Agronomía Mesoamericana

UNIVERSIDAD DE COSTA RICA

Scientific article

Volumen 36: Artículo jh40ks60, 2025
e-ISSN 2215-3608, https://doi.org/10.15517/jh40ks60

Unveiling the genetic structure of Costa Rican bovines through genetic admixture models*

Develando la estructura genética de los bovinos costarricenses mediante modelos de mezcla genética

Bernardo Vargas-Leitón[1], Johnny Núñez-Cárdenas[2], Anthony Valverde[3], Bernal León-Rodríguez[4], Alejandro Saborío-Montero[2]

* Reception: June 27, 2025. Acceptance: August 19, 2025. This study was developed as part of the project 737-C3747 Vinculación interinstitucional para la generación conjunta de productos de investigación en producción animal, Estación Experimental Alfredo Volio Mata (EEAVM), Universidad de Costa Rica.

[1] Universidad Nacional. Heredia, Costa Rica. bernardo.vargas.leiton@una.ac.cr (corresponding author, https://orcid.org/0000-0002-1778-9672).

[2] Universidad de Costa Rica. San José, Costa Rica. johnny.nunezcardenas@ucr.ac.cr (https://orcid.org/0009-0009-9685-3318); alejandro.saboriomontero@ucr.ac.cr (https://orcid.org/0000-0002-9840-0058).

[3] Instituto Tecnológico de Costa Rica, Escuela de Agronomía, Laboratorio de Reproducción Animal. San Carlos, Costa Rica. anvalverde@itcr.ac.cr (https://orcid.org/0000-0002-3191-6965).

[4] Servicio Nacional de Salud Animal. Heredia, Costa Rica. bleonr@senasa.go.cr (https://orcid.org/0000-0002-0070-6746).

Abstract

Introduction. The cattle population in Costa Rica is highly diverse due to the variety of production systems, ongoing crossbreeding, and the increase in genetic imports. Under these circumstances, admixture models provide a more suitable approach to describe the genetic composition of this population than phylogenetic trees based on predefined populations. **Objective.** To model the genetic structure of the Costa Rican cattle population by means of supervised and unsupervised genetic admixture models. **Materials and methods.** Hair samples from 1412 randomly selected bovines from 744 herds across 8 regions of Costa Rica were collected in 2015 and genotyped for 18 microsatellite markers. Two approaches that make use of admixture genetic models were compared: an unsupervised scenario, based exclusively on genotype data; and a supervised scenario, which relied on genetic data assisted by prior information on phenotype and production purpose. **Results.** Analysis of genetic data under both scenarios provided similar results when the number of clusters (K) was lower than five, although estimates from the supervised model were more homogeneous with lower standard deviations. A priori defined subpopulations were distributed consistently among clusters in both scenarios. The most probable subpopulation clustering was observed at K = 3, which mainly separated Bos indicus breeds, Jersey and other Bos taurus breeds. Breed types clustered concordantly with breed clustering, shedding light on the genetic structure of the population. **Conclusions.** The combination of admixture genetic models under a supervised approach yielded the most consistent results, which reveals the importance of considering genetic-environmental interrelationships to achieve a more precise description of the genetic structure of the Costa Rican cattle population.

Keywords: cattle breeds, microsatellites, phylogeny, genetic clustering.

Resumen

Introducción. La población ganadera de Costa Rica es altamente diversa debido a la variedad de sistemas de producción, el cruzamiento continuo y el incremento de las importaciones genéticas. Bajo estas circunstancias, los modelos de mezcla podrían ser más eficientes para describir la estructura genética de esta población, a diferencia de árboles filogenéticos basados en poblaciones predefinidas. **Objetivo.** Modelar la estructura genética de la población bovina costarricense mediante modelos de mezcla genética supervisados y no supervisados. **Materiales y métodos.** Se recolectaron muestras de pelo de 1412 bovinos seleccionados al azar de 744 hatos en 8 regiones de Costa Rica en el año 2015 y genotipados para 18 marcadores microsatélites. Se compararon dos enfoques que hacen uso de modelos de mezcla genética: un escenario no supervisado, basado exclusivamente en datos de genotipo, y un escenario supervisado, que se basó en datos genéticos asistidos por información previa sobre fenotipo y propósito de producción. **Resultados**. El análisis de datos genéticos bajo ambos escenarios proporcionó resultados similares cuando el número de conglomerados (K) fue menor a cinco, aunque las estimaciones del modelo supervisado fueron más homogéneas y con menores desviaciones estándar. Las subpoblaciones definidas *a priori* se distribuyeron consistentemente entre los conglomerados en ambos escenarios. La agrupación de subpoblaciones más probable se obtuvo para K = 3, que separó principalmente las razas *Bos indicus*, Jersey y otras razas *Bos taurus*. Los tipos de raza se agruparon de manera concordante con la agrupación de razas, lo que arrojó luz sobre la estructura genética de la población. **Conclusiones**. La combinación de modelos de mezcla genética bajo un enfoque supervisado proporcionó los resultados más consistentes, lo que revela la importancia de considerar las interrelaciones genético-ambientales para lograr una descripción más precisa de la estructura genética de la población bovina de Costa Rica.

Palabras clave: razas bovinas, microsatélites, filogenia, conglomerados genéticos.

Introduction

The history of cattle in Costa Rica dates back to the late 16th century (Cordero-Solórzano et al., 2015; Martínez et al., 2012). These first cattle introduced were mainly Iberian taurine animals, which later gave rise to local Creole cattle. The first imports of modern breeds occurred in the mid-19th century, namely Devonshire and Durham cattle from England, followed by subsequent imports of other dairy breeds, such as Holstein, Jersey, and Ayrshire, by the end of the same century (Cordero-Solórzano et al., 2015).

Bos indicus cattle were introduced into the Americas approximately 150 years ago (Utsunomiya et al., 2019) and arrived in Costa Rica in the early 20th century (Vásquez-Loaiza & Molina-Coto, 2020). The Brahman breed is now the most common in the country, along with other zebu breeds such as Nelore, Gyr, Guzerá, and Indubrasil. Due to this diverse origin the current cattle population is genetically admixed.

