BAG | Journal of Basic & Applied Genetics

Vol XXXV (1): July 2024

Journal of the Argentine Society of Genetics

RANDOM FOREST IN PLANT GENETICS AND BREEDING: AN APPLICATION IN TOMATO AS A MODEL CROP

RANDOM FOREST EN GENÉTICA Y MEJORAMIENTO GENÉTICO DE PLANTAS: UNA APLICACIÓN EN TOMATE COMO CULTIVO MODELO

Faviere G.[1], Vitelleschi M.S.[2], Pratta G.R.[3]

[1] Instituto de Investigaciones Teóricas y Aplicadas en Estadística (IITAE), Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario, Rosario, Santa Fe, Argentina.

[2] IITAE, Consejo de Investigaciones de la Universidad Nacional de Rosario (CIUNR), Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario, Rosario, Santa Fe, Argentina..

[3] Instituto de Investigaciones en Ciencias Agrarias de Rosario (IICAR), Facultad de Ciencias Agrarias, Universidad Nacional de Rosario-CONICET, Zavalla, Santa Fe, Argentina

Corresponding author:
Guillermo R. Pratta
gpratta@unr.edu.ar

ORCID 0000-0002-3682-0946

Cite this article as:
Faviere G., Vitelleschi M.S., Pratta G.R. 2024. *RANDOM FOREST* IN PLANT GENETICS AND BREEDING: AN APPLICATION IN TOMATO AS A MODEL CROP. BAG. Journal of Basic and Applied Genetics XXXV (1).

Received: 07/13/2023
Revised version received: 03/13/2024
Accepted: 03/13/2024
General Editor: Elsa Camadro
DOI: 10.35407/bag.2024.35.01.03
ISSN online version: 1852-6233

Available online at
www.sag.org.ar/jbag

ABSTRACT

Random Forest approaches have been used in phenotyping at both morphological and metabolic levels and in genomics studies, but direct applications in practical situations of plant genetics and breeding are scarce. Random Forest was compared with Discriminant Analysis for its ability in classifying tomato individuals belonging to different breeding populations, exclusively based on phenotypic fruit quality traits. In order to take into account different steps in breeding programs, two populations were assayed. One was composed by a set of RILs derived from an interspecific tomato cross, and the other was composed by two of these RILs and the corresponding F_1, F_2 and backcross generations. Being tomato an autogamous species, the first population was considered a final step in breeding programs because promising genotypes are being evaluated for putative commercial release as new cultivars. Meanwhile, the second one, in which new variation is being generated, was considered as an initial step. Both Random Forest and Discriminant Analysis were able to classify populations with the aim of evaluating general variability and identifying the traits that most contribute to this variability. However, overall errors in classification were lower for Random Forest. When comparing the adequacy of classification between populations, errors of both statistical analyses were greater in the second population than in the first one, though Random Forest was more precise than Discriminant Analysis even in this initial step of plant breeding programs. Random Forest allowed breeders to get a reliable classification of tomato individuals belonging to different breeding populations.

Key words: discriminant analysis, Machine Learning, parametric and non-parametric classification techniques, phenotype identification, traits categorization

RESUMEN

Los enfoques de Random Forest se han utilizado en la fenotipificación, tanto a nivel morfológico como metabólico, y en estudios de genómica, pero las aplicaciones directas en situaciones prácticas de fitomejoramiento y genética son escasas. Random Forest se comparó con el Análisis Discriminante por su capacidad en la clasificación de individuos de tomate pertenecientes a diferentes poblaciones de mejoramiento, exclusivamente en función de los rasgos fenotípicos de calidad de la fruta. Para tener en cuenta los diferentes pasos en los programas de mejoramiento, se ensayaron dos poblaciones. Una estaba compuesta por un conjunto de RILs derivadas de un cruce interespecífico de tomate, y la otra estaba compuesta por dos de estas RILs y las correspondientes generaciones F_1, F_2 y retrocruzas. Siendo el tomate una especie autógama, la primera población se consideró un paso final en los programas de mejoramiento porque se están evaluando genotipos prometedores para su lanzamiento comercial putativo como nuevos cultivares. Mientras tanto, la segunda, en la que se está generando nueva variación, se consideró como un paso inicial. Tanto Random Forest como Análisis Discriminante pudieron clasificar poblaciones con el objetivo de evaluar la variabilidad general e identificar los rasgos que más contribuyen a esta variabilidad. Sin embargo, los errores generales en la clasificación fueron menores para Random Forest. Al comparar la adecuación de la clasificación entre poblaciones, los errores de ambos análisis estadísticos fueron mayores en la segunda población que en la primera, aunque Random Forest fue más preciso que el Análisis Discriminante incluso en este paso inicial de los programas de fitomejoramiento. Random Forest permitió a los criadores obtener una clasificación fiable de individuos de tomate pertenecientes a diferentes poblaciones de cría.

Palabras clave: análisis discriminante, Aprendizaje Automático, técnicas de clasificación paramétricas y no paramétricas, identificación de fenotipos, categorización de rasgos.

INTRODUCTION

In plant breeding, different populations are voluntarily created by crossing selected genotypes to obtain hybrids. Then, the genetic structure of these artificial populations may be predicted, which allows estimating the components of genetic mean values and variances underlying the traits to be improved (Kearsey and Pooni, 1996). Hence, various groups (families, generations) of objects (individuals) are available for evaluation, enabling the application of supervised classification to assess the generated genetic variability (Stephan *et al.*, 2015).

One of the first challenges that biological sciences must deal with is classification (Duda *et al.*, 2000), an inherent process in most human activities that consists in accurately and efficiently assigning a class or a type to a given object under study (Trainor *et al.*, 2017). Objects are considered as factors that are evaluated by a series of variables or attributes with the goal of constructing groups according to their similarities (Hastie *et al.*, 2008). Two principal approaches are distinguished in this common challenge: supervised and unsupervised classification. In the first one, a priori known groups of objects are assessed aiming to establish objective criteria through data analysis for predicting with low uncertainty the belonging of new objects to any of those groups (Alhusain and Hafez, 2017). In unsupervised classification, instead, the belonging of studied objects to a given group is unknown and the goal is to find the underlying structure of data according to similarities found during the assessment (Larose and Larose, 2015).

Tomato (*Solanum lycopersicum* L.) is one of the most important horticultural crops worldwide (FAOSTAT, 2017). Also, it is a model species for plant genetics and breeding by means of both conventional and advanced strategies (Gerszberg *et al.*, 2015). Phenotypic evaluation is essential at different steps of a breeding program, especially when variability for quantitative agronomic traits is increased by crosses to wild germplasm (Dempewolf *et al.*, 2017).

Discriminant Analysis is a parametric method widely used for classifying in biological and agronomic applications (Alhusain and Hafez, 2017) while the non-parametric classification techniques, such as Random Forest, becomes necessary in many studies (Singh *et al.*, 2016).

The objective of this research was to assess the use of Random Forest to classify populations with different genetic structure according to phenotypic variability for fruit quality traits in two different usual situations of plant breeding. The accuracy and robustness of Random Forest in identifying the desired genotypes and the proportional contribution of measured traits in defining such genotypes were compared with the results obtained by Discriminant Analysis.

