CROP BREEDING AND APPLIED BIOTECHNOLOGY

NOTE

SSR-based genetic analysis of sweet corn inbred lines using artificial neural networks

Fernando Ferreira[1], Carlos Alberto Scapim[1], Carlos Maldonado[2] and Freddy Mora[2*]

Crop Breeding and Applied Biotechnology
18: 309-313, 2018
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332018v18n3n45

Abstract: *Studies on genetic diversity and population structure provide basic information at the molecular level, which is a key input for breeding programs of crop species. This study evaluated the genetic diversity of 12 elite lines of sweet corn, using 20 microsatellite markers. To determine the genetic differentiation among lines, we used an artificial neural network with the self-organizing map (SOM) algorithm. This algorithm identified three genetically differentiated groups and produced relatively more accurate results than UPGMA, according to the indices of Davies-Bouldin and RMSSTD (Root Mean Square Standard Deviation). The expected heterozygosity was high (He>0.5) for 90% and the polymorphism information content high (PIC>0.6) for 40% of the SSR loci, indicating their potential to detect genetic differences among lines. The high genetic differentiation, detected by the neural network procedure, would allow the selection of promising divergent sweet corn genotypes.*

Key words*: Unsupervised learning, self-organizing map, clustering.*

INTRODUCTION

Maize (*Zea mays* L.) is the most cultivated cereal in the world, due to its importance for human and animal nutrition. In addition, corn has many industrial applications such as the production of ethanol, oil and high-amylose starch. Consequently, several breeding programs have been undertaken to achieve genetic gains in several traits of interest (e.g., Kulka et al. 2018). Studies of diversity and genetic structure allow plant breeders to investigate the population variability and thus provide basic information at the molecular level (Ballesta et al. 2015, Mora et al. 2015, Contreras-Soto et al. 2017); a key aspect in maize breeding programs (Saavedra et al. 2013, Amaral et al. 2016).

Information on population structure and genetic diversity provides crucial inputs for breeding of crop species including corn. A key molecular marker type in genetic studies are the microsatellite markers (single sequence repeats, SSR) because of their high levels of both polymorphism and number of alleles per locus (Mora et al. 2017). In corn, SSR markers have been used in the analysis of genetic diversity (Lopes et al. 2015), studies of population structure and mapping of quantitative trait loci.

Analyses of genetic clustering based on molecular markers are a simple and powerful tool to determine the population structure (Peña-Malavera et al. 2014). Currently, the main clustering techniques use the Markov Chain Monte Carlo (MCMC) algorithms to fit a model to molecular data. For instance, Pritchard et al.

***Corresponding author:**
E-mail: morapoblete@gmail.com

Received: 24 May 2016
Accepted: 01 May 2017

[1] Universidade Estadual de Maringá, Departamento de Agronomia, Av. Colombo, 5790, 87.020-900, Maringá, PR, Brazil
[2] Universidad de Talca, Instituto de Ciencias Biológicas, 2 Norte 685, 3.460-000, Talca, Chile

(2000) proposed a Bayesian clustering method, which assumes that the populations are in Hardy-Weinberg equilibrium (HWE). On the other hand, Gao et al. (2007) proposed an alternative method to analyze population structure that relaxes the assumption of HWE in the underlying populations. Other methods have been used as an alternative to MCMC, such as principal component and artificial neural network analysis (ANN, e.g. Barbosa et al. 2011). The objectives of this study were to examine the genetic diversity of 12 elite sweet corn lines using 20 microsatellite markers and to determine their genetic differentiation using ANN.

MATERIAL AND METHODS

Twelve parental lines of sweet corn of an elite line group of the Maringá State University and Syngenta Seeds Ltd. (Werle et al. 2014) were genetically evaluated with 20 SSR markers obtained from the Maize Genetic Data Bank (http://www.maizegdb.org) (Table 1). The genomic DNA was extracted using the protocol described by Gawel and Jarret (1991) with minor modifications, in young leaves from agricultural fields in Cascavel and Mauá da Serra, Paraná State of Brazil. The polymerase chain reaction (PCR) amplification was performed by the Touchdown PCR program (Don et al. 1991), using volumes of 20 µL, containing 25 ng of DNA, with 2.0 µL of 10 × reaction buffer, 2.5 mM MgCl2, 0.1 mM of each dATP, dGTP, dCTP, dTTP, and 0.3 mL of each primer (F and R primers) and 1 U Taq-DNA-Polymerase (Invitrogen). After amplification, 20 µL per sample (a total of 120 aliquots) were separated by electrophoresis on 10% (w/v) denaturing polyacrylamide gel. All 120 samples amplified per SSR primer were run with 1X TBE at 80 V for 18 h. A low range DNA ladder (Thermo Scientific) was used as a molecular weight marker reference. Gels were visualized under ultraviolet transilluminator and photographed using the Kodak 1D 3.5 program. The numbers of alleles per locus were determined based on their relative position on the polyacrylamide gel.

