Comparison of two geo-evolutionary analysis methods using local and cross-border bovine breeds

Áxel Villalobos-Cortés a, Amparo Martínez b, José Vega-Pla c and Juan Vicente Delgado b

aLaboratorio de Análisis y Biología Molecular Aplicada, Instituto de Investigación Agropecuaria, Panamá, Panamá; bDepartamento de Genética, University of Córdoba, Córdoba, Spain; cLaboratorio de Investigación de las Fuerzas Armadas, Córdoba, Spain

ABSTRACT
New clustering algorithms have emerged in population genetics and molecular ecology areas as important tools to infer the genetic structure of a given population. Currently, methodologies based on Bayesian models are used for this purpose. These methodologies consider the geographical area, whereas other studies explore the use of geographic coordinates based in Voronoi mosaics. The aim of this study was to use a multi-locus genotype database of bovine populations to compare the populations with two programmes that based their analysis on Bayesian models with different focuses. A database of 1053 animals from 33 bovine populations was analysed (native, Creole and cross-border breeds) using 27 microsatellite markers. The STRUCTURE version 2.2.3 software was selected for the identification and calculation of the proportion of the mixture of individuals within each population as recommended by the software using different K calculations (n=2, 3, …, 33). Additionally, TESS, which uses the Bayesian model with clustering algorithms, was used for the analysis. We concluded that both models showed different results; the discrepancies probably occurred because some of the studied populations might have shown the panmixia principle, thereby altering the final result. This study draws attention to the selection of the most appropriate analysis software for different populations with different selection and mating models based on basic genetic principles, such as the Hardy–Weinberg and linkage disequilibrium, because the results can be misinterpreted, leading to incorrect conclusions. Population behaviour, including reproduction and genetic origins, should be taken into account before using a given quantitative analysis method.

Introduction
Landscape genetics is a new field of study that has emerged with the recent improvement of molecular techniques combined with new statistical tools, such as geo-statistics, maximum likelihood and Bayesian estimations, and more rapid computers with a higher analysis capacity (Guillot et al. 2005). Landscape genetics is defined as the study of the interaction between spatial patterns and ecological processes. This field aims to provide information about the interaction between the natural landscape and micro-evolutionary processes, such as genetic flow, genetic erosion and selection (Holderegger & Wagner 2006). It also aids in the identification of hidden frontiers that cause breaks in the genetic flow along a population with no evident cause. The two keys of landscape genetics are the detection of genetic discontinuities and their correlation with the landscape and environmental features, such as mountains and moisture gradients (Manel et al. 2003). This focus is useful because it enables different disciplines, such as evolutionary and ecological biology, to understand how the movement of individuals or gametes influences the genetic structure of a population (Manel et al. 2003). Clustering and similarity methods have the potential to group individuals into population units and to detect migrants without the need for an a priori definition of limits (Pritchard et al. 2000; Falush et al. 2003). However, these cluster models methods do not explicitly take into account the spatial nature of the problem of detecting and locating genetic discontinuities. In other words, the methods are based on the assumption that the a priori assignment of a population is performed independently and is identically distributed among individuals. As a consequence, these methods do not make use of spatial coordinates of the sampled individuals except in some ad hoc post-process schemes, like those...
consisting of drawing by hand the spatial convex hull, on each of the inferred populations (Guillot et al. 2005). Other analysis models, such as the model proposed by Guillot et al. (2005), assume that space dependencies frequently exist among individuals. Based on this assumption, hierarchical models were developed in which a priori information about the spatial organisation of the individuals was introduced. In addition to the detection of genetic discontinuities, the model takes into account the following points: removal of confusing coordinates from the individual sample, estimation of the number of populations in the study area, quantification of the spatial dependency of the data, assignment of the individual to the population of origin and detection of migrants between populations.

The group of sampled individuals is viewed as a representation of one or several panmictic populations separated by geographical frontiers across space or other causes of isolation (i.e. historical reproductive, social, religious or cultural). First, the spatial organisation of the population is specified, followed by the genetic properties of each population conditioned to their spatial organisation.

The aim of this study was to compare two analysis methodologies based on Bayesian models using multilocus genotypes of bovine populations under the assumptions that they have different geographical and breed origins.

**Materials and methods**

A methodological study was performed using samples obtained from the allele database of the *Biobovis Consortium*, which is a project established more than ten years ago that has led to several publications (Villalobos-Cortés et al. 2010; Delgado et al. 2012; Martínez et al. 2012; Ginja et al. 2013; Villalobos-Cortés et al. 2015). Using carefully acquired laboratory data, the statistical basis of this study allows testing of the efficiency of the methods used for genetic diversity analyses of domestic breeds.

