Euphytica (2021) 217:195
https://doi.org/10.1007/s10681-021-02926-5

Relative performance of cluster algorithms and validation indices in maize genome-wide structure patterns

María Eugenia Videla · Juliana Iglesias · Cecilia Bruno

Received: 3 June 2021 / Accepted: 16 September 2021 / Published online: 29 September 2021

Abstract A number of clustering algorithms are available to depict population genetic structure (PGS) with genomic data; however, there is no consensus on which methods are the best performing ones. We conducted a simulation study of three PGS scenarios with subpopulations k = 2, 5 and 10, recreating several maize genomes as a model to: (1) compare three well-known clustering methods: UPGMA, k-means and, Bayesian method (BM); (2) asses four internal validation indices: CH, Connectivity, Dunn and Silhouette, to determine the reliable number of groups defining a PGS; and (3) estimate the misclassification rate for each validation index. Moreover, a publicly available maize dataset was used to illustrate the outcomes of our simulation. BM was the best method to classify individuals in all tested scenarios, without assignment errors. Conversely, UPGMA was the method with the highest misclassification rate. In scenarios with 5 and 10 subpopulations, CH and Connectivity indices had the maximum underestimation of group number for all cluster algorithms. Dunn and Silhouette indices showed the best performance with BM. Nevertheless, since Silhouette measures the degree of confidence in cluster assignment, and BM measures the probability of cluster membership, these results should be considered with caution. In this study we found that BM showed to be efficient to depict the PGS in both simulated and real maize datasets. This study offers a robust alternative to unveil the existing PGS, thereby facilitating population studies and breeding strategies in maize programs. Moreover, the present findings may have implications for other crop species.

Keywords Unsupervised learning · Population genetic structure · Multivariate technique · Outcome misclassification · SNPs · Maize

M. E. Videla · C. Bruno (✉)
Estadística y Biometría. Facultad de Ciencias Agropecuarias (FCA), Universidad Nacional de Córdoba, Córdoba, Argentina
e-mail: cebruno@agro.unc.edu.ar

M. E. Videla · C. Bruno
Unidad de Fitopatología y Modelización Agrícola, Consejo Nacional de Investigaciones Científicas y Tecnológicas (UFyMA -CONICET), Córdoba, Argentina

M. E. Videla
Universidad Nacional de Villa María, Córdoba, Argentina

J. Iglesias
Estación Experimental Pergamino, INTA, Instituto Nacional de Tecnología Agropecuaria, CC 31, B2700WAA Pergamino, Buenos Aires, Argentina

J. Iglesias
UNNOBA, Universidad Nacional del Noroeste de La Provincia de Buenos Aires, Monteagudo 2772, B2700WAA Pergamino, Buenos Aires, Argentina

Introduction

The genetic diversity of a group of individuals can be exhaustively characterized in different species (Becerra and Paredes 2000) using new technologies that allow us to evaluate thousands of genomic variants simultaneously (González-recio et al. 2014; Baloch et al. 2017). In a group of individuals that have been molecularly characterized, those more similar in their genetic profile are expected to have some degree of relatedness and, therefore, to be able to group, defining populations or genetic groups (Peña-Malavera et al. 2014; Vittorazzi et al. 2018). Single Nucleotide Polymorphism (SNP) markers have gained importance to explain a great proportion of the variance among individuals, and are the markers most widely used to identify genetic similarity patterns because they are very abundant in the genome (Baloch et al. 2017). This variability among individuals of a single population, generating internal groups or subgroups, may be due to very diverse causes, including gene flow, dispersion, introgression or mutations (Dutheil 2020). Thus, genotypes with the same genetic similarity are expected to form groups representing patterns or a population genetic structure (PGS).

From a more general perspective, the existence of a PGS implies the occurrence of different relatedness levels or genetic similarity among some subgroups within a single sample or population. Regardless of the drivers of its formation, it is necessary to identify the PGS and quantify its magnitude because the obtained information can be incorporated in subsequent statistical data analyses. Specifically in the context of phenotype-genotype association studies, knowing the PGS is important so it can be included in Genome Wide Association Studies (GWAS) models, since it has been proven that its presence reduces false positive rates (Malosetti et al. 2007). False positives occur when a molecular marker is mistakenly associated with the variation of a phenotype. For instance, in the genome selection context, PGS is a key factor affecting predictions of genetic values. For this reason, neglecting it might lead to unrealistic assessments of precision (Windhausen et al. 2012; Riedelsheimer et al. 2013) and to preferential selection of individuals within a single subpopulation (Isidro et al. 2015). Exploring the number of genetic groups within a set of individual genotypes and assigning individuals to groups has become an essential task in population genetics studies (Beugin et al. 2018) as well as in other areas, such as plant breeding, in which the phenotypic information is complemented with genotypic data (Thorwarth et al. 2017; Haile et al. 2018; Yuan et al. 2020).

In order to exploit the genetic relationship among individuals, a large number of multivariate methods have been proposed for the automatic identification of subgroups within populations. However, providing tools that can accurately identify those patterns is a methodological-statistical challenge in the context of massive genomic data with thousands of individuals. Multivariate analyses have been used for decades to obtain diverse types of information from genetic data (Bruno and Balzarini 2010; Jombart et al. 2010). In particular, geometric clustering algorithms, which group genotypes based on the pairwise genetic distances, have been used without assuming a specific population model (Bruno and Balzarini 2010; Beugin et al. 2018). These methods are usually fast and produce comparably accurate results in a variety of simulation scenarios (Peña-Malavera et al. 2014). On the other hand, there are other proposals to find PGS based on genetic models. Jombart et al. (2010) evaluated the discriminant analysis of principal component (DAPC) using four simulated genetic population models: an island model, a hierarchical islands model, a one-dimension hierarchical stepping stone, and a standard one-dimension stepping stone. This geometric approach method that considers population genetic models was found to be as accurate as Bayesian models, and computationally faster and more suitable to disentangle the underlying structure in complex population genetic models. However, it was difficult to interpret from the biological perspective, since it did not provide probabilities of group membership (Legendre and Legendre 2012), i.e., it did not allow the differentiation between a well-defined population structure and one with weak separation between groups (Jombart et al. 2010).

On the other hand, clustering is a multivariate technique used to classify objects or cases into relative groups named clusters. This technique is also known as unsupervised learning or exploratory data analysis. Cluster analysis aims to group similar observations into a number of clusters based on the observation values obtained from several variables for each individual. Clustering goal is ubiquitous in pattern recognition and similar in concept to discriminant analysis; however, in the discriminant analysis, the group to which an individual belongs is known a priori, whereas in a cluster analysis, that information is unknown. Thus, the discriminant analysis is called supervised learning cluster analysis, unsupervised learning. Cluster analyses can be divided into hierarchical and non-hierarchal. Hierarchical methods are frequently used because they are available in many software tools and can be applied directly to molecular data by selecting a suitable distance metric with no need to know the existence of groups of individuals a priori (Bruno and Balzarini 2010; Odong et al. 2011). Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Sokal 1958) and Ward's method (Ward 1963) are the hierarchical clustering methods most widely used with molecular marker data (Balzarini et al. 2011), whereas the k-means and k-medoids methods are regularly used non-hierarchical algorithms, which are frequently applied to detect PGS (Lee and Tracy 2009). Besides hierarchical and non-hierarchical algorithms, there are "model-based" approaches that estimate the probability that a set of genotypes are related by descent from a common ancestor. These approaches consider a population genetic model to calculate the probabilities that an individual belongs to a group. These methods are usually more computing-demanding but of easier biological interpretation than the geometric approach method, since they estimate the probabilities that an individual belongs to different groups simultaneously on the basis of its genotypes. Then, the highest value of probability of an individual to belong to a particular group genuinely reflects the probability that the individual belongs to that group (Beugin et al. 2018). However, for some individuals, the probability of belonging to a group can be ambiguous. An individual may have similar values of probability of belonging to different groups, and the researcher has to make the decision of assigning it to one or another group. Pritchard et al. (2000) proposed a clustering method based on Bayesian models and Markov chains, which is implemented in the software STRUCTURE (Pritchard et al. 2000) and in the Landscape and Ecological Association studies (LEA) package in R software (Frichot and François 2015). Later, Raj et al. (2014) proposed an upgrading of the software STRUCTURE, which they named fastSTRUCTURE, with the aim of providing efficient algorithms for proximal inference of the underlying model, using a faster variational Bayesian frame than STRUCTURE. These variational algorithms are almost two orders of magnitude faster than STRUCTURE and achieve accuracies comparable to the method proposed by Alexander et al. (2009), included in the software ADMIXTURE. In ADMIXTURE, genetic structure is obtained using a Bayesian approach, based on a Markov Chain Monte Carlo (MCMC) approach for a posteriori sampling distribution. It uses the same probability model as STRUCTURE, but is focused on obtaining a maximum likelihood rather than on sampling the distribution a posteriori. Since high dimensional optimization is much faster than MCMC, this maximum likelihood approach can be more efficient in a context of high density of molecular markers such as SNP (Alexander et al. 2009).