MATERIALS AND METHODS

Plant populations and traits under study

Two populations were evaluated with the aim of considering two different plant breeding activities. Both of them represent genomic recombination among the same parental genotypes (cv. Caimanta of *S. lycopersicum* and LA0722 of *S. pimpinellifolium*) in extreme situation of linkage disequilibrium, genotypic composition and inbreeding level. The first population comprised eight of the 18 RILs obtained by Rodriguez *et al.* (2006), hereafter named as L1, L5, L6, L8, L9, L15, L17, and L18 (total N=396 plants, because some individuals were lost during the transplant. The final number of individuals per RIL is given in Table 1). These 8 RILs were selected for adequately representing the total variability. In this population, linkage disequilibrium is low (<0.01), inbreeding level is high (>0.99) and genotypes are homozygous, representing potential new tomato commercial cultivars derived after several cycles of artificial selection over both early and advanced generations of selfing from a cross between cultivated and exotic germplasms. Data analyzed in this research are the mean values over six years of agronomic evaluation because its genetic structure is stabilized and this population was considered as a final step in a breeding program.

The second population comprised two selected RILs (L1 and L18), their F_1 (L18 x L1) and its segregating generations F_2 (L18 x L1), obtained by selfing, and both backcrosses F_1 (L18 x L1) x L18 and F_1 (L18 x L1) x L1, hereafter named as F_1, F_2, BC_1 and BC_2, respectively (N=218 plants). In this population of basic breeding generations (Kearsey and Pooni, 1996), linkage disequilibrium is high (>0.20), inbreeding level is relatively low (coefficient F=0.5). Genotypes are both homozygous and heterozygous, representing early generations from a cross among elite genotypes. Gene segregation occurring from the meiosis in F_1 gives new opportunity of recombining and selecting over the genotypes resulting from the previous breeding actions that allowed deriving the parental RILs. In fact, L1 and L18 are registered in the Argentinean National Registry of Cultivars). Data analyzed in this research were measured just in one year of agronomic evaluation because the genetic structure of the segregating generations (F_2 and both BCs) varies with each cycle of selfing, and this population was considered as an initial step in a breeding program.

Both sets of population were assayed under greenhouse conditions at the experimental field station "José F. Villarino", Universidad Nacional de Rosario, Argentina (Latitude: 33.02° S, Longitude: 60.88° O, Altitude: 50 masl) according to a completely randomized design with six replications. Following Mahuad *et al.*

(2013), 11 quantitative traits were evaluated, five of them in fruits harvested at breaker stage (when carotenoids accumulation becomes visible) and the other six in fruits harvested at red ripe stage (with approximately 90% of red surface). In 10 fruits per plant at breaker stage, weight (W, in g), diameter (D, in cm), height (H, in cm), shape index (SI, ratio between H and D), and shelf life (SL, number of days from harvest until the fruit stored at 25 ± 3 ºC loses commercial value due to, for instance, excessive softening), were measured. In fruits at red ripe stage, the following traits were evaluated: soluble solids content (SS, in Brix degrees, as the percentage of sucrose in the fruit juice), pH and titratable acidity (TA, in g of citric acid per 100 g of homogenate) of the fruit juice, firmness (F, measured on two opposite equatorial sides with a digital firmness type Shore A tester Durofel, DFT 100, with a 0.10 cm^2 cap), ratio a/b or chroma index (parameter related to color tone, being "a" the absorbance at 540 nm wavelengths and "b" the absorbance at 675 nm wavelengths), and L value or reflectance percentage (L, parameter related to color intensity, presenting values that range from +100 for white to 0 for black). Values "a", "b" and L were determined with a Chroma Meter CR 400. The color parameters and F were determined in five intact fruits per plant, whereas the SS and the pH were measured in the juice obtained by homogenizing a variable number of three to eight fruits per plant, which depended on the fruit size. In the first set of populations, the mean locule number per fruit (LN) was also measured in five fruits per plant.

Statistical Analyses

Random Forest is a non-parametric classification technique of Machine Learning proposed by Breiman (2001). It is a classifier that generates a big number of decision trees, and each tree is grown from a bootstrap sample of the response variable. The best split is selected from a random subset of variables at each node of the tree, and then the tree grows to the maximum extent without pruning. Each individual is classified by each tree and the most common outcome is used as the final classification. For this classification, contribution of each variable to form the groups is assessed (Breiman, 2001). In plant genetics and breeding, the quantification of this contribution could be considered as a description of phenotypes according to the importance of traits and it could objectively assist in the identification of phenotypes in their belonging to a given breeding population. Random Forest applies a built-in cross-validation, which in this research consisted of a first training step with 2/3 of the data and a validation step with the remaining 1/3, according to Breiman (2001), to estimate set error via the use of Out-Of-Bag (OOB) samples. Firstly, based on the training algorithm, data that did not take part at a given iteration in the bootstrap sample (the so called OOB data) are predicted using the tree grown with the bootstrap sample to be assigned at a given group. This process is known as validation step, and finally, once each individual has been assigned to a given group according to the OOB predictions, the error rate, known as the OOB estimate of error rate, is calculated. In other words, low values of the OOB estimate of the error rate indicate a high precision in the classification (Hastie *et al.*, 2008). Furthermore, two measures on the importance of variables are given by Random Forest (Hastie *et al.*, 2008): the Mean Decrease Accuracy (MDA), which is obtained from the OOB error estimation, and the Mean Decrease Gini (MDG), based in the Gini Index. Hyperparameter tuning of RF greatly influences its accuracy but there is a large scientific discussion on how to accomplish it (Bernard *et al.*, 2009). In this research, the number of tree (NT), the percentage of errors in global classification (GC) associated to NT and the OOB were taken as metrics for tuning hyperparametrization (Probst *et al.*, 2019).

Machine Learning techniques have not been yet widely used in plant genetics and breeding with a similar goal to that of this research. Consequently, Discriminant Analysis, the most common multivariate technique (Lapins and Nash, 1957, Lynch *et al.*, 1987, Sivakumar *et al.*, 2017, Abu-Ellail *et al.*, 2020), was used as a control for group assignment and for measuring the importance of variable contribution detected by Random Forest. Every trait was included in all cases, i.e., no selection of variables was accomplished neither by Random Forest nor by Discriminant Analysis. Random Forest was performed by the statistical package RStudio, version 1.0 ("randomForest" library, randomForest function), while SAS software, version 9.0, was used for Discriminant Analysis, through the procedure "proc discrim". Significance of all statistical tests was assigned at a limit p-value of 0.05.