In recent decades, numerous *Bos taurus* and *Bos indicus* breeds have been introduced into the country due to the continuous importation of genetics from different regions worldwide (Vargas & Van Arendonk, 2004). Crossbreeding has also become a common practice as a strategy to improve adaptability for production under tropical conditions. Currently, the number of cattle in Costa Rica is nearly one third of the country's human population, and these cattle are used for beef (63.2 %), dairy (16.2 %), dual-purpose (20.5 %), or labor (0.1 % corresponding to oxen) (Instituto Nacional de Estadística y Censos [INEC], 2021; United Nations, 2024).

Interest in characterizing livestock in Costa Rica has turned to molecular analysis, which uses microsatellite markers to distinguish genetically between breeds (Martínez et al., 2015) and locations (Cordero-Solórzano et al., 2015). Microsatellites are composed of 1 to 10 nucleotides and are a subcategory of tandem repeats that constitute repetitive genomic regions (Carneiro Vieira et al., 2016). Microsatellites are useful for a broad spectrum of genomic

determinations, such as kinship relation or parenthood (Webster & Reichart, 2005), criminology (Wickenheiser, 2002), genetic mapping (Ma et al., 2022), genetic diversity (Martínez et al., 2015) and population clustering from genetic structure (Dufresnes et al., 2023), among others. In cattle, microsatellites have been widely used to estimate genetic distance among populations (Agung et al., 2019).

In 2015, a representative sample (n = 1412) of the Costa Rican bovine population was genotyped using microsatellite markers (Martínez et al., 2015). Animals were assigned *a priori* to 16 breed types based on pedigree, phenotypic traits, and production purpose. Based on these data, several genetic diversity indexes were obtained, genetic distances between assigned breed types were measured, and clusters were formed.

An alternative approach for the genetic analysis of these genotypes is the use of genetic admixture models. Genetic admixture models represent the genetic makeup of an individual as a combination of contributions from different ancestral populations (Skotte et al., 2013). This can be inferred solely from genotype data, or by making use of additional prior information that might be available on the populations under study. Hence, admixture models offer a suitable and adaptable method for the genetic analysis of populations that comprise a significant number of individuals with varied genetic compositions.

Understanding the history and evolution of the local cattle population can provide insight into the genetic similarities between so-called pure breeds and their crosses, as well as their association with different production systems and geographical contexts within the country. Therefore, the aim of this study was to model the genetic structure of the Costa Rican cattle population by means of supervised and unsupervised genetic admixture models.

Materials and methods

Sampling and grouping

The sampling strategy was designed to obtain a representative sample of the cattle population present in different regions within the country. This approach differs from most of the previous studies describing genetic structure of cattle populations (Agung et al., 2019; Brasil et al., 2013; Edea et al., 2015; Egito et al., 2007; Kumar et al., 2003; Martínez et al., 2012; Martinez et al., 2023; Ocampo et al., 2021), which have focused mainly on characterizing specific breeds. The sampling strategy was based on the framework provided by the national Sistema Integrado de Registro de Establecimientos Agropecuarios (Integrated Registry System for Agricultural and Livestock Establishments, SIREA), which is managed by the Servicio Nacional de Salud Animal (Costa Rican Animal Health National Service, SENASA).

A total of 1,412 purebred and crossbred animals from 744 herds across Costa Rica were randomly sampled for tail hair follicles by 25 field technicians. The samples were sent to the Laboratorio de Bioseguridad (Biosafety Laboratory, SENASA) for DNA extraction and genotyping of 18 microsatellite markers. The specific details of the sampling procedure can be found in Cordero-Solórzano et al. (2015). To keep inbreeding low, the number of herds to be sampled was maximized, and samples were restricted to two animals per herd, unless several breed types were present within the same productive system.

Because the animals were randomly selected, it was not always possible to accurately determine their specific breed composition. Therefore, an approximate classification of breed types was performed based on farm records (if available), phenotype (skin, coat color, hump, ears) and the production purpose of the animal on the farm (milk, beef or dual-purpose). The number and proportion of animals sampled according to predetermined breed types are shown in Table 1, as well as the nomenclature used for this study.

Based on this classification, breeds and crossbreeds with significant representation were assigned to seven distinct subpopulations (Table 1, subpopulations 1, 11-16). The identified breeds or cross breeds that were present

Table 1. *A priori* defined subpopulations with associated breed types, nomenclature, sample size, and proportion of sampled bovines. Costa Rica, 2015.

Cuadro 1. Definición *a priori* de subpoblaciones con tipos raciales asociados, nomenclatura, tamaño de muestra y proporción entre las muestras bovinas. Costa Rica, 2015.

Subpopulation	Breed type	Nomenclature	Sample size	%
1	*B. indicus*	B_indicus	97	6.9
2	Brahman	Brahman	263	18.6
3	*B. taurus*	B_taurus	22	1.6
4	*B. taurus*×*B. indicus*	B_tau×B_ind	43	3.0
5	Undefined beef	[a]Undef_Beef	103	7.3
6	Undefined dual-purpose	[a]Undef_Dual	189	13.4
7	Undefined milk	[a]Undef_Milk	39	2.8
8	Cross beef	[b]Cross_Beef	68	4.8
9	Cross dual-purpose	[b]Cross_Dual	95	6.7
10	Cross milk	[b]Cross_Milk	64	4.5
11	Guernsey	Guernsey	49	3.5
12	Holstein	Holstein	89	6.3
13	Holstein×Jersey	Hol×Jer	34	2.4
14	Jersey	Jersey	158	11.2
15	Brown Swiss	Brown_Swiss	49	3.5
16	Simmental	Simmental	50	3.5
Total			1412	100

[a]Undef: Undefined breeds or cross breeds grouped according to production purpose. [b]Cross: Identified cross breeds present in minor quantities and grouped according to production purpose. / [a]Undef: Razas o cruces no definidos agrupados según el propósito de producción. [b]Cross: Cruces identificados presentes en cantidades menores y agrupados según el propósito de producción.

in minor quantities were grouped in six subpopulations (Table 1, subpopulations 1, 3, 4, 8, 9, 10) according to genus and/or production purpose. Finally, undefined breeds or cross breeds were classified into three subpopulations according to production purpose (Table 1, subpopulations 5, 6, 7). This classification enabled the use of STRUCTURE's functionality to incorporate prior information in the clustering process.