In summary, while in the unsupervised algorithms (i.e. hierarchical method) the number of groups is not a prerequisite, many other clustering algorithms (non-hierarchical, model-based methods) do require that the number of clusters be known beforehand. To overcome this problem, various cluster validation indices have been proposed from several disciplines to select the optimum number of groups (Peng et al. 2012). Some examples include the Dunn index (Dunn 1974), CH (Caliński and Harabasz 1974), the H statistic (k) (Hartigan 1975), the Silhouette statistic (Kaufman and Rousseeuw 1990), the gap statistic (Tibshirani et al. 2001), the Clest resampling method (Dudoit and Fridlyand 2002), the L method (Salvador and Chan 2004), and the Connectivity index (Handl and Knowles 2005). These internal validation indices allow us to measure the quality of a clustering. Hence, a clustering algorithm can be run several times, with a different number of clusters in each run, and the clustering that optimizes the considered index is selected as the final result (Günter and Bunke 2003). The aim of our work was to evaluate the relative behavior of three clustering methods that work under different computational algorithms to identify a PGS and four internal validation indices that determine the optimum number of clusters in high-dimensional genomic data. Thus, the following clustering methods were tested: Unweighted Pair Group Method with Arithmetic Mean (UPGMA) hierarchical method, k-means non-hierarchical method, and the Bayesian Method (BM) approach. In addition, to evaluate the goodness of a clustering structure based on intrinsic information of the dataset rather than on external

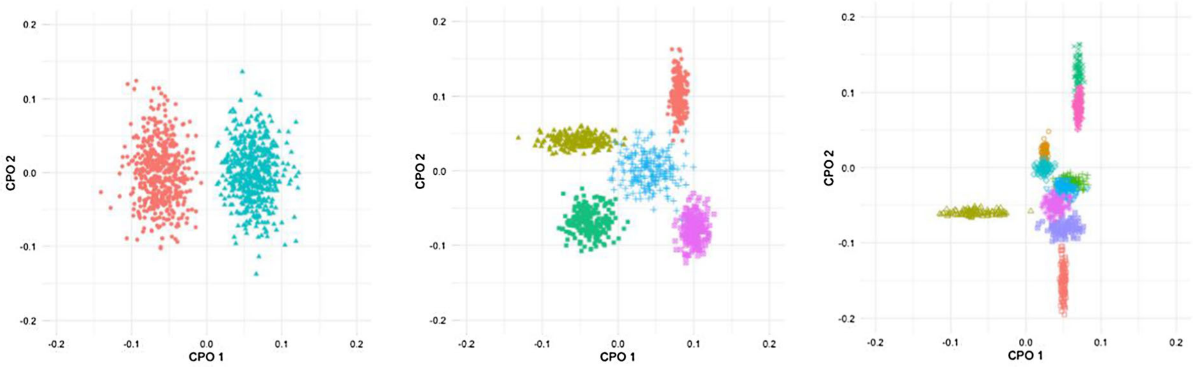

Fig. 1 Scatter plot of the Principal Coordinate Analysis of a molecular data simulation of 1000 individuals genotyped with 80 K SNPs for three simulation scenarios differing in the number of k-groups: k = 2 (left), k = 5 (center), and k = 10 (right). Each individual is represented by a dot. Individuals belonging to the same group are represented with the same symbols

information, the following internal validation indices were selected: CH, Connectivity, Dunn and Silhouette. All of these indices can also be used on binary data from SNPs. In order to assess the impact of the number of subpopulations on the performance of the different proposed methods and indices, a large number of simulations were performed using X-breed software on R. Then, we provided an illustration based on a publicly available maize dataset from Mazaheri et al. (2019a, b) to depict the outcome obtained by our simulation study in a real dataset.

Materials and methods

Simulation dataset design

The simulated datasets used for clustering were obtained using the package “Xbreed” in R, which was developed for performing genomic and phenotypic simulation (Esfandyari and Sørensen 2019). We simulated an SNP database for diploid individuals using a historical population. With the aim to achieve a desired level of linkage disequilibrium (LD), the following genetic parameters introduced in the simulation code were set to recreate a maize population as an example: number of individuals of the initial population, number of molecular markers, number of generations, mutation rate, and narrow-sense heritability. Based on these parameters, Xbreed allowed us to simulate the PGS of the historical population according to the phenotypic performance of individuals. The Xbreed simulation routine performs random crossings between individuals from a historical population to create new generations. Different phenotypic variances were set up with the aim of distinguishing subpopulations in a recent population. Then, individuals were randomly selected from the different simulated subpopulations, and different datasets (scenarios) with different numbers of subpopulations were generated. We obtained 300 datasets of 80 K SNPs, each containing 1000 individuals, arranged in three PGS scenarios defined by different numbers of population groups: Scenario 1, with two subpopulations or k = 2 (S1); Scenario 2, with k = 5 (S2), and Scenario 3, with k = 10 (S3). Each scenario consisted of 100 replicates. The level of genetic differentiation (Fst) obtained between subpopulations of each scenario was 0.03, which is regarded as a low (Latch et al. 2006). The Principal Coordinate Analysis was run to order the individuals in the optimal space defined by the first two principal coordinates and shows the level of differentiation achieved between populations for each scenario. Figure 1 shows one replicate per scenario of the result of the Principal Coordinate Analysis. SNP databases were coded according to the minor allele as binary data, i.e., the most frequent homozygous allele was coded as 0, the heterozygous allele as 1, and the least frequent homozygous allele as 0.

Table 1 Genetic divergence value (F_{ST}) among 11 subpopulations of a dataset published by Mazaheri et al. (2019a) for 942 inbred lines of maize genotyped with 899,784 SNP molecular markers

	Broad origin-public	IDT	NSS-Mo17	NSS-Oh43	Popcorn	SS-B13	SS-B37	SS-B73	SS-BSSSC0	Sweet corn
IDT	0.019									
NSS-Mo17	0.018	0.037								
NSS-Oh43	0.009	0.025	0.019							
Popcorn	0.008	0.030	0.024	0.016						
SS-B13	0.020	0.040	0.042	0.030	0.032					
SS-B37	0.014	0.033	0.032	0.022	0.023	0.025				
SS-B73	0.023	0.044	0.047	0.036	0.038	0.035	0.027			
SS-BSSSC0	0.010	0.028	0.029	0.019	0.019	0.021	0.016	0.022		
Sweet corn	0.007	0.028	0.026	0.016	0.011	0.029	0.022	0.035	0.018	
Tropical	0.007	0.026	0.024	0.015	0.012	0.027	0.020	0.031	0.017	0.013

Real maize dataset used to illustrate the simulation outcome

The compared algorithms, the clustering validation methods, and a real dataset published by Mazaheri et al. (2019a, b) were used for illustration purposes. The aim of that work was to detect candidate genes associated with the stalk biomass (plant height and stalk diameter) and the stalk anatomy (rind thickness, vascular bundle density and area) of maize; for that purpose, it was necessary to characterize the underlying PGS. We used that PGS information to illustrate our results obtained by simulation and compare them with the ones obtained by Mazaheri et al. (2019b).

For that study, the authors used a panel of 942 inbred lines (WiDiv-942), which included a diverse set of public, expired plant variety protection (exPVP), and germplasm enhancement of maize (GEM)-derived inbreds. This panel is representative of the main North American field corn heterotic groups, including stiff stalk, non-stiff stalk, and Iodent, as well as sweet corn, popcorn, and tropical inbreds (Mikel and Dudley 2006). A total of 899,784 SNPs were identified from whole seedlings of each member of the WiDiv-942 panel. The software program Admixture 1.23 (Alexander et al. 2009) was used by the authors to classify the WiDiv-942 panel into subpopulations using a subset of 93,991 SNPs that were pruned based on a pairwise LD threshold of r^2 = 0.1. No additional information on the choice of this set of SNPs was provided by the authors. Furthermore, each subpopulation was labeled based on the pedigree of the majority of the inbreds within each subpopulation. According to this classification, the panel was divided into a total of 11 subpopulations (k = 11), with four subpopulations matching with stiff stalk (SS) heterotic patterns, two matching with non-stiff stalk subpopulations (NSS), one subpopulation having the same pattern as that of broad origin-public lines, one matching with Iodent subpopulation (IDT), one subpopulation matching with sweet corn, one matching with popcorn subpopulation, and one subpopulation matching with tropical inbreds. A total of 201 inbreds with less than 0.5 of probability of belonging to any of the subpopulations were classified as a "mixed" group. Thus, we considered the existence of 11 clusters as gold standard to illustrate the result of the simulations. The average genetic diversity among subpopulations that make up this set of public data was characterized by estimating the Fst statistic of Wright (1949). On average, the 11 subpopulations were differentiated by an Fst of 0.0239 ± 0.009. The least divergent populations were Sweet corn and the broad-origin public lines, with an Fst of 0.007, whereas the most differentiated ones were SS-B73 and NSS-Mo17, with an Fst of 0.047 (Table 1). In this study, we used the whole panel of SNPs provided by Mazaheri et al. (2019a, b), i.e., the 899,784 SNPs, to characterize the PGS.

