Review Article

Multivariate statistical methods used in population genetics

Abstract

Several multivariate statistical methods are used in population genetics but there are very few studies that have revealed the strengths and weaknesses of different methods. Thus, this study aims to reveal the strengths and weaknesses of the different multivariate statistical methods used in population genetics through the world. This synthesis is carried out according to the methodology "*Preferred Reporting Items for Systematic Reviews and Meta-Analyzes*" (PRISMA). This study shown that various statistical methods or combination of multivariate statistical methods are used in population genetics. It emerges that there is no a priori a better method, so it is necessary to determine the method adapted to both the data collected and the research objective. This study identified the most commonly used multivariate statistical methods in genetics such as: ordination methods (52.50%) are methods that summarize the information contained in the data matrix by minimizing wastage. This are: principal components analysis (by 32.0% of the articles), principal coordinates analysis (by 7.50% of the articles), discriminant analysis of principal component, factorial correspondence analysis, factorial discriminant analysis, factorial analysis on distance table. Clustering methods (35%) that aim to form groups of individuals that are as similar as possible, including the hierarchical ascending clustering (17.50% of articles), neighbor-joining, and Bayesian clustering model (by 15% of the articles). The analysis of the molecular variance (7.50%) which consists of studying the intra and inter-population variation and the Mental test (5%) which aims to test the correlation between the matrix of genetic distances and other distance matrices (environmental causes of genetic variability).

Key words: genetics; multivariate statistical methods; ordination methods; classification methods.

1. Introduction

Genetic diversity is defined as the level of similarity or difference in the genetic makeup of individuals, populations and species [1]. This genetic diversity is extremely important because it represents the basic survival material and on which selection can act [2]. It can represent a direct advantage for a population, the adaptive value of a trait being generally higher for a gene with several allelic states or for a population formed by different individuals. Various types of markers are used for the study of genetic diversity. Morphological markers are initially used for the study of genetic diversity. Later, cytological and biochemical markers are used. Then with the advent of molecular tools, molecular markers became the method of choice for the analysis of genetic diversity [2]. This

use of very specific markers means that the study of genetic variability generates data or results in the form of complex tables with a set of many variables and often of diverse typology [3]. Consequently, the summary of information has always been of interest in population genetics as in several fields such as biology, physics, computer science, economics, and social sciences. Thus, several multivariate statistical methods have been developed with the aim of reducing the number of variables, or even the number of individuals without considerable loss of information, in order to make interpretation easier and graphical representation easier [4]. That's why the study of diversity and genetic structuring requires the use of adequate tools and techniques and suitable statistical methods. Indeed, the search for multivariate statistical methods used in genetics can be problematic. The question arises as to whether there is a method which is a priori defined as the most relevant for the study of genetic diversity. This study aims to achieve a synthesis of multivariate methods used in population genetics and specifically to: (i) identify the different multivariate statistical methods used for the study of genetic diversity; (ii) bring out the main criteria for using these methods, which can improve the quality of analyzes of the population genetic structure.

2. Methodology

The inclusion and exclusion method used in this study is based on the PRISMA method "Preferred Reporting Items for Systematic Reviews and Meta-Analyzes" [5]. Although the PRISMA diagram was created with the aim of improving the quality of the way of reporting systematic reviews, it is a standard relevant to any review of the literature [6].

2.1. Research strategy

The electronic literature search was performed on the "Google" search engine, in order to find articles that used multivariate statistical methods to study the diversity and genetic structure of populations. The key words combinations were used in both French and English for the publications collection. The following key words were used: "*Multivariate statistical methods, genetic diversity, populations*", "*Multivariate statistical methods, genetic diversity, and populations*".

2.2. Eligibility of study and data extraction

A study is eligible for this review if:

- the study deals with the analysis of the diversity genetic or population genetic structure;
- a multivariate statistical method was used;
- the study is written in French or English.

2.3. Inclusion criteria for data collection

During the consultation of the database, all the publications found online were selected from the titles, available summaries and full texts. Thus, only those which are complete and meet the eligibility criteria were included. Thus, for data extraction, in each study the multivariate statistical method used was identified and noted.

3. Result and discussion

3.1. Result

From data collection we found two hundred and ninety (290) studies of which, sixteen (16) duplicates were found and excluded, one hundred twenty-two (122) titles not concerned were excluded, eleven (11) presenting only the title and or abstract also excluded and fifty-four (109) non-eligible studies were excluded, to finally retain 32 studies including three (5) theses and twenty (27) articles. Table 1 presents the multivariate statistical methods identified in the bibliographic research of this study which are grouped into 04 categories: ordination methods (52.50%), classification methods (35%), molecular analysis of variance (7, 50) and the correlation test between matrices (5%).

