Usefulness of the clustering methodologies to discriminate between purebred and crossbred individuals

Silvia Teresa Rodríguez-Ramilo, M. A. Toro, J. Fernandez

Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)
Available online at www.inia.es/sjar

Spanish Journal of Agricultural Research 2010 8(2), 347-355
ISSN: 1695-971-X

Usefulness of the clustering methodologies to discriminate between purebred and crossbred individuals

S. T. Rodríguez-Ramilo[1,2]*, M. A. Toro[1,3] and J. Fernández[1]

[1] *Departamento de Mejora Genética Animal. Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA). Ctra. A Coruña, km 7,5. 28040 Madrid. Spain*

[2] *Departamento de Bioquímica, Genética e Inmunología. Facultad de Biología. Campus Universitario de Vigo. 36310 Vigo. Spain*

[3] *Departamento de Producción Animal. ETS Ingenieros Agrónomos. Universidad Politécnica de Madrid (UPM). Ciudad Universitaria. 28040 Madrid. Spain*

Abstract

Molecular markers have been successfully used to distinguish between livestock species and breeds not closely related, for example through the clustering methodology. However, the differentiation between purebred and crossbred individuals would be an appealing purpose that has been little explored. In this study three clustering approaches are tested for their ability to detect crossbred individuals and to separate them from pure ones. Real microsatellite data from Iberian and Duroc breeds were utilised as an example. Simulated F1, Iberian and Duroc backcrossed individuals obtained from the real microsatellite were also assessed. The results of this study indicate that the clustering methods showed a reduced ability to detect the original subpopulations (Iberian breed, Duroc breed, F1, Iberian backcross and Duroc backcross). Reasons for such performance could be the absence of Hardy-Weinberg and linkage equilibrium within the subpopulations and the fact that the Iberian group was compound by individuals belonging to different strains. To test the influence of these factors an allele randomisation procedure was performed within each subpopulation. After that, none of the methods recovered the five groups, but the algorithm implemented in BAPS (Bayesian analysis of population structure) gave a partition where pure Iberian individuals were separated for the rest. It can be concluded that the lack of homogeneity within groups is the main cause of the reduced accuracy of the clustering methods in the separation of pure and crossed individuals.

Additional key words: Bayesian method, breed, Iberian and Duroc pigs, microsatellite loci, Nei's minimum distance, simulated annealing.

Resumen

Utilidad de las metodologías de agrupamiento para discriminar entre individuos puros y cruzados

Los marcadores moleculares se han empleado satisfactoriamente para distinguir especies ganaderas y razas no muy relacionadas, por ejemplo a partir de la metodología de agrupamiento. Sin embargo, la diferenciación entre individuos puros y cruzados podría ser un objetivo interesante que ha sido escasamente explorado. En este estudio se evalúan tres métodos de agrupamiento para detectar individuos cruzados y separarlos de los individuos puros. Se utilizaron datos reales de microsatélites de Ibérico y Duroc como ejemplo. También se evaluaron individuos simulados F1 y retrocruces de Ibérico y de Duroc obtenidos a partir de los microsatélites reales. Los resultados de este estudio indican que los métodos de agrupamiento presentaron una capacidad reducida para detectar las subpoblaciones originales (raza Ibérica, raza Duroc, F1, retrocruce de Ibérico y retrocruce de Duroc). Las razones de este comportamiento pueden ser la ausencia de equilibrio de Hardy-Weinberg y de ligamiento dentro de subpoblaciones y el hecho de que el grupo Ibérico está formado por individuos pertenecientes a distintas variedades. Para evaluar la influencia de estos factores se realizó un procedimiento de aleatorización de alelos dentro de cada subpoblación. Después de este procedimiento, ninguno de los métodos proporcionó los cinco grupos, pero el algoritmo implementado en BAPS (Bayesian analysis of population structure) proporcionó una partición en la que los individuos Ibéricos puros eran separados de los demás. Se puede concluir que la ausencia de homogeneidad dentro de grupos es la causa principal de la reducida precisión de los métodos de agrupamiento en la separación de individuos puros y cruzados.

Palabras clave adicionales: cerdo Ibérico y Duroc, distancia mínima de Nei, loci microsatélites, método Bayesiano, raza, templado simulado.

* Corresponding author: silviat@uvigo.es
Received: 02-04-09; Accepted: 15-03-10.