The mean number of alleles per locus (A), and the observed (Ho) and expected heterozygosis (He) were determined using GenAlEx 6.5 (Peakall and Smouse 2012). The polymorphic information content (PIC) was calculated with Cervus 3.0 (Kalinowski et al. 2007) and the fixation index (Fst) estimated with FSTAT 2.9.3 (Goudet 1995).

The population differentiation was inferred based on an ANN approach of the Self-Organizing Map algorithm (SOM, Kohonen 1998). Additionally, SOM results were compared with: 1) principal coordinate analysis (PCoA) implemented in GenAlex 6.5; and 2) the Unweighted Pair Group Method with Arithmetic Mean (UPGMA), according to the default

Table 1. SSR locus information, number of alleles per locus, expected heterozygosity (He), polymorphic information content (PIC) and coefficient of fixation (Fst) in 12 sweet corn inbred lines evaluated by microsatellite markers.

SSR locus	Bin position	Chromosome	No. of alleles	He	PIC value	Fst
bnlg1367	7	7	4	0.723	0.671	0.952
bnlg1371	6.01	6	6	0.806	0.774	0.946
bnlg1927	4.07	4	4	0.699	0.639	0.944
bnlg2190	10.06	10	4	0.627	0.576	0.985
bnlg2191	6.02	6	3	0.44	0.394	0.998
mmc0001	3.09	3	4	0.653	0.588	0.612
mmc0111	2.02	2	5	0.777	0.736	0.949
mmc0181	8.06	8	4	0.628	0.559	0.998
umc1069	8.08	8	4	0.613	0.564	0.752
umc1071	1.01	1	3	0.654	0.577	0.929
umc1137	9.07	9	6	0.739	0.699	0.889
umc1152	10.01	10	4	0.54	0.495	0.937
umc1549	7.02	7	3	0.638	0.561	0.431
umc1636	9.02	9	3	0.572	0.477	0.998
umc1757	4.01	4	4	0.752	0.702	0.933
umc2047	1.09	1	3	0.59	0.521	0.942
umc2165	6.07	6	4	0.714	0.661	0.97
umc2214	2.1	2	4	0.719	0.664	0.826
umc2292	5	5	2	0.279	0.239	0.997
umc2308	5.08	5	3	0.572	0.505	0.998

settings in the PHYLIP program (Felsenstein 1989). Principal coordinate analysis was based on standardized covariance of genetic distances calculated for codominant markers (option DISTANCE, sub-option GENETIC), according to Mora et al. (2015). Nei's genetic distances between inbred lines were used to create the UPGMA dendrogram, and their reliability was assessed by bootstrapping. The allelic frequency (calculated in GenAlEx) was used to start the learning process of SOM.

The SOM is an unsupervised learning algorithm able to reduce very high dimensional data into patterns that can be usefully interpreted (Kohonen 1998). This method consists of two layers of artificial neurons (or nodes): an input layer (data) with "p" = 1, 2, ..., "r" (one for each molecular marker) and an output layer consisting of a two-dimensional map with "a" neurons, established in a hexagonal grid (Paini et al. 2010). The procedure implemented in this study can be summarized in the following steps: (A) Starting weight vectors "w", taking random values from the input vectors "p". (B) Calculating the Euclidean distance between p and w. (C) Assigning each p with the closest w, based on the distance results. (D) Updating w from the assigned p. (E) Repeating steps B, C and D until achieving convergence (Kohonen 1998). The Root Mean Square Standard Deviation (RMSSTD, Grover and Vriens 2006) and the Davies-Bouldin index (DB; Davies and Bouldin 1979), both computed using functions from the ClusterSim library of the R project, were used to test the procedure accuracy.

