question.

Ho is the proportion of heterozygous individuals observed in a studied population, calculated by Eq. 7:

$$Ho = \frac{Nx}{N} \tag{7}$$

in which: Ho = observed heterozygosity; Nx = number of heterozygotes; N = total number of individuals in the sample.

He can be defined as an estimated sum of all individuals that could be heterozygous for a locus (Eq. 8):

$$He = 1 - \sum P_i^2 \tag{8}$$

in which: He = expected heterozygosity; P_i = frequency of allele i.

The fixation index (F), which can range from -1 to +1, estimates the mean coefficient of inbreeding, given by Eq. 9:

$$F = \frac{He - Ho}{He} \tag{9}$$

in which: Ho = observed heterozygosity, which is the proportion of N samples that are heterozygous at a given locus; He = proportion of heterozygosity expected under random mating.

Analysis of the genetic structure of the population

To access the structure of the 96 genotypes, analyses were performed using the Bayesian method in Structure software version 2.3.4 (Pritchard et al. 2000).

Considering that the present study was carried out with a population from plants obtained from controlled crosses, we adopted the "no admixture" model, correlated with the allelic frequencies of the population (Cerqueira-Silva et al. 2014, Silva et al. 2016).

The burn-in period and iteration number were set to 25,000 and 75,000, respectively, for each run. The number of groups (K) was varied systematically from 1 to 5, and 20 simulations were performed to estimate each K.

The ad hoc ΔK method described by Evanno et al. (2005), implemented in the online tool StructureHarvester (Earl and Vonholdt 2012), was used to estimate the most likely K for the population.

The threshold value of 0.60 was used as the maximum probability of association between the subgroups. Based on the posterior probability of association (q) of a given genotype belonging to a given group relative to the total number of groups (K), we classified individuals with $q > 0.60$ as members of a given cluster, whereas for clusters with an association (q) with values ≤ 0.60 the genotype was classified as mixed (Cerqueira-Silva et al. 2014).

RESULTS AND DISCUSSION

Diversity parameters via SSR

Genetic variability was detected between the genotypes evaluated for the 45 markers used. The number of alleles per locus ranged between two and three, averaging 2.13, with a total of 96 alleles for all evaluated loci.

In a genetic characterization study of guava accessions from different municipalities in Pakistan for germplasm formation, Kareem et al. (2018) found 85 alleles from 18 SSR primers, with an average of 4.7 alleles per locus–a high number when compared with the one found in this study.

Costa and Santos (2013) analyzed genetic variability through 13 SSR in Psidium accessions from the Embrapa Semiárido germplasm bank and found the total of 183 alleles. The number of alleles per locus ranged from seven to 22, averaging 14.07.

Because the full-sib population evaluated in this study originates from previous selections, a low number of alleles is expected, and the high number of alleles observed in other studies is expected when evaluating accessions from germplasm banks.

The Ne ranged from 1.02 to 1.99 (Table 4). Ho values ranged from zero to 0.97, and He values ranged from 0.02 to 0.50 (Table 4).

Table 4. Diversity parameters for 45 microsatellite markers evaluated in 96 guava genotypes: number of alleles per locus (Na), number of effective alleles (Ne), information index (I), observed heterozygosity (Ho), expected heterozygosity (He), and fixation index (F).

Locus	Na	Ne	I	Ho	He	F
mPgCIR027	2.000	1.951	0.681	0.816	0.488	-0.673
mPgCIR029	2.000	1.837	0.648	0.000	0.456	1.000
mPgCIR030	2.000	1.977	0.687	0.000	0.494	1.000
mPgCIR035	2.000	1.830	0.646	0.674	0.454	-0.485
mPgCIR037	2.000	1.169	0.275	0.000	0.145	1.000
mPgCIR038	2.000	1.067	0.143	0.000	0.062	1.000
mPgCIR040	2.000	1.264	0.363	0.000	0.209	1.000
mPgCIR042	2.000	1.067	0.144	0.000	0.063	1.000
mPgCIR094	2.000	1.162	0.267	0.000	0.139	1.000
mPgCIR096	2.000	1.749	0.620	0.621	0.428	-0.450
mPgCIR099	2.000	1.149	0.253	0.000	0.130	1.000
mPgCIR102	2.000	1.046	0.108	0.000	0.044	1.000
mPgCIR106	2.000	1.998	0.693	0.968	0.499	-0.938
mPgCIR111	3.000	1.150	0.275	0.083	0.130	0.360

continue...