Table 1. Multivariate statistical methods used in genetics

Statistical methods	N	Valide %	Cumulated %
PCoA	3	7.50	7.50
PCA	13	32.50	40.00
DAPC	1	2.50	42.50
FCA	1	2.50	45.00
FDA	1	2.50	47.50
FADT	2	5.00	52.50
AMOVA	3	7.50	60.00
AHC	7	17.50	77.50
Neighbour-joining	1	2.50	80.00
Bayesian Clustering Model	6	15.00	95.00
Mantel Test	2	5.00	100.00

PCoA: Principal Coordinate Analysis, PCA: Principal Component Analysis, CHA: Ascending Hierarchical Classification, AMOVA: Molecular Analysis of Variance, DACP: Discriminant Analysis of Principal Component, AFC: Factorial Correspondence Analysis, DFA: Discriminating Factorial Analysis, FADT : Factorial Analysis on Distance Table.

3.2. Discussion

In this section, we discussed our results by making a comparison with other similar studies available. Then we presented a summary of the different methods identified. This study addressed multivariate statistical methods used in population genetics in published research. This study made it possible to identify the most commonly used multivariate statistical methods in genetics: ordination methods are methods that summarize the information contained in the data matrix while minimizing loss. These are principal component analysis (PCA), principal coordinate analysis (PCoA), discriminant analysis of principal component (DAPC), factorial correspondence analysis (AFC), discriminant factorial analysis (AFD), factor analysis on a distance table (AFTD); classification methods which aim to form groups of individuals as similar as possible which are among others the ascending hierarchical classification

(AHC), neighbor-joining and Bayesian clustering model. The analysis of molecular variance (AMOVA) which consists of studying intra and inter population variation and the Mental test which aims to test the correlation between the genetic distance matrix and other distance matrices (environmental causes of genetic variability). This study showed that multivariate statistical methods, simultaneously analyzing several variables on an individual, are widely used in population genetics regardless of the type of phenotypic or molecular data. The methods identified in this review are nowadays the most used for the study of population genetic analysis and seem particularly useful and are similar to those found by [7], unlike the identification of new techniques: ADPC, AFD , FADT, AMOVA and Bayesian clustering model, which could be explained by the extraction of new types from complex genetic data resulting from the use of modern tools in recent years, in particular the types of markers which make the development of new specific methods . In most cases, the statistical analysis was carried out by combining different methods and in the majority it is the PCA associated with one of the classification methods, which corroborates the work of [7]. This study showed that the most used methods are: principal component analysis (32.50%), simple to carry out and that it exists in almost all statistical software and does not require any hypothesis on the distribution of the original variables ; ascending hierarchical classification (17.50%), no preliminary assumptions, simple and easy to interpret result; Bayesian clustering model (15%), constitutes a new efficient method taking into account the a priori information and that contained by the data with a low rate of assignment error and the analysis in principal coordinates (7.50 %) used because it does not require any hypothesis on the distribution of the origin variables and is less used than PCA because it is more complex taking place on distance or dissimilarity tables and not on the original variables.

Choice and overview of multivariate statistical methods for genetics

Genetic data generated using various analytical techniques can be analyzed using specific multivariate statistical methods or a combination of these methods. The choice of methods to be used depends on the achievable objectives defined in the studies and on the nature of the data depending on whether the data involve exclusively quantitative, Qualitative variables and in many cases, mixed data.

Ordination methods

Principal component analysis (PCA) is defined as a data reduction technique, applicable to quantitative data. The PCA transforms the multi-correlated variables into another set of uncorrelated variables for further study. These new sets of variables are linear combinations of original variables. It consists of a representation of a point cloud which corresponds to a data matrix with N individuals and P variables in a subspace of the P dimensional space absorbing the maximum of the total variance of the cloud, chosen so as to optimize a certain criterion. Intuitively, we will look for the subspace giving the best possible visualization of point cloud, this leads us to look for a rotation of the initial system of axes allowing to better see the cloud [8]. For a better visualization of the data, their analysis consists mainly in establishing the relations existing between the observations, between the variables and

between the observations and the variables and transforming them thereafter into new variables called principal components, the axes generated by these variables are called main axes. These axes are those of the ellipsoid of a multidimensional normal distribution. The representation in a space of reduced dimension, makes it possible to highlight possible structures within the data [9]. Several criteria make it possible to obtain the components, the inertia criterion is the oldest, the principle of which consists in considering each line of the data table as a point in a P-dimensional space, to visualize the positioning of individuals by compared to the others, it is necessary to project the point cloud of the space with P dimensions on a space of smaller dimension [8]. This method requires at least the stability of the variances to avoid distortions during projections on the factorial plane [10]. The distance calculated between individuals is the Euclidean distance provided with the metric on the reduced centered data, the diagonal of which is the inverse of the variances.

This method is based on the development of mutually independent eigenvalues and eigenvectors (main components) classified in decreasing order of size of the variance. Such components provide scatterplots of observations with optimal properties for studying the underlying variability and correlation.