RESULTS AND DISCUSSION

Genetic diversity of markers

The microsatellites used in this study yielded 228 alleles, with a mean value of 3.85 alleles per locus. The number of alleles per locus ranged from two to six. Only 5% of all marker data were missing due to amplification failure or null alleles. The mean Ho value in the SSR loci was low (0.088), a result expected in pure lines, and reached a maximum in locus umc1549 (0.75). The genetic diversity was equivalent to the expected heterozigocity for diploid data, which is defined as the probability that two chosen randomly haplotypes (alleles) are different in the sample (Sserumaga et al. 2014). The expected heterozigocity (He) ranged from 0.279 (umc2292) to 0.806 (bnlg1371), with an average of 0.637 (Table 1). These values were significantly correlated with the number of alleles (r = 0.764), though the highest He values had two loci (bnlg1371 and umc1137) with six alleles. Ninety percent of the loci had high He (>0.5), indicating their adequacy to differentiate sweet corn inbred lines.

The PIC ranged from 0.239 to 0.774, with a mean of 0.580 (Table 1). The PIC values of eight loci (bnlg1367, bnlg1927, mmc0111, bnlg1371, umc2165, umc1757, umc2214, and umc1137) exceeded 0.6, indicating their information potential to detect differences among the sweet corn inbred lines (Sserumaga et al. 2014).

A high level of differentiation was found in the 12 sweet corn lines (Fst = 0.897, Table 1), where the Fst per locus ranged from 0.431 (umc1549) to 0.998 (mmc0181, bnlg2191, umc1636 and umc2308). The Fst values indicated that

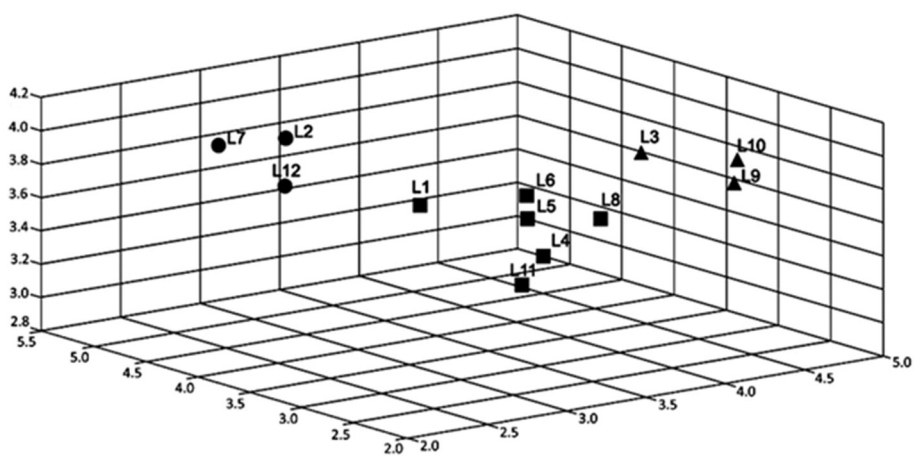

Figure 1. Results from the clustering procedure based on artificial neural network analysis of 20 sweet corn inbred lines, which evidenced three genetically differentiated groups.

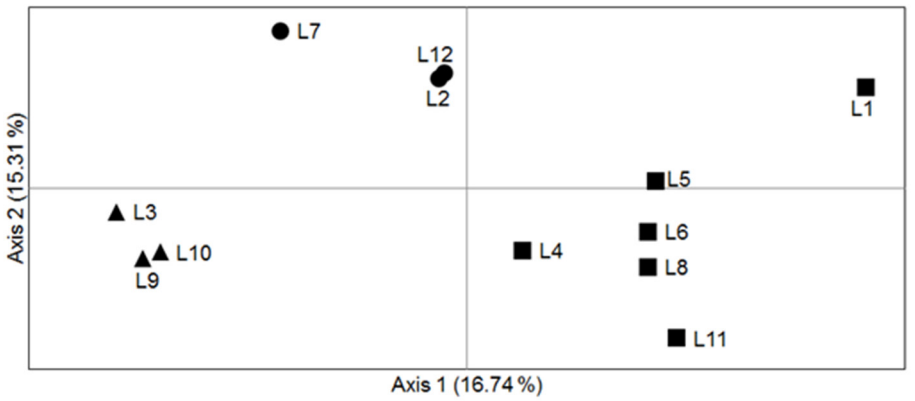

Figure 2. Principal coordinate analysis of genetic distances of 20 sweet corn inbred lines. Symbols denote inbred lines belonging to a particular genetically differentiated cluster, according to the artificial neural network model.