Genetic data analysis

All animals in the sample were genotyped for the following 18 microsatellite markers: BM1818, BM1824, BM2113, CSRM60, CSSM66, ETH10, ETH225, ETH3, ILSTS006, INRA23, MGTG4B, RM067, SPS113, SPS115, TGLA122, TGLA126, TGLA227 and TGLA53. Fifteen of these markers are among those recommended by the Food and Agriculture Organization of the United Nations [FAO], (2011) for use in genetic molecular characterization of cattle. Data analysis was performed using STRUCTURE software, which is an analysis program for investigating population structure using multilocus genotype data allele frequencies to characterize population groups and assign individuals to these groups (Pritchard et al., 2000).

STRUCTURE can detect individuals that are genetically admixed, and the genetic proportions from each population, which makes it suitable for crossbred populations. The software uses a clustering method based on

simulated models from a Markov Chain Monte Carlo (MCMC) approach to estimate the posterior distribution of every individual admixture coefficient. The mean of this distribution represents an estimate of the amount of an individual's genome that is derived from one of the inferred populations (Kumar et al., 2003).

Two alternative scenarios, unsupervised and supervised, were also compared in this analysis. The unsupervised scenario strictly relied on information provided by the microsatellite data to generate the clusters. The supervised scenario made use of the option LOCPRIOR available within STRUCTURE software. This option can be used when there is additional sample-characteristic data available to the user, including linguistic, geographical, cultural, or phenotypic information (Porras-Hurtado et al., 2013).

In this study, options LOCPRIOR and LOCISPOP were combined, which made possible the use of previously defined breed types as prior information to assist the clustering procedure, when the signal of genetic structure was relatively weak. An additional Boolean parameter (POPFLAG) was used to indicate which of the subpopulations (breeds or crossbreeds) were identified with certainty (Table 1, subpopulations 2, 11, 12, 13, 14, 15, 16). If the data suggested that these prior breed types were informative, the LOCPRIOR models from STRUCTURE used that information as "learning samples" to perform the clustering, assuming that individuals from the same group came from the same population.

For both scenarios, alternative numbers of clusters (K) were explored, ranging from a minimum of 2 to a maximum of 16. Following the recommendations of Porras-Hurtado et al. (2013), twenty replicates were simulated for each K, with a burn-in period length of 100 000 and 200 000 MCMC repeats. Allele frequencies were assumed to be correlated, as this configuration is considered best in cases of subtle population structure (Falush et al., 2003). Other parameters, such as the relative admixture level between populations (*alfa*), were inferred from the data, and *lambda*, the parameter of the distribution of allelic frequencies, was set to one, as advised by Pritchard et al. (2000). The additional parameters implemented in STRUCTURE were set to default values.

Due to computational demands, STRUCTURE analyses were conducted using a high-performance computing cluster (14 nodes, each with Xeon E5-2650v3, 64 Gb RAM, 1TB HDD) from the Research Center on Science and Engineering of Materials (CICIMA) at the Universidad de Costa Rica.

The Evanno method (Evanno et al., 2005) was used to identify the population model that best fit the genetic data as implemented by Structure Harvester software (Earl & vonHoldt, 2012). This method estimates the most likely genetic structure of the population from all the evaluated approaches, based on two criteria: LnProb(K), which provides an estimate of the posterior log probability of the data for a given K, and Delta K, which is an ad hoc quantity related to the second order rate of change of the log probability of data with respect to the number of clusters (Earl & vonHoldt, 2012), proposed by the authors as a better predictor of the real number of clusters.

The results obtained from Structure Harvester for both scenarios were further processed with CLUMPP software (Jakobsson & Rosenberg, 2007). This software deals with label switching and multimodality problems arising in population genetic cluster analyses, permuting the cluster outputs obtained from independent runs (replicates), and thus providing optimal assignment of individuals to clusters by obtaining an aligned consensus cluster.

As the number of individuals was too large (n = 1412), permutational analysis was based on the *population* Q-matrix in both scenarios. This matrix gives the proportion of membership of each predefined population in each of the K generated clusters. Permutational analysis was performed by using the *Large K Greedy* algorithm (Jakobsson & Rosenberg, 2007), with 10000 repeats, each representing a randomly selected input order. Although breed types were not used as prior information in the unsupervised scenario, the population Q-matrix was also obtained as an output, for comparison against the supervised scenario.

Optimized consensus clusters were obtained from CLUMPP software and further processed with the *ggplot2* package in R to produce stacking graphs for a better perspective of similarities/differences between scenarios. To further explore the associations between a priori breed types and clusters, as well as between breed types

themselves, InfoStat software was used to construct a Minimum-Spanning Tree (MST) based on the consensus *individual* Q-Matrix obtained from CLUMPP for K = 3, under both scenarios.

Results

Hair samples were collected in the main cattle-raising regions of the country (Figure 1). Samples from purebred dairy cattle predominated in dairy farms located in the highlands of the central region, those from crossbred cattle mainly came from dual-purpose herds frequently found in the Northern and South Pacific regions, while those from purebred *B. indicus* breeds were more evenly distributed across the country, mainly coming from beef or dual-purpose herds. Based on its size and wide distribution, it is possible to assert that this sample represented the structure of the national livestock population at the time of the study.