PCA has various advantages such as: (1) these results can also be used for subsequent analyzes; (2) does not require any statistical model or any hypothesis on the distribution of the original variables, (3) A discriminating factor analysis can be performed on well-distinguished individuals on the PCA. PCA also presents difficulties such as: (1) more appropriate when the different variables have the same unit of measure, which can be avoided by standardizing all the variables and this normalization is done by dividing each variable by its standard deviation valued ; (2) when the original variables are not or weakly correlated (which would indicate a low probability that the variance of the elements can be explained by common features) or the variables are strongly correlated (in which case there would be danger of collinearity), it is not necessary to apply this method. Recently, a surge has been reported in the use of PCR in genetic diversity studies [2]. It is advisable to be very careful with regard to the objectives of this method: indeed, the individuals and the variables are presented on different spaces: if a variable defines a direction of the space of the individuals it cannot be summarized at a point and one cannot interpret a proximity between point-variables and point-individuals [11].

Principal coordinates analysis (PCoA)

This is another ordination method, somewhat similar to PCA, developed by [12]. PCoA regularly finds the eigenvalues and eigenvectors of a matrix containing the distances between all the data points, measured with the Gower distance or the Euclidean distance (Example Figure 1). It produces a 2 or 3 dimensional dispersion diagram of the samples, so that the distances between the samples reflect the genetic distances with minimal distortion. This method has drawbacks such as: (1) not providing a direct link between the components and the original variables and (2) Having complex functions of the original variables.

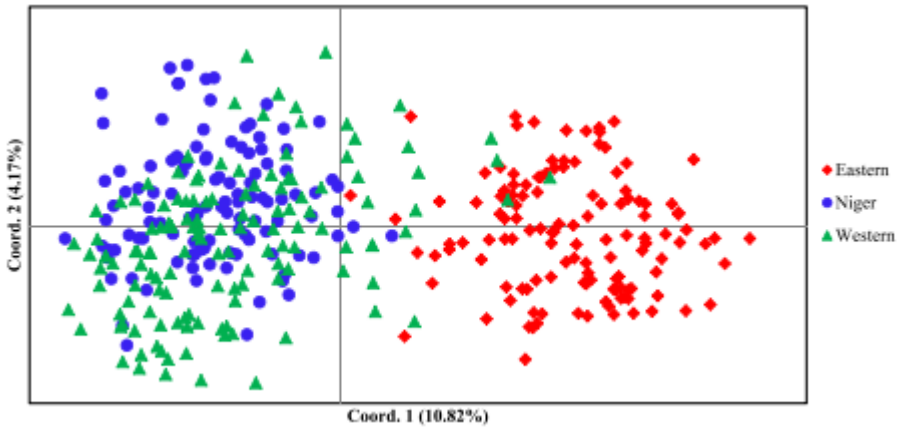

Fig.1. Principal Coordinate Analysis (PCoA) of date palm accessions from the eastern genetic pool (Iraq, Pakistan, United Arab Emirates, and Oman) in red, western genetic pool (Mauritania, Morocco, Tunisia, Libya, Sudan, and Egypt) in green, and the populations of South Niger in blue [13].

Factorial correspondence analysis (AFC)

Like the principal component analysis, the correspondence analysis can be presented from various sense. Since the work of [10], we have mainly used the algebraic and geometric properties of the descriptive tool that is the analysis. This method is not a particular case of the analysis in principal components although one can be reduced to this technique by making appropriate changes of variables (on the condition of treating each space separately). We can finally show that this is the search for the best simultaneous representation of two sets constituting the rows and columns of a data table. Correspondence analysis has a different field of application than principal component analysis. While the latter is reserved for tables of potentially heterogeneous measures and for the treatment of continuous numerical variables, the factorial analysis of correspondences is a method adapted to contingency tables and makes it possible to study the possible existing relationships between two nominal variables. The factorial analysis of correspondences amounts to carrying out the general analysis of a weighted point cloud in a space provided with the metric χ^2. We will therefore refer to the general analysis with metrics. Here we have a table X crossing two qualitative variables (P = 2), comes down to looking for the eigenvalues and vectors of the products of the two profile tables associated with X.

Factor analysis on distance table (AFTD)

AFTD [14] makes it possible to determine, in reduced space, the main directions of dispersion of a centered Euclidean representation, from a square, symmetrical and semi-defined positive matrix. To avoid the "double zero" problem, the data table can be recoded using similarity coefficients which do not consider the double absence to be significant in the comparison between species or varieties. The similarity coefficients can be reduced to dissimilarity coefficients by a simple linear relation, and allow to obtain an inter-distance matrix. The use of AFTD therefore induces a loss of variables during the analysis, since the distance matrix is presented as an array with two identical inputs (symmetric matrix of diagonal 0). These methods have the same objective as PCA: to find a configuration of

individuals in a small space, but the initial data are different; here we only know the $(n\,(n-1))/2$ distances or dissimilarity between objects, and not the variables describing them. In cases where a true Euclidean distance between individuals is available is only a version of the PCA [11].