Introduction

Livestock populations have been subjected to a variety of evolutionary forces during their histories. The cumulative effects of genetic drift, caused by founder effects and small population size, together with natural and artificial selection has led to the formation of distinct breeds. Studies of breed relationships, based on genetic markers, have found that breeds are significantly differentiated at the genetic level. Genetic markers could, therefore, provide a potentially powerful way of identifying the breed to which an individual belongs, when pedigree information is not available (Oldenbroek, 2007).

The identification of crossbred animals is essential in some situations. Oldenbroek (2007) indicated that attention should be given to the conservation of local breeds, taking the introduction of crossbreeding as an example. First, breeding schemes should guarantee the maintenance of viable populations of the local breed through a sound pure breeding scheme. In addition, the breed might be used for the production of commercial crosses with a high performance breed. The commercial crosses might benefit from higher input production systems, while the local breed should be maintained in its original production environment to maintain its adaptation characteristics. The use of the local breed as a female populations (instead of a male population, which might be more profitable) may be advisable to guarantee the maintenance of a large population of the local genotype adapted to the production environment. Finally, the use of a high performance breed that will produce crosses that can not be distinguished from the local breed is not advisable, because of the risk of involuntary introduction of different genotypes into the local breed. In these circumstances it would be interesting to separate purebred from crossbred individuals based on the molecular information available for them.

During the last years several clustering methods have been proposed to separate a set of individuals into different populations if their genetic origin is unknown beforehand or to study the correspondence between the inferred genetic clusters and known predefined populations (Pritchard *et al*., 2000; Falush *et al*., 2003; Corander *et al*., 2003, 2004; Corander and Tang, 2007). The procedures generally involve Markov chain Monte Carlo (MCMC) approaches. These clustering methods are useful when genetic data for potential source populations are not available (that is, there are not accessible data on reference populations), in opposition to assignation methods (where data on reference populations are needed to assign the problem individuals to those reference populations). Thus, clustering methodologies offer a powerful tool to answer questions of ecological, evolutionary or conservation relevance (Manel *et al.*, 2005).

Fully Bayesian clustering methods have been proposed to estimate hidden population substructure. In these approaches both the allele frequencies of the molecular markers and the number of genetically divergent populations are processed as random variables (BAPS: Bayesian analysis of population structure; Corander *et al*., 2003, 2004; Corander and Tang, 2007). These methods operate by searching the number of inferred clusters (*K*) and the classification of individuals to those clusters that minimise Hardy-Weinberg and linkage disequilibrium (HWD and LD) within those subpopulations with no prior information on the population sampling design.

A not fully Bayesian method (STRUCTURE: Pritchard *et al*., 2000; Falush *et al*., 2003), have also gained popularity for the clustering analysis. In this approach the number of clusters is an input parameter and the analysis should be carried out with different *K*s to find out the number of clusters providing the highest likelihood. As in the previous methodology, the underlying assumption is that the inferred subpopulations are in Hardy-Weinberg and linkage equilibrium (HWE and LE).

Two main features distinguish BAPS from STRUCTURE. First, in BAPS the number of populations is treated as an unknown parameter that could be estimated from the data set. Second, after BAPS version 2 a stochastic optimisation algorithm is implemented to infer the posterior mode of *K* instead of the MCMC algorithm also utilised in STRUCTURE. Notwithstanding, the most widely used genotypic clustering method is that implemented in the program STRUCTURE.

Methods that do not make any assumption about HWE and LE have been also developed to infer subpopulation hidden structure (Dupanloup *et al*.,

Abbreviations used: BAPS (Bayesian analysis of population structure), HWD (Hardy-Weinberg disequilibrium), LD (linkage disequilibrium), MCMC (Markov chain Monte Carlo), *MC1R (melanocortin receptor 1* gene), MGD (maximisation of the genetic distance method), *OCA2 (pink-eyed dilution* gene).

2002). Recently, Rodríguez-Ramilo *et al*. (2009) proposed a new approach to estimate the number of clusters and to assign individuals to the inferred subpopulations. The implemented criterion is the maximisation of the averaged genetic distance between subpopulations using a *simulated annealing* algorithm.

In the present study, the above three clustering approaches were tested for their ability to detect crossbred individuals and to separate them from pure ones. To deal with this objective, real microsatellite data from Iberian and Duroc breeds were used as an example. Simulated F1, Iberian and Duroc backcrossed individuals obtained from the real microsatellite were also assessed.