Table 4. Continuation...

Locus	Na	Ne	I	Ho	He	F
mPgCIR131	3.000	1.022	0.067	0.011	0.021	0.496
mPgCIR135	2.000	1.586	0.556	0.383	0.370	-0.036
mPgCIR138	2.000	1.675	0.593	0.488	0.403	-0.211
mPgCIR140	3.000	1.709	0.630	0.580	0.415	-0.397
mPgCIR143	2.000	1.212	0.318	0.000	0.175	1.000
mPgCIR147	3.000	1.199	0.336	0.181	0.166	-0.088
mPgCIR149	2.000	1.991	0.691	0.587	0.498	-0.179
mPgCIR153	2.000	1.170	0.276	0.074	0.145	0.493
mPgCIR160	2.000	1.875	0.659	0.355	0.467	0.240
mPgCIR161	2.000	1.815	0.641	0.000	0.449	1.000
mPgCIR163	2.000	1.112	0.208	0.000	0.101	1.000
mPgCIR164	2.000	1.865	0.657	0.043	0.464	0.907
mPgCIR166	2.000	1.530	0.530	0.315	0.346	0.090
mPgCIR169	2.000	1.996	0.692	0.957	0.499	-0.918
mPgCIR180	2.000	1.959	0.683	0.536	0.489	-0.095
mPgCIR181	2.000	1.316	0.404	0.000	0.240	1.000
mPgCIR184	2.000	1.830	0.646	0.427	0.454	0.059
mPgCIR187	2.000	1.240	0.344	0.000	0.194	1.000
mPgCIR188	2.000	1.180	0.287	0.000	0.153	1.000
mPgCIR191	2.000	1.870	0.658	0.458	0.465	0.015
mPgCIR193	2.000	1.115	0.211	0.000	0.103	1.000
mPgCIR200	2.000	1.338	0.420	0.121	0.253	0.522
mPgCIR202	2.000	1.762	0.624	0.000	0.433	1.000
mPgCIR204	2.000	1.999	0.693	0.923	0.500	-0.847
mPgCIR207	2.000	1.965	0.684	0.578	0.491	-0.177
mPgCIR209	2.000	1.112	0.208	0.000	0.101	1.000
mPgCIR210	3.000	1.249	0.417	0.152	0.199	0.237
mPgCIR221	3.000	1.771	0.653	0.012	0.435	0.972
mPgCIR227	2.000	1.418	0.471	0.000	0.295	1.000
mPgCIR230	2.000	1.838	0.648	0.297	0.456	0.348
mPgCIR249	2.000	1.997	0.693	0.250	0.499	0.499
Mean	2.133	1.536	0.476	0.242	0.312	0.416

The observed mean heterozygosity (0.24) was lower than the expected mean heterozygosity (0.31), possibly suggesting the presence of null alleles. When a mutation occurs in the primer-binding site, preventing allele amplification, the number of supposed homozygotes in plants heterozygous for the allele increases (Carvalho et al. 2010).

The information index (I) was used to indicate the genetic diversity of the population and ranged from 0.07 (mPgCIR131) to 0.69 (mPgCIR106, mPgCIR149, mPgCIR204 and mPgCIR249), averaging 0.47 (Table 4). This result suggests moderate diversity in the population, and that primers mPgCIR106 and mPgCIR204 were the most efficient in discriminating genotypes with greater genetic diversity. These same primers obtained high Ho values, confirming the index information (Lacerda et al. 2001).

This index, also known as the Shannon index, started to be more commonly used in genetic analysis with the advent of bioinformatics (Sherwin et al. 2006). It varies from 0 to 1, with values closer to 0 denoting lower genetic diversity (Moura et al. 2005).

The fixation index (F), which corresponds to the inbreeding coefficient, was estimated for the entire population and averaged 0.42, ranging from -0.94 to 1 between loci (Table 4). Only 13 loci obtained negative values, which is expected in random mating and indicates excess heterozygosity. While substantial positive values indicate inbreeding or undetected null alleles, the presence of null alleles is a problem in microsatellite data analysis, as they can lead to a false interpretation of results (Souza et al. 2008).