This database consists of a sample of 1053 animals taken from 33 bovine populations (the selected ones are part of the authorised for the study, but selecting representations of autochthonous races of Europe, creole of America and crossbred races of Indian and European origin) under the following assumptions: origin-adapted breeds (A1: Berrenda en Colorado; A2: Berrenda en Negro; A3: Mostrenca; A4: Pajuna; A5: Retinta; A6: Negra Andaluza; A7: Toro de Lidia; A8: Rubia Gallega; A9: Vaca Canaria; A10: Palmera), breeds adapted in sites that were not their origins (C1: Guabala; C2: Guaymi; C3: Longhorn; C4: Criollo Poblano; C5: Criollo Baja California; C6: Criollo de Chihuahua; C7: Criollo de Nayarit; C8: Criollo de Chiapas; C9: Criollo Ecuatoriano; C10: Criollo Casanare; C11: Criollo Uruguayo; C12: Paraguayo; C13: Pampa Chaqueno; C14: Criollo Patagónico; C15: Criollo Argentino) and commercial cross-border breeds (T1: Holstein; T2: Hereford; T3: Brown Swiss; T4: Gyr; T5: Brahman; T6: Red Sindhi; T7: Guzerat; T8: Nelore). These populations were analysed using 27 microsatellite markers internationally recommended by the Food and Agricultural Organisation (FAO) and the International Society for Animal Genetics (ISAG). For these genotypes, we used the STRUCTURE version 2.2.3 software (Pritchard et al. 2000) to model the structure of the K populations within the database. The first model used was Bayesian and was based on the Markov Chain Monte Carlo test (MCMC), which uses models based on the clustering of individuals. The clustering criteria for the individuals were as follows: reduce the Hardy–Weinberg equilibrium and keep the disequilibrium phase between loci within groups to a minimum. The DISTRACT version 1.1 software (Rosenberg 2004) was also used for the graphical representation of the population’s structure due to its larger colour palette availability. The population’s structure was evaluated using the whole sample group \( (n = 1053) \) with the assumption that the samples belonged to an unknown K number of genetically different groups. The mixture proportion was calculated using population information after the optimal K value was obtained. The mean individual assignation \( (q) \), which was defined as the average proportion of each genotype inferred from each of the K groups \( (K = 2–33) \), was determined.

The second Bayesian model used a clustering algorithms that was applied in population genetic studies where the geographical distribution based in coordinates was included to localise genetic and geographical discontinuities (Guillot et al. 2005; Francois et al. 2006). This model is based on a hierarchical model of mixture proportions where the a priori distribution of the groups is defined as ‘hidden Gaussian random fields’ located on a spatial distribution network or mosaic. In this spatial organisation model using Voronoi mosaics, the value \( \Delta \) is taken as the geographical region under study. The K different populations \( (2–33) \) are considered present in the spatial domain under study, and these populations occupy some sub-domain. Each sub-domain is assumed to be joined using convex polygons. These polygons are localised as random variables with uniform distributions on the whole spatial domain.
The membership of any point of the domain in the population can be considered a colour that is often referred to as the Voronoi mosaic.

In this mode, the probability that two individuals belong to the same population decreases with the increase in the geographical distance between them or with reproductive, geographical or other types of isolation. Mixture models, such as Pritchard et al. (2000), give equal importance to all partitions, whereas this model tends to favour spatially organised partitions. The programme used for this aim was TESS (Chen et al. 2007; Durand et al. 2009a, 2009b). This software searches for the structure of individual multilocus genotype populations sampled in different geographical regions in which no predefined population is assumed. The software returns membership probabilities and geographical clustering of the individuals. To determine the optimal K, we used the Kmax value and the deviance information criterion (DIC). The deviance information criterion (DIC) was introduced by Spiegelhalter et al. (2002) to compare the relative fit of a set of Bayesian hierarchical models. It is similar to Akaike’s information criterion (AIC) in combining a measure of goodness of fit and measure of complexity, both based on the deviance; the model with the smallest DIC is estimated to be the model that would best predict a replicate dataset which has the same structure as that currently observed. These criteria were obtained in a first analysis round to initially determine the K values in which the optimal population was found with lower DIC values; then, a second analysis round was performed to find the optimal K, similar to the STRUCTURE programme.