Material and methods

Real data

The microsatellite data set from Alves *et al*. (2006) was evaluated. The individuals' genotype for 36 microsatellite markers, two on each autosome, was available for one hundred and seventy Iberian individuals with the following distribution across different strains: 31 Torbiscal, 32 Guadyerbas, 50 Retinto, 30 Lampiño and 27 Entrepelado. The equivalent molecular information on a total of 64 Duroc pigs was also used.

Simulated data

From the real microsatellite data of the 170 Iberian animals (the five strains together) and the 64 Duroc ones, 100 individuals were generated simulating a filial generation (F1). To obtain the genotypic data of the F1 population, one individual was randomly chosen from each breed (Iberian and Duroc) to be the parents. Then, one allele at random was chosen from each of the selected parental for the first locus, constituting the genotype of the F1 individual for that particular locus. The same procedure was followed independently with the remaining loci, as free recombination between markers was assumed. The obtained F1 individuals were also crossed with the 170 Iberian and the 64 Duroc pigs respectively, to generate another 100 individuals for each corresponding backcross (Iberian and Duroc backcrosses, respectively). The procedure to get the genotypes of the individuals of the Iberian and Duroc backcrosses was the same as explained before. From the real data 10 replicates of the F1 and the backcrosses were generated to evaluate with the three clustering methodologies.

Allele randomisation

To ascertain the influence of the existence (or not) of HWE and LE in the microsatellite data set within the initial subpopulations (Iberian breed, Duroc breed, F1, Iberian Backcross and Duroc Backcross) an allele randomisation procedure was also implemented. The reason is that HWD and/or LD could compromise the accuracy of the Bayesian methodologies being evaluated (Kaeuffer *et al.*, 2007). The randomisation also promotes the homogenisation within subpopulations, removing the differentiation between subgroups of the same breed.

The randomisation procedure within subpopulations was implemented as follows. Consider a subpopulation genotyped for L loci where the matrix of genotypes (G) for $i = 1,\ldots, N$ individuals can be illustrated as

$$G = \begin{bmatrix} g_1^{(1,1)} & g_1^{(2,1)} & g_1^{(1,2)} & g_1^{(2,2)} \cdots & g_1^{(1,L)} & g_1^{(2,L)} \\ \vdots & & & & & \\ g_i^{(1,1)} & g_i^{(2,1)} & g_i^{(1,2)} & g_i^{(2,2)} \cdots & g_i^{(1,L)} & g_i^{(2,L)} \\ \vdots & & & & & \\ g_N^{(1,1)} & g_N^{(2,1)} & g_N^{(1,2)} & g_N^{(2,2)} \cdots & g_N^{(1,L)} & g_N^{(2,L)} \end{bmatrix} \quad [1]$$

where $g_i^{(1,L)}$ is the allele 1 at locus L of individual i. The allele randomisation implies to move alleles by chance within each column independently with the aim of breaking up the combinations of alleles (*i.e.* randomising the genotypes within loci), removing linkage disequilibrium (as process is independent for each locus) and erasing the possible substructure of the defined group.

GENEPOP software version 4.0.6 (Raymond and Rousset, 1995) was used for the linkage and Hardy-Weinberg equilibrium analyses in each replicate data set. To compute LE the option of *exact test for genotypic disequilibrium* was selected with the suboption of *test for each pair of loci in each subpopulation*. A P-value for each pair of loci is computed for all subpopulations (Fisher's method), and the high (or reduced) proportion of loci pairs with significant linkage ($P < 0.05$) is a measure of the LD (or LE). Regarding HWE, the option *F_{ST} and other correlations, isolation by distance* has been chosen with the suboption of *all populations*. The Wright's F statistic (Wright, 1931) F_{IS} (inbreeding

Table 1. Mean percentage of loci pairs with significant linkage and mean F_{IS} (± standard error) before and after the allele randomisation procedure in each subpopulation