Discriminant factorial analysis (AFD)

AFD is a data analysis technique that aims to describe, explain and predict an individual's membership in predefined groups. Originally, this method was studied by Fisher in 1936. It should also be noted that the discriminant analysis technique gives rise to two main approaches. in the first hand, the discriminating factor analysis (or descriptive discriminating analysis), which is a factorial or descriptive method, which like the PCA and the AFC [15], which aims to propose a new system of representation, latent variables formed from linear combinations of the predictive variables, which make it possible to discern groups of individuals as much as possible and in the other hand, the linear discriminant analysis, which is a predictive method consisting in constructing a classification function (assignment rule) used to predict the class in which an individual belongs based on the values taken by the predictor variables. In this sense, it belongs to the second family of classification methods as underlined by [4]. We place ourselves within the framework of the modeling of a Y qualitative variable with K modalities from explanatory variables $X1,\ldots,Xp$ quantitative. We assume that we have a sample of size n for which the explanatory variables and the variable to be explained were measured simultaneously. We therefore place ourselves in a so-called supervised framework, where each modality of Y represents a class (group) of individuals that we seek to discriminate. It is a question of seeking which are the linear combinations of the quantitative variables which make it possible to separate the K modalities as best as possible and to give a graphic representation (for factorial methods such as the analysis of principal components) which gives the best account of this operation. This visualization on the factorial space also makes it possible to describe the links between the variable to be explained Y and the P explanatory variables. We will therefore have to define: how to measure so that Y discriminates well and how to find u so that $Y = X.u$ discriminates at best [11]. Genomic data obviously poses problems for discriminant analysis, because the large number of genes (variables) in relation to the number of individuals makes it impossible to invert the matrix of intra-class covariances, which constitutes the main limitation of the application of discriminating factor analysis in the context of studying population genetic diversity.

Discriminant Analysis of Principal Components (DAPC)

The discriminant analysis of principal component (DAPC) is based on discriminant analysis not of the data itself, but of the main axes of a prior principal component analysis (PCA) [16]. This grouping method is of interest for highlighting the structuring of diversity in addition to an analysis by the Bayesian classification model under STRUCTURE [17]. The function find.clusters is used in DAPC to determine the number of possible groups in the population to be studied by successively executing K-means according to the rank of the numbers of possible groups (K) and calculates the Bayesian

information criterion (BIC) [18] of the corresponding model for each value of *K*. *find.clusters* and the *dapc* function are used for any set of quantitative data and have a specific implementation for genetic data (Example Fig. 2). DAPC is a very efficient ordination method for studying the genetic structuring of populations, it combines PCA, K-means and classification.