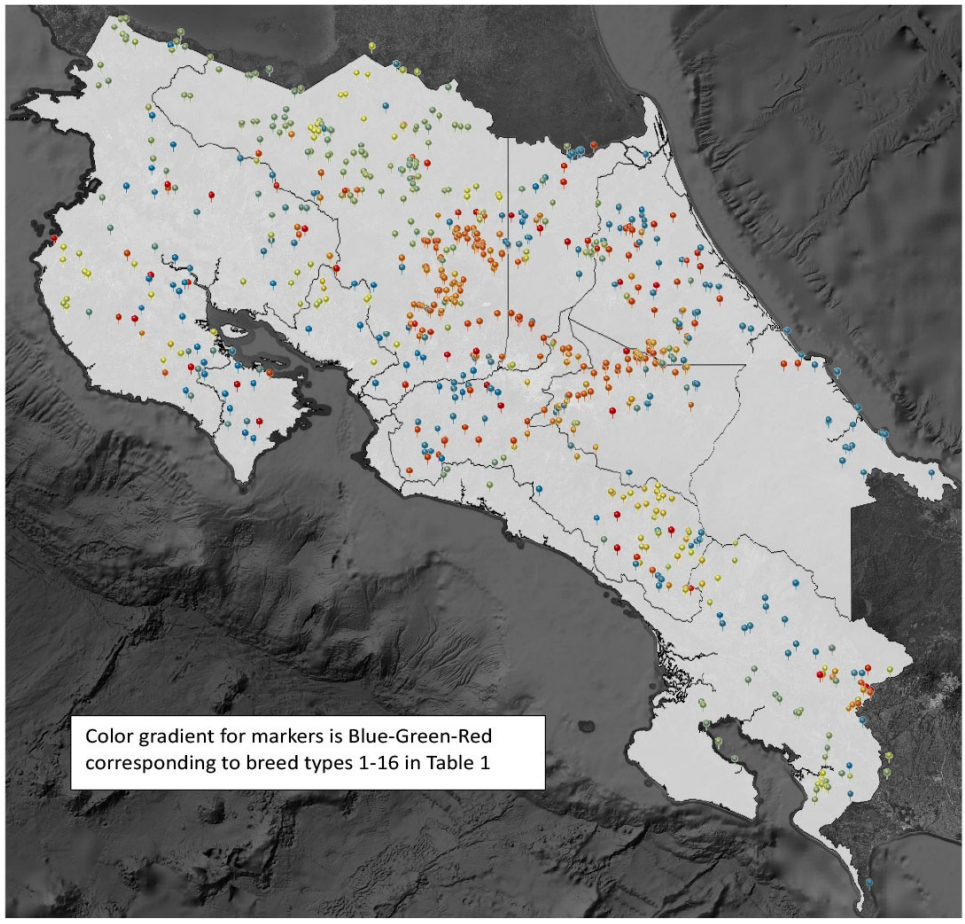

Figure 1. Geographic distribution of collected cattle hair samples (*n*= 1412). Costa Rica, 2015.

Figura 1. Distribución geográfica de las muestras de cabello de ganado (*n*= 1412). Costa Rica, 2015.

ISSN 2215-3608 https://doi.org/10.15517/jh40ks60

Structure Harvester

Structure Harvester results revealed differences between the unsupervised and supervised models (Figure 2). For the supervised model, there was a consistent increase in the mean of Log (K) from K = 2 to K = 7 (Figure 2A). For the unsupervised model, a similar trend was observed up to K = 5. Beyond this point, mean LnP(K) values became more variable, with consistently higher standard deviations compared to the supervised model.

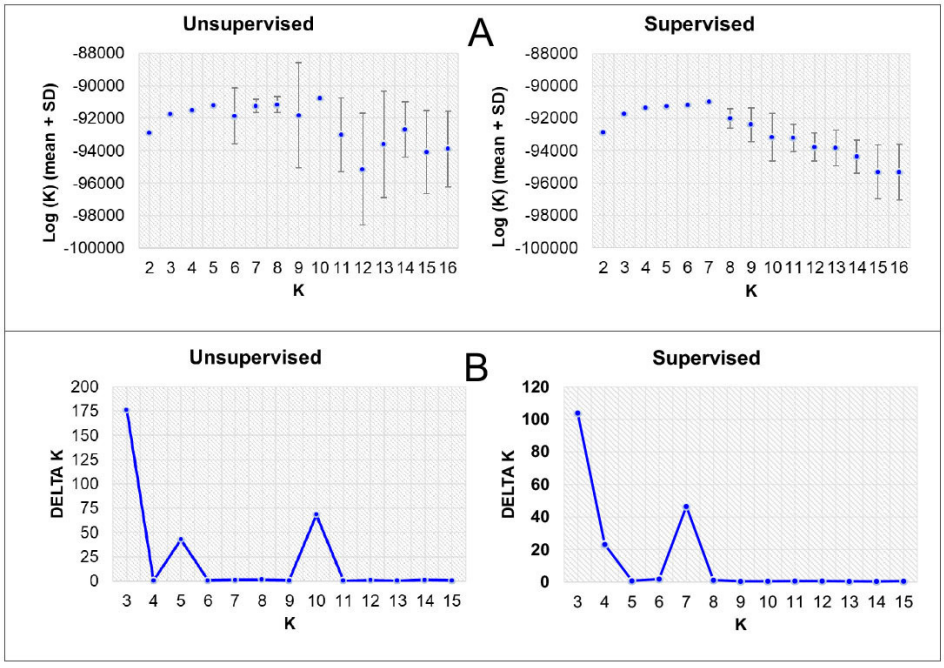

Figure 2. Estimates of Log (K) (mean ± SD) (A) and Delta K criterion [mean(|ln(K)|/sd(ln(K)] (B) obtained from the Structure Harvester procedure for unsupervised and supervised admixture genetic models. Costa Rica, 2015.

Figura 2. Estimados de Log (K) (media ± DE) (A) y criterio Delta K [media(|ln(K)|/sd(ln(K)] (B) obtenidos del procedimiento de Structure Harvester para los modelos de mezcla genética no supervisado y supervisado. Costa Rica, 2015.

In both scenarios, when K > 6, the standard deviations were greater than those at lower K values, indicating reduced cluster stability. According to the Delta K criterion proposed by the Evanno method (Figure 2B) the optimal number of clusters was three, which showed the highest Delta K value in both models. Additional lower peaks of Delta K were observed for K = 5 and K = 10 in the unsupervised model, and for K = 7 in the supervised model. Based on previous results, subsequent analysis focused on the results obtained for K < 6.

Genetic structure and clustering

For K < 6, an exploration of the solutions provided by STRUCTURE for different replicates of the same K, revealed frequent problems of "label switching"; these are replicates with similar membership coefficient estimates but with different permutations of the cluster labels. To address this issue, CLUMPP software was used to produce a consensus cluster for each K level in both unsupervised and supervised scenarios.

The genetic structure of the population when K = 2 is depicted in Figure 3. The results from both approaches (unsupervised and supervised) were similar, generating a first cluster (CL1) that was mostly represented for *B. indicus* and Brahman breed types, as well as for crossbred or undefined breed types used for beef or dual-purpose. The second cluster (CL2) was mainly composed of dairy breeds, dairy cross breeds, and other minor *B. taurus* breeds, as well as undefined breed types for dairy purposes.