The loci with negative values are the same whose Ho was greater than expected, which indicates that the alleles for these loci are not being fixed by inbreeding. The remaining loci, with positive values, have excessive homozygosity, which may mean failure in allele amplification, since the population evaluated originates from cross-pollination, and negative fixation indices would be expected. However, it is important to emphasize that, because the population originates from previous selections, it is possible that positive values predominate. However, it is important to emphasize that, since the population comes from crosses between relatives and from previous selections, positive values may be predominant.

Dissimilarity by neighbor joining

Genetic dissimilarity was detected between the studied genotypes, and genotype 5 was the most divergent. The total mean dissimilarity was 0.26.

Once obtained the dissimilarity matrix, the evaluated genotypes were clustered into three distinct groups using the neighbor joining method. This method was chosen because it shows greater similarity with the Bayesian analysis than the UPGMA method, with both having a close and high CCC.

The cutoff point in the dendrogram was determined using the criterion of Mojena (1977), with cuts at 73 to 80% dissimilarity and k = 1.25. This is a statistical criterion in which the calculation is based on the relative size of the distance levels in the dendrogram, dispensing with prior knowledge of the conformation of the groups (Faria et al. 2012).

Genotypes 5, 14, 31, 38, 44 and 64 were clustered in the first group, highlighted by the green color (Fig. 1), which was the most distant. The greatest dissimilarity found (0.83) was between genotypes 5 and 85, which had 15 alleles in common in 31 of the 45 analyzed loci. All genotypes in this group exhibited greater dissimilarity with genotype 85.

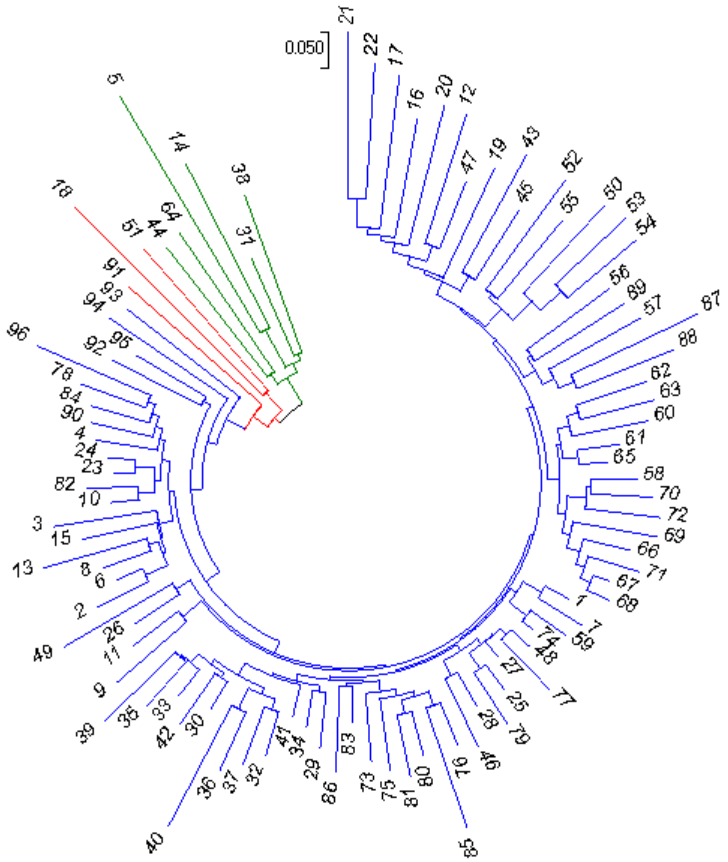

Figure 1. Dendrogram generated by the neighbor joining method (cophenetic correlation coefficient = 0.92), from the distance matrix by the weighted index obtained by microsatellite markers, listing the selected guava genotypes.

Group II, in red (Fig. 1), consisted of genotypes 18, 51 and 91. Like the genotypes in group I, genotype 18 was the most distant from genotype 85, with 0.74 dissimilarity and 22 shared alleles. Genotypes 51 and 91 were the most dissimilar to genotype 5, with dissimilarity values of 0.81 and 0.72, respectively.

Group III, in blue (Fig. 1), contained the largest number of individuals, with 87 genotypes in total (90.6%). This number of genotypes in the same group indicates that these individuals share the greatest number of alleles for the evaluated loci. Individuals 67 and 68 were the least dissimilar, that is, the closest, with 0.93 similarity and 84 alleles in common.