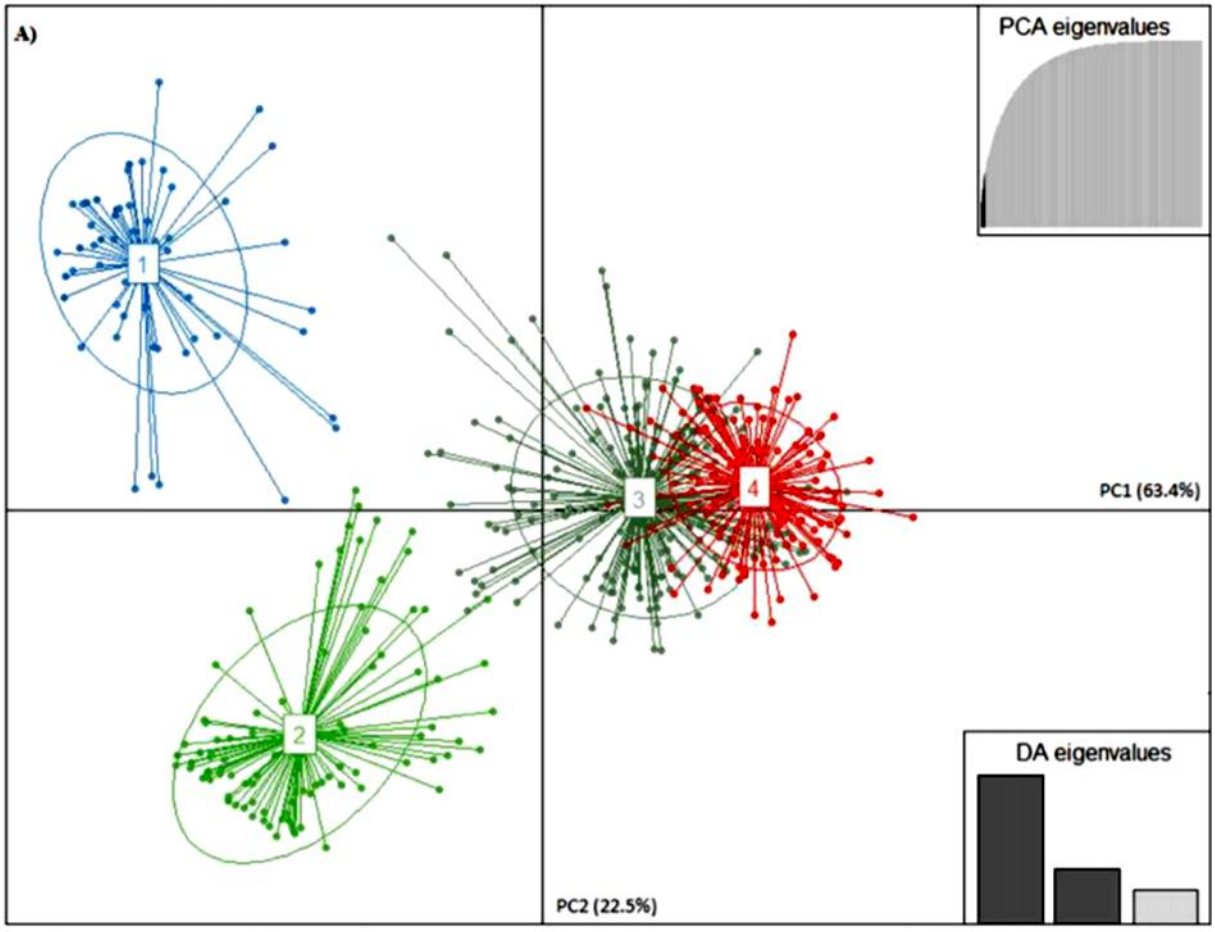

Fig.2. Genetic structure estimated by DAPC of 509 multilocus genotypes of almond cultivated in Lebanon on the first two axes 63.4% (PC 1) and 22.5% (PC 2). Individuals are represented by dots and groups as ellipses [19].

• Classification Methods

Ascending hierarchical classification (AHC)

The purpose of classification methods is to group individuals with common traits into an optimal number of distinct, homogeneous classes. The algorithmic criterion linking individuals by their resemblance, then determines a classification tree or dendrogram obtained in an ascending manner, starting from an individual towards all the objects. The criterion used is generally that of Ward [11]. Two groups will be all the more distinct as they are respectively homogeneous (low intra-class inertia) and as far apart as possible from each other (significant inter-class inertia). The merging of two classes increases inertia (dispersion of information); the grouping criterion then consists of bringing together the two classes for which the variation in inertia will be minimal (minimization the intra-class variation). At each step of the program, individuals are aggregated into classes, and step by step hierarchical tree will be performed. Cutting the tree at a given level partitions the dendrogram into an

optimal number of groups defined by the most characteristic species-modalities. The link between the factorial analyzes and the methods of AHC results in the classification of the objects according to their coordinates on the first factorial axes and thus makes it possible to clarify the structures. Used in conjunction with ordination methods, the hierarchical classification method, according to Ward's criterion, is an excellent complement to factorial analysis since it accurately reproduces the relationships between the pairs of closest objects, from a fairly natural inertia optimization criterion [20]. This method therefore remains the most complementary with regard to the effects of spatial contraction due to ordination methods [21]. In addition, the classification procedure makes it possible, no longer visually but by calculation, to assign each individual to a class. Hierarchical classifications have advantages and disadvantages: Hierarchical classifications allow the determination of optimal number of classes by reading the tree, and that the number of groups is not a priori known, unlike supervised approaches and it is a method flexible, but very costly in computation time.

Neighbor-joining

Neighbor-joining (Figure 3) is a method of phylogenetic reconstruction, developed by geneticist and evolutionary biologists Kimura and Nei based on the concept of the theory of evolution. This method is based on the matrix of genetic distances between individuals or species. The question to which this method gives an answer is: what is the relative similarity between individuals or species [22]. This method has drawbacks, the main one being the lack of a clear optimality criterion, although some applications try to approximate the minimum evolution. In addition, the requirement for a mutation rate more or less close to the biological clock and the fact that only a small part of available data is used after the transformation into a distance matrix has certain disadvantages. On the other hand, there are situations where the analysis of the rapprochement of neighbors would be the tool of choice. These include cases where the data cannot be considered hierarchical (a preliminary for other methods), for example, analyzes within a species and very large data sets when the more complex methods will not be completed within a reasonable time [23].