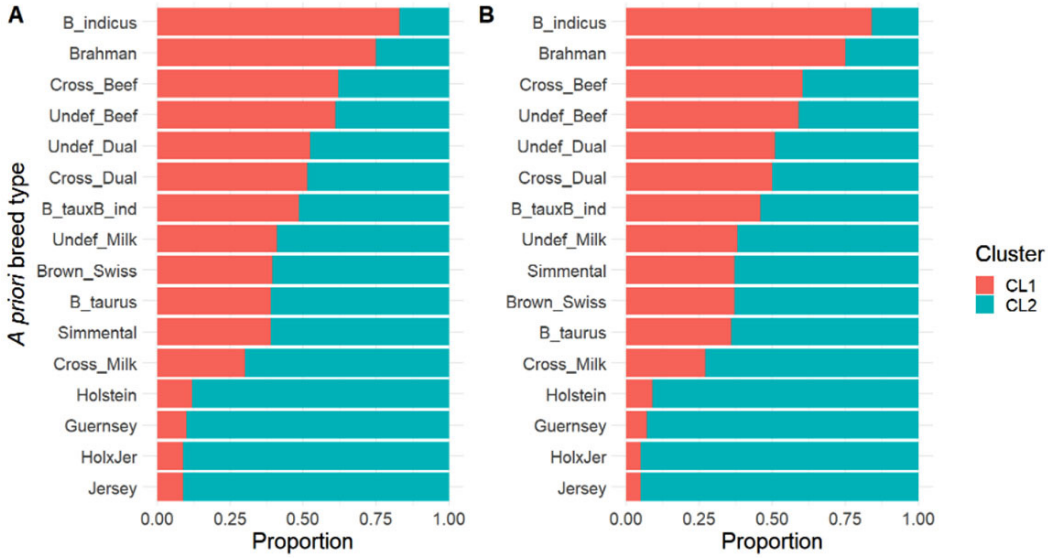

Figure 3. Genetic structure of the Costa Rican bovine population, associated with breed types for K = 2, according to unsupervised (A) and supervised (B) admixture genetic models, and sorted decreasingly by proportion in CL1. Costa Rica, 2015.

Figura 3. Estructura genética de la población bovina de Costa Rica, asociada con tipos raciales para K = 2, de acuerdo con modelos de mezcla genética no-supervisados (A) y supervisados (B), en orden descendente según proporción en CL1. Costa Rica, 2015.

A clear distinction was observed between beef (CL1) and dairy (CL2) clusters; however, it is important to note that no breed type was completely excluded from either cluster, indicating that the genetic structure of all breed types is constituted by an admixture, to lesser or greater degree.

The optimal genetic structure of the population according to the Structure Harvester was obtained when K = 3 (Figure 4). For both scenarios, a similar clustering pattern was obtained. CL1 for K = 3 remained very similar to that for K = 2. The CL2 and CL3 in K = 3 seem to be a decomposition of the previous CL2 when K = 2, where CL2 in K = 3 is highly represented by Holstein and Guernsey breeds, dairy cross breeds, Simmental and to a lesser extent, Brown Swiss and other minor *B. taurus* breed types. CL3 was strongly associated with Jersey and Holstein×Jersey crosses. The combined contribution of the latter mentioned crosses might be related to crossbred animals with greater proportion of one or the other breeds.

The genetic structure for the unsupervised and supervised models for K = 4 also maintained similar patterns of clustering (Figure 5). CL1 was mainly related to Brahman and other minor *B. indicus* breeds, beef crosses, undefined beef, and dual purpose cattle, either for crosses or undefined; CL2 was more associated with Jersey and Holstein×Jersey crosses; CL3 was more represented by Simmental and Brown Swiss breeds and in a leaser way by other minor *B. taurus* breeds and dairy crosses; and CL4 was represented by Guernsey and Holstein breeds.

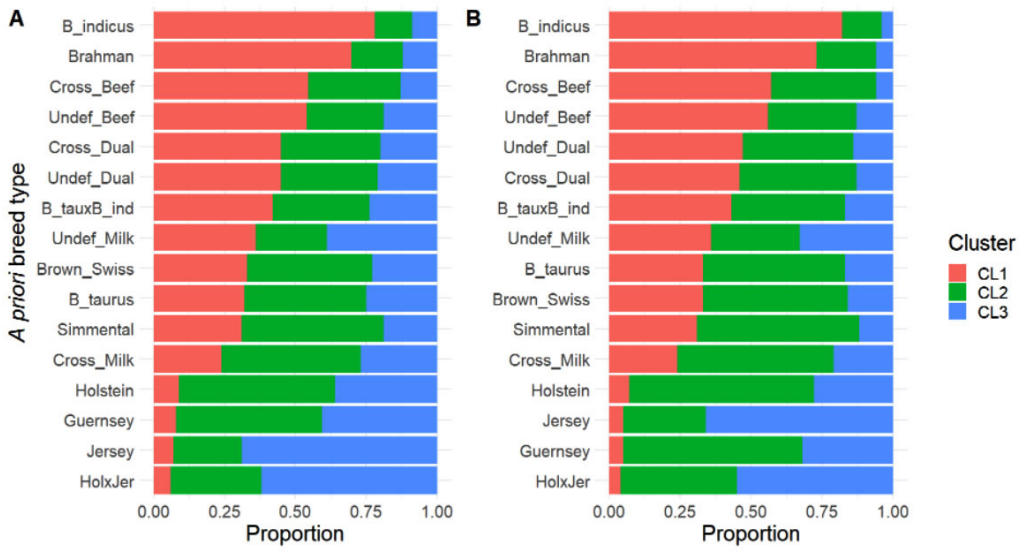

Figure 4. Genetic structure of the Costa Rican bovine population, associated with breed types for K = 3, according to unsupervised (A) and supervised (B) admixture genetic model approaches, and sorted decreasingly by proportion in CL1. Costa Rica, 2015.

Figura 4. Estructura genética de la población bovina de Costa Rica, asociada con tipos raciales para K = 3, de acuerdo con modelos de mezcla genética no-supervisados (A) y supervisados (B), en orden descendente según proporción en CL1. Costa Rica, 2015.