Subpopulation	Alelle randomisation			
	Before		After	
	Significant linkage (%)	F_{IS}	Significant linkage (%)	F_{IS}
Iberian	83.38 ± 0.05	0.17 ± 0.00	5.37 ± 0.31	–0.06 ± 0.00
Duroc	20.44 ± 6.29	0.15 ± 0.00	5.17 ± 0.28	–0.08 ± 0.00
F1	15.73 ± 0.87	–0.09 ± 0.02	5.10 ± 0.30	–0.09 ± 0.00
Iberian backcross	10.17 ± 0.56	–0.02 ± 0.01	5.03 ± 0.35	–0.03 ± 0.01
Duroc backcross	10.21 ± 0.27	–0.02 ± 0.00	4.89 ± 0.38	–0.02 ± 0.00

coefficient describing the divergence of observed heterozygosis from expected heterozygosis within populations assuming panmixia) is provided. Table 1 shows the mean percentage of loci pairs with significant linkage and the mean F_{IS} (± standard error) before and after the allele randomisation procedure in each subpopulation.

Clustering analysis

Bayesian methods

The analyses were performed with STRUCTURE version 2.1 (Pritchard *et al*., 2000; Falush *et al*., 2003) and BAPS version 4.14 (Corander *et al*., 2003, 2004; Corander and Tang, 2007).

Parameters for the implementation of STRUCTURE comprise a burn-in of 10,000 replicates following 50,000 replicates of MCMC. Specifically, the admixture model (each individual may draw some fraction of its genome from each of the available populations) and the option of correlated allele frequencies between populations (in the different populations allele frequencies are likely to be similar probably due to migration or shared ancestry) were selected, as this configuration is considered the best by Falush *et al*. (2003). Similarly, the degree of admixture has been inferred from the data. Lambda, the parameter of the Dirichlet distribution of allelic frequencies, was set to one, as the manual of STRUCTURE advices.

The range of tested *K*s was set from 2 to 20 as the number of simulated populations was five (Iberian breed, Duroc breed, F1, Iberian Backcross and Duroc Backcross) but the Iberian group was made of five different strains. Five runs were carried out for each of the ten data sets and for each possible number of clusters (from 2 to 20) in order to quantify the variation in the likelihood of the data for a given *K*.

The criterion implemented in STRUCTURE to determine *K* is the likelihood of the data for a given *K*, *L*(*K*). The number of populations is identified using the maximal value of this likelihood returned by STRUCTURE. However, it has been observed that once the real *K* is reached the likelihood at larger *K*s plateaus or continues increasing slightly, and the variance between runs increases (Evanno *et al*., 2005). Consequently, the distribution of *L*(*K*) did not show a clear mode for the true *K*. Notwithstanding, an *ad hoc* quantity based on the second order rate of change of the likelihood function with respect to *K* (Δ*K*) did show a clear peak at the true value of *K*. Evanno *et al*. (2005) suggested to estimate Δ*K* as:

$$\Delta K = \left| avg\left[L\left(K+1\right)\right] - 2\times avg\left[L\left(K\right)\right] + avg\left[L\left(K-1\right)\right]\right| / sd\left[L\left(K\right)\right] \quad [2]$$

where *avg* is the arithmetic mean across replicates and *sd* is the standard deviation. The value of *K* selected will correspond to the modal value of Δ*K*. Once *K* was estimated (using the Δ*K*), a classification test was also performed on the replicate with the maximal value of the likelihood for that particular *K*. Thus, we obtained the highest percentage of individuals corresponding to the predefined subpopulations that are grouped together by each clustering method. Averaging across the five replicates of the estimated *K* is not possible as the groups are not equivalent between replicates.

Another Bayesian clustering method, the one implemented in BAPS, was also used to infer population structure based on multilocus genotypes. This program estimates the hidden population substructure by testing

whether the inferred subpopulations are in Hardy-Weinberg and linkage equilibrium. A major advantage compared to most other methods is that the number of populations is treated here as an unknown parameter that can be estimated from the data set. The maximum number of clusters was set to 20 and the clustering of individuals option was the elected one. A further advantage is the computing time required to run the analysis that is extremely short compared to other available methods.

Maximisation of the genetic distance method (MGD)

The rationale behind this approach (Rodríguez-Ramilo *et al*., 2009) is that it is expected that highly differentiated populations show a high genetic distance between them. This distance can be calculated from the molecular markers information without assumptions concerning to HWE or LE. The approach utilises a *simulated annealing* algorithm (Kirkpatrick *et al*., 1983) to find the partition which shows the maximal average genetic distance between populations. From all the genetic distances previously published in the literature (Laval *et al*., 2002), one of the most used is the Nei's minimum distance (Nei, 1987). One of the advantages of Nei's minimum distance is that it can be calculated through the pairwise coancestry between individuals (Caballero and Toro, 2002). In this approach, Nei's minimum distance between two subpopulations can be expressed as:

$$D_{AB} = \left[\left(f_{AA} + f_{BB}\right)/2\right] - f_{AB} \qquad [3]$$

where f_{AA} is the average coancestry between individuals of subpopulation A and f_{AB} is the average pairwise coancestry between all possible couples of individuals, one from subpopulation A and other from subpopulation B. The molecular coancestry can be easily calculated from molecular information.