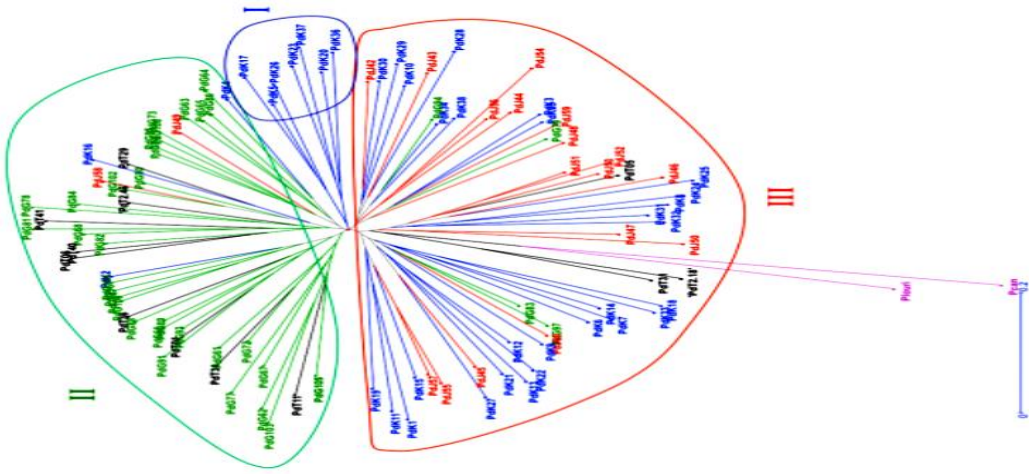

Fig.3. NJ dendrogram of 105 Tunisian accessions of the date palm constructed with DAS genetic distances based on 125 allele microsatellites. Bootstrap values are calculated under 1000 repetitions. In green Gabe's accessions, in blue Kerkennah, in red Djerba and in black Tozeur [24].

Bayesian Clustering Model

Bayesian classification model is implemented in STRUCTURE [25] is used to study the genetic structure of populations (Figure 3). This spatial grouping model belongs to two categories with or without admixture. This model is based on the calculation of the posterior probability that a given individual belongs to a group from a priori information (geographic origin) and likelihood (based on the genetic value of the individual). Different approaches have been proposed to estimate the number of groups in each model. To estimate the number of groups, STRUCTURE relies on a statistical criterion, noted $Ln\ P\ (D|K)$, which calculates the logarithm of the probability of the data for each run. From a statistical outlook, this criterion is a penalized adjustment measure based on a Gaussian approximation of the deviance of the model. As a general rule, STRUCTURE is performed for several values of K (number of groups) and $Ln\ P\ (D|K)$ is calculated for each case. In practice, it is recommended to plot $Ln\ P\ (D\mid K)$ as a function of K and to choose the value of K which corresponds to a plateau on the curve $Ln\ P\ (D|K)$. The ΔK criterion [26] aims to automate this process. The ΔK makes it possible to determine the optimal number that can be found in the population, the greatest variance corresponds to the optimal value of K. The ΔK is determined from logarithm of likelihood $Ln\ P\ (D|K)$.

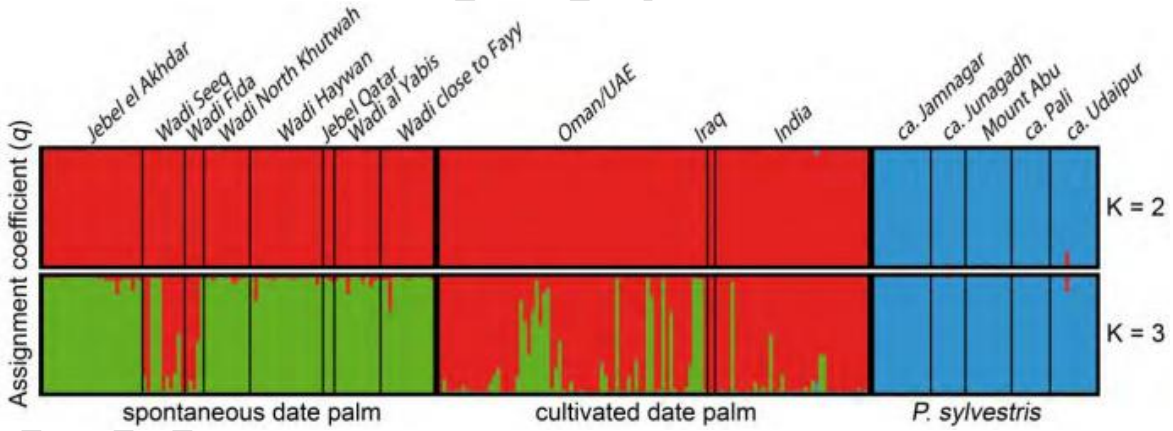

Fig. 4. Genetic structure of population studied with 272 individuals of the Phoenix genotyped for K = 2 and 3. Each individual is represented by a vertical bar partitioned in the colored segments representing the assignment coefficients [27].

Analysis of Molecular Variance (AMOVA)

This method makes it possible to study genetic diversity within and between populations. AMOVA [28] consists of a molecular study by analysis of variance. The analysis of variance makes it possible to assess the effect of one or more qualitative variables (the factors) on a quantitative variable (of gene expression). Before detailing the models used, it should be noted at once that the tests implemented in AMOVA imply independent observations. However, the expression of the different genes are not

independent, because of the regulations between genes and it is however difficult to take into account this structure of dependence between genes.