RESULTS

Population of RILs

Mean values and standard deviations for evaluated traits across six years of agronomic evaluation are presented in Figure 1. A wide phenotypic variability among RILs was found for all traits, though, as expected, the absolute range of variation depended on the scale of each variable. In respect to Random Forest application, three variables were randomly selected for determining each node in the iterative construction of each tree, this sub-conjunct of traits was used for choosing the best partition. The percentages of errors in global classification and in

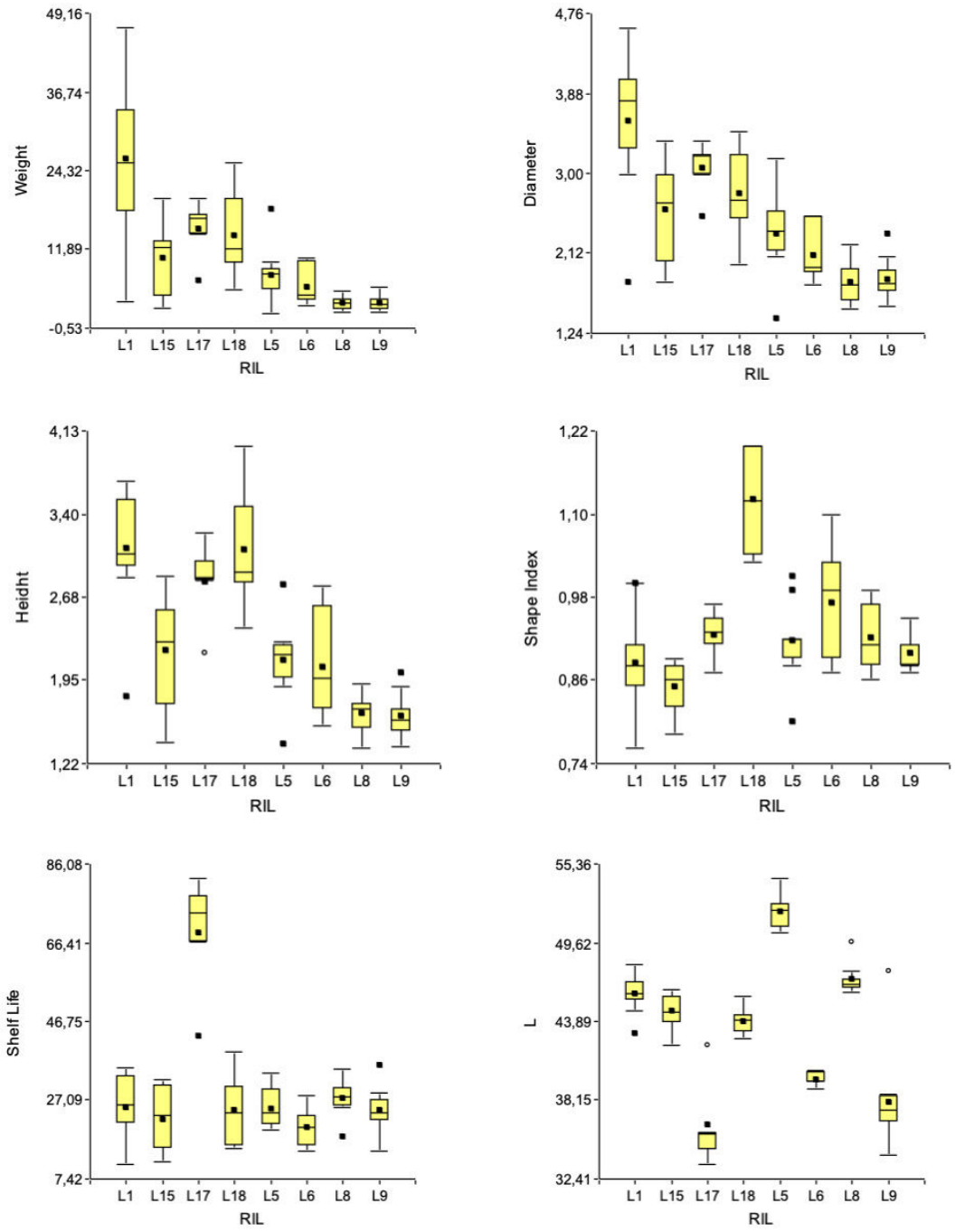

Figure 1 (continues). Values of the fruit traits weight (W, in g), diameter (D, in cm), height (H, in cm), shape index (SI, ratio H/D), shelf life (SL, in days), reflectance percentage (L, in %), chroma index (ratio a/b, being "a" the absorbance at 540 nm wavelengths and "b" the absorbance at 675 nm wavelengths), locule number (LN), soluble solids content (SS, in ºBrix), pH, titratable acidity (TA, in g of citric acid per 100 g of homogenate juice), and firmness (F, in %) in eight tomato RILs obtained by Rodriguez *et al.* (2006).

the classification per group (RIL) are shown in Table 1. From 200 constructed trees onwards, both classification errors were stabilized in a null value, hence there are none misclassifications after this number. As the most frequent number of trees constructions in the literature is at least 500, the 100% of plants were perfectly classified into the expected group, the OOB error being 0%. Therefore, as 0% OOB error has been obtained with just 200 constructed trees (Table 1), this RIL population was accurately classified by Random Forest. The most important traits to obtain this excellent classification were L, TA, pH, SS and F, according to their MDA and MDG values (Table 2).