Comparison of cluster altgorithms

Three clustering methods belonging to different families of methods and differing notably in their computational algorithms were selected: Unweighted Pair Group Method with Arithmetic Mean (UPGMA) hierarchical method (Sokal 1958), the non-hierarchical k-means (MacQueen 1967), and the Bayesian Method (BM) approach (Frichot and François 2015). To evaluate the performance of the clustering methods, each one was implemented to identify, in each scenario, the number of subpopulations expected assumed in the simulation. As a criterion for comparing methods, the misclassification rate (MCR) was calculated for each replicate of each simulation scenario. Misclassifications occur when the method classifies an individual in a different subpopulation from the simulation one. The MCR was obtained from the confusion matrices between the cluster of origin (simulated subpopulation) and the classifier vector obtained as a result of each method. Then, summary statistics (mean and standard deviation) of MCR were calculated in each scenario for each dataset.

The hierarchical clustering analysis is one of the classification techniques most widely used to analyze sampling data based on various loci. In this analysis, each object first belongs to an individual cluster; then, after successive iterations, groups are merged until stop conditions are reached. These hierarchical algorithms do not require the number of groups as initial parameter; instead, a distance metric and a clustering method need to be selected. In this sense, there are different metrics for the different types of data. In this work, we used the complement to one function of the Jaccard similarity index $(1-J)$ to transform the similarity index into a distance between pairs of individuals i and j (d_{ij}^{J}). Thus, the distances between individuals i and j can be estimated as $d_{ij}^{J}=1-J$, where J is the Jaccard similarity index based on number of matches of heterozygous alleles (1) between individuals. Then, the Jaccard similarity index is expressed as $J=a/(a+b+c)$, with a, b and c representing the cell counts of a crossed contingency table that collects the information through the genomic profiles of individuals i and j, where a is the number of copresences of SNP markers and b and c are the number of times the marker was present in one individual and absent in the other, according to the coding of the SNP marker for this work. The Jaccard index is appropriate and recommended for estimating similarity/dissimilarity between two individuals when their characteristics have been measured with binary data (Bruno et al. 2003). Thus, the UPGMA algorithm is applied to a distance matrix. Selection starts with two elements that are at the shortest distance (the nearest ones), forming a class that will continue together in the following steps of the algorithm. If we consider an initial partition in which each individual is a class, then we can express that partition as $P_1=\{x_1\},\ldots,\{x_n\}$, then, if $\mathrm{IJ}=\{x_i,x_j\}$such that$d(x_i,x_j)$is minimun$\forall i,j=1,\ldots,n|i\neq j$, we have a new partition $P_2=\{x_1\},\ldots\{x_i,x_j\},\ldots,\{x_n\}$. Distances between the new class and the remaining observations are calculated as the mean of the distances between all the pairs of observations of two different classes before the merge:

$$d(IJ;x_k)=\frac{d(x_i,x_k)+d(x_j,x_k)}{2}\qquad k=1,\ldots,n$$

The algorithm stops when it achieves the final partitioning, which has all the observations $(P_r=\{N\})$. This procedure allows us to classify the objects under study and their grouping into clusters, such that the objects within a single group are more similar to one another than the objects belonging to different groups (Bruno et al. 2003). The clusters are visualized using a dendrogram that allows us to identify the clustering structure. However, when a high number of individuals are to be grouped, interpreting this diagram may be difficult. For this work, we used the function *hclust* of the package *stats* with the method *average* implemented in the algorithm UPGMA and the *Jaccard* distance in R (R Core Team 2019).

On the other hand, the partitioning approaches, such as the k-means algorithm, have been widely used (Rendón and Abundez 2016). Unlike UPGMA, K-means does require the number of groups (k) and the distance metric as initial parameters. First, each one of the p measurements taken from the sample of n observations: $x_{ij}(i=1,\ldots,n;\ j=1,\ldots p)$ is associated with one of the k groups $(k\leq n)$ according to the distance of each point to the centroid of each cluster. That is, let $x_{11},x_{12},\cdots,x_{np}$ be a random sequence of points (vector) in the sample (X_N), each point was selected independently of the preceding one using a fixed probability measure *pr*. Thus, $Pr(x_{11} \in A) = pr(A)$ and $Pr\left(x_{n+1} \in A|x_{11,}x_{12}, \ldots, x_{np,}\right) = pr(A)$, $np = 1, 2, \cdots$, for any measurable A value set in X_N. Relative to a given k-group $z = (z_1, z_2, \cdots, z_k)$, $z_i \in X_N$, $i = 1, 2, \cdots, k$, we define a minimum distance partition $S(z) = \{S_1(z), S_2(z), \ldots, S_k(z)\}$ of X_N, by $S_k^{(t)}(z) = \left\{z_i z_i - \hat{\mu}_{z_i}^t < z_i - \hat{\mu}_{z_j}^t \forall 1 \leq j \leq k\right\}$, where each z_i that belongs only to one $S_k^{(t)}(z)$, even if it could go in two of them. The set $S_k^{(t)}(z)$ contains the points in X_N nearest z_i, with tied points being assigned arbitrarily to the set of lower index. Then, in the successive iterations of the algorithm, new points are randomly selected and added to the group with nearest mean distance between the centroid and the new point. In each iteration, new centroids are calculated in order to take account of the new point or group of points. Then, the classification of the observations is assigned to a group as a function of the minimum distance to the new centroid. Thus, at each step or stage of iteration, the k-means are, in fact, the means of the groups they represent. The process is repeated until no significant changes in the position of the centroid are observed in the successive steps, minimizing $SSE = \sum_{i=1}^{n} x_i \hat{\mu}_{y_t^i}^{t2}$. The variance within the cluster can be estimated as SSE/np. The a priori assignment of the number of clusters is the main limitation of the k-means algorithm; the final classification may strongly depend on the selection of the centroid (Oliva et al. 2001). To implement this method in R, we considered the *kmeans* function of the *stats* package implemented by the algorithm proposed by MacQueen (1967).

Finally, the Bayesian Method (BM) was also compared; in this method, the genotypes of a collection or population are assigned probabilistically to groups. This model-based method assumes that observations within each cluster were randomly drawn from some parametric (theoretical statistical) model with known distribution. In the other words, BM assumes that each individual is originated from one of k populations based on the genetic information obtained from its own characteristic set of allelic frequencies that determines the probability of distribution. Then, the inference for the parameters corresponding to each cluster is made jointly with the inference for the cluster membership of each individual, using Bayesian statistical methods (Pritchard et al., 2000). If the genotypes indicate that individuals are admixed, then they are assigned jointly to two or more subpopulations. In this method, as in the diffuse clustering Bayesian methods, the a posteriori probabilities indicate the uncertainty of the assignment of individuals to clusters. Thus, one of the main challenges when applying the BM is to specify the appropriate statistical distribution model for the observations (individuals) from each cluster. The BM assumes that there is a Hardy–Weinberg equilibrium within each population as well as complete linkage equilibrium between loci within population, i.e., it assumes that segregation occurs independently. Thus, let X be a vector of individuals (genotypes) and assuming that each locus in each genotype, $\left(x_l^i\right)$ represents a random and independent sample of a population whose independently sampled alleles represent an appropriate population frequency distribution, it is possible to estimate the population frequency distribution as $\Pr\left(x_l^{(i,a)} = j|Z, P\right) = p_{z(i)lj}$, where x_l^i represents the genotype of the *i*-th individual for the *l*-th locus, with i = 1,…, N and l = 1,…,L. Let Z be an unknown vector representing the population of origin of the individuals whose z(1) elements represent the original population that gave rise to the *i*-th individual and P is an unknown vector of the allele frequencies in the population whose $p_{i|l}$ element represents the *l*-th locus in the k population, where $k = 1, 2, \ldots, K$ and $l = 1, 2, \ldots, J_l$. Note that J_l is the number of distinct alleles observed at locus l. Given that there is no information on the population of origin of each individual and that the probability that individual *i* originates from population *k* is the same for all k, such probability can be estimated as $\Pr\left(z^{(i)} = k\right) = 1/k$. The Dirichlet distribution is used to specify the probability of a particular set of allelic frequencies p_{kl} for the population k in the locus l, $p_{kl} \sim \mathcal{D}(\lambda_1, \ldots, \lambda_{J_l})$, independently for each k population and each l locus. The expected frequency of the allele j is proportional to λ_j, and the variance of this frequency decreases as the sum of allelic frequencies increases. The BM uses Markov Chain Monte Carlo (MCMC) methods to infer the probability that an individual belongs to a cluster. In the first step, the algorithm estimates the allelic frequencies for each population, assuming that the original population of