Silva et al. (2021) evaluated this same population for the traits of soluble solids content, fruit weight, pulp weight, number of fruits per plant and yield per plant. The individual with the highest yield per plant was 53. With this information, we may recommend its cross with individual 5, which, in addition to being the most genetically divergent, was also one of the individuals with high yield value per plant. The dissimilarity between them is 0.73, with 24 alleles in common.

Population genetic structure

Bayesian analysis suggested the formation of two groups (Fig. 2). This is a more rigorous analysis that allows observing the population structure in more defined groups, as it is less subjective than hierarchical methods, such as the neighbor joining method.

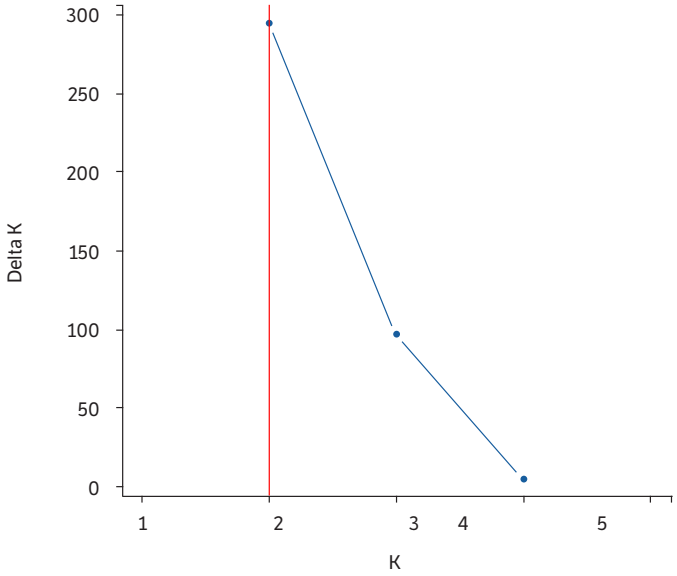

Figure 2. Delta K (ΔK) for the respective number of groups (K).

Based on Evanno et al. (2005), the optimal delta K was observed when K = 2, suggesting that maximum structuring was observed when the sample was divided into two well-structured groups (Fig. 3).

A 60% probability of adhesion for belonging to a certain group was adopted. Thus, the evaluated population was separated as follows: group I, in red (Fig. 3), was formed by the majority of genotypes, with 88 in total (91.6%). This group contained most of the genotypes that belonged to the same group (III) in the analysis by the neighbor joining method (Tables 5 and 6 of the Supplementary Material).

Most genotypes belonging to group I have 100% probability of adhesion. However, some genotypes have mixed probability. For example, genotype 20 has 80% probability of adhesion to the red group, but 20% adherence to the green group, indicating alleles shared with the genotypes in this group (Fig. 3, Tables 5 and 6 of the Supplementary material).

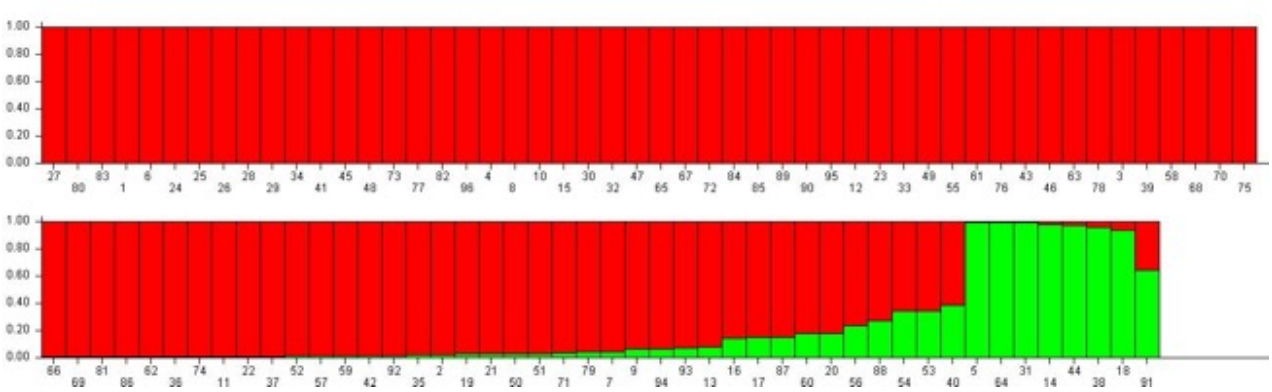

Figure 3. Clustering and genetic structuring of the 96 genotypes of *Psidium guajava* L. by Bayesian inference (K = 2).