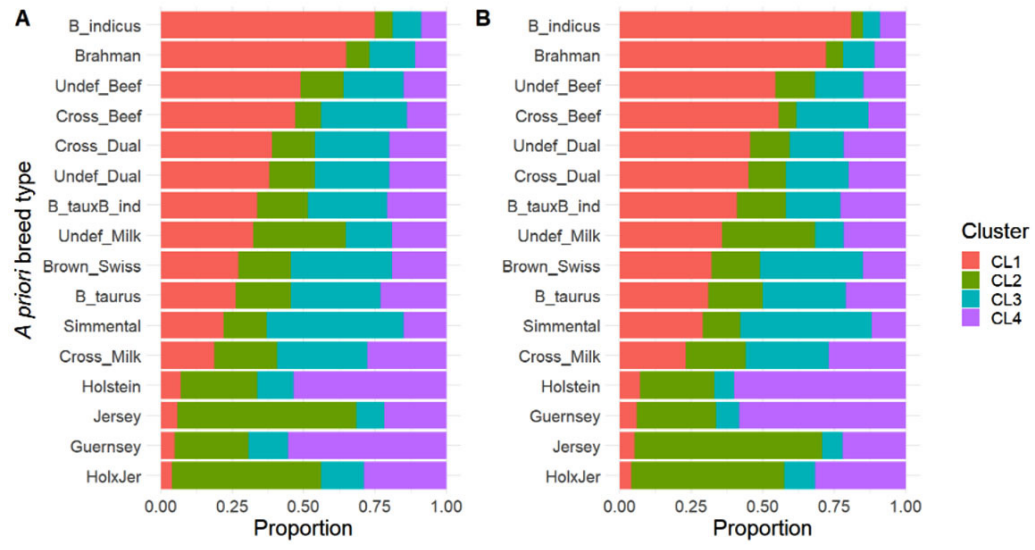

Figure 5. Genetic structure of the Costa Rican bovine population, associated with breed types for K = 4, according to unsupervised (A) and supervised (B) admixture genetic model approaches, and sorted decreasingly by proportion in CL1. Costa Rica, 2015.

Figura 5. Estructura genética de la población bovina de Costa Rica, asociada con tipos raciales para K = 4, de acuerdo con modelos de mezcla genética no-supervisados (A) y supervisados (B), en orden descendente según proporción en CL1. Costa Rica, 2015.

For K = 5, the clustering patterns obtained for the unsupervised and supervised approaches began to differ (Figure 6). An evident difference between clustering patterns in K = 5 was that, in the unsupervised scenario, the main representation for *B. indicus*, Brahman, beef cross and undefined breeds, as well as dual purpose (cross and undefined) were divided into two clusters (CL1 and CL2) with comparable proportions of individuals, while in the supervised model those classifications remained mostly represented in CL1, with CL2 not showing a large association with any of the breed types. CL3 was associated with Jersey and Holstein×Jersey breed types, CL4 was related to Simmental and Brown Swiss breeds, and CL5 was mostly composed of Holstein and Guernsey breeds.

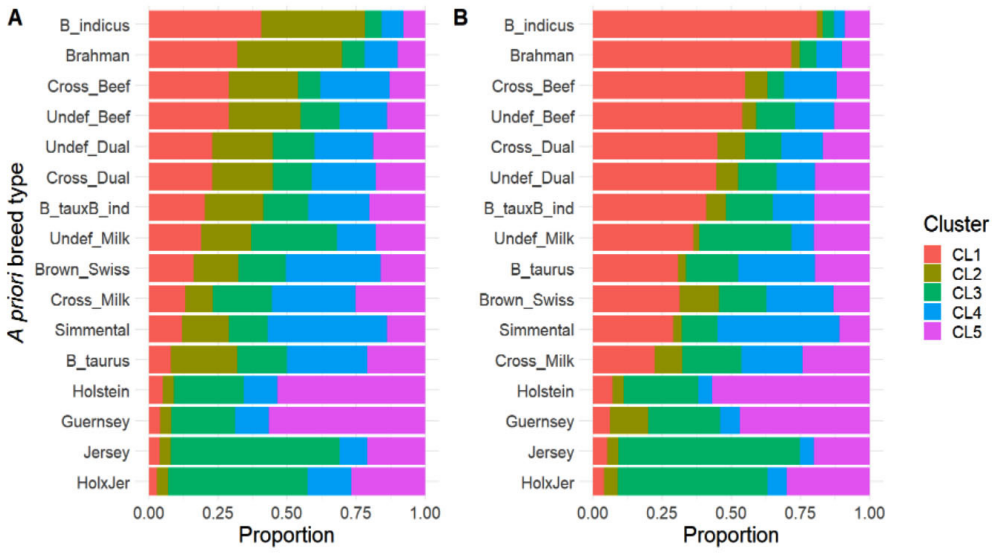

Figure 6. Genetic structure of the Costa Rican bovine population, associated with breed types for K = 5, according to unsupervised (A) and supervised (B) admixture genetic model approaches, and sorted decreasingly by proportion in CL1. Costa Rica, 2015.

Figura 6. Estructura genética de la población bovina de Costa Rica, asociada con tipos raciales para K = 5, de acuerdo con modelos de mezcla genética no-supervisados (A) y supervisados (B), en orden descendente según proporción en CL1. Costa Rica, 2015.

For K > 5, newly formed clusters lacked clear association with predefined breed types. Preidentified breeds such as Holstein, Jersey, Guernsey, and Simmental remained heavily associated with specific clusters in both scenarios, although with significant proportions from other clusters as well. Redundant clusters were also observed, as two clusters might have almost identical genetic structure regarding the relative contribution from the predefined breed types.

This pattern is also consistent with the decrease in Log(K) observed in Figure 2A. For the supervised scenario, another major difference was that the cluster conformed by Brahman and other minor *B. indicus* breeds (CL1 in Figures 2 to 6) remained almost unchanged from K = 1 to 16, while for the unsupervised scenario, this cluster was divided from K = 5 onward.