Table 2 Summary statistics (mean, standard deviation, minimum and maximum) of the misclassification rates (MCR) of three clustering methods evaluated in three simulation scenarios (100 replicates of molecular data each) differing in the number of subpopulations (k) or population genetic structure (PGS)

MCR [%]	UPGMA			k-means			BM		
	k = 2	k = 5	k = 10	k = 2	k = 5	k = 10	k = 2	k = 5	k = 10
Mean	0.496	0.558	0.650	0.000	0.066	0.066	0.000	0.000	0.000
SD	0.247	0.229	0.188	0.000	0.125	0.125	0.000	0.000	0.000
Min	0.000	0.000	0.302	0.000	0.000	0.000	0.000	0.000	0.000
Max	0.503	0.803	0.900	0.000	0.553	0.553	0.000	0.000	0.000

Table 3 Overestimation error rate (E III^+) of the number of groups suggested by four selection indices to three clustering methods and population genetic structure (PGS) simulated under two groups. Each index was evaluated for k number of groups (k = 2 to k = 15)

Clustering methods	Selection indices	Evaluated group number (*k*)														E III^+
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	
UPGMA	CH	42	41	14	2	1	0	0	0	0	0	0	0	0	0	0.58
	Connectivity	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
	Dunn	98	2	0	0	0	0	0	0	0	0	0	0	0	0	0.02
	Silhouette	98	2	0	0	0	0	0	0	0	0	0	0	0	0	0.02
K-means	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
	Connectivity	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
	Silhouette	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
BM	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
	Connectivity	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
	Silhouette	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00

The column 2 depicts the true number of simulated groups or simulated PGS

each individual is known; in the next step, it estimates the probability that each individual belongs to that original population, assuming that the allelic frequencies of that population are known. The Bayesian algorithm was implemented using the package *Landscape and Ecological Association studies* (LEA) in the software R (Frichot and François 2015).

To compare methods within each dataset, summary statistics, i.e. standard deviation, minimum and maximum of the MCR, were obtained. These summary statistics were calculated from a confusion matrix. A confusion matrix, also called error matrix, is a specific table layout used to visualize the performance of an algorithm. We constructed a confusion matrix to compare the classification of an individual known by simulation with the classifications achieved with the clustering method. Thus, we compared the matching between the simulated classifier vector (group of origin) and the classifier vector (vector whose elements are the result of the subpopulation to which an individual belongs) obtained by each clustering method applied to the number of simulated subpopulations. Whenever an individual was assigned to a group other than the one that was simulated, it was considered a classification error. Finally, we estimated the mean of this classification error over 100 replicates and compared the variation of this misclassification for each algorithm.

Validation of cluster number using validation indices

Each clustering algorithm was run several times with different numbers of groups, from k = 2 to k = 15. Four internal validation indices of the optimal group number were implemented: CH (Caliński and Harabasz 1974), Connectivity (Handl and Knowles 2005), Dunn (Dunn 1974) and Silhouette (Rousseau 1987). Each validation index has its own optimization criteria, from which a given number of clusters is proposed. CH compares the deviation within the group with the dispersion among groups, considering the average compactness. Connectivity is related to the distance between neighboring observations within a cluster. Dunn index is the ratio of the minimum distance between two observations of different groups to the maximum distance between two observations of a single group; thus, the index seeks to maximize the inter-cluster distance while minimizing the intra-cluster distance. Silhouette measures the degree of confidence in a clustering assignment of an observation. For the Connectivity index, the lowest value indicates the optimal number of groups, whereas for the remaining indices, the optimal number of groups is selected according to the highest index value.

The internal validation indices were evaluated through the clustering algorithms. An accurate algorithm must provide reasonable results, even when it assumes an incorrect number of clusters. Therefore, we used different k numbers of groups for each algorithm. Then, we compared the number of groups suggested by the validation index with the value that should have been suggested by the index according to the simulation. Thus, we counted the number of times the index suggested a number of incorrect groups as a classification error rate (type III error (E III)). The classification error might occur either because the number of estimated groups is higher or lower than the simulated one. The correct number of groups is usually unknown beforehand in a real dataset so that we used a simulated dataset to calculate the error rate of group number selection (E III) for each method. We discriminated between overestimation and underestimation of the number of groups. Thus, overestimation (E III$^+$) occurred when the index selected a k higher than the simulated one, and underestimation (E III$^-$) occurred when the index determined a lower k than the simulated one.

Results

Evaluation of the performance of the compared algorithms using simulated datasets

Regarding the clustering methods, BM had a null MCR in all the simulations of the three evaluated scenarios. K-means had a lower MCR than UPGMA for the three scenarios. The MCR of the latter method increased with increasing number of subpopulations (Table 2).

We evaluated the performance of the cluster algorithms for all simulated datasets, as described in the *Simulation dataset design* section. For the first simulation scenario (S1) including two subpopulations (k = 2), the four validation indices had null type III error (0%) of overestimation of the number of groups when k-means and BM were used. Hence, the four indices indicated two clusters for all replicates. By contrast, when UPGMA was implemented, CH overestimated the number of groups 58% of the times, Connectivity had null E III + , whereas Dunn and Silhouette overestimated only 2% of the times (Table 3). The number of times that CH index applied to UPGMA indicated the right simulated k was equal to the number of times that it failed. Indeed, in 42 of the 100 replicates, CH indicated k = 2 and in 41 replicates, it indicated k = 3 (instead of k = 2). In order to understand this result, each cluster from each dataset was analyzed in depth. The analysis consisted of varying the number of k groups, counting the number of individuals within each cluster and estimating the deviation within (B(k)) and among groups (W(k)). To verify the statistical significance among groups, an AMOVA test (Excoffier 1992) and t-distributed stochastic neighbor embedding (t-SNE) (Hinton and Roweis 2002) were run. The results of the AMOVA indicated that the mean values and the variability among the k = 3 groups were statistically significant. However, the scatter plot of the ordination data with t-SNE showed two major groups (Fig. 1). In agreement with the result obtained through the simulation dataset, the smallest group was completely nested inside one of the major groups (Fig. 2).

In the second simulation scenario (S2), including five subpopulations (k = 5), the indices for the validation of the selected number of groups also had low EIII$^-$ and E III$^+$ when BM was used for classification. CH index matched the true number of groups 86% of

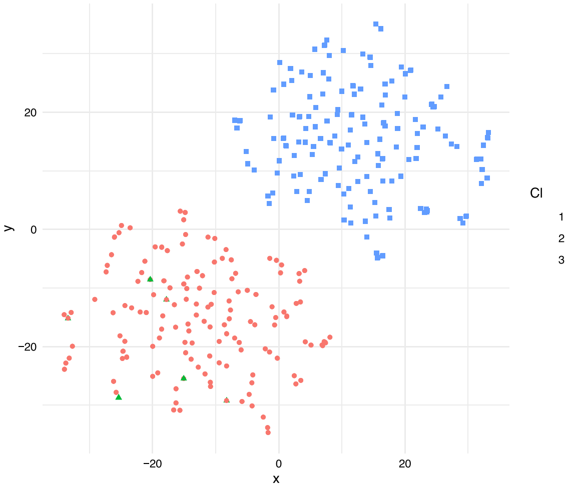

Fig. 2 Scatter plot of the t-SNE (t-distributed stochastic neighbor embedding) for a dataset of simulation scenario 1 (S1) with k = 2 clusters. The represented number of groups (k = 3) was suggested by the CH index with UPGMA algorithm; groups are identified with different symbols (dots, squares and triangles). Individuals belonging to the same group are represented with the same symbol

the times. Dunn and Silhouette indices indicated the correct number of subpopulations (k = 5) in the 100 replicates. By contrast, Connectivity index wrongly indicated two groups in all replicates. In UPGMA, CH overestimated that number, Connectivity index failed in all replicates (100% –of times) and Dunn and Silhouette underestimated the number of groups. K-means also indicated the correct number of groups, i.e. k = 5 in most cases, with the following percentages for each of the indices: CH 71%, Dunn 78% and Silhouette 76%. Again, Connectivity index misclassified the numbers of groups in all replicates. (Table 4).

In the third simulation scenario (S3), with k = 10, CH underestimated the number of groups by 17% and connectivity did so by 100%. Dunn and Silhouette indices had null overestimation and underestimation type III errors when clustering was performed via BM. All the indices underestimated the number of groups obtained via UPGMA and k-means, with a type III error of 100%, except for CH, which overestimated the number of clusters in all the simulations using UPGMA, $EIII^+ = 1$(Table 5).