Eight genotypes (8.33%) were allocated to group II, in green (Fig. 3), namely 5, 14, 18, 31, 38, 44, 64 and 91. Genotypes 5, 14, 18, 31, 38, 44 and 64 showed near 100% adhesion to the green group. The adhesion of genotype 91, which belonged to the green group, was just above 60%, but with alleles shared with the red group.

This structuring means that group I has a set of alleles that differentiates it from group II for the set of markers used, which are genomic markers that are not related to phenotypic traits.

A low number of groups was formed. This happens when individuals share most of the genomic regions analyzed, which can be explained by the genetic structure of the population. These individuals are related and structured as full-sib families, in addition to having been previously selected based on their superior agronomic traits.

Obtaining two groups is enough to help direct the next crosses between guava genotypes belonging to distinct groups. Thus, the 96 genotypes clearly comprise a defined set of genetic structure.

Considering the different types of analysis, preferential crosses are indicated for genotypes that are in different groups, as they are more genetically distant, which will increase vigor. Additionally, the number of alleles evaluated and shared among them should also be observed, by selecting the most divergent ones.

Furthermore, for selection purposes, these results should be combined with the agronomic data of the evaluated population, by selecting the individuals with the greatest agronomic potential and, thereby, increasing gains and the efficiency of the selection process.

CONCLUSION

The SSR used in this study were efficient in discriminating guava full-sib genotypes.

There was variability in the evaluated population, which was structured in three distinct groups by the neighbor joining hierarchical method and in two groups by Bayesian analysis.

The Shannon index indicated that the primers used in the study were efficient to estimate genetic divergence in the population, which was moderate.

It is possible to indicate the most genetically divergent genotypes, that is, those allocated to different groups, to be used in the guava breeding program to obtain segregating populations. Accordingly, crosses between individual 5 and genotypes 85, 89, 45, 51, 8, 16, 19, 96, which were the most dissimilar to each other, are recommended.

AUTHORS' CONTRIBUTION

Conceptualization: Santos, E. A., Viana, A.P. and Oliveira, J. A. V. S.; **Methodology:** Santos, E. A., Viana, A.P. and Oliveira, J. A. V. S. **Investigation:** Santos, E. A., Viana, A.P., Oliveira, J. A. V. S. and Walter, F.H.B.; **Writing – Original Draft:**

Oliveira, J. A. V. S.; **Writing – Review and Editing:** Oliveira, J. A. V. S. and Santos, E. A.; **Funding Acquisition:** Viana, A. P.; Resources: Viana, A.P.; **Supervision:** Santos, E. A. and Viana, A. P.

DATA AVAILABILITY STATEMENT

All dataset were generated and analyzed in the current study.

FUNDING

Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro
https://doi.org/10.13039/501100004586
Grant No: E-26/202.786/2017.

ACKNOWLEDGMENTS

Not applicable.

REFERENCES

Almeida, E. J., Santos, J. M. and Martins, A. B. G. (2009). Resistance of guava and araça to *Meloidogyne mayaguensis*. Pesquisa Agropecuária Brasileira, 44, 421-423. https://doi.org/10.1590/S0100-204X2009000400014

Bezerra, C. M., Ribeiro, R. M., Viana, A. P., Cavalcante, N. R., Silva, F. A., Ambrósio, M. and Amaral Junior, A. T. (2020). Guava breeding via full-sib family selection: conducting selection cycle and divergence between parents and families. Crop Breeding and Applied Biotechnology, 20, e256520112. https://doi.org/10.1590/1984-70332020v20n1a12

Carvalho, A. C. M., Freitas, M. L. M., Moraes, S. M. B., Moraes, M. L. T., Stranghetti, V., Alzate-Marin, A. L. and Sebbenn, A. M. (2010). Genetic diversity, inbreeding and gene flow in a small fragmented population of *Copaifera langsdorffii*. Brazilian Journal of Botany, 33, 599-606. https://doi.org/10.1590/S0100-84042010000400008

