USING TWO DIFFERENT APPROACHES TO INFER THE GENETIC STRUCTURE OF POPULATIONS WITH COMPLEX RELATIONSHIPS: THE CASE OF THE AVILEÑA-NEGRA IBÉRICA

D. Martin-Collado[1,2], K.J. Abraham[3], S.T. Rodriguez-Ramilo[1], M.A. Toro[4], M.J. Carabaño[1] and C. Diaz[1]

[1]Dpto. Mejora Genética Animal. INIA. Ctra. de la Coruña km.7.5 28040, Madrid, Spain
[2]AbacusBio Limited, PO Box 5585, Dunedin, New Zealand
[3]Dpto. de Biologia Celular e Molecular, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo Ribeirão Preto SP, Brazil
[4]Dpto. Producción Animal, E.T.S.I. Agrónomos, Universidad Politécnica de Madrid. Ciudad Universitaria, 28040, Madrid, Spain

SUMMARY

The inference of the genetic structure of domestic animal populations has important implications in the design of breeding programs. In this paper, we assessed the utility of a graphical clustering algorithm (GCA) to identify the genetic structures of real livestock populations with complex relationships comparing it to a Bayesian clustering algorithm (STRUCTURE). The genetic structure of the Spanish cattle breed Avileña-Negra Ibérica was inferred by the analysis of 13,343 animals from 70 herds genotyped for 17 microsatellites. We compared the results of GCA and STRUCTURE regarding the ability to restore Hardy-Weinberg equilibrium in each subpopulation and the average coancestry within and between subpopulations. Both approaches described a similar structure for the ANI breed, which was found to have three genetic subpopulations and a pool of individuals that cannot be assigned without ambiguity to any of the subpopulations. This structure is coherent with the history of the breed. The GCA showed to be a much faster method to infer genetic structure with high ability to determine the core hidden structure of populations with complex relationships.

INTRODUCTION

The demographic and correlated genetic structure of livestock populations has important implications for the design of breeding and conservation programs. In addition, there is a renovated interest in studying the genetic structure of livestock populations since population stratification could bias the prediction of genomic breeding values as well as the results of GWAS (Janss *et al.* 2013). Genetic structures can be analysed using molecular information and several methodologies have been developed for this aim. The bayesian methods and in particular the STRUCTURE software (Pritchard *et al.* 2000) has become very popular. However, STRUCTURE might show difficulties in assesing the genetic structure of populations with a complex pedigree structure. In this paper, we introduce the use of graphical clustering algorithms (GCA) for the inference of the genetic structure of livestock populations. We used a new GCA (Abraham *et al.* forthcoming), that make use of the population molecular coancestry matrix to determine its genetic structure. The Avileña-Negra Ibérica (ANI) breed is an example of a population with complex relationships, where herds are subpopulations with different degrees and patterns of connection among them. Vasallo and Díaz (1986) determined that ANI breed had a pyramidal structure with the majority of herds recurrently buying bulls from the same leading herds. Therefore, the genetic status of the breed was highly dependent on the genetic management of those few leading herds. Since then the ANI population might have evolved. The aim of this paper is twofold: first, to assess the usefulness of GCA to identify the genetic structure of real livestock populations by comparing its performance to Bayesian clustering algorithms (STRUCTURE); and second, to study the current genetic structure of the ANI beef cattle breed using microsatellites (MS) genotypes to assess the genetic relationships among individuals.

MATERIAL AND METHODS

Material. The ANI breed is a Spanish beef cattle breed reared under extensive conditions. We analysed a data set of 13,343 individuals from 70 herds genotyped for 17 MS.

Methods. We compared the performance of a GCA with the model based algorithm implemented by STRUCTURE.

STRUCTURE algorithm. The Bayesian clustering algorithm implemented in STRUCTURE can assign either the individuals or a fraction of their genome (a proportion of inferred ancestry) to a number of clusters (K) based on multilocus genotypes (Pritchard *et al.* 2000). To determine K we used Evanno's *et al*. (2005) criterion that was found to perform better than the one initially proposed by Pritchard *et al.* (2000) to detect the more likely number of subpopulation (K) when the pattern of dispersal of individuals was not homogeneous.