Mantel Test

The Mantel test is the most commonly used method to study the relationships between environmental variables and genetic structure. It makes possible to calculate the correlation between two matrices and to assess whether this correlation is significant by comparing it to the distribution of values obtained following permutations within the matrices [29]. This completely general method was quickly used to test the correlation between a matrix containing the genetic distances for each pair of individuals or subpopulations and a matrix containing the geographical distances between these same individuals or subpopulations (Fig. 5). Developed more recently, the partial Mantel test makes it possible to evaluate the effect of one variable on another, while controlling the effect of a third [30]. Since then, it has often been applied in population genetics, in particular to human population data to test the correlation between genetic distances and linguistic distances, taking into account geographic distances ([31]; [32]). Several difficulties are nevertheless present when using Mantel tests. First, the choice of genetic or environmental distance can influence the outcome. [31] show that the conclusions differed according to whether they applied their tests to genetic distances measured by the Fst genetic differentiation index (the most commonly used measure) or by the Rst index [33]. [34] showed that it is sometimes well to use $Fst/(1-Fst)$ and the natural logarithm of the geographic distances. Still in the same study, the choice of linguistic distance was essential. Finally, this method requires choosing which permutation technique will make it possible to obtain the free distribution of the correlations [35].

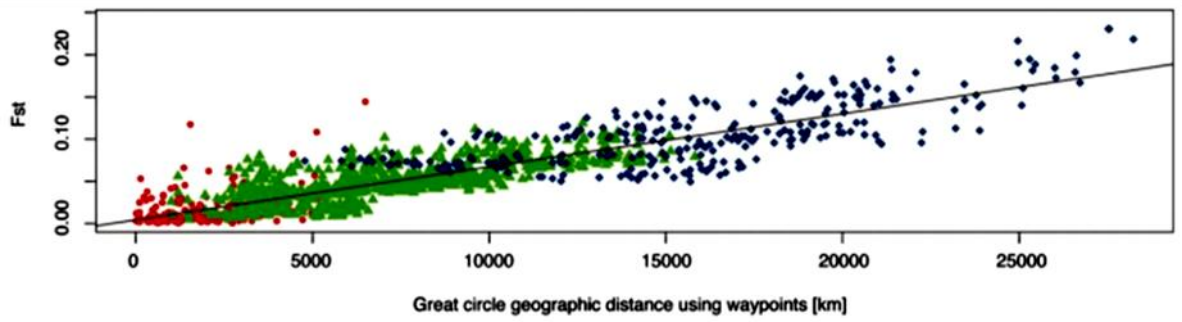

Fig.5. Mantel test used to assess the correlations between genetic distances and geographic distances calculated for each pair of populations from the HGDP. In this graph, the genetic differentiations (*Fst*) are plotted as a function of geographic distances taking into account the supposed crossing points of human migrations. The correlation between *Fst* and geographic distance is 0.8881 (P-value of the Mantel test = 0.000) [36].

Conclusion

This study addressed multivariate statistical methods used in population genetics. These methods have the advantage of making it possible to draw the main information contained in a matrix with several variables. It is important to remember, however, that there may be other multivariate methods not

presented in this study. Ordination methods are the most used in population genetics followed by classification methods. This study led to the conclusion that there is no single method for processing genetic data. Indeed there is not, a priori, a better approach; we have to find the right strategy combining exploration and modeling, adapted both to the data and to the desired objective.

References

1. Parizeau ME., (1997). La biodiversité. Edition de boeck, Bruxelles.

2. Bhandari HR, Bhanu AN, Srivastava K, Singh MN, Shreya et al., (2017). Assessment of Genetic Diversity in Crop Plants: An Overview. *Adv Plants Agric Res* 7(3): 00255. DOI: 10.15406/apar.2017.07.00255.

3. Singh S. and Pawar IS., (2005). Theory and Application of Biometrical Genetics. CBS Publishers, India.

4. Kamingu G., (2016). Analyse factorielle discriminante. Lareq One Pager, Vol. 11, no. 2, 7682.

5. Moher D., Liberati A., Tetzlaff J, Altman DG., Group P., (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA, statement. PLoS Med 6: e1000097.

6. Bérard C., Tanguay C., Bussières JF., (2014). Revue de la littérature reproductible. Annales de l'Unité de recherche en pratique pharmaceutique, p.1-9. http://urppchusj.wordpress .com.

7. Osawaru M. E., Ogwu M. C. and Aiwansoba R. O., (2015). Hierarchical Approaches to the Analysis of Genetic Diversity in Plants: A Systematic Overview. University of Mauritius. *Research journal* – Volume 21, 36 p.

8. Ding S., (2010). On the Application of PCA Technique to Fault Diagnosis. Tsinghua Science & Technology 15 (2) 138–144.