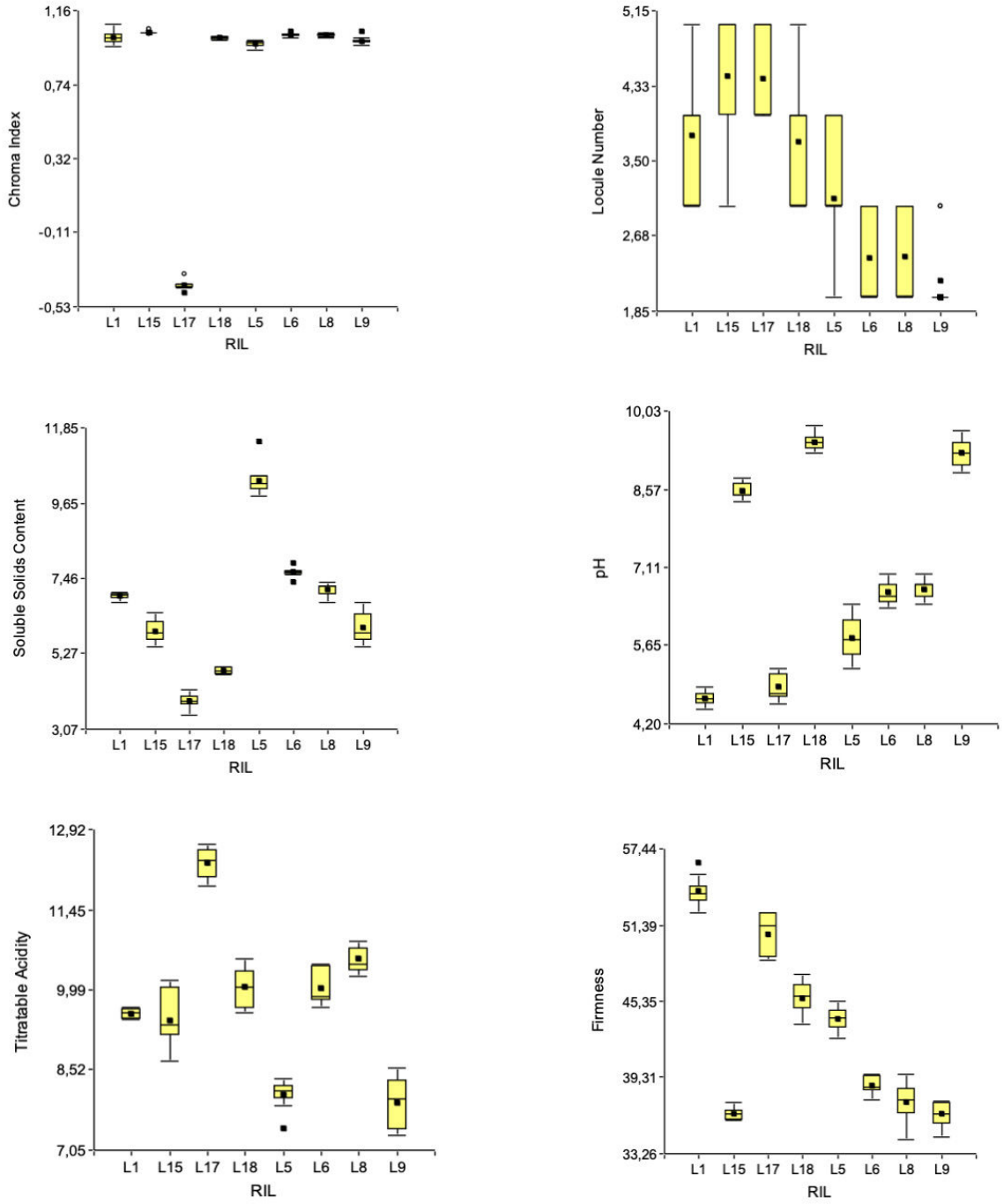

Figure 1 (continuation). Values of the fruit traits weight (W, in g), diameter (D, in cm), height (H, in cm), shape index (SI, ratio H/D), shelf life (SL, in days), reflectance percentage (L, in %), chroma index (ratio a/b, being "a" the absorbance at 540 nm wavelengths and "b" the absorbance at 675 nm wavelengths), locule number (LN), soluble solids content (SS, in ºBrix), pH, titratable acidity (TA, in g of citric acid per 100 g of homogenate juice), and firmness (F, in %) in eight tomato RILs obtained by Rodriguez *et al.* (2006).

With the control technique, Discriminant Analysis, six linear discriminant functions (LDF) were obtained that allowed classifying RILs and measuring the contribution of each trait to the total variability (Table 3, LDF 1 and LDF 2 are shown as footnotes). However, when it was contrasted with Random Forest, errors in misclassifying were greater in Discriminant Analysis (Table 3). This misclassification made by Discriminant Analysis could be explained by two reasons. Firstly, Rao's F test (p<0.0001, Table 3) indicated that there were at least two groups of RILs which have different average vectors or LDF, i.e., the eight RILs are not univocally different among them but they could be clustered in either two, three, four, five or six groups, all these groups being

Table 1. Percentage of errors in global classification (GC) and per RIL (indicated with L and the respective number, these genotypes are eight tomato RILs obtained by Rodriguez *et al.*, 2006) considering different number of trees (NT) in each applied Random Forest

NT	GC	L1 (54)*	L15 (42)	L17 (30)	L18 (42)	L5 (66)	L6 (42)	L8 (54)	L9 (66)
50	3.03	0	14.29	0	0	0	14.29	0	0
100	1.52	0	14.29	0	0	0	0	0	0
150	1.52	0	14.29	0	0	0	0	0	0
200	0	0	0	0	0	0	0	0	0
300	0	0	0	0	0	0	0	0	0
400	0	0	0	0	0	0	0	0	0
500	0	0	0	0	0	0	0	0	0

* Numbers in parenthesis indicate the final number of plants per RIL evaluated in the research

Table 2. Contribution of each fruit trait to the classification by Random Forest of eight tomato RILs obtained by Rodriguez *et al.* (2006), according to Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG). Fruit traits: weight (W, in g), diameter (D, in cm), height (H, in cm), shape index (SI, ratio H/D), shelf life (SL, in days), reflectance percentage (L, in %), chroma index (ratio a/b, being "a" the absorbance at 540 nm wavelengths and "b" the absorbance at 675 nm wavelengths), locule number (LN), soluble solids content (SS, in ºBrix), pH, titratable acidity (TA, in g of citric acid per 100 g of homogenate juice), and firmness (F, in %).

Fruit Trait	MDA	MDG
D	5.02	1.74
H	6.27	2.37
SI	6.37	3.01
W	6.69	2.53
SL	5.31	2.01
L	15.11	8.34
a/b	7.12	2.92
LN	4.39	1.14
SS	13.53	8.64
pH	13.86	8.31
F	13.25	7.20
TA	14.92	8.18

statistically significant. In fact, in Table 3 it is shown sequentially that all the obtained LDF were significant, i.e. classification RILs according to these six LDF is not robust. Secondly, traits which were identified as the most important in their contribution to general variability, varied in respect to those identified by Random Forest. For instance, a/b was identified by Discriminant Analysis as having the most important contribution to LDF 1 (footnote to Table3) while this trait had low contribution to general variability in the analysis with Random Forest (Table 1). Hence a lower robustness of classification is accomplished with Discriminant Analysis compared to Random Forest. In fact, differences in identifying traits contribution to RILs classification between Random Forest and Discriminant Analysis resulted in a wrong predicted assignation to RIL 15 of six plants actually belonging to RIL 9. This misclassification caused a global apparent error of 1.51% in Discriminant Analysis, while the global apparent error was null in Random Forest.

Population of Basic Generations

Mean values and standard deviations for the six basic generations evaluated are presented in Figure 2. Though some difference due to environmental influences were detected on the mean values of parental lines between both databases, general tendencies for morphological traits were observed since L1 had flattened fruits with higher weight and size than L18, whose fruits were elongated. Also, the F_1 phenotype agreed to gene actions reported by Pereira da Costa *et al.* (2014). For instance, the lower weight of F_1 fruits was explained by negative dominance of exotic alleles early contributed by LA0722 that are segregating in dispersion among L1 and L18 (Cabodevila *et al.*, 2021). In agreement, individuals from backcross to L1 (BC_2) had slightly heavier fruits than those of backcross to L18 (BC_1). Regarding variances, and as expected, the F_2 generation had a larger dispersion than both backcrosses for all traits; the F_1, genetically uniform, and the parents were the least variable generations.

The different genetic structure among generations in this set of populations, in contrast to the previous set in which all populations were homozygous, provoked some not unexpected effects on applying classification

Table 3. Significance of the Sequential Test (ST) for the Linear Discriminant Functions (LDF) in the Discriminant Analysis applied to eight tomato RILs obtained by Rodriguez *et al.* (2006)

Steps of ST	Null Hypothesis	Canonical Correlation (CC)	Square CC	F-value	*p*-value
1	None LDF is significant	0.99	0.98	103.01	<0.0001
2	Only LDF 1 is significant	0.99	0.98	60.71	<0.0001
3	Only LDF 1 and 2 are significant	0.98	0,96	41.61	<0.0001
4	LDF 1, 2, and 3 are significant	0.96	0,92	27.33	<0.0001
5	LDF 1, 2, 3, and 4 are significant	0.94	0.88	20.34	<0.0001
6	LDF 1, 2, 3, 4, and 5 are significant	0.83	0.66	12.00	<0.0001
7	LDF 1, 2, 3, 4, 5, and 6 are significant	0.72	0.52	9.46	<0.0001

The first two LDF, with standardized coefficients and traits, were:
LDF 1 = - 0.41 D + 0,89 H – 0.32 SI – 0.57 W + 0,17 SL – 0.03 L – 1.01 a/b + 0.09 LN – 0.34 SS – 0.30 pH + 0.02 F + 0.05 TA
LDF 2 = - 1.04 D + 0.84 H – 0.37 SI – 0.03 W + 0,003 SL – 0.04 L + 0.15 a/b + 0.13 LN + 0.37 SS – 0.80 pH + 0.40 F + 0.11 TA

techniques. Firstly, when including the F_2 generation either in Random Forest or in Discriminant Analysis, plants from all other generations were misclassified as belonging to the F_2 (data not shown). Even though differences in population size, common to plant breeding process, could explain this observation, the actual cause of this misclassification is the segregation of genes observed in the F_2, BC_1, and BC_2 generations while both parents and the F_1 are uniformly homozygote and heterozygote, respectively. Then the higher level of segregation and recombination among the F_2 in comparison to both BC_1 and BC_2 explains that even for not genetically uniform generation, all plant were classified as belonging to the F_2. Hence, data from F_2 generation were not taken into account for applying both classification techniques.