The resulting files from STRUCTURE and TESS were processed in CLUMPP® version 1.1.2 (Jakobsson & Rosenberg 2007) to calculate the $H^\prime$ value, which is the estimator of the alignment between both matrices ($Q_{\text{STRUCTURE}}$ vs. $Q_{\text{TESS}}$). Then, the matrices were plotted in DISTRUCT (Rosenberg 2004).

**Results and discussion**

In the first model, the analysis was performed using the K value that maximised the value of the data determined by the calculation of $\text{Ln}(P/D)$ according to Pritchard et al. (2000) with K = 28 as the optimal value. Figure 1 shows the slope change from $K = 28$ to $K = 29$ in a $\text{Ln}(P/D)$ vs. K curve with more than 900 points. This change indicates that the maximum number of significant sub-divisions of breeds or highly related breed groups has been reached (Pritchard et al. 2000). This analysis method was performed by Canón et al. (2006) in a study performed in 45 European caprine populations where K = 25–30 was reported. When many highly related populations are included in the study, working first with the lowest K to obtain groups of associated breeds is recommended, followed by an analysis of each related group within a geographical zone until the optimal K is reached (Rosenberg et al. 2003).

The analysis results obtained using the TESS spatial clustering models showed an optimal value of K = 28, which was similar to the value obtained with the model used in STRUCTURE. Therefore, both methods are assumed to have a similar analysis power regarding the calculation of the K populations. Figure 2 shows the representative graph of DIC vs $K_{\text{max}}$ with a value of 155,183.

We proceeded to calculate the aligning similarity value between the $Q_{\text{STRUCTURE}}$ vs. $Q_{\text{TESS}}$ matrices. A value of $H^\prime = 0.5053$ was observed, resulting in differences in the distribution pattern of the individuals within the studied populations between models. When performing a more detailed analysis using two

![Figure 1. Slope change from K = 28 to K = 29 in the Ln(P/D) vs. K curve (Pritchard et al. 2000).](image-url)
populations from group C from the same region (C1 and C2), the STRUCTURE analysis showed that both populations were identified as genetically homogeneous and different populations, similar to the other analysed populations. The STRUCTURE analysis also identified six animals in the individual \( q \) matrix within population C1 that shared the genome of breed groups T4 to T8 (Figure 3).

In contrast, both breeds (C1 and C2) were identified as genetically homogeneous using TESS, although the same six individuals from group T4-T8 sharing the genome with breed C1 analysed in STRUCTURE were observed. If we observe the five T breeds (T4-T8) analysed in TESS, one (T8) appears with a different colour pattern compared to the other previously evaluated populations. In contrast, STRUCTURE interprets these breeds as a single genetically homogenous cluster (Figure 3).

Other incongruities observed between the two software suites occurred between populations A4 and A5 and between A8 and A9. In TESS (Figure 4), both population groups were observed as genetically homogenous, whereas in STRUCTURE, they were observed as different populations. Similar behaviour was observed in the remaining populations.

Unlike STRUCTURE, the spatial model used in TESS did not find any differences between individuals of populations C1 and C2 and assigned them the same mixture proportion values. Although population T8 was interpreted as a different genetic group than breeds T4 to T7 in the TESS, another result was found in STRUCTURE models, these populations remained in the same homogeneous genetic group as seen in Figure 3. Additionally, an influence of some populations on others was observed. The differences observed by the spatial model (TESS) could be due to the following assumptions stated by (Guillot et al. 2005), which were frequently assumed to exist by the model: space dependency among entities and viewing the group of sampled individuals as the representative one or several panmictic populations separated by geographical barriers across space. In this model, the probability of two individuals belonging to the same population decreases with the increase in their geographical distance. Mixture models, such as the Pritchard et al. (2000) model used in STRUCTURE, assign equal importance to all partitions, whereas the TESS model tends to favour spatially organised partitions. Several authors have noted that different clustering algorithms can infer dissimilar solutions for the
optimal partition from a set of data; for instance, differentiations have been observed between BAPS, GENELAND and STRUCTURE. An example of this discrepancy is the analysis of genetic data from the toad species *Bufo calamita* in England (Rowe & Beebee 2007). In one study, the variable geographical dispersion had a small effect on the clustering capacity when the other variables were set as constant (Helmer et al. 2003).