89.7% of the total variation in the locus allele frequency was due to genetic differences among the lines under study (Eloi et al. 2012). The genetic diversity found in this study was similar to that reported by Eloi et al. (2012) and Sserumaga et al. (2014). Forty percent of the loci had high PIC values (>0.6), indicating a great potential to detect differences among pure lines, confirming previous studies (Sserumaga et al. 2014).

Population structure analysis

The SOM algorithm, based on the allelic frequency of SSR loci, showed that the 12 parent lines formed three genetically differentiated groups (Figure 1). The first included half of the lines (6/12), and the second and third group contained three lines each. Similar to the findings reported by Kohonen (1998), the clustering results from the neural network agreed with those of the PCoA analysis (Figure 2). Groups 1 and 2 in the UPGMA dendrogram were strongly supported by bootstrapping, while group 3 had low bootstrap values, indicating a low level of confidence. Lines L1 and L5 were grouped differently by the UPGMA (Figure 3) than by the former methods. However, the RMSSTD value for SOM clustering was relatively lower than the RMSSTD value of UPGMA (0.36 and 0.38, respectively), indicating higher homogeneity in the SOM clusters. Similarly, the DB index was higher by the UPGMA method (DB=1.93), indicating lower precision than by the SOM algorithm (DB=1.82). These findings agree with a previous study of Peña-Malavera et al. (2014), who reported higher error rates of the UPGMA method than of SOM and PCoA, because it produces highly unbalanced clusters.

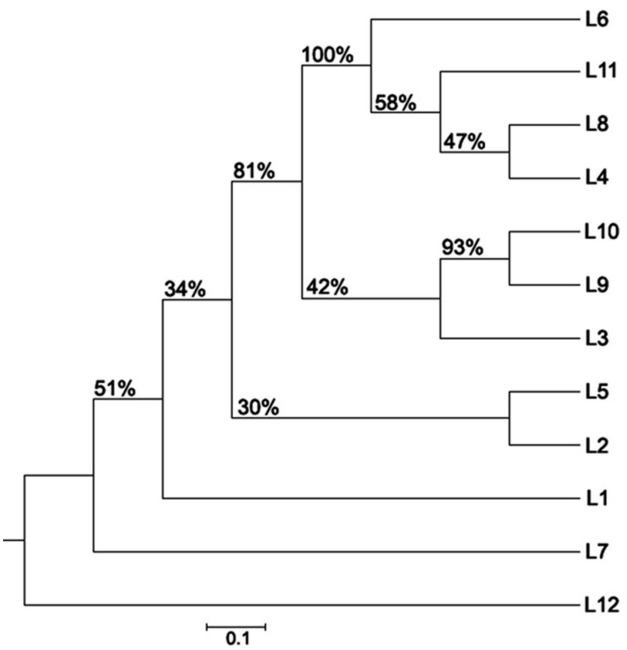

Figure 3. Cluster dendrogram of 20 sweet corn inbred lines, evaluated with SSR markers. The tree was constructed using the unweighted pair group method with arithmetic average (UPGMA) based on Nei's genetic distances. Values at the nodes indicate a percentage of 10,000 bootstrap runs supporting a particular node.

The clustering analysis using neural networks (via SOM) offers a faster alternative of identifying genetic clustering than the MCMC methods, as highlighted by Nikolic et al. (2009). As similarly found in previous studies (Barbosa et al. 2011, Peña-Malavera et al. 2014), neural networks have good adaptation to multi-allelic data and provide precise results in the identification of genetically differentiated groups. Finally, the high genetic differentiation detected among maize lines would allow the selection of promising divergent genotypes in the current breeding program of sweet corn.

REFERENCES

Amaral ATJ, Freitas ILJ, Guimarães AG, Maldonado C, Arriagada O and Mora F (2016) Bayesian analysis of quantitative traits in popcorn (*Zea mays* L.) through four cycles of recurrent selection. **Plant Production Science 19**: 574-578.

Ballesta P, Mora F, Ruiz E and Contreras-Soto R (2015) Marker-trait associations for survival, growth, and flowering components in *Eucalyptus cladocalyx* under arid conditions. **Biologia Plantarum 59**: 389-393.