As the values for the averaged distance reached no maximum in a sensible range, a similar procedure to the proposed in Evanno *et al*. (2005) was implemented. It was based on the rate of change in the averaged genetic distance between successive K values (ΔK) calculated as:

$$\Delta K = \left| D\left(K+1\right) - 2D\left(K\right) + D\left(K-1\right) \right| \qquad [4]$$

where D is the averaged genetic distance in the optimal solution for a given K. The inferred number of clusters corresponds to the value with the highest ΔK.

Accuracy of the clustering analysis

To determine the performance of each method, the number of inferred clusters (K) was evaluated through the modal value over replicates. A more detailed measure can be obtained as the proportion of individuals correctly grouped with their predefined population. This parameter was evaluated by averaging over clusters the highest proportion of each subpopulation (*i.e.* larger group of individuals) located at the same cluster. This mean value was also averaged over replicates.

Results

No method inferred the partition with five clusters, both before and after the allele randomisation procedure, in any of the replicates. The number of inferred clusters before the allele randomisation was as follows. BAPS differentiated from 11 to 16 clusters, denoting a higher sensibility against differences within the predefined groups. Iberian pigs were divided into up to 8 different clusters and Duroc into 2-5 groups. The modal value was 13. Both STRUCTURE and MGD yielded three groups as the best partition. For these two methods, the general pattern was as follows. Each backcross grouped with its predominant pure line (*i.e.* the Iberian backcross with the Iberian and the Duroc backcross with the Duroc), while F1 individuals were placed into both groups. Although with similar performances, MGD method was more powerful because the proportion of backcrosses which grouped with the opposite pure breeds was lower than in the STRUCTURE partition.

However, in both approaches (STRUCTURE and MGD) an additional cluster was formed with a reduced proportion of individuals of different subpopulations. When the precise individuals included in this group were assessed, it could be observed that they were mostly Guadyerbas pigs and their corresponding descendants (F1 and backcrosses), although in some replicates there were also Torbiscal individuals (data not shown). This performance was also observed for BAPS where, in all replicates, Guadyerbas and Torbiscal individuals appeared grouped in separates clusters together with some backcrossed pigs (never F1 ones).

The number of inferred clusters after the allele randomisation was always three even in the BAPS approach. The results of STRUCTURE were worse after randomisation, because, although this method performed

Table 2. Mean proportion of each subpopulation classified in each cluster (± standard error) with BAPS (Bayesian analysis of population structure), STRUCTURE and MGD (maximisation of the genetic distance method) approaches before and after the allele randomisation procedure

	Iberian	Duroc	F1	Iberian backcross	Duroc backcross
Before allele randomisation					
BAPS	0.38 ± 0.03	0.62 ± 0.03	0.41 ± 0.03	0.51 ± 0.03	0.58 ± 0.03
STRUCTURE	0.63 ± 0.02	0.93 ± 0.03	0.49 ± 0.02	0.49 ± 0.01	0.73 ± 0.03
MGD	0.68 ± 0.03	1.00 ± 0.00	0.58 ± 0.03	0.63 ± 0.03	0.98 ± 0.00
After allele randomisation					
BAPS	1.00 ± 0.00	1.00 ± 0.00	0.96 ± 0.01	0.89 ± 0.02	0.81 ± 0.03
STRUCTURE	0.50 ± 0.02	0.95 ± 0.00	0.50 ± 0.00	0.38 ± 0.01	0.74 ± 0.00
MGD	0.98 ± 0.01	1.00 ± 0.00	0.85 ± 0.02	0.59 ± 0.02	0.84 ± 0.02

generally as before, there were more mixing of some subpopulations across clusters. Contrarily, MGD showed some improvement, as F1 individuals appeared now separated from the pure breeds. Notwithstanding, still some Iberian backcrossed individuals were allocated in the same group as pure Iberian individuals. BAPS, yielded groups with 100%, 50% or more and less than 50% of Iberian genome.