Graphical clustering algorithm. The GCA works on a symmetric matrix whose off-diagonal elements are the values of the correlation between the corresponding elements to be clustered. In our case, the matrix contained the molecular coancestry values among the 13,343 ANI individuals analysed, which were obtained from the information on the frequencies of the markers following Caballero and Toro (2002). The matrix was calculated by Metapop software (Pérez-Figueroa *et al.*, 2009). The GCA used comprises two algorithms that are run one after another. The first one identifies all possible independent (or less related) individuals using a modification of a method shown in Abraham and Fernando (2012) and the second builds the clusters around these independent animals, as described in detail in Williams *et al*. (2011). Two thresholds (molecular coancestry values) have to be set to determine which individuals are considered as independent and which are defined as closely related. The thresholds were established according to the percentiles of the distribution of the molecular coancestry values. The percentile chosen depends on the expected genetic differentiation of the subpopulations. We did not expected ANI population to have a simple genetic structure; therefore, the threshold that defined independent animals was set to be very conservative. Several thresholds were used to define independent animal however the number of independent animals was similar. The case we present correspond with the percentile 1.25 of the molecular coancestry matrix. The second threshold (closely related individuals) corresponded with the percentile 75.

Genetic contribution of herds. We analysed 209,694 animals of 732 herds included in ANI breed Herdbook to complement the result of the genetic analysis. We determined the contribution of the different herds to the genetic composition of the ANI breed with ENDOG software (Gutiérrez and Goyache, 2005) by calculating the probability of gene origin of the ancestors and then summing the contribution values of the ancestors belonging to each herd.

RESULTS

None of the genotyped MS was in Hardy-Weinberg equilibrium (HWE) when considering the population as a whole. The lack of HWE might be an indicator of the presence of a stratified genetic structure. The F_{ST} differentiation index among herds was on average 0.074. The average molecular coancestry within (f_{ii}) and among (f_{ij}) herds was 0.329 and 0.278, respectively. We estimated the different statistics with Metapop software (Perez-Figueroa *et al*. 2009) except for the HWE which was calculated with Genepop (Rousset 2008).

STRUCTURE algorithm. STRUCTURE inferred the existence of three genetic clusters and assigned a proportion of ancestry coming from each cluster for all participating individuals. Animals were assigned to a certain cluster when at least 90% of their genome (the proportion of inferred ancestry given by STRUCTURE) was coming from that cluster. According to this definition of clusters, there were 1134, 1054 and 1015 animals in the first, second and third clusters, respectively. Those animals that were not included in the clusters were grouped together in a pool (ST-Pool). The number of loci in HWE within the clusters increased with respect to the whole population; 16, 9 and 11 MS were found to be in HWE for clusters 1, 2 and 3, respectively. Only 2 MS were in HWE in the ST-Pool. Clusters f_{ii} were higher than f_{ij}, as expected, while the f_{ii} of the ST-Pool was the same as the f_{ij} of herds (Table 1).

We analysed the distribution of herds across the clusters and found two types of herds, those whose individuals were mostly associated to one specific cluster and those whose individuals were distributed among different clusters. It should be noted that the majority of individuals in most herds were assigned to the ST-Pool. However, there were nine herds with a majority of individuals not assigned to the pool but to a specific cluster.

Graphical clustering algorithm. The GCA also identified three clusters. Cluster sizes were 257, 534 and 458 for clusters 1, 2 and 3 respectively. Those animals not assigned to any cluster were also grouped in pool (GCA-Pool). In this case, 15, 16 and 15 MS were in HWE in clusters 1, 2 and 3, respectively. Only 1 microsatellite was in HWE in the GCA-Pool. As expected, f_{ii} was higher than f_{ij} (Table 1). Both the f_{ii} and the f_{ij} of the GCA clusters were larger than the ones of the clusters inferred by STRUCTURE.

Table 1. Molecular coancestry within (f_{ii} in diagonal*)* and across (f_{ij} off-diagonal) the genetic clusters of the Avileña-Negra Ibérica population inferred by STRUCTURE software (values on the left of the slash) and the graphical clustering algorithms -GCA- (on the right of the slash)

Cluster	1	2	3	ST-Pool/CGA-Pool
1	0.343/0.427	0.244/0.327	0.241/0.306	0.271/0.280
2	0.244/0.327	0.311/0.409	0.239/0.366	0.276/0.316
3	0.241/0.306	0.239/0.366	0.343/0.417	0.271/0.301
ST-Pool/CGA-Pool	0.271/0.280	0.276/0.316	0.271/0.301	0.278/0.269

In line with STRUCTURE, GCA described the same two types of herds; those whose animals are mostly associated to one cluster and those associated to different ones. Furthermore, the distribution of herds across clusters was very similar. However, in this case a higher percentage of individuals of a herd were assigned to the GCA-Pool. In general, many herds had individuals assigned to clusters 2 and 3 as expected from the f_{ii} and f_{ij} values (Table 1). This connection between clusters 2 and 3 is also observed in the analysis of herds of STRUCTURE.