A minimum spanning tree was built to explore the relationships between predefined breeds and clusters in a two-dimensional plane (Figure 7). The results for both scenarios were practically identical. CL1 was located next to beef animals (B_ind, Brahman, Cross_Beef, Undef_Beef and Undef_Dual). CL2 was close to Holstein and Guernsey breeds, dairy cross breeds, Simmental, Brown Swiss and other minor *B. taurus* breeds, while CL3 was oriented in the same direction as Jersey, Holstein×Jersey and dairy cross breeds.

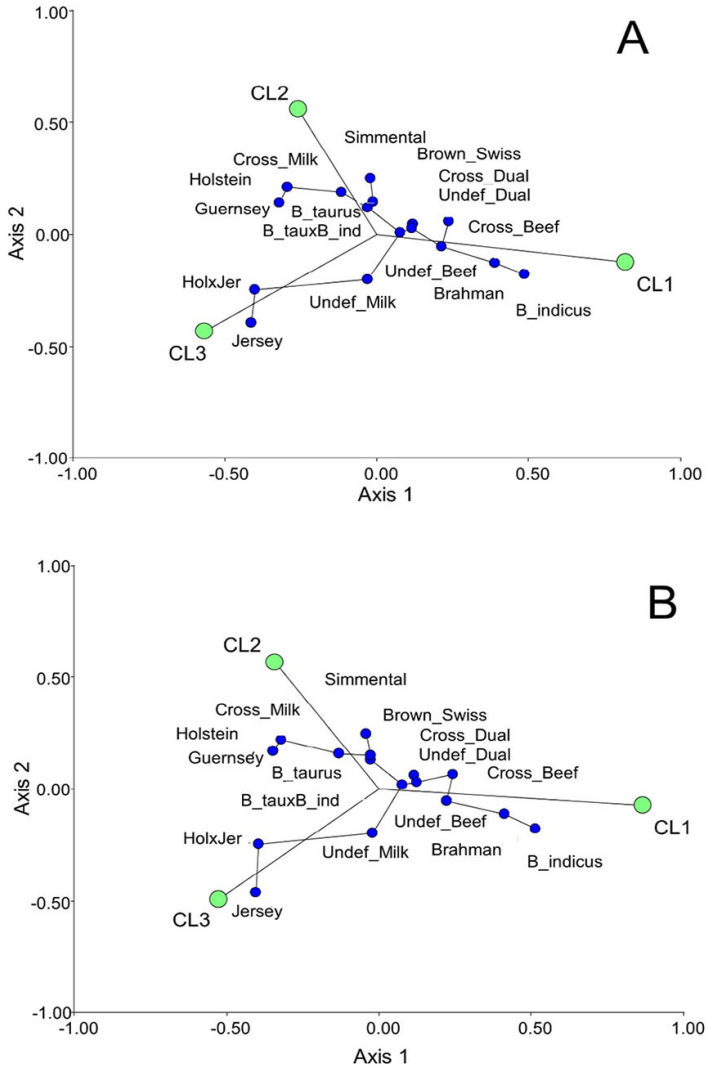

Figure 7. Minimum spanning tree based on breed types for the unsupervised (A) and supervised (B) models with K = 3. Costa Rica, 2015.

Figura 7. Árbol de recorrido mínimo basado en tipos raciales para modelos no-supervisados (A) y supervisados (B) con K = 3. Costa Rica, 2015.

The horizontal axis mainly separates *B. taurus* breeds, located mostly on the left, from *B. indicus* breeds, located to the right. The intermediate position of the crossbred type of cattle is also evident. The vertical axis also provides further differentiation between Jersey and Holstein×Jersey cross, located below, compared to Holstein and other *B. taurus* and dairy cross breeds, located above. It is likely that the Holstein×Jersey group is composed of animals with a greater influence of Jersey breed, as this breed is more common within the country.

Discussion

The genetic characterization of cattle populations in Costa Rica provides essential information for guiding national breeding and conservation strategies. By identifying the genetic diversity, population structure, and unique genetic resources within and between breeds, this study supports informed decisions to enhance productivity while preserving valuable genetic variation. Such knowledge can help prioritize the conservation of rare or locally adapted breeds, optimize crossbreeding programs, and ensure the long-term resilience of the national herd in the face of environmental and market challenges.

Most studies describing the genetic structure of cattle have focused on specific breeds of interest (Agung et al., 2019; Brasil et al., 2013; Edea et al., 2015; Egito et al., 2007; Kumar et al., 2003; Martínez et al., 2012; Martinez et al., 2023; Ocampo et al., 2021), with only a few considering crossbred cattle (Gebrehiwot et al., 2020). In contrast, real cattle populations are often diverse, due to historical events of admixture or recurrent crossbreeding. In these circumstances, the term "breed" becomes abstract and uninformative when it is desired to describe the genetic structure of a real population.

In this study, a representative sample of cattle taken from many herds distributed throughout the country was analyzed. This sample included a large proportion (43 %) of crossbred or undefined breed cattle. In this context, the use of admixture genetic models to describe the genetic structure of a population is conceptually more appropriate, especially when compared to phylogenetic trees based on predefined populations, as performed in previous studies on the same population (Cordero-Solórzano et al., 2015; Martínez et al., 2015). This analysis yielded additional information about the genetic structure and relative composition of local cattle, revealing that there is substantial genetic admixture between the *a priori* defined breed types, and thus providing a more complete and in-depth analysis, which is also more consistent with the historical development of local cattle.

None of the breeds was exclusively assigned to a specific single cluster. This phenomenon provides evidence of an admixture population, regardless of the breed type, which is reasonable, due to the widespread origins of the cattle population analyzed. This pattern was observed in results obtained from both, the unsupervised and supervised models.

Although *B. indicus* and *B. taurus* were separately domesticated, a small number of European cattle breeds still display shared ancestry with indicine cattle (Upadhyay et al., 2019). In fact, the *B. indicus* population expanded in the Americas due to the extensive use of European *B. taurus* cows, particularly the Creole breeds, which descended from Iberian taurine animals brought to the continent after 1492, followed by repeated backcrossing to *B. indicus* bulls (Utsunomiya et al., 2019).

For low values of K the study revealed strong agreement between the genetic structure obtained from unsupervised and supervised scenarios. At this level, there was also consistency between *a priori* breed types and the actual determined genetic structure. This provides evidence to support that prior classification based on information such as phenotypes and production purpose was useful for differentiating, at a large scale, among genetic subpopulations present within the country.