Results of clustering algorithm behavior in real maize data

Another step of this experiment to validate the outcome obtained with the simulation dataset was to compare algorithms and indices with a real dataset published by Mazaheri et al. (2019a, b). In that work, the authors determined the PGS using ADMIXTURE procedure, and found k = 11 subpopulations. To

Table 4 Overestimation error rate (E III +) and underestimation error rate (E III−) of the number of groups automatically determined for four selection indices under three clustering methods and a population genetic structure (PGS) simulated with five populations. Each index was evaluated for k number of groups (k = 2 to k = 15)

Method	Indexes	Group number (*k*)														$E\ III^-$	$E\ III^+$
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	11	0	0	2	11	24	16	17	11	2	5	1	0	0	0.11	0.87
	Connectivity	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0.00
	Dunn	95	1	0	4	0	0	0	0	0	0	0	0	0	0	0.96	0.00
	Silhouette	97	0	0	3	0	0	0	0	0	0	0	0	0	0	0.97	0.00
K-means	CH	3	5	21	71	0	0	0	0	0	0	0	0	0	0	0.29	0.00
	Connectivity	98	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0.00
	Dunn	0	6	16	78	0	0	0	0	0	0	0	0	0	0	0.22	0.00
	Silhouette	0	0	3	76	18	3	0	0	0	0	0	0	0	0	0.03	0.21
BM	CH	4	0	10	86	0	0	0	0	0	0	0	0	0	0	0.14	0.00
	Connectivity	52	24	20	4	0	0	0	0	0	0	0	0	0	0	0.96	0.00
	Dunn	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0.00	0.00
	Silhouette	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0.00	0.00

The column 5 depicts the true number of simulated groups or simulated PGS

Table 5 Overestimation error rate (E III +) and underestimation error rate (E III-) of the number of groups suggested by four selection indices applied to molecular data whose population genetic structure (PGS) was simulated with 10 populations and subjected to clustering with three methods. Each index was evaluated for k number of groups (k = 2 to k = 15)

Method	Indexes	Group Number (*k*)	E III^{-}	E III^{+}													
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA																	
	CH	0	0	0	0	0	0	0	0	0	0	0	6	17	77	0.00	1.00
	Connectivity	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0.00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0.00
	Silhouette	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0.00
K-means	CH	3	5	21	71	0	0	0	0	0	0	0	0	0	0	1.00	0.00
	Connectivity	98	2	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0.00
	Dunn	0	6	16	78	0	0	0	0	0	0	0	0	0	0	1.00	0.00
	Silhouette	0	0	3	76	18	3	0	0	0	0	0	0	0	0	1.00	0.00
BM	CH	0	0	0	0	0	0	0	17	83	0	0	0	0	0	0.17	0.00
	Connectivity	17	17	0	0	16	17	33	0	0	0	0	0	0	0	1.00	0.00
	Dunn	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0.00	0.00
	Silhouette	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0.00	0.00

The column 10 depicts the true number of simulated groups or simulated PGS

compare the results in terms of the number of populations (groups) suggested by each validation index, the index values were standardized. Standardization consisted of subtracting the general mean of each index from each observed value. Then, the difference between the observed and the expected value (mean) was divided by the standard deviation. Standardization was performed for all values of the validation indices obtained in all the clustering methods.

In the standardized value, the CH, Dunn and Silhouette indices are expected to present their highest (maximum) value when the number of groups is 11, i.e., the number of reference groups published by Mazaheri et al. (2019b). By contrast, the Connectivity index value is expected to be the lowest for k = 11. Thus, in the standardized scale, the optimization criterion of each validation index is conserved. Figure 3 shows the behavior of each validation index for each the clustering methods used. The CH, Connectivity and Dunn indices indicated k = 2 groups, independently of the clustering method used. Silhouette index proposed 5 subpopulations with UPGMA, 13 with k-means and 15 with BM. Thus, k-means and BM were closer to the number of groups published by Mazaheri et al. (2019b) than UPGMA. None of the indices reached its optimum value for k = 11 with UPGMA and k-means. Nevertheless, the Dunn index yielded the second highest value for the expected number of groups (k = 11) in BM (Fig. 3).

We also compared the performance of the clustering method by estimating the percentage of mismatch between the expected classification value of a genotype and the classification made by the clustering algorithm evaluated in this work. Regarding the genotype classification made by the methods, according to the published number of subpopulations (k = 11), UPGMA had the highest number of mismatches in the classification (58%), i.e., less than half (42%) of the genotypes were assigned to the expected group. The highest percentage of mismatch between the expected classification and the one reported in the published work was obtained with BM (18%). The mismatch rate of BM was approximately three times lower than that of UPGMA, whereas k-means had half of the mismatch of UPGMA and twice as high as that of BM, which was 31%.

Regarding the classification of genotypes to each subpopulation set by Mazaheri et al. (2019b), Fig. 4 shows a heatmap illustrating the matching (diagonal)

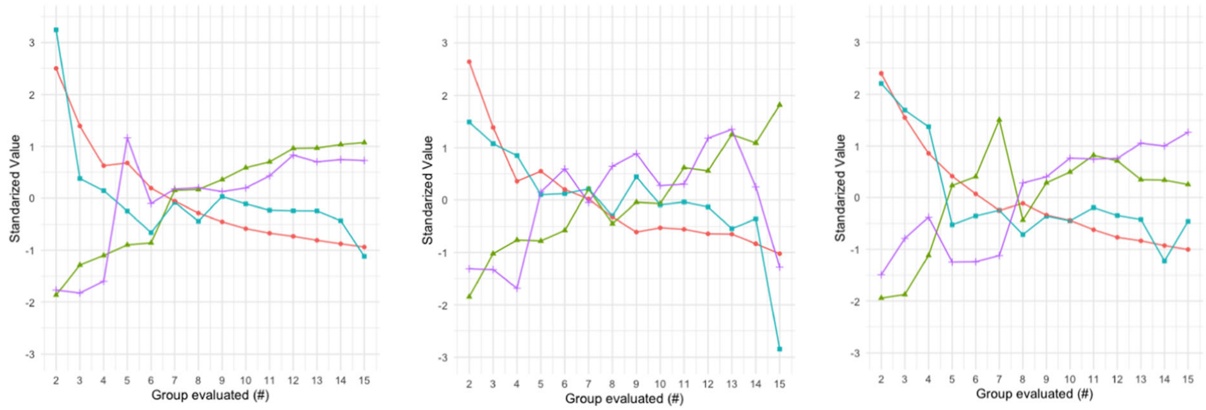

Fig. 3 Scatter plots of the standardized value of four validation indices, CH (–●–), connectivity (–▲–), Dunn (–■–) and Silhouette (–+–), as a function of the number of evaluated groups, ranging from k = 2 to k = 15, for the real dataset published by Mazaheri et al. (2019a). All the validation indices were evaluated for each of the three clustering methods: UPGMA (left), k-means (center) and Bayesian Method (right). For the indices CH, Dunn and Silhouette, the highest number indicates the optimum number of groups, whereas for Connectivity, the lowest value is the one indicating the optimum number of groups

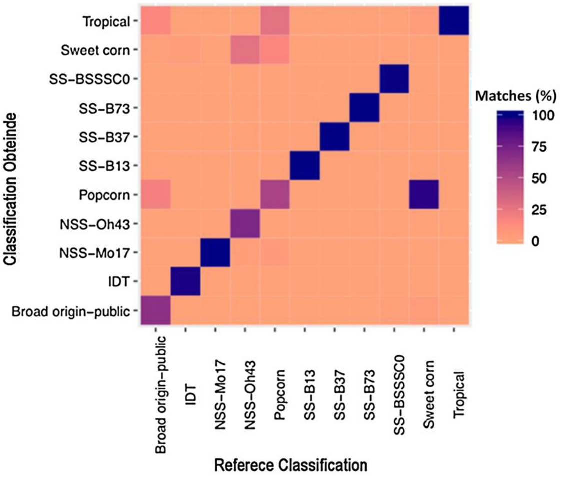

Fig. 4 Heatmap of the confusion matrix of the matching percentage between reference classification of the real data set published by Mazaheri et al. (2019a) and that obtained by the Bayesian Method (BM). A value of 100 indicates exact match (100%) between the reported classification and that obtained by BM, whereas a value of zero indicates null matching (0%)

between the classification published by Mazaheri et al. (2019b) and that obtained by BM in the present work. The classification mismatch between methods is observed outside the diagonal. Darker colors indicate a greater degree of matching. The results show that BM had a perfect classification of the four stiff stalk populations (SS-B13, SS-B37, SSB73 and SS-BSSSC0), of the tropical subpopulation, and of one of the non-stiff stalk subpopulations (NSS-Mo17). The other non-stiff stalk subpopulation (NSS-Oh43) was correctly classified 71.2% of the times by BM, whereas the remaining genotypes were clustered with sweet corn. The genotypes within Iodent subpopulation were classified correctly 97.1% of the times, whereas 2.9% of the times they were assigned to the sweet corn group. The group composed of broad origin-public lines was correctly classified 65.3% of the times. The rest of times (34.5%) they were grouped with the popcorn and tropical lines. The popcorn subpopulation was correctly classified 58.3% of the times, and was confused mostly with tropical lines, sweet corn and NNS-Mo17. Finally, the sweet corn population was confused almost entirely with popcorn (Fig. 4).