Cerqueira-Silva, C. B. M., Cardoso-Silva, C. B., Conceição, L. D. H. C. S., Nonato, J. V. A., Oliveira, A. C. and Corrêa, R. X. (2009). Comparison of coefficients and distance measurements in passion fruit plants based on molecular markers and physicochemical descriptors. Genetics and Molecular Research, 8, 870-879. https://doi.org/10.4238/vol8-3gmr614

Cerqueira-Silva, C. B. M., Santos, E. S. L., Jesus, O. N., Vieira, J. G. P., Mori, G. M., Côrrea, R. X. and Souza, A. P. (2014). Molecular genetic variability of commercial and wild accessions of passion fruit (*Passiflora* spp.) targeting ex situ conservation and breeding. International Journal of Molecular Sciences, 15, 22933-22959. https://doi.org/10.3390/ijms151222933

Costa, S. R. and Santos, C. A. F. (2013). Allelic database and divergence among Psidium accessions by using microsatellite markers. Genetics and Molecular Research, 12, 6802-6812. https://doi.org/10.4238/2013.December.16.6

Cruz, C. D. (2013). Genes: a software package for analysis in experimental statistics and quantitative genetics. Acta Scientiarum, 35, 271-276. https://doi.org/10.4025/actasciagron.v35i3.21251

Cruz, C. D., Ferreira, F. M. and Pessoni, L. A. (2011). Biometria Aplicada ao Estudo da Diversidade Genética. Viçosa.

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

Earl, D. A., and VonHoldt, B.M (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources, 4, 359-361. https://doi.org/10.1007/s12686-011-9548-7

Doyle, J. J. and Doyle, J. L. (1990). Isolation of plant DNA from fresh tissue. Focus, 12, 13-15.

Faria, P. N., Cecon, P. R., Silva, A. R., Finger, F. L., Silva, F. F., Cruz, C. D. and Sávio, F. L. (2012). Clustering methods in a study of genetic diversity of peppers. Horticultura Brasileira, 30, 428-432. https://doi.org/10.1590/S0102-05362012000300012

Gomes Filho, A., Oliveira, J. G., Viana, A. P., Siqueira, A. P. O., Oliveira, M. G. and Pereira, M. G. (2010). RAPD molecular markers and morphological descriptors in the evaluation of genetic diversity of guava (*Psidium guajava* L.). Acta Scientiarum, Agronomy, 32, 627-633. https://doi.org/10.4025/actasciagron.v32i4.4720

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. and Tatham, R. L. (2009). Análise multivariada de dados. Porto Alegre: Bookman.

Kareem, A., Jaskani, M. J., Mehmood, A., Khan, I. A., Awan, F. S. and Sajid, M. W. (2018). Morpho-genetic profiling and phylogenetic relationship of guava (*Psidium guajava* L.) as genetic resources in Pakistan. Revista Brasileira de Fruticultura, 40, e-069. https://doi.org/10.1590/0100-29452018069

Kumar, S., Nei, M., Dudley, J. and Tamura, K. (2008). Mega: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings in Bioinformatics, 9 299-306. https://doi.org/10.1093/bib/bbn017

Lacerda, D. R., Acedo, M. D. P., Lemos Filho, J. P. and Lovato, M. B. (2001). Genetic diversity and structure of natural populations of *Plathymenia reticulata* (Mimosoideae), a tropical tree from the Brazilian Cerrado. Molecular Ecology, 10, 1143-1152. https://doi.org/10.1046/j.1365-294X.2001.01264.x

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: an evaluation. The Computer Journal, 20, 359-363. https://doi.org/10.1093/comjnl/20.4.359

Moura, E. F., Pinto, J. E. B. P., Santos, J. B., Lameira, O. A. (2005). Genetic diversity in a jaborandi (*Pilocarpus microphyllus* Stapf.) germplasm bank assessed by RAPD markers. Revista Brasileira de Plantas Medicinais, 7, 1-8.