The finding that K = 3 is the optimal clustering level, according to the Delta K criterion, might seem somewhat low. However, genetic clustering algorithms are parsimonious in the sense that they choose the smallest number of ancestral populations that can explain the most salient variation in the data (Alexander & Lange, 2011; Lawson et al., 2018). Parsimony is desirable because it leads to more easily interpretable and probably more realistic parameter estimates (Alexander & Lange, 2011). In fact, for real data the assumption that there is an optimal true value of K is always incorrect, rather, the question is whether the model is a good enough approximation to be practically useful (Lawson et al., 2018).

The assignment of clusters at low levels of K was readily biologically interpretable, as the divergence between indicine cattle and taurine cattle is estimated to have occurred approximately 250 000 years ago (Upadhyay et al.,

2019). Most studies describing the genetic structure of cattle populations also reported clear distinctions between *B. indicus* and *B. taurus* breeds (Brasil et al., 2013; Edea et al., 2015; Egito et al., 2007; Gebrehiwot et al., 2020; Martínez et al., 2012; Martinez et al., 2023).

For K = 3, the partitioning of *B. taurus* breeds into two groups can be explained by the presence of several dairy breeds within the sample, which were declared "identified" (POPFLAG = 1) for the supervised scenario. These are genetically more homogeneous and therefore can be readily differentiated from other subpopulations. Interestingly, these breeds were separated almost identically by unsupervised models, which supports the efficiency of the algorithm in confirming their genetic identity.

B. indicus remained separate in both models (unsupervised and supervised), which might indicate less genetic variability within *B. indicus* than within *B. taurus*. A possible reason for this clustering behavior might be the more recent introduction of *B. indicus* into the country (Cordero-Solórzano et al., 2015).

As the number of groups (K) increased, some differences became evident between supervised and unsupervised scenarios, which are likely due to the presence of a large proportion of individuals with highly heterogeneous genetic compositions, mainly those identified in the present study as crossbred or undefined. At higher levels of clustering, within-group variance becomes large compared to between-group variance and putative subtle differences among groups become difficult to detect by clustering algorithms. The consistent reduction in Log(K) values for $K > 6$ (Figure 1) for both scenarios also suggested that the genetic differences among these newly conformed groups were not as obvious.

Sample size can also play a role in these results; specifically, groups that contain fewer samples or have undergone little population-specific drift of their own are likely to be fit by STRUCTURE as mixes of multiple drifted groups, rather than assigned to their own ancestral population (Lawson et al., 2018). In addition, when differences are less obvious, newly conformed clusters are based on an approximately equal distribution of unassigned samples, which gave rise to redundant clusters that were observed in this study.

The use of supervised clustering procedures in the current study contributed to more stable clustering patterns, especially for higher values of K. This is important because, under real circumstances, some subpopulations are known and several reference individuals from each population are available (Alexander & Lange, 2011). In these cases, ancestry estimates can be estimated more accurately because there is less uncertainty in allele frequencies (Alexander & Lange, 2011; Gebrehiwot et al., 2020). This can be clearly appreciated in Figure 1A, where the means of log (K) for the supervised model showed a more homogeneous pattern and consistently lower standard deviations.

Crossbred and undefined breeds had a weaker relation with clusters, compared to breed types defined as pure breeds. This becomes evident from their distant position from each cluster in the minimum spanning tree (Figure 6), placing crossbred and undefined breeds far away from the clusters. Despite the distance of those categories from clusters, orientations in the minimum spanning tree of undefined or crossbred cattle according to purpose (beef or milk) were consistent with clusters containing breeds typically used for those purposes.

Microsatellites are one of the best marker techniques for the analysis of genetic structure in a population (Abdul Muneer, 2014); however, there have been reports of troubles inferring admixture populations when individuals from the founding subpopulations were not included in the analysis and the microsatellite considered too few loci (eleven) (Vaughan et al., 2009). The admixture models implemented in STRUCTURE allow the detection of admixture at any time point (Anderson & Thompson, 2002), which makes this software suitable for the analysis of admixed populations. Many real population histories, however, are not neatly separable into divergence and admixture phases but the methods can be applied to any data set, producing ancestry bar plots (Lawson et al., 2018).

There are more recent genetic tools, including single nucleotide polymorphisms (SNP), which have shown similar patterns to microsatellite markers (Coates et al., 2009), but might be more precise in clustering analysis of populations (Zimmerman et al., 2020) and more accurate in representing the distribution of genetic diversity among individuals (Pérez-González et al., 2023). In addition to the aim of the study, the selected genetic tool will

depend mainly on economic solvency, with a greater density of markers (i.e. SNPs) being more expensive than microsatellite markers (Puckett, 2017).

Conclusions

The use of admixture models provided additional insights on the genetic structure of cattle in Costa Rica, compared to previous studies on the same population. The models were able to describe the complex genetic structure of the local population. This approach constitutes a valuable resource in the study of genetically admixed populations, which is the prevalent case in commercial livestock populations in the tropics.

This methodology makes also possible to use prior information, such as phenotypes or production purpose, which in this study contributed to obtaining more stable clustering patterns. This reveals the importance of considering genetic-environmental interrelationships when describing genetic structure of cattle population.

All the *a priori* breed types showed some degree of genetically admixed composition, revealing that the so-called pure breeds are nothing but a consensus term of breed types showing a greater proportion of microsatellite markers that set them into a defined cluster with similar purposes and phenotypes, without entirely pertaining to that group, due to the nature of the mixed composition of their genome.

Acknowledgement

Authors acknowledge the collaboration from SENASA, which was responsible for sample and data collection. AV thanks the Costa Rica Institute of Technology (ITCR) and Vice-Chancellor's office of Research and Extension; VIE (Project VIE-5402-2151-1019)].

Interests conflict

The authors declare no conflict of interest.

References

Abdul Muneer, P. M. (2014). Application of microsatellite markers in conservation genetics and fisheries management: recent advances in population structure analysis and conservation strategies. *Genetics Research International, 2014*(1), Article 691759. https://doi.org/10.1155/2014/691759

Agung, P. P., Saputra, F., Zein, M. S. A., Wulandari, A. S., Putra, W. P. B., Said, S., & Jakaria, J. (2019). Genetic diversity of Indonesian