Discussion

Modern maize hybrids are the result of crossing an inbred line from one heterotic pattern with an inbred line from a different heterotic pattern (Lee and Tracy 2009). The concept of heterotic patterns, or heterotic groups, was proposed by breeders as a means of maximizing the amount of hybrid vigor (Reif et al. 2005; Schnable and Springer 2013; Meena et al. 2017). Classifying the elite germplasm from different heterotic groups is an important task in any breeding program to enhance crossings (Meena et al. 2017). Classification of heterotic patterns is generally based on several criteria, such as pedigree, molecular marker-based associations, and performance in hybrid combinations (Lee and Tracy 2009; Vittorazzi et al. 2018). In this sense, molecular markers have shown to be very useful to classify inbreds in heterotic groups (Lu and Bernardo 2001; Li et al. 2002; Schnable and Springer 2013) and elucidate the underlying PGS. Finding the number of genetic groups in a set of individual genotypes and assigning individuals to the groups have become essential tasks in population genetics (Beugin et al. 2018) as well as in other areas, such as plant genetic breeding, where the phenotypic information is complemented with the genotypic one (Thorwarth et al. 2017; Haile et al. 2018; Yuan et al. 2020).

The new technologies have improved the genetic selection process. The PGS search in a high-dimensional data collection, such as those generated by SNP molecular markers, implies an increased complexity in the management of massive databases. While the growth of the dataset size has been accompanied with enhanced computing capabilities, population genomics is more than a simple “big data” of population genetics. The targets of study are “genomes” rather than just “multiple genes”. A large number of multivariate methods have been proposed for the automatic identification of subgroups within populations. However, implementing analytical techniques that take into account genetic structures and accurately identify patterns in the context of massive genomic data of thousands of individuals poses a challenge both at the biological and statistical-methodological levels (Dutheil 2020). In this sense, cluster analysis is a widely used tool to classify genotypes when there is no previous information of a pattern or underlying structure that generates groups. Several cluster algorithms do not require previous information about the groups, making the clustering analysis a widely used tool that has received great attention in several areas of application (Hedrick 2005).

However, the use of clustering algorithms involves a number of likely complex decisions, such as metric selection and, in the case of hierarchical algorithms, the clustering method. Metric selection depends strongly on the nature of the variables (Bruno and Balzarini 2010). For this reason, selecting the number of groups beforehand may pose a problem to the researcher dealing with real data. One way of obtaining an automatic response of the number of groups or PGS is by applying validation indices. In this work, based on simulated PGS scenarios, we evaluated the quality of the partitions generated by cluster algorithms. Additionally, we evaluated the goodness of a clustering structure based on internal validation indices. When the correct partition is available by simulation it is possible to estimate its quality by measuring how closely each simulation situation is related to the cluster and how well separated a cluster is from other clusters. Thus, we intend to answer the question of whether there is one clustering method and/or validation index that is most robust to elucidate the underlying PGS in a panel of maize genotypes.

Our results showed that the BM approach was the best performing one for genotype classification in the three evaluated scenarios. UPGMA was the worst performing one, with the highest MCR, and with the misclassification increasing as the number of simulated groups increased. UPGMA is based on the mean of the differences between groups for the estimation of the number of simulated groups. Since the mean is a measure of central tendency, it tends to “approximate” or homogenize the differences. Then, by shortening the distances between individuals, UPGMA tends to generate unbalanced groups, with a few groups with a large number of individuals and several small groups with very few individuals. These clusters with so few individuals generate an overestimation of the true number of groups. For a deeper understanding of the MCR of UPGMA, we analyzed the classifier vectors of the genotypes generated by this algorithm for S1 to calculate the number of individuals within each group for k = 2 and k = 3, in each of the replicates where the number of selected groups was different than the simulated one. For instance, when we used UPGMA and the CH index, the latter indicated k = 3 instead of k = 2 in 41/100 replicates (Table 3).

For the first replicate, although the number of simulated subpopulations in the PGS was k = 2, UPGMA generated a group with 997 individuals (here identified as group A) and another group with 3 individuals (here identified as group B). When the same dataset was set to form k = 3 groups, the number of individuals of group B also had three individuals and group A (the major group of individuals) was split into two new groups of 497 and 500 individuals. This behavior, consisting of generating a group with very few individuals and another one with a large number, was observed in the 100 replicates of S1. Due to the great imbalance of these groups, the AMOVA (Excoffier et al. 1992) indicated statistically significant differences between groups, which is explained by the great difference between individuals of the small group and individuals of the other groups. However, when t-SNE multivariate technique was used to visualize groups in high-dimensional data, the algorithm indicated two groups, with the group with fewest individuals being completely nested in the other group (Group A). This result matched with the simulated PGS of k = 2 (Fig. 1). UPGMA showed the same behavior when we analyzed in depth a classifier vector for k = 4, again forming two major groups, with $n_A = 500$ individuals and $n_B = 491$ in 40% of the 100 replicates, and two groups with a lower number of individuals ($n_C = 6$ and $n_D = 3$) (data not shown). In other words, although UPGMA formed more clusters than expected, two of them were large and two had < 10 ‰ individuals of the total of individuals to be classified.

Thus, although the number of groups suggested by UPGMA was not that expected by the simulation (k = 2), two clusters comprised 99% of the data. The differences between simulated groups was estimated through an Fst = 0.03 for the 100 replicates, which is assumed to be low. Given the characteristics of UPGMA, individuals with a greater divergence, achieved only by random, are classified as different and assigned to separate groups. This behavior reinforces the need to describe each cluster; i.e., as a minimum, the number of individuals that fall within each cluster should be identified, rather than having a totally automated procedure to make the decision on the number of clusters. Finally, the k-means method had an intermediate behavior between those of UPGMA and BM. This result may be due to the fact that k-means minimizes the within-group distances and assigns an individual to a group when the distance between that individual and the centroid is minimum. Therefore, the groups formed by k-means are more compact, with clearer divisions than those obtained by UPGMA.

Internal validations indices are based on the clustering partition as input to assess the quality of the clustering. Hence, we selected measures that reflected the compactness, connectedness and separation of the cluster partition. Connectedness indicates to what extent genotypes are placed in the same cluster as their nearest neighbors in the data space, and is here measured by connectivity (Handl and Knowles 2005). Compactness assesses cluster homogeneity, usually by looking at the intra-cluster variance, while separation quantifies the degree of separation between clusters (usually by measuring the distance between cluster centroids). Since compactness and separation demonstrate opposing trends –with increasing number of clusters, compactness increases, whereas separation decreases–, popular methods combine the two measures into a single score. The Dunn and Silhouette indices are examples of non-linear combinations of compactness and separation.

In our results, the CH did not always detect the same simulated PGS, regardless of the clustering method used. When the distances between groups are small, the numerator of the CH index tends to be lower and, consequently, the index value decreases. Likewise, if compactness within a group is low due to high intra-group variability, the distances between individuals within a group will be similar to distances between groups. Consequently, the coefficients of distances between groups can also be similar between other groups. A smaller denominator of CH yields a lower number of groups; the latter fact may lead to the establishment of a number of groups similar, but not equal, to the simulated one, as it occurred in our results. Since the optimization criterion of CH is the maximum index value, the greater the dispersion between groups, the greater the CH. Thus, CH index applied to UPGMA tended to overestimate the number of groups, since this method generates unbalanced groups. Groups with a few individuals have low internal dispersion but generate a greater dispersion among groups. At the same time, this generates a higher CH coefficient, indicating a higher number of groups than expected. By contrast, CH applied to the k-means tended to underestimate the number of groups in scenarios with $k > 2$. This can be explained by the capacity of the method to form highly compacted groups, reducing the denominator of CH and increasing its value. BM and non-hierarchical k-means method showed a similar behavior, since each individual was assigned to a group according to its highest probability of belonging, and highly compacted groups were formed.

Regarding Connectivity index, in scenarios with a group number other than k = 2, it had the highest underestimation error of the number of groups, regardless of the clustering method used. This index estimates the distance between each individual and its nearest neighbors. If the neighbor belongs to the individual's group, then the index assigns a zero value; by contrast, if the neighbor belongs to another group, then it adds 1/j, where *j* indicates the *j*-th nearest neighbor. Thus, the lower the number of underlying groups of PGS, the lower the value of Connectivity coefficient. As a consequence, our results show the high rate of underestimation of the number of groups. For instance, in the S2, when the k-means method was applied, the connectivity index proposed two groups (k = 2) in 98% of the replicates, when it had to indicate k = 5.