Table 2 shows the mean proportion of correct groupings over replicates for each predefined subpopulation before and after the allele randomisation procedure. Before the allele randomisation procedure BAPS showed a reduced percentage of correct groupings (from 0.38 ± 0.03 to 0.62 ± 0.03) due to the large number of cluster the method yields. The highest proportion corresponded to the MGD method (1.00 ± 0.00 in the Duroc subpopulation). After the allele randomisation procedure (*i.e.* when HWE and LE had been established and subpopulations homogenised) the best performance corresponded to BAPS (from 0.81 ± 0.03 to 1.00 ± 0.00). However, the accuracy of STRUCTURE, in general, was reduced, due to the higher mixing of some subpopulations across clusters.

Discussion

Molecular markers have been successfully used to distinguish between livestock species and breeds not closely related (Toro *et al*., 2009). However, the differentiation between purebred and crossbred individuals would be a challenging objective that would be useful (*e.g.* Delgado and Martínez, 2007).

In this study three clustering methodologies were tested to detect crossbred individuals and to separate them from pure ones. Real microsatellite data from Iberian and Duroc breeds and simulated F1, Iberian and Duroc backcrossed individuals obtained from the real data were assessed as a practical example. The results indicate that the evaluated methods present a reduced ability to detect the five predefined initial subpopulations (Iberian breed, Duroc breed, F1, Iberian backcross and Duroc backcross). Regarding the main objective of the study, it is especially relevant the fact that pure populations (either Iberian or Duroc) were never classified alone but «mixed» individuals were also included in the same cluster, despite the high genetic differentiation between both breeds (F_{ST} = 0.16). None of the clustering methods was able to separate pure Iberian animals from their corresponding backcrosses, but some times even F1 individual could not be segregated.

Analysing the composition of the clusters inferred by the different methods it seems that one of the causes for such a poor performance could be the lack of homogeneity within breeds. Clustering methods tended to separate more clearly different strains within the Iberian pigs than these from the crossbreed animals. Moreover, BAPS, which presents a higher sensibility to genetic differences, separate in different cluster even animals from the same strain. For these reasons, it could be an interesting purpose to test the clustering performance of the methodologies using an experimental validation with real F1 and backcrossed individuals.

Fabuel *et al.* (2004) already indicated that there is a lack of homogeneity within the Iberian breed. Therefore, the study could consider that the real number of subpopulations was greater than five, accounting for all the strains within the Iberian breed, the subpopulation of Duroc, and all possible F1 and backcrosses

arising from the mixing of any of the five strains and the Duroc pigs. However, Fabuel *et al.* (2004) also found that for all Iberian subpopulations, the genetic distance to the Duroc breed is greater than that to any of the other subpopulations of the Iberian breed. In addition, when these authors used STRUCTURE algorithm within the Iberian breed forcing the same number of clusters and subpopulations (five), only Torbiscal genomes and 99.5% of Guadyerbas genomes were classified as two separated clusters. However, the results were less clear for the other subpopulations (Retinto, Entrepelado and Lampiño strains), whose genomes were attributed to diverse clusters. Consequently, the real number of clusters is not easy to determine and, in this study, was kept to five instead of a higher number.

Other factors that have been claimed to affect the performance of Bayesian clustering methods are the presence of Hardy-Weinberg and linkage disequilibrium within subpopulations. Kaeuffer *et al.* (2007) showed that a high LD between loci increases the probability of detecting spurious clustering with STRUCTURE. Rodríguez-Ramilo *et al*. (2009) also found biased estimates when using BAPS in simulated scenarios comprising HWD and/or LD. For this reason, other method that does not take into account Hardy-Weinberg and/or linkage equilibrium was included in this study to carry out the analyses.