Barbosa CD, Viana AP, Quintal SSR and Pereira MG (2011) Artificial neural network analysis of genetic diversity in *Carica papaya* L. **Crop Breeding and Applied Biotechnology 11**: 224-231.

Contreras-Soto RI, Mora F, de Oliveira MAR, Higashi W, Scapim CA and Schuster I (2017) A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. **PLoS ONE 12**: e0171105.

Davies DL and Bouldin DW (1979) A cluster separation measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence 2**: 224-227.

Don RH, Cox PT, Wainwright BJ, Baker K and Mattick JS (1991) 'Touchdown' PCR to circumvent spurious priming during gene amplification. **Nucleic Acids Research 19**: 4008.

Eloi IBO, Mangolin CA, Scapim CA, Gonçalves CS and Machado MFPS (2012) Selection of high heterozygosity popcorn varieties in Brazil based on SSR markers. **Genetics and Molecular Research 11**: 1851-1860.

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). **Cladistics 5**: 164-166.

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

Gawel NJ and Jarret RL (1991) A modified CTAB DNA extraction procedure for Musa and Ipomoea. **Plant Molecular Biology Reporter 9**: 262-266.

Goudet J (1995) FSTAT (vers. 1.2): a computer program to calculate F-statistics. **Journal of Heredity 86**: 485-186.

Grover R and Vriens M (2006) **The handbook of marketing research: uses, misuses, and future advances**. Sage Publ, London, 705p.

Kalinowski ST, Taper ML and Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. **Molecular Ecology 16**: 1099-1106.

Kohonen T (1998) The self-organizing map. **Neurocomputing 21**: 1-6.

Kulka VP, Da Silva TA, Contreras-Soto RI, Maldonado C, Mora F and Scapim CA (2018) Diallel analysis and genetic differentiation of tropical and temperate maize inbred lines. **Crop Breeding and Applied Biotechnology 18:** 31-38.

Lopes AD, Scapim CA, Machado MFPS, Mangolin CA, Silva TA, Cantagali LB, Teixeira FF and Mora F (2015) Genetic diversity assessed by microsatellite markers in sweet corn cultivars. **Scientia Agricola 72**: 513-519.

Mora F, Arriagada O, Ballesta P and Ruiz E (2017) Genetic diversity and population structure of a drought-tolerant species of Eucalyptus, using microsatellite markers. **Journal of Plant Biochemistry and Biotechnology 26**: 274-281.

Mora F, Castillo D, Lado B, Matus I, Poland J, Belzile F, Von Zitzewitz J and Del Pozo A (2015) Genome-wide association mapping of agronomic traits and carbon isotope discrimination in a worldwide germplasm collection of spring wheat using SNP markers. **Molecular Breeding 35**: 1-12.

Nikolic N, Park YS, Sancristobal M, Lek S and Chevalet C (2009) What do artificial neural networks tell us about the genetic structure of populations? The example of European pig populations. **Genetics Research 91**: 121-132.

Paini DR, Worner SP, Cook DC, De Barro PJ and Thomas MB (2010) Using a self-organizing map to predict invasive species: sensitivity to data errors and a comparison with expert opinion. **Journal of Applied Ecology 47**: 290-298.

Peakall R and Smouse PE (2012) GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research – an update. **Bioinformatics 28**: 2537-2539.

Peña-Malavera A, Bruno C, Fernandez E and Balzarini M (2014) Comparison of algorithms to infer genetic population structure from unlinked molecular markers. **Statistical Applications in Genetics and Molecular Biology 13**: 391-402.

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

Saavedra J, Silva TA, Mora F and Scapim CA (2013) Bayesian analysis of the genetic structure of a Brazilian popcorn germplasm using data from simple sequence repeats (SSR). **Chilean Journal of Agricultural Research 73**: 99-107.

Sserumaga JP, Makumbi D, Ji H, Njoroge K, Muthomi JW, Chemining'wa GN and Kim H (2014) Molecular characterization of tropical maize inbred lines using microsatellite DNA markers. **Maydica 59**: 267-274.

Werle AJK, Ferreira FRA, Pinto RJB, Mangolin CA, Scapim CA and Gonçalves LSA (2014) Diallel analysis of maize inbred lines for grain yield, oil and protein content. **Crop Breeding and Applied Biotechnology 14**: 23-28.