9. Baba K., Lahcen B., Latifa O., Choukri C., (2014). Application des méthodes d'analyses statistiques multivariées à la délimitation des anomalies de Sidi Chennane. Journal of Materials and Environmental Science. 5 (4) 1005-1012.

10. Benzécri J.-P., (1973). L'Analyse des Données. Tome 1 : la taxinomie. Tome 2 *:* L'Analyse des Correspondances (2de. Édition. 1976). Dunod, Paris.

11. Saporta G., (2011). Probabilités analyse des données et statistique. 3e édition révisée, Edition Technip 25 rue Ginoux, 75015 Paris : 155-266 p.

12. Schoenberg IJ., (1935). Remarks to Maurice Frchet's article "Sur la définition axiomatique d'une classe d'espaces distanciés vectoriellement applicable sur l'espace de Hilbert." Ann Math 38(3): 724-732.

13. Zango O., Emira C., Nathalie C., Zehdi-Azouzi S., Muriel G-B., Summar A. N., Alain L., Hervé R., Yacoubou B., Frédérique A., (2017). Genetic diversity of Southeastern Nigerien date palms reveals a secondary structure within Western populations. T*ree Genetics & Genome*s 13: 75, DOI 10.1007/s11295-017-1150-z.

14. Gower J.C., (1966). Sorne distance properties of latent root and vectors methods used in multivariate analysis. *Biometrika* 53: 325-338.

15. Mavita Y., (2013). Analyse factorielle des correspondances de Benzécri. Une illustration à l'aide de la métrique de Chi – deux, LAREQ One pager (décembre), 8 (11) : 101.

16. Jombart T., Devillard S. and Balloux F., (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. BMC Genetics 11: 94.

17. Arnaud JF., Cuguen J., Fenart S., (2011). Metapopulation structure and fine-scaled genetic structuring in crop-wild hybrid weed beets. *Heredity* 107 (5): 395-404.

18. Schwarz, G., (1978). Estimating the dimension of a model. The annals of statistics: 461–464.

19. Bariaa H., (2018). Structure et dynamique de la diversité génétique de l'amandier cultivé au Liban : facteurs biologiques et anthropiques. Sciences agricoles. Université Montpellier. Français. NNT : 2018MONTG049. Tel-01993151.

20. Saporta G., (1990). Probabilités. Analyse des données et statistique. *Edition Technip. Paris :* 493 p.

21. Legendre L. et Legendre P., (1984). Ecologie numérique. *Masson Ed, Paris,* 2 volume: 260 et 335p.

22. Anders B., (2010). Development and modification of bioactivity, Neighbor joining, in comprehensive natural products II.

23. Michael W. G., (2011). Neighbor Joining, in the Yeasts (Fifth Edition), Volume 1.

24. Zehdi-Azouzi S., Emira C., Karim G., Ahmed Ben A., Aymen B., Soumaya R., Mohamed Ben S., Sylvain S., Jean C. P., Frédérique A. B., Amel S. H., (2016). Endemic insular and coastal Tunisian date palm genetic diversity. Genetica 144:181–190. DOI 10.1007/s10709-016-9888-z.

25. Pritchard, J. K., M. Stephens, et al., (2000). Inference of population structure using multilocus genotype data. Genetics 155(2): 945-959.

26. Evanno, G., Regnaut S., (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14(8): 2611-2620.

27. Gros-Balthazard M., (2012). Les origines, l'histoire évolutive et biogéographique du palmier-dattier (*Phœnix dactylifera* L.) : l'apport de la génétique et de la morphométrie. Discipline : Evolution, Ecologie, Ressources génétiques, paléontologie. Université Montpellier II.

28. Kerr K., Martin M., Churchill G., (2000). Analysis of variance for gene expression microarray data, *Journal of Computational Biology*, 7, 819-837.

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

30. Smouse P. E., Long J. C., and Sokal R. R., (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Biol*. 35: 627–632.

31. Belle, E. and Barbujani G., (2007). Worldwide analysis of multiple microsatellites: language diversity has a detectable in influence on DNA diversity. Am. J. Phys. Anthropol. 133 **:** 1137–1146.

32. Wang, S., Lewis C. M. Jr., Jakobsson M., Ramachandran S., Ray N., (2007). Genetic variation and population structure in Native Americans. PLoS Genet. 3: e185.

33. Slatkin M., (1995). A measure of population subdivision based on microsatellite allele frequencies. Genetics 139: 457.

34. Rousset F., Raymond M., (1997). Statistical analyses of population genetic data: New tools, old concepts. *Trends Ecol. Evol*., 12 : 313-317.

35. Legendre, P., (2000). Comparison of permutation methods for the partial correlation and partial mantel tests. *Journal of Statistical Computation and Simulation* 67: 37–74.

36. Ramachandran, S., Deshpande O., Roseman C. C., Rosenberg N. A., Feldman M. W., (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. USA 102: 15942–15947.