To test if the above (HWD and/or LD) were the reasons for the inaccuracy of the Bayesian evaluated methodologies, the analyses were also carried out after an allele randomisation procedure within each subpopulation to test the methodologies in the most favourable situations. This procedure assured the equilibria (HWE and LE) breaking the possible substructure of the breeds (*i.e.* differences between strains disappeared). The performance of methods changed radically (especially BAPS) now grouping all pure individuals together. Notwithstanding, separation of pure from crossbred animals was still not attained except for BAPS almost yielding private clusters for Iberian individuals. In fact, STRUCTURE provided worst results with a higher proportion of crossbred animals grouped with pure ones. The randomisation procedure is convenient to clarify that the failure of the clustering methods is due to the lack of HWE and LE, especially in the pure breeds (Iberian and Duroc), because one can expect in other examples that the purebreds satisfy both conditions. In this sense, the randomisation of the Iberian and Duroc subpopulations is more realistic than the randomisation of F1 and backcrosses, as these subpopulations may never be in HWE and LE (but it will depend on differences in allelic frequencies). García *et al*. (2006) also described how methods such as the Bayesian-based model proposed by Pritchard *et al*. (2000) may lead to population mixes when using a panel of 25 microsatellite markers. However, they found that STRUCTURE clustering algorithm estimated the ancestry of the simulated populations with a reasonable accuracy (see Table 4 in García *et al.,* 2006).

Separation of populations via clustering methods relays on the differences in frequency for alleles of the genotyped loci. The greater of such difference the higher the power of the methodology to separate groups. The increase in the number of genotyped markers could increase the accuracy of methods even when available markers exhibit not extreme opposed frequencies in different populations. But this enhanced power will have as a side effect the higher probability of the methods of detecting within populations differences. The present study has proven that differences between Iberian and Duroc, although large (F_{ST} = 0.16), are not enough to completely separate pure from crossbred pigs using clustering methodologies. The problem could be reduced when trying to identify the presence of white-coat pig genes (from Landrace or Large White) as the F_{ST} values are 0.15 and 0.22 between the Iberian population and the Landrace and Large White, respectively (García *et al*., 2006).

The accuracy of the methods is also affected by the particular combination of molecular markers evaluated. That is, the precision of a particular combination of markers is valid for this particular population but not for all populations. Due to the increasing availability of SNPs (single nucleotide polymorphisms) for the livestock species it will be possible, in the future, to look for a set of markers where the frequencies are extremes and opposite in different breeds to use as a panel for the classification of pure and crossbred individuals, instead of using a large number of little informative ones. In this scenario the calculation of exclusion probabilities may be a better tool than the clustering methodology.

In the extreme situation of an allele absent from one of the populations we would have a diagnostic allele/loci to perform exclusion analysis, detecting the crossbred animals if the «foreing» allele appear in its genome. In the particular case of the Iberian pig Fernández *et al.* (2004) and Dalvit *et al.* (2007) indicated the usefulness of the two colour genes *MC1R* and *OCA2* to identify breeds in Iberian versus Duroc individuals.

Allele 4 of the *MC1R* gene is not present in Iberian pigs and, thus, is an indicator of the presence of Duroc genome in the individual. This approach has the advantage of avoiding the problem of substructuring of the Iberian breed (*i.e.* genetically differentiated strains) as none of them carry the diagnostic allele. Notwithstanding, it should be noticed that, as Fernández *et al.* (2004) indicated, the calculation of the exclusion probabilities originally assume absence of linkage between markers and requires a precise knowledge of the frequencies of the specific alleles in the introgressed breed. Moreover, Dalvit *et al.* (2007) highlighted that research should be extended to a greater number of individual samples to verify the exclusiveness of the detected markers. In situations where a specific breed haplotype is not available, more general approaches, such as the clustering ones, could be implemented as a tool to differentiate between purebred and crossbred individuals, with the limitations the present study highlight.

Assignation methodology is another available tool to classify individuals based on their molecular information. The idea is comparing the genotype of the unknown animal with the allelic frequencies of an already defined group of individuals and calculating the probability of the particular genotype of belonging to that population (notice that HWE and LE is also assumed). The ability of such an approach to detect crossbred individuals is very dubious because, to get reliable results, the comparison groups (pure and crossbred individuals) have to be the same as the tested individual or, at least, have the same frequencies as its population of origin. Therefore, the accuracy of the analysis is dependent on the structure of the populations.

The present study has shown that the usefulness of the clustering methods is generally low and highly dependent on the homogeneity of the sampled subpopulations, being this factor more important than the effect of HWD and LD. This is especially relevant for the Iberian breed, as this comprises quite differing subpopulations (strains). The detection of introgression seems more powerful through exclusion techniques using diagnostic alleles or a panel of a set of markers with extreme opposite frequencies.