Finally, Dunn and Silhouette indices had the best performance with the BM, showing a MCR of zero in all cases (Tables 2, 3 and 4). Nevertheless, this result can be misleading for Silhouette index. When this index is applied in the BM context, the number suggested should be carefully examined. Indeed, given the way Silhouette works in the BM context, it is very likely that this index is emulating the same confidence as BM. This possibility may be explained by the fact that BM estimates the probability that an individual belongs to a group, whereas the Silhouette index determines the number of groups based on the degree of confidence of the membership of an individual to that group. In this sense, the higher the probability of an individual to belong to a group, the higher the confidence. Thus, Silhouette index could be highly correlated with BM; therefore, despite its zero error rate, it would not be the most appropriate validation index combined with BM. Moreover, Silhouette index applied to UPGMA indicated two groups in all cases, regardless of the simulated configuration. Indeed, it recommended two groups instead of five in 97% of the replicates, and two groups instead of 10 in 100% of the times. Regarding k-means, notably, the Dunn and Silhouette indices presented the same behavior when the simulation was set with more than two groups, i.e. both indices indicated five groups 78% of the times when the simulated PGS was of 10 groups. Therefore, we conclude that Silhouette index is not suitable for these studies and that Dunn works well with BM. Finally, when PGS simulation was made considering k = 10, none of the indices yielded the correct number of groups for k-means and UPGMA. Thus, it is necessary to characterize the groups obtained by these methods and not to rely on automatic mechanisms.

The BM method also showed the best performance to classify the genotypes with the real dataset used for illustration, whereas UPGMA suggested groupings that differed greatly from expected. In this case, none of the validation indices indicated the established number of subpopulations, with Dunn being the only one to show a relative maximum for k = 11 with the classification obtained by BM. This k value would be in agreement with the biological information provided by Mazaheri et al. (2019a) using ADMIXTURE. The k = 11 subpopulations reported by the authors was based on a subset of 93,991 SNPs and knowledge of pedigree, molecular marker-based associations, and performance in hybrid combinations. We worked with the whole SNP panel because we did not have information on the cleaning made by the authors and therefore, any selection would have resulted in a different molecular dataset. However, it is known that the higher the number of SNP markers, the greater the molecular information (Gao et al. 2012). Thus, we decided to use all the available markers rather than selecting a subset of the same size (93,991 SNP), but of a probably different combination.

The fact the results of BM were closest to those of ADMIXTURE is not surprising and is very likely because ADMIXTURE and BM are based on the same estimation method. Both estimate the probability of individuals to belong to a group; however, the former does so based on maximum likelihood estimations and the latter, through a Bayesian model. However, each model has unique characteristics that may contribute to its efficiency in final classification. For example, BM can incorporate data on linked loci (Falush et al. 2003), whereas ADMIXTURE uses maximum likelihood to assign the probability of an individual to belong to a group (Lawson et al. 2018). Yet, both methods are equally capable of differentiating groups whose allelic frequency distributions are not extremely different. This feature makes them suitable for searching PGS in several research works aiming at both species conservation and genetic improvement.

In our simulation study, BM yielded perfect classifications (with no distance between the expected and the observed value) at a genetic divergence of Fst = 0.03. The dataset used for illustration had mean divergence values between populations close to 0.03. Latch et al. (2006) evaluated the performance of BM using the software STRUCTURE in wild populations, i.e., populations not subjected to genetic breeding; the method was capable of making good classification when the genetic divergence values were close to Fst = 0.1 (Latch et al. 2006). The behavior of BM with the software STRUCTURE was evaluated by Evanno et al. (2005) in a simulation study in several dispersal scenarios; they found a better performance of BM in more complex structures than that of an island model (almost no genetic flow) (Latch et al. 2006). On the contrary, in our study we did not assume any genetic model. The dataset used for illustration consisted of a panel of genotypes representing the main heterotic groups used in North America. The genetic divergence values estimated among 11 subpopulations was Fst = 0.029, a value similar to the simulated ones. On the other hand, we were able to compare the illustrated dataset and the simulated scenarios because of their similar genetic divergence among subpopulations. The aim of the simulation was to represent a broad spectrum of possible combinations of inbred lines used to obtain maize hybrids that will originate different PGS. In our work, the divergence values estimated between the subpopulations of the real data panel were 30% smaller than the values reported by Latch et al. (2006). Our results indicate that even at low divergence levels, the compared methods were able to identify groups with a high degree of matching with the PGS reported by Mazaheri et al. (2019b). However, in some cases, the MCR was high, as in the case of genotypes of the Sweet corn population, which were clustered with genotypes of the Popcorn population 95% of the times (Fig. 4). The genetic divergence value estimated between these populations in this work was 0.011. Latch et al. (2006) reported that for an Fst = 0.02, the software STRUCTURE was not able to identify the true number of subpopulations and suggested that around Fst = 0.02 the software fails to detect the number of existing groups or PGS. This phenomenon would explain the high MCR between these populations, which are not genetically related and whose heterotic groups are different, but which do share characteristics that make them different from other improved maize varieties, such as early flowering, different plant and tassel architecture, low kernel row number and low germination rate. Some inbreds of Sweet corn and Popcorn population may have been submitted to selection process to improve agricultural traits that made them more similar to each other and more different than the maize grown for grain (Lee and Tracy 2009).

The genotypes present in these subpopulations may be involved in evolutionary processes of genetic fixation or in differentiation processes that can still not be mathematically differentiated by computational algorithms. Quantifying genetic differentiation between individuals and establishing different subpopulations with the obtained values have been the pursuit of geneticists since the rise of population genetics (Wright 1949). Latch et al. (2006) compared the relative performance of three Bayesian methods to search for genetic structure, including BM, and concluded that, despite its best performance, this method assigns correctly individuals to their subpopulation of origin at a minimum Fst of 0.05. Moreover, the authors suggested that for Fst values below 0.03, BM does not identify a clear genetic structure and that at Fst below 0.02, the algorithms fail to identify the correct number of subpopulations; they suggested that these software tools provide a classification with low probability when genetic divergence between groups (Fst) is low. In this work, while the studied population was not wild, BM showed a similar behavior to that reported by Latch et al. (2006), although we worked with smaller values than those proposed in that work. Plant genetic breeding programs intend to obtain promising genotypic lines that have better plant health traits, adaptation to abiotic stress and high production. In this sense, unlike wild populations, crossings are targeted, reducing the genetic base, i.e., commercial hybrids usually have a common ancestor (Acosta 2009) and, therefore, their levels of divergence are expected to be lower than those of a wild population.

ADMIXTURE was proposed as an alternative to BM based on its estimation of the degree of relatedness due to common ancestry between individuals. That estimation is performed by maximum likelihood and makes the process computationally more efficient than STRUCTURE software (Alexander and Lange 2011). ADMIXTURE uses an algorithm based on ancestry in unrelated individuals and adopts the likelihood model embedded in genetic structure. The approach is similar to that of BM; both programs model the probability of genotype membership to a group using ancestry proportions and frequencies of population alleles. In addition, they estimate allele frequencies of populations simultaneously with ancestry proportions. The constant development of programs that optimize computing time with algorithms that can process large databases have led us to explore possible differences in performance among methods. The performance of a method depends on the nature of data structure. Here, we compared algorithms designed to work with large databases from SNP markers with other clustering methods, like UPGMA and k-means, used in diverse contexts. Our results confirm that BM has a good capacity to infer PGS. The information provided in this study offers a robust alternative to unveil PGS, thereby facilitating future population studies and genetic improvement strategies in maize breeding programs. Additionally, our results may help maize breeders to incorporate the identified genetic variation into hybrid breeding programs. The present findings might also have broad implications for other crop species.

Acknowledgements We thank the Ministerio de Ciencia y Tecnología de la Provincia de Córdoba (MinCyT) for funding the GRFT- 0279-011245/2019 and PIODO 2018 projects, and the National Institute of Agricultural Technology of Argentina (INTA) for supporting this work through the project 2019-PE-E6-I114. This work was also supported by the National Scientific and Technical Research Council of Argentina (CONICET) and INTA. EV is a fellow of CONICET and professor of the Universidad Nacional de Villa María, JI is a professor of Universidad Nacional del Noroeste de la Provincia de Buenos Aires (UNNOBA) and researcher of INTA, and CB is a professor of the Universidad Nacional de Córdoba and researcher of CONICET.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by MEV and CB. The first draft of the manuscript was written by MEV and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the Ministerio de Ciencia y Tecnología de la Provincia de Córdoba (MinCyT) through the projects GRFT- 0279–011245/2019 and PIODO 2018, and by the National Institute of Agricultural Technology of Argentina (INTA).through 2019-PE-E6-I114.