Genetic contribution of herds. Three herds were the origin of the majority (56.5%) of the genes in the population in 2012. These three herds are among those nine herds whose animals are mostly associated to one specific cluster. Furthermore, each of them appeared assigned to a different cluster both in STRUCTURE and GCA solutions.

DISCUSSION

STRUCTURE and the GCA inferred similar genetic structures suggesting that the results are robust. However, there were important differences in terms of the computational time. When using the K determination criterion suggested by Evanno *et al.* (2005), the STRUCTURE algorithm needs to be run several times per K to calculate a standard deviation of the replicates. In our case, we tested the algorithm from K= 1 to 50 which took several weeks to run. GCA took less than an hour to obtain the whole set of solutions. According to the results of both analyses, the ANI population can be divided in three genetic clusters and one pool of animals that could not be assigned to any of the clusters and that grouped the majority of animals. We set very strict criteria in both approaches for animals to be assigned to a cluster, aiming to get the core ANI population structure, given its expected complexity. Thus, we expect that the size of the pool would be reduced if the criteria were looser or, as observed in human population, once a first level of stratification is identified then new stratification levels may appear. The average HWE across MS increased within the clusters, validating the clustering, while the pool was in HW disequilibrium indicating that there may be secondary genetic structures within it. The distribution of individuals in different herds among the three clusters was quite consistent when comparing GCA and STRUCTURE inferences. Both approaches gave the same solution regarding the herds that are mostly associated to one cluster; each of the three herds that, according to the pedigree analysis, have had a major genetic contribution to the breed was found in each of the three clusters. However, there were some differences among those herds that cannot be clearly associated to one cluster. The results of the analysis indicate that currently ANI is stratified in three linages with a number of herds mainly related to one of the lineages and then, another group of herds whose individuals are distributed across lineages. Due to the large number of herds and individuals included in this last group and the limited number of herds with a major contribution to the breed, it will be of great importance to understand the composition of the pool and see how it is related with the genetic variability of the breed.

ACKNOWLEDGEMENTS

We thank ANI breeders Association for providing the data and helpful comments for the results interpretation. DMC was funded by the Collaboration Agreement (CC09-009) between INIA and FEAGAS supported by Spanish Ministry of Agriculture, Food and Environment. KJA wished to acknowledge the support of Capes-Brazil (Programa PVE) .

REFERENCES

Abraham K.J. and Fernando R. (2012) In ‘New Frontiers in Graph Theory’, pp. 307-322, editor Y. Zhang, InTech publishers, New York.

Abraham K.J., Martin-Collado D., Toro M.A., Carabaño M.J, Rodriguez-Ramilo S. and Diaz C. (forthcoming) A graph theoretical approach to estimating the number of constituent populations in an admixed population.

Caballero A. and Toro M.A. (2002). *Conserv. Genet.* **3**: 289.

Evanno G., Regnaut S. and Goudet J. (2005) *Mol. Ecol.* **14**: 2611.

Gutiérrez J.P. and Goyache F. (2005) *J. Anim. Breed. Genet.* **122**: 172.

Janss L., de los Campos G., Sheehan N. and Sorensen D. (2012) Genetics, **192**: 693.

Pérez-Figueroa A., Saura M., Fernández J., Toro M.A. and Caballero A. (2009). *Conserv. Genet.* **10**: 1097.

Pritchard J.K., Stephens M. and Donelly P. (2000) *Genet.* **155**: 945.

Rousset F. (2008) *Mol. Ecol. Resour.* **8**: 103.

Vasallo J.M. and Díaz C. (1986) *Livest. Prod. Sci.* **15**: 285.

Williams T.D., Turan N., Diab A.M., Wu H., Mackenzie C., Bartie K., Hrydziuszko, O., Lyons B.P., Stentiford G.D., Herbert J.M., Abraham J.K., Katsiadaki I., Leaver M.J., Taggart J.B., George S., Viant M.R., Chipman K.J. and Falciani F. (2011) *Plos Comput. Biology* **7**: e1002126.