Acknowledgments

We thank A. Caballero and two anonymous referees for helpful comments on the manuscript. Authors want to thank also L. Silió and the Pig-Breeding Group from the Animal Breeding Department of the INIA for microsatellite genotypes of pigs. This work was funded by the Ministerio de Educación y Ciencia and Fondos Feder (CGL2006-13445-C02/BOS, CGL2004-03920/B0S), Plan Estratégico del INIA (CPE03-004-C2), and Xunta de Galicia.

References

ALVES E., FERNÁNDEZ A.I., BARRAGÁN C., ÓVILO C., RODRÍGUEZ C., SILIÓ L., 2006. Inference of hidden population substructure of the Iberian pig breed using multilocus microsatellite data. Span J Agric Res 4, 37-46.

CABALLERO A., TORO M.A., 2002. Analysis of genetic diversity for the management of conserved subdivided populations. Conserv Genet 3, 289-299.

CORANDER J., TANG J., 2007. Bayesian analysis of population structure based on linked molecular information. Math Biosci 205, 19-31.

CORANDER J., WALDMANN P., SILLANPAA M.J., 2003. Bayesian analysis of genetic differentiation between populations. Genetics 163, 367-374.

CORANDER J., WALDMANN P., MARTTINEN P., SILLANPAA M.J., 2004. BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics 20, 2363-2369.

DALVIT C., DE MARCHI M., CASSANDRO M., 2007. Genetic traceability of livestock products: a review. Meat Sci 77, 437-449.

DELGADO J.V., MARTÍNEZ A., 2007. Método molecular para verificar el origen genético de cerdos y carnes ibéricos. Cárnica 2000 288, 89-91. [In Spanish].

DUPANLOUP I., SCHNEIDER S., EXCOFFIER L., 2002. A simulated annealing approach to define the genetic structure of populations. Mol Ecol 11, 2571-2581.

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.

FABUEL E., BARRAGÁN C., SILIÓ L., RODRÍGUEZ C., TORO M.A., 2004. Analysis of genetic diversity and conservation priorities in Iberian pigs based on microsatellite markers. Heredity 93, 104-113.

FALUSH D., STEPHENS M., PRITCHARD J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567-1587.

FERNÁNDEZ A., FABUEL E., ALVES E., RODRÍGUEZ C., SILIÓ L., 2004. DNA tests based on coat colour genes for authentication of the raw material of meat products from Iberian pigs. J Sci Food Agric 84, 1855-1860.

GARCÍA D., MARTÍNEZ A., DUNNER S., VEGA-PLA J.L., FERNÁNDEZ C., DELGADO J.V., CAÑÓN J., 2006. Estimation of the genetic admixture composition

of Iberian dry-cured ham samples using DNA multilocus genotypes. Meat Sci 72, 560-566.

KAEUFFER R., RÉALE D., COLTMAN D.W., PONTIER D., 2007. Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. Heredity 99, 374-380.

KIRKPATRICK S., GELATT C.D., VECCHI M.P., 1983. Optimization by simulated annealing. Science 220, 671-680.

LAVAL G., SANCRISTOBAL M., CHEVALET C., 2002. Measuring genetic distances between breeds: use of some distances in various short term evolution models. Genet Sel Evol 34, 481-507.

MANEL S., GAGGIOTTI O.E., WAPLES R.S., 2005. Assignment methods: matching biological questions with appropriate techniques. Trends Ecol Evol 20, 136-142.

NEI M., 1987. Molecular evolutionary genetics. Columbia University Press, USA. 512 pp.

OLDENBROEK K. (ed), 2007. Utilisation and conservation of farm animal genetic resources. Wageningen Acad Publ, The Netherlands. 232 pp.

PRITCHARD J.K., STEPHENS M., DONNELLY P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945-959.

RAYMOND M., ROUSSET F., 1995. GENEPOP (Version 1.2): population genetics software for exact tests and ecumenicism. J Hered 86, 248-249.

RODRÍGUEZ-RAMILO S.T., TORO M.A., FERNÁNDEZ J., 2009. Assessing population genetic structure via the maximisation of genetic distance. Genet Sel Evol 41, 49.

TORO M.A., FERNÁNDEZ J., CABALLERO A., 2009. Molecular characterization of breeds and its use in conservation. Lives Sci 120, 174-195.

WRIGHT S., 1931. Evolution in mendelian populations. Genetics 16, 97-159.