Paiva, C. L., Viana, A. P., Santos, Eileen A., Freitas, J. C. O., Amaral Júnior, A. T. (2016). Genetic gain estimated by different selection criteria in guava progenies. Bragantia, 75, 418-427. https://doi.org/10.1590/1678-4499.477

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. Bioinformatics, 28, 2537-2539. https://doi.org/10.1093/bioinformatics/bts460

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155, 945-959. https://doi.org/10.1093/genetics/155.2.945

Ramalho, M. A. P., Toledo, F. H. R. B., Souza, J. C. and Teixeira, R. A. (2010). Competências em melhoramento genético de plantas no Brasil. Viçosa: Arka.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular and Biology Evolution, 4, 406-425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Sherwin, W. B., Jabot, F., Rush, R. and Rossetto, M. (2006). Measurement of biological information with applications from genes to landscapes. Molecular Ecology, 15, 2857-2869. https://doi.org/10.1111/j.1365-294X.2006.02992.x

Silva, F. A., Viana, A. P., Corrêa, C. C. G., Carvalho, B. M., Sousa, C. M. B., Amaral, B. D., Ambrósio, M. and Glória, L. S. (2020). Impact of Bayesian Inference on the Selection of Psidium guajava. Scientific Reports, 10, 1999. https://doi.org/10.1038/s41598-020-58850-6

Silva, F. A., Viana, A. P., Corrêa, C. C. G., Santos, E. A., Oliveira, J. A. V. S., Andrade, J. D. G., Ribeiro, R. M. and Glória, L. S. (2021). Bayesian ridge regression shows the best fit for SSR markers in *Psidium guajava* among Bayesian models. Scientific Reports, 11, 13639. https://doi.org/10.1038/s41598-021-93120-z

Silva, F. H. L., Muñoz, P. R., Vincent, C. I. and Viana, A. P. (2016). Generating relevant information for breeding *Passiflora edulis*: genetic parameters and population structure. Euphytica, 208, 609-619. https://doi.org/10.1007/s10681-015-1616-8

Silva, S. N., Silva, M. A., Marçal, T. S., Ferreira, A., Fontes, M. M. P. and Ferreira, M. F. S. (2017). Genetic parameters of pollen viability in guava (*Psidium guajava* L.). Australian Journal of Crop Science, 11, 1-8. https://doi.org/10.21475/ajcs.2017.11.01.pne91

Souza, G. A., Carvalho, M. R. O., Martins, E. R., Guedes, R. N. C. and Oliveira, L. O. (2008). Genetic diversity estimated through ISSR markers in populations of *Zabrotes subfasciatus*. Pesquisa Agropecuária Brasileira, 43, 843-849. https://doi.org/10.1590/S0100-204X2008000700008

Turchetto-Zolete, A. C., Turchetto, C., Zanella, C. M., e Passaia, G. (2017). Marcadores Moleculares na Era Genômica: Metodologias e Aplicações. Ribeirão Preto: Sociedade Brasileira de Genética.

SUPPLEMENTARY MATERIAL.

Adaptations to the Doyle and Doyle Protocol (1990) to Genomic DNA extraction to *Psidium guajava* L.

Young leaf samples of selected genotypes were collected and macerated in liquid nitrogen, and genomic DNA extraction was performed using the procedure proposed by Doyle and Doyle (1990), with some adaptations. Here's the protocol: 900 µL of extraction buffer containing 2% CTAB, 2 mol L-1 NaCl, 20 mmol L-1 EDTA, and 100 mmol L-1 Tris-HCl (pH 8) were added to the tubes containing the macerated samples, as well as 2% PVP and 2% mercaptoethanol, the latter two necessary for the removal of phenolic compounds. The material was incubated at 65°C for 40 minutes and gently homogenized by inversion every 10 minutes.

After the samples reached room temperature, the tubes were centrifuged for 8 minutes at 14,000 rpm and poured into a new 2-mL tube. 800 µL chloroform:isoamyl alcohol (24:1) was added to carry out the deproteinization. This material underwent gentle inversions for approximately 10 minutes, until it became cloudy.

The organic phase was separated by centrifugation at 14,000 rpm for 8 minutes. The supernatant was transferred to a properly identified 2-mL tube. These steps were repeated three more times, and the added chloroform must be 100 uL more than the volume of the supernatant in each step. Nucleic acids were precipitated by adding two-thirds (500 µL) of the volume of ice-cold isopropanol, and incubated for 30 minutes at -70°C or for 3 hours at -20°C.

The precipitate was sedimented by centrifugation at 14,000 rpm for 10 minutes. The supernatant was discarded, and the precipitate washed twice with 500 µL of chilled 70% ethanol to remove the salt present and twice more with 500 µL of chilled 95% ethanol. After each wash, the material was centrifuged for 5 minutes at 14,000 rpm.