Similar to the previous study in RIL population, randomly selecting based on three variables was chosen for dividing the nodes in the iteratively construction of the trees during training the algorithm. However, Table 4 shows that, even considering 500 trees, errors in global classification and in classification per groups are not eliminated. Given that in 200 trees errors are stabilized in minimum values, this number is retained for continuing the analysis. However, the robustness of classification in this population is lower than in the RILs population previously analyzed according to this data respecting the number of trees construction. In fact, of the 120 evaluated plants, only 78 (65%), all belonging to the backcrosses generations (Table 5), could be adequately assigned to their respective group. Of these two backcrosses, BC_1 plants were better assigned (91% classified as BC_1 and 9% as BC_2) than BC_2 plants (76.6 % classified as BC_2, 17% as BC_1, and 6.4% as L1). All parental and F_1 plants were misclassified as either BC_1 or BC_2. The importance of the variables to classification algorithm explain these observations, given that SI (particularly higher in L18 and its BC_1) had the greatest MDA and MDG, i.e., the greatest contribution to construct the decision trees (Table 6). Other important traits were SS, D, and H, though their MDA and MDG values were low compared to SI.

In respect to the control technique Discriminant Analysis, the existence of at least two generations with different mean vectors or LDF was contrasted with Rao's F test (p<0.0001). Though four LDF were obtained in the analysis, the latter 2 were no significant in this population of basic generations. In fact, just the two first LDF, whose composition is shown as a footnote in Table 7, had strictly statistical significance. Traits mostly contributing to LDF 1 were D (1.97), H (-1.14), and a/b (-0.72), while again D (-1.04) and H (0.84) together with pH (-0.80) were the most important in LDF 2. Interestingly, SI was not detected as a highly important trait by Discriminant Analysis, possibly due to heteroscedasticity. Accordingly, Discriminant Analysis had lower ability than Random Forest for classifying plants, and only 67 over 120 plants (56%) were correctly assigned to their respective group (Table 8). However, though both techniques got a best classification for both backcrosses (72.7% in BC_1 and 53.9% in BC_2), Discriminant Analysis, in opposition to Random Forest could adequately assign some F_1 and L1 plants. Despite this observation, its estimated apparent error is very high (43.48%). Though many tomato plants from these basic breeding generations were misclassified by both

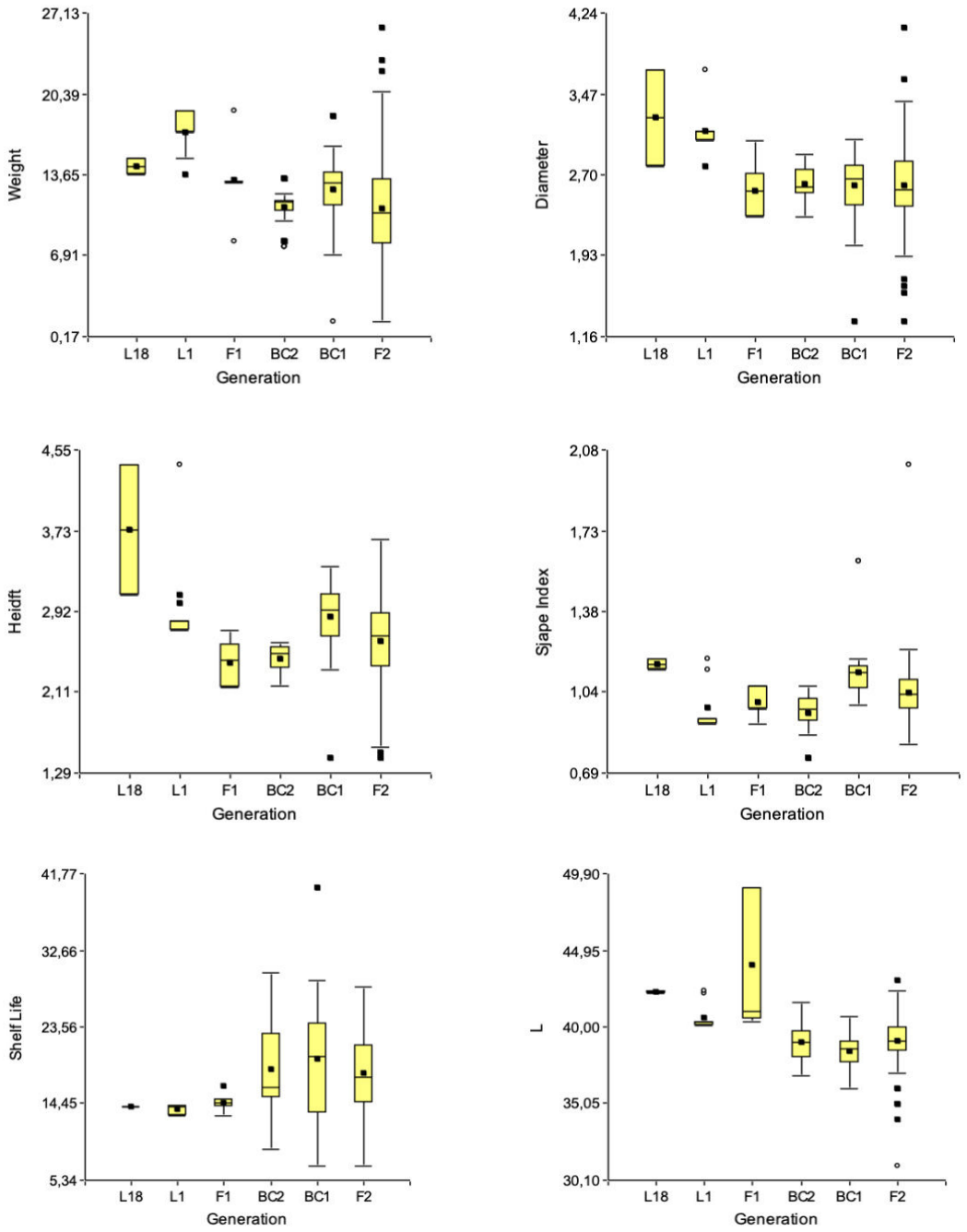

Figure 2 (continues). Values of the fruit traits weight (W, in g), diameter (D, in cm), height (H, in cm), shape index (SI, ratio H/D), shelf life (SL, in days), reflectance percentage (L, in %), chroma index (ratio a/b, being "a" the absorbance at 540 nm wavelengths and "b" the absorbance at 675 nm wavelengths), soluble solids content (SS, in ºBrix), pH, titratable acidity (TA, in g of citric acid per 100 g of homogenate juice), and firmness (F, in %) in the population composed by two parental tomato RILs (L18 and L1; obtained by Rodriguez *et al.* 2006), their F_1 (second cycle hybrid L18 x L1), and the segregating generations F_2 (obtained by selfing the F_1), BC_1 (F1 x L18) and BC_2 (F1 x L1).

techniques, it is noteworthy that just plant 7 from RIL 18 was assigned to a different group, concretely to BC_2 by Random Forest and to BC_1 by Discriminant Analysis. In all other cases, misclassification on individuals was to the same erroneous group (data not shown).