Availability of data and material The datasets and/or analyses generated during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Acosta R (2009) Reseña. El cultivo de maíz, su origen y clasificación. El Maíz En Cuba Cultiv Trop 30:53–54

Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformat. https://doi.org/10.1186/1471-2105-12-246

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664. https://doi.org/10.1101/gr.094052.109

Baloch FS, Alsaleh A, Shahid MQ et al (2017) A whole genome DArTseq and SNP analysis for genetic diversity assessment in durum wheat from central fertile crescent. PLoS One 12:1–18. https://doi.org/10.1371/journal.pone.0167821

Balzarini M, Teich I, Bruno C, Peña A (2011) Making genetic biodiversity measurable: a review of statistical multivariate methods to study variability at gene level. Rev La Fac Ciencias Agrar 43:261–275

Becerra VV, Paredes MC (2000) Use of biochemical and molecular markers in genetic diversity studies. Agric Técnica 60:270–281

Beugin MP, Gayet T, Pontier D et al (2018) A fast likelihood solution to the genetic clustering problem. Methods Ecol Evol 9:1006–1016. https://doi.org/10.1111/2041-210X.12968

Bruno C, Balzarini M (2010) Distancias genéticas entre perfiles moleculares obtenidos desde marcadores multilocus multialélicos. Rev La Fac Ciencias Agrar 41:11

Bruno C, Balzarini M, Di Rienzo J (2003) Comparación de medidas de distancais entre perfiles RAPD. J Basic Appl Genet 15:29–32

Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat Methods 3:1–27

Dudoit S, Fridlyand J (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol 3:1–21

Dunn JC (1974) Well-separated clusters and optimal fuzzy partitions. J Cybern 4:95–104

Dutheil JY (2020) Statistical population genomics. Springer Nature, New York

Esfandyari H, Sørensen AC (2019) xbreed: an R package for genomic simulation of purebred and crossbred populations

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491. https://doi.org/10.3354/meps198283

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567–1587. https://doi.org/10.1093/genetics/164.4.1567

Frichot E, François O (2015) LEA: An R package for landscape and ecological association studies. Methods Ecol Evol 6:925–929. https://doi.org/10.1111/2041-210X.12382

Gao Z, Luo W, Liu H et al (2012) Transcriptome analysis and SSR/SNP markers information of the blunt snout bream (*Megalobrama amblycephala*). PLoS One 7:1–10. https://doi.org/10.1371/journal.pone.0042637

González-recio O, Rosa GJM, Gianola D (2014) Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livest Sci 166:217–231. https://doi.org/10.1016/j.livsci.2014.05.036

Günter S, Bunke H (2003) Validation indices for graph clustering. Pattern Recognit Lett 24:1107–1113. https://doi.org/10.1016/S0167-8655(02)00257-X

Haile JK, N'Diaye A, Clarke F et al (2018) Genomic selection for grain yield and quality traits in durum wheat. Mol Breed 38:1–18. https://doi.org/10.1007/s11032-018-0818-x

Handl J, Knowles J (2005) Exploiting the trade-off—the benefits of multiple objectives in data clustering. International conference on evolutionary multi-criterion optimization. Springer, Berlin, Heidelberg, pp 547–560

Hartigan JA (1975) Clustering algorithms. Wiley, Hoboken, New Jersey

Hedrick P (2005) Large variance in reproductive success and the Ne/N ratio. Evolution (n y) 59:1596–1599. https://doi.org/10.1007/BF01515409

Hinton G, Roweis S (2002) Stochastic neighbor embedding. Adv Neural Inf Process Syst 15:883–840

Isidro J, Jannink JL, Akdemir D et al (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128:145–158. https://doi.org/10.1007/s00122-014-2418-4

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:1–15

Kaufman L, Rousseeuw P (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York

Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE (2006) Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conserv Genet 7:295–302. https://doi.org/10.1007/s10592-005-9098-1

Lawson DJ, van Dorp L, Falush D (2018) A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nat Commun 9:1–11. https://doi.org/10.1038/s41467-018-05257-7

Lee EA, Tracy WF (2009) Modern maize breeding. Handbook of maize. Springer, New York, pp 141–160

Legendre P, Legendre L (2012) Numerical ecology. Elsevier, Oxford

Li Y, Du J, Wang T et al (2002) Genetic diversity and relationships among Chinese maize inbred lines revealed by SSR markers. Maydica 43:93–101

Lu H, Bernardo R (2001) Molecular marker diversity among current and historical maize inbreds. Theor Appl Genet 103:613–617. https://doi.org/10.1007/PL00002917

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. pp 281–297

Malosetti M, Van Der LCG, Vosman B, Van EFA (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to phytophthora infestans in potato. Genetics 889:879–889. https://doi.org/10.1534/genetics.105.054932

Mazaheri M, Heckwolf M, Vaillancourt B et al (2019a) Data from: genome-wide association analysis of stalk biomass and anatomical traits in maize. Dataset. https://doi.org/10.5061/dryad.n0m260p

Mazaheri M, Heckwolf M, Vaillancourt B et al (2019b) Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC Plant Biol 19:1–17. https://doi.org/10.1186/s12870-019-1653-x

Meena AK, Gurjar D, Patil SS, Kumhar BL (2017) Concept of heterotic group and its exploitation in hybrid breeding. Int J Curr Microbiol Appl Sci 6:61–73

Mikel MA, Dudley JW (2006) Evolution of North American dent corn from public to proprietary germplasm. Crop Sci 46:1193–1205. https://doi.org/10.2135/cropsci2005.10-0371

Odong TL, van Heerwaarden J, Jansen J et al (2011) Determination of genetic structure of germplasm collections: Are traditional hierarchical clustering methods appropriate for molecular marker data? Theor Appl Genet 123:195–205. https://doi.org/10.1007/s00122-011-1576-x

Oliva F, Cáceres M, Font X, Cuadras C (2001) Contribuciones desde una perspectiva basada en proximidades al Fuzzy K-means Clustering. Dissertation, XXVI Congreso Nacional de Estadística e Investigación Operativa

Peña-Malavera A, Bruno C, Fernandez E, Balzarini M (2014) Comparison of algorithms to infer genetic population structure from unlinked molecular markers. Stat Appl Genet Mol Biol 13:391–402. https://doi.org/10.1515/sagmb-2013-0006

Peng Y, Zhang Y, Kou G, Shi Y (2012) A multicriteria decision making approach for estimating the number of clusters in a data set. PLoS One. https://doi.org/10.1371/journal.pone.0041713

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–59

R Core Team (2019) R: a language and environment for statistical computing. In: R A Lang. Environ Stat Comput https://www.r-project.org/

Raj A, Stephens M, Pritchard JK (2014) FastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197:573–589. https://doi.org/10.1534/genetics.114.164350

Reif JC, Hallauer AR, Melchinger AE (2005) Heterosis and heterotic patterns in maize [Zea mays L.; USA; Europe; Japan; China]. Maydica (Italy)

Rendón EL, Abundez IM (2016) RENTOL: un algoritmo de agrupamiento basado en K-means. Res Comput Sci 128:149–157

Riedelsheimer C, Endelman JB, Stange M et al (2013) Genomic predictability of interconnected biparental maize populations. Genetics 194:493–503. https://doi.org/10.1534/genetics.113.150227

Rousseau P (1987) Silhouettes: a gaphical aid to the interpretation and validation of custer analysis. J Comput Appl Math 20:53–55

Salvador S, Chan P (2004) Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In: 16th IEEE international conference on tools with artificial intelligence. IEEE. pp 576–584

Schnable PS, Springer NM (2013) Progress toward understanding heterosis in crop plants. Annu Rev Plant Biol 64:71–88

Sokal RR (1958) A statistical method for evaluating systematic relationships. Univ Kansas, Sci Bull 38:1438

Thorwarth P, Ahlemeyer J, Bochard AM et al (2017) Genomic prediction ability for yield-related traits in German winter barley elite material. Theor Appl Genet 130:1669–1683. https://doi.org/10.1007/s00122-017-2917-1

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B (Statistical Methodol) 63:411–423

Vittorazzi C, Júnior ATA, Guimarães AG et al (2018) Research article evaluation of genetic variability to form heterotic groups in popcorn. Genet Mol Res 17:1–17. https://doi.org/10.4238/gmr18083

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Windhausen VS, Atlin GN, Hickey JM et al (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 Genes Genomes, Genetics 2:1427–1436. https://doi.org/10.1534/g3.112.003699

Wright S (1949) The genetical structure of populations. Ann Eugen 15:323–354

Yuan J, Wang X, Zhao Y et al (2020) Genetic basis and identification of candidate genes for salt tolerance in rice by GWAS. Sci Rep 10:1–9. https://doi.org/10.1038/s41598-020-66604-7

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.