After discarding the last supernatant, the material was taken to the dry bath apparatus, until all ethanol was removed. Then, the material was resuspended in 100 µL of TE solution (Tris-EDTA – 10 mmol L-1 Tris-HCl, 1 mmol L-1 EDTA, pH 8) with RNase at a final concentration of 10 µg mL-1 and incubated in a water bath at 37°C for 40 minutes. The material was then stored at 20 °C until use.

Information on the origin of each genotype of the 17 full-sib families

Table 5. Coding of the *Psidium guajava* samples evaluated in this study.

Genotyped samples	Field coding*	Genotyped samples	Field coding*
1	F13 P3 B2	51	F2 P5 B1
2	F15 P11 B2	52	F2 P8 B1
3	F6 P11 B2	53	F15 P12 B1
4	F6 P6 B2	54	F15 P10 B1
5	F2 P7 B2	55	F15 P11 B1
6	F6 P3 B2	56	F10 P1 B2
7	F2 P12 B2	57	F10 P6 B2
8	F1 P10 B2	58	F10 P12 B1
9	F3 P11 B2	59	F10 P10 B1
10	F6 P11 B1	60	F17 P7 B2
11	F10 P3 B2	61	F17 P2 B2
12	F11 P11 B1	62	F17 P4 B2
13	F15 P2 B2	63	F17 P5 B2
14	F6 P5 B1	64	F17 P6 B2
15	F8 P1 B2	65	F17 P8 B2
16	F12 P4 B1	66	F10 P7 B1
17	F8 P4 B2	67	F10 P1 B1
18	F10 P5 B1	68	F10 P7 B2
19	F12 P11 B2	69	F10 P2 B1
20	F7 P9 B1	70	F10 P8 B1
21	F12 P12 B2	71	F10 P11 B1
22	F12 P9 B2	72	F10 P8 B2
23	F12 P3 B1	73	F10 P5 B2
24	F12 P10 B2	74	F10 P4 B2
25	F12 P5 B1	75	F1 P11 B2
26	F12 P8 B2	76	F1 P12 B2
27	F12 P6 B1	77	F1 P8 B2
28	F12 P1 B2	78	F1 P5 B2
29	F8 P10 B1	79	F1 P4 B2
30	F8 P2 B1	80	F1 P2 B1
31	F8 P1 B1	81	F1 P6 B1
32	F8 P5 B2	82	F1 P3 B2
33	F8 P12 B2	83	F1 P11 B1
34	F8 P7 B2	84	F1 P7 B1
35	F8 P6 B2	85	F1 P9 B2
36	F8 P8 B2	86	F13 P8 B2
37	F8 P3 B1	87	F13 P1 B1
38	F8 P11 B2	88	F13 P4 B1
39	F8 P10 B2	89	F13 P2 B2
40	F8 P12 B1	90	F11 P10 B2
41	F8 P9 B2	91	F6 P10 B1
42	F8 P3 B2	92	F6 P2 B2
43	F2 P4 B1	93	F6 P12 B2
44	F2 P8 B2	94	F6 P10 B2
45	F2 P5 B2	95	F6 P8 B1
46	F2 P10 B2	96	F6 P4 B2
47	F2 P9 B2		
48	F2 P2 B2		
49	F2 P10 B1		
50	F2 P9 B1		

*The genotype 1 is the plant 3 of the block 2 and from the family 13; F: family; P: plant; B: block.

Table 6. Separation of genotypes and their respective families in each group suggested by Bayesian analysis.

	Genotypes in group I	Genotypes in group II
Family 1	8, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85	-
Family 2	7, 43, 45, 46, 47, 48, 49, 50, 51, 52	5, 44
Family 3	9	-
Family 6	3, 4, 6, 10, 92, 93, 94, 95, 96	14, 91
Family 7	20	-
Family 8	15, 17, 29, 30, 32, 33, 34, 35, 36, 37, 39, 40, 41, 42	31, 38
Family 10	11, 56, 57, 58, 59, 66, 67, 68, 69, 70, 71, 72, 73, 74	18
Family 11	12, 90	-
Family 12	16, 19, 21, 22, 23, 24, 25, 26, 27, 28	-
Family 13	1, 86, 87, 88, 89	-
Family 15	2, 13, 53, 54, 55	-
Family 17	60, 61, 62, 63, 65	64