Therefore, we can state that the populations in this study are not natural or wild populations, although some wild-like populations are observed, such as A3. Because the populations in this study may be under selection by human manipulation and the animals are located in several geographical regions, the use of the spatial clustering model in TESS will not be the most appropriate in cases where these type of populations are present. The model that Pritchard et al. (2000) proposed exhibited better behaviour when populations with genetic differentiation indices equal or higher than 0.05 were compared, such as most breeds in this study (Latch et al. 2006; Chen et al. 2007). A study performed by Chen et al. (2007) comparing the performance of GENELAND, TESS, STRUCTURE and GENECLUST found that STRUCTURE achieved better results when identifying populations in high-geographical admixture scenarios, also detecting clinal variation and better computational speed. Additionally, other analysis tools based on data generated by STRUCTURE have been used to identify the number of genetically homogeneous groups with higher precision through the determination of the modal distribution of $\Delta K$ (Evanno et al. 2005) using HARVESTER STRUCTURE (Dent & von Holdt 2012), especially with microsatellites. This approach has the advantage of allowing better classification with fewer data due to previous information and a direct inference regarding the ranks and their maximum limits. However, with large data for a given K, the subsequent assignation of individuals must be the same under both models. In spatial clustering models the coloured Poisson–Voronoi tessellation model consists in assuming that there is an unknown number of polygons that approximate the true pattern of population spread across space. These polygons are centred around spatial points and each polygon belongs to one of the K population which is coded by an integer (or colour for graphical representation), the parameter lambda ($\lambda$) controls the number of polygons in $\Delta$ and therefore the strong dependence of hidden spatial clustering. Low $\lambda$ values correspond to a low partition of $\Delta$ and therefore a strong dependency of hidden spatial organisations of the populations, whereas high $\lambda$ values correspond to a high fragmentation and weak spatial dependency. When the number of points is large, each mosaic contains only one individual sample and the mosaic model behaves similar to the Pritchard et al. (2000) or Corander et al. (2008) models.

Different authors have noted that different clustering algorithms can infer different solutions for the optimal partition from a set of data. For instance, differences were observed between BAPS, GENELAND and STRUCTURE when analysing genetic data from a toad species in England (Rowe & Beebee 2007). This phenomenon is due to differences in the subjacent models, the statistical estimators or the algorithms used to calculate the estimators. It is important to consider that all previously used programmes were based on the MCMC method and hence tended to converge. Thus, the results of these programmes were not the exact solution in several cases but instead were an approximation in which some categories remained unknown. One efficient strategy is to perform a large number of tests and review the tests with similar results. Given that these differences in algorithms can produce differences in solutions, executing the analysis of genetic data using more than one approach is appropriate. The results that coincide suggest the presence of a strong genetic signal. If one or several results do not coincide, the assumptions the analysis may be based on interact with the usual convergence flow of the MCMC. Importantly, the convergence symptoms should not be ignored to avoid selecting the clustering *a priori* (Frantz et al. 2009). In a study
where the influence of five variables was evaluated on the clustering of individuals using markers such as the loci number, sample size, group number, geographical dispersion of the sample and assumptions about the correlation of the studied allelic frequencies. Rosenberg et al. (2005) observed that each of these variables had an effect on the clustering. When the geographical dispersion was used as a variable and the other variables as constants, the geographical dispersion had a very small effect on the clustering capacity (Helmer et al. 2003). Unlike the geographical dispersion, the loci number and sample size showed a strong correlation with the clustering capacity.

Tools are needed to select from the increasing number of models and programmes. This issue is a notable problem in statistics and genetic and ecological models because these models are essentially descriptive and no obvious chosen set of criteria exists to use when comparing models. This study draws attention to the selection of the most appropriate selection programme for different populations with different selection and mating models based on basic genetic principles, such as the Hardy–Weinberg and linkage equilibrium, because the results can be misinterpreted, leading to incorrect conclusions. We have to consider how the population behaves in terms of reproduction and the population’s genetic origins before using a given quantitative method. In this particular study it is suggested that STRUCTURE was more consistent than the TESS model.

Conclusions

Given the fact that the cattle breeds in the study show a geographical dispersion and intensities in selection and heterogeneous crossbreeding such as the T1, T2, T5, T8 crossborder breeds that are subject to high-intensity selection and are located in different regions of the world, unlike the indigenous and native breeds that are deeply rooted in a region with low-selection intensity. For example, the C1 y C2 that are located in specific regions in Panama or C10 in Colombia. This feature limits the analysis in the TESS programme which sets the geographics coordinates in the analysis model. The use of STRUCTURE, an a priori model without geographical coordinates, allows for analysing the data with greater accuracy.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Funding

The authors would like to acknowledge, the Biobovis Consortium, Cordoba University, the Agricultural Research Institute of Panama (Instituto de Investigación Agropecuaria de Panamá – IDIAP) and the National Secretary of Science and Technology (Secretaría Nacional de Ciencia y Tecnología) for funding this study.