DISCUSSION

Random Forest and other Learning Machine approaches have been used in phenotyping at both morphological and metabolic level (Amit and Geman, 1997; Singh *et*

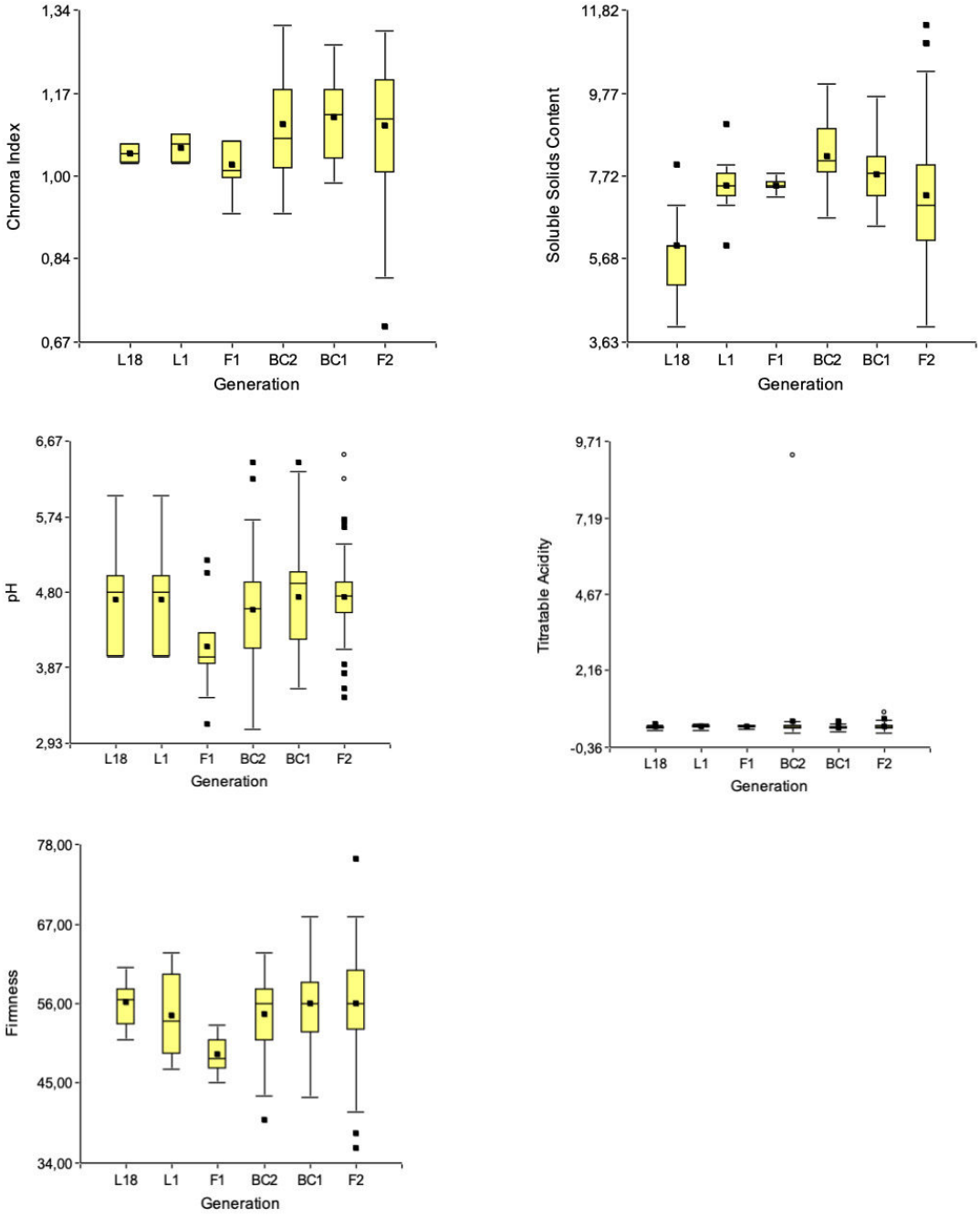

Figure 2 (continuation). Values of the fruit traits weight (W, in g), diameter (D, in cm), height (H, in cm), shape index (SI, ratio H/D), shelf life (SL, in days), reflectance percentage (L, in %), chroma index (ratio a/b, being "a" the absorbance at 540 nm wavelengths and "b" the absorbance at 675 nm wavelengths), soluble solids content (SS, in ºBrix), pH, titratable acidity (TA, in g of citric acid per 100 g of homogenate juice), and firmness (F, in %) in the population composed by two parental tomato RILs (L18 and L1; obtained by Rodriguez *et al.* 2006), their F_1 (second cycle hybrid L18 x L1), and the segregating generations F_2 (obtained by selfing the F_1), BC_1 (F1 x L18) and BC_2 (F1 x L1).

al., 2016; Zhao *et al.*, 2016, Trainor *et al.*, 2017) and in genomic studies (Chen and Ishwaran, 2012). However, the direct applications in practical situations of plant genetics and breeding, as were reported in this paper, have been infrequent. Though Random Forest was used in studies on wide genomic associations, detection of correlation among phenotypic traits and molecular markers and identification of different fruits, classification of breeding populations exclusively based in phenotypes is a vacant application (Biau, 2012; Chen

Table 4. Percentage of errors in global classification (GC) and per generation (L1: parental RIL 1, L18: parental RIL 18, F_1: second cycle hybrid L18 x L1, BC_1: backcross F_1 x L18, BC_2: backcross F_1 x L1, these genotypes are the six basic generations derived from RILs 1 and 18 obtained by Rodriguez *et al*, 2006, to initiate a breeding program) considering different number of trees (NT) in each applied Random Forest

NT	GC	F_1 (10)*	L1 (9)	L18 (8)	BC1 (46)	BC_2 (47)
50	41.30	100	100	100	18.18	47.06
100	32.61	100	100	100	9.09	35.29
150	30.43	100	100	100	9.09	29.41
200	28.26	100	100	100	9.09	23.53
300	28.26	100	100	100	9.09	23.53
400	30.43	100	100	100	9.09	29.41
500	30.43	100	100	100	9.09	29.41

* Numbers in parenthesis indicate the final number of plants per generation evaluated in the research

Table 5. Predicted classification of plants into groups (basic breeding generations L1: parental RIL 1, L18: parental RIL 18, F_1: second cycle hybrid L18 x L1, BC1: backcross F_1 x L18, BC_2: backcross F_1 x L1) in the training and validation of Random Forest (Method Out of Bag, OOB). L1 and L18 are tomato RILs obtained by Rodríguez *et al.* (2006)

Actual Group	Predicted Group					Total (%)
	F_1	L1	L18	BC_1	BC_2	
F_1	0 (0)	0 (0)	0 (0)	5 (50)	5 (50)	**10 (100)**
L_1	0 (0)	0 (0)	0 (0)	0 (0)	9 (100)	**9 (100)**
L18	0 (0)	0 (0)	0 (0)	4 (50)	4 (50)	**8 (100)**
BC_1	0 (0)	0 (0)	0 (0)	42 (91)	4 (9)	**46 (100)**
BC_2	0 (0)	3 (6.4)	0 (0)	8 (17.0)	36 (76.6)	**47 (100)**

and Ishwarab, 2012). Regarding this vacancy, a common but not desirable situation in plant breeding programs is the loss of identification of plant material in a given plot or genotype's mix when manipulating seeds, especially when exotic germplasm was introgressed (Dempewolf *et al.*, 2017). Hence the availability of reliable classification methods based on fast and easy to evaluate phenotypic traits and generated by supervised tools would be greatly advantageous. However, due to the different genetic structure of the various breeding populations coexisting in the same program, it is necessary to evaluate the adequacy of developing classification tools exclusively based on phenotypic variability using Random Forest for each types of population. Classification of individuals for improving their management in hybridization, recombination and selection, taking into account not only general variability but also traits mostly contributing to its conformation is a key step in breeding programs, as well as in defining the best phenotype for each situation (Niazian and Niedbała, 2020). According to our results, Random Forest was a better classifying technique than the most widely applied Discriminant Analysis. However, classification was better by either technique in final stages than in early stages of breeding programs. In final stages, when variability is dispersed and fixed among pure lines or other uniform population by effect of both artificial selection and inbreeding, best phenotypes appear to be more precisely classified than in early stages, when variability is created by crosses and recombination. Both the greater level of gene segregation and particularly the higher linkage disequilibrium of the population of basic generations, account for the less reliable classification obtained in it by both statistical methods. In consequence, the different genetic structure between both populations hinders the establishment of a robust algorithm for

Table 6. Contribution of each fruit trait to the classification with Random Forest of five breeding basic generations according to Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG). The genotypes are the basic generations derived from RILs 1 and 18 obtained by Rodriguez *et al*, 2006, to initiate a breeding program. Fruit traits: weight (W, in g), diameter (D, in cm), height (H, in cm), shape index (SI, ratio H/D), shelf life (SL, in days), reflectance percentage (L, in %), chroma index (ratio a/b, being "a" the absorbance at 540 nm wavelengths and "b" the absorbance at 675 nm wavelengths), soluble solids content (SS, in ºBrix), pH, titratable acidity (TA, in g of citric acid per 100 g of homogenate juice), and firmness (F, in %).

Traits	MDA	MDG
D	2.99	3.62
H	2.49	2.23
SI	5.26	6.08
W	0.18	2.69
SL	-0.79	1.65
L	0.84	2.28
a/b	0.59	1.78
SS	3.82	2.82
pH	-1.27	1.84
F	-1.52	1.19
TA	1.41	2.01

Table 7. Significance of the Sequential Test (ST) for the Linear Discriminant Functions (LDF) in the Discriminant Analysis applied to the population of basic generations derived from RILs 1 and 18 obtained by Rodriguez *et al*, 2006, to initiate a breeding program

Steps of ST	Null Hypothesis	Canonical Correlation (CC)	Square CC	F-value	*p*-value
1	None LDF is significant	0.86	0.73	3.06	<0.0001
2	Only LDF 1 is significant	0.73	0.53	2.19	0.0022
3	Only LDF 1 and 2 are significant	0.67	0.45	1.80	0.0632
4	LDF 1, 2, and 3 are significant	0.43	0.18	0.94	0.4968

The first two LDF, with standardized coefficients and traits, were:

LDF 1 = 1.97 D – 1.14 H + 0.14 SI – 0.50 W – 0.07 SL – 0.14 L – 0.72 a/b – 0.57 SS – 0.22 pH - 0.72 F + 0.15 TA

LDF 2 = 1.67 D + 0.14 H + 0.59 SI – 1.16 W – 0.26 SL + 0.83 L + 0.41 a/b - 0.44 SS – 0.06 pH + 0.26 F + 0.18 TA

categorizing earlier breeding step generations. However it allows a better classification in the final step (the population of RILs) since they represent different gene associations from the same single cross. In fact, RILs also had a high level of segregation (Pratta *et al*., 2011) but their linkage disequilibrium was low (Cambiaso *et al*., 2019). Though imbalance of data could also partially explain a less robust classification, in this research the number of individuals composing the unities of classification was certainly more noticeable in the population of basic generations than in the population of RILs but this is a common fact in plant breeding assays due to the different genotypic constitution of parents, F_1 and segregating F_2 and BCs.

The identification of variables most important for classification was more accurate and robust with Random Forest than with Discriminant Analysis, hence description of phenotypes was more precise using the first technique. Additionally, the genetic structure of a population may be assessed by multivariate methods, as Principal Component Analysis, whose main application is related to the characterization of general variability. Though this method is often used for classification, it is not adequate enough for categorizing. However, it can be used in a preliminary approach to reduce the data dimensionality and then apply classification methods such as Random Forest or Discriminant Analysis to obtain an appropriate classification (Hastie *et al*., 2008). Finally, it is interesting to point out that Discriminant Analysis is a parametric statistical technique while Random Forest is a non parametric one. One of the assumptions in Discriminant Analysis is that the variables come from a multivariate normal distribution. However, this assumption is not a requirement for Random Forest application, which becomes an additional advantage.

Random Forest was more accurate and robust than

Table 8. Predicted classification of plants into groups (breeding basic generations: L1: parental RIL 1, L18: parental RIL 18, F_1: second cycle hybrid L18 x L1, BC_1: backcross F_1 x L18, BC_2: backcross F1 x L1) in the Discriminant Analysis, RILs 1 and 18 were obtained by Rodriguez et al, (2006)

Actual Group	Predicted Group					Total (%)
	F_1	L1	L18	BC_1	BC_2	
F_1	0 (0)	0 (0)	0 (0)	5 (50)	5 (50)	**10 (100)**
L_1	0 (0)	0 (0)	0 (0)	0 (0)	9 (100)	**9 (100)**
L18	0 (0)	0 (0)	0 (0)	4 (50)	4 (50)	**8 (100)**
BC_1	0 (0)	0 (0)	0 (0)	42 (91)	4 (9)	**46 (100)**
BC_2	0 (0)	3 (6.4)	0 (0)	8 (17.0)	36 (76.6)	**47 (100)**

Discriminant Analysis for classifying tomato genotypes by phenotypic fruit quality traits at two different usual situations of plant breeding. Though a specific application such as identification of eventual unknown group of plants, was approached in the present research, a wide use of this technique in plant genetics and breeding can be proposed from these results. For instance, the evaluation of general variability, the identification of traits that most contribute to this variability, and even the definition of the best phenotype at different steps in breeding programs, are potential areas of application for Random Forest.

BIBLIOGRAPHY

Abu-Ellail F.F.B., Hussein E.M.A., El-Bakry A. (2020) Integrated selection criteria in sugarcane breeding programs using discriminant function analysis. Bull. Natl. Res. Cent. 44: 21-35. https://doi.org/10.1186/s42269-020-00417-6

Alhusain L., Hafez A.M. (2017) Cluster ensemble based on Random Forests for genetic data. BioData Min. 10: 101-125. https://doi.org/10.1186/s13040-017-0156-2

Amit Y., Geman D. (1997) Shape quantization and recognition with randomized trees. Neural Comput. 9: 1545-1588. https://doi.org/10.1162/neco.1997.9.7.1545

Bernard S., Heutte L., Adam S. (2009) Influence of hyperparameters on Random Forest accuracy. In: Benediktsson J.A., Kittler J. and Roli F. (Eds.) Multiple Classifier Systems. MCS 2009. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 171-180. https://doi.org/10.1007/978-3-642-02326-2_18

Biau G. (2012) Analysis of a Random Forests model. J. Mach. Learn. Res. 13: 1063-1095. https://dl.acm.org/doi/10.5555/2188385.2343682

Breiman L. (2001) Random Forests. Mach. Learn. 45: 5-32. https://doi.org/10.1023/A:1010933404324

Cabodevila V.G., Cambiaso V., Rodríguez G.R., Picardi L.A., Pratta G.R., Capel C., Lozano R., Capel J. (2021) A segregating F_2 population from a tomato second cycle hybrid allows the identification of novel QTL for fruit quality traits. Euphytica. 217: 453-461. https://doi.org/10.1007/s10681-020-02731-6

Cambiaso V., Giménez M.D., Pereira da Costa J.H., Vazquez D.V., Picardi L.A., Pratta G.R., Rodríguez G.R. (2019) Selected genome regions for fruit weight and shelf life in tomato RILs discernible by markers based on genomic sequence information. Breed. Sci. 69: 447–454. https://doi.org/10.1270/jsbbs.19015.

Chen X., Ishwaran H. (2012) Random Forests for genomic data analysis. Genomics. 99: 323-329. https://doi.org/10.1016/j.ygeno.2012.04.003

Dempewolf H., Baute G., Anderson J., Kilian B., Smith C., Guarino L. (2017) Past and future use of wild relatives in Crop Breeding. Crop Sci. 57: 1070-1082. https://doi.org/10.2135/cropsci2016.10.0885

Duda R., Hart P., Stork D. (2000) Pattern Classification. Wiley, Hoboken NJ, USA.

FAOSTAT (2017) https://www.fao.org/faostat/en/#data/QCL (accessed July 2021).

Gerszberg A., Hnatuszko-Konka K., Kowalczyk T., Kononowicz A.K. (2015) Tomato (*Solanum lycopersicum* L.) in the service of biotechnology. Plant Cell Tissue Organ Cult. 120: 881–902. https://doi.org/10.1007/s11240-014-0664-4

Hastie T., Tibshirani R., Friedman J. (2008) The elements of Statistical Learning. Data Mining, Inference, and Predictions. Springer, New York, NY, USA.

Kearsey M.J., Pooni H.S. (1996) The Genetical Analysis of Quantitative Traits. Chapman and Hall, London, UK.

Lapins K., Nash S.W. (1957) Discriminant function analysis in identification of peach varieties in nursery trees. Can. J. Plant Sci. 37: 12-25. https://doi.org/10.1007/BF02853700

Larose D., Larose C. (2015) Data Mining and Predictive Analytics. Wiley, Hoboken NJ, USA.

Lynch D.R., Schaalje G.B., Tai G.C.C., Young, D.A. (1987) Use of canonical discriminant analysis in assessing the merit of crosses in terms of breeding goals. Am. Potato J. 64: 385–395. https://doi.org/10.1007/BF02853700

Mahuad S.L., Pratta G.R., Rodriguez G.R., Zorzoli R., Picardi L.A. (2013) Preservation of *Solanum pimpinellifolium* genomic fragments in recombinant genotypes increased tomato fruit quality. J. Genet. 92: 195-203. https://doi.org/10.1007/s12041-013-0245-z

Niazian M., Niedbała G. (2020) Machine Learning for Plant Breeding and Biotechnology. Agriculture. 10: 615-640. https://doi.org/10.3390/agriculture10100436

Pereira da Costa J.H., Rodríguez G.R., Pratta G.R., Picardi L.A., Zorzoli R. (2014) Pericarp polypeptides and SRAP markers associated with fruit quality traits in an interespecific tomato backcross. Genet. Mol. Res. 13: 2539-2547. https://doi.org/10.4238/2014.January.24.10

Pratta G.R., Rodriguez G.R., Zorzoli R., Valle E.M., Picardi L.A. (2011) Phenotypic and molecular characterization of selected tomato recombinant inbred lines derived from a cross *Solanum lycopersicum* x *S. pimpinellifolium*. J. Genet. 90: 229-237. https://doi.org/10.1007/s12041-011-0063-0

Probst P., Wright M.N., Boulesteix A.L. (2019) Hyperparameters and tuning strategies for random forest. Data Min. Knowl. Discov. 9: 1-15. https://doi.org/10.48550/arXiv.1804.03515

Rodriguez G.R., Pratta G.R., Zorzoli R., Picardi L.A. (2006) Recombinant lines obtained from an interspecific cross among *Lycopersicon* species selected by fruit weight and fruit shelf life. J. Am. Soc. Hortic. Sci. 131: 651-656. https://doi.org/10.21273/JASHS.131.5

Singh A., Ganapathysubramanian B., Singh A.K., Sarkar S. (2016) Machine Learning for high-throughput stress phenotyping in plants. Trends Plant Sci. 21: 110-124. https://doi.org/10.1016/j.tplants.2015.10.015

Sivakumar V., Celine V.A., Venkata Ramana C. (2017) Discriminant function method of selection in vegetable cowpea genotypes. Int J. Curr. Microbiol. Appl. Sci. 10: 4954-4958. https://doi.org/10.20546/ijcmas.2017.610.469

Stephan J., Stegle O., Beyer A. (2015) A random forest approach to capture genetic effects in the presence of population structure. Nat. Commun. 6: 7432-7442. https://doi.org/10.1038/ncomms8432

Trainor P.J., De Filippis A.P., Rai S.N. (2017) Evaluation of classifier performance for multiclass phenotype discrimination in untargeted metabolomics. Metabolites. 21: 742-762. https://doi.org/10.3390/metabo7020030.

Zhao J., Bodner G., Rewald B. (2016) Phenotyping: using Machine Learning for improved pairwise genotype classification based on root traits. Front. Plant Sci. 7: 1083-1100. https://doi.org/10.3389/fpls.2016.01864

ACKNOWLEDGEMENTS

Authors are grateful to Drs. Sabina Mahuad and Victoria Cabodevila for generating phenotypic data during their respective Doctoral Thesis, and to Agencia Nacional de Promoción Científica y Tecnológica, Ministerio de Ciencia y Tecnología, Argentina, for financial supporting.

AUTHOR CONTRIBUTION STATEMENT

Conceptualization: Pratta G.R. Data curation: Vitelleschi M.S., Pratta G.R. Formal analysis: Vitelleschi M.S., Faviere G. Funding acquisition: Pratta G.R. Investigation: Vitelleschi M.S., Faviere G., Pratta G.R. Methodology: Vitelleschi M.S. Project administration: Vitelleschi M.S., Pratta G.R. Resources: Vitelleschi M.S., Faviere G. Supervision: Vitelleschi M.S., Pratta G.R. Writing-original draft: Pratta G.R. Writing-review & editing: Vitelleschi M.S., Pratta, G.R.