ORCID

Áxel Villalobos-Cortés (http://orcid.org/0000-0003-4223-0560)
Amparo Martínez (http://orcid.org/0000-0002-6944-0501)
José Vega-Pla (http://orcid.org/0000-0003-0916-9876)
Juan Vicente Delgado (http://orcid.org/0000-0003-1657-8838)

References

Cañón J, García D, García-Atance MA, Obexer-Ruff G, Lenstra P, Ajmone-Marsan P, Dunner S, Consortium E. 2006. Geographical partitioning of goat diversity in Europe and the Middle East. Anim Genet. 37:327–334.
Chen C, Durand E, Forbes F, François O. 2007. Bayesian clustering algorithms as certaining spatial population structure: a new computer program and a comparison study. Mol Ecol Notes. 7:747–756.
Corander J, Siren J, Arjas E. 2008. Bayesian spatial modelling of genetic population structure. Comput Stat. 23:111–129.
Delgado JV, Martínez AM, Acosta A, Alvarez LA, Armstrong E, Camacho E, Cañón J, Cortés O, Dunner S, Landi V, et al. 2012. Genetic characterization of Latin-American Creole cattle using microsatellite markers. Anim Genet. 43:2–10.
Dent E, von Holdt B. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour. 4:359–361.
Durand E, Chen C, Francois O. 2009a. TESS 2.3: Bayesian Clustering using tessellations and Markov models for spatial population genetics [Internet]. [cited 2009 Sep 5]. Available from: http://membres-timc.imag.fr/Olivier.Francois/tess.html.
Durand E, Jay F, Gaggiotti OE, Francois O. 2009b. Spatial inference of admixture proportions and secondary contact zones. Mol Biol Evol. 26:1963–1973.
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.
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.
Francois O, Ancelet S, Guillot G. 2006. Bayesian clustering using hidden Markov random fields in spatial population genetics. Genetics. 174:805–816.
Frantz AC, Cellina S, Krier A, Schley L, Burke T. 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? J Appl Ecol. 46:493–505.
Ginja C, Gama LT, Cortes O, Delgado JV, Dunner S, García D, Landi V, Martín-Burriel I, Martínez-Martínez A, Penedo MC, et al. 2013. Analysis of conservation priorities of Iberoamerican cattle based on autosomal microsatellite markers. Genet Select Evol. 45:35.

Guillot G, Estoup A, Mortier F, Cosson JF. 2005. A spatial statistical model for landscape genetics. Genetics. 170:1261–1280.

Helmer S, Neumann T, Moerkotte G. 2003. Estimating the output cardinality of partial preaggregation with a measure of clusteredness. Proceedings of the 29th international conference on Very large data bases - Volume 29. VLDB Endowment, Berlin, Germany.

Holdereger R, Wagner H. 2006. A brief guide to landscape genetics landscape ecology. Land Ecol. 21:793–796.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 23:1801–1806.

Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE Jr. 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.

Manel S, Schwartz MK, Luikart G, Taberlet P. 2003. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol. 18:189.

Martínez AM, Gama LT, Cañón J, Ginja C, Delgado JV, Dunner S, Landi V, Martín-Burriel I, Penedo MT, Rodellar C, et al. 2012. Footprints of Iberian cattle in America 500 years after the arrival of Columbus. PLoS One. 7:e49066.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2003. Genetic structure of human populations. Science. 298:2381–2385.

Rosenberg NA. 2004. DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes. 4:137–138.

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 1:e70.

Rowe G, Beebee T. 2007. Defining population boundaries: use of three Bayesian approaches with microsatellite data from British natterjack toads (Bufo calamita). Mol Ecol. 16:785–796.

Spiegelhalter D, Best N, Carlin B. 2002. Bayesian measures of model complexity and fit. JR Statist Soc B. 64:583–639.

The Biobovis Consortium [Internet]. Available from: http://biobovis.jimdo.com/.

Villalobos-Cortés AI, Martínez AM, Escobar C, Vega-Pla JL, Delgado JV. 2010. Study of genetic diversity of the Guaymi and Guabala bovine populations by means of microsatellites. Livest Sci. 131:45–51.

Villalobos-Cortés AI, Martínez AM, Vega-Pla JL, Landi V, Quiroz V, Marques JR, Delgado JV. 2015. Genetic relationships among five zebu breeds naturalized in America accessed with molecular markers. Ital J Anim Sci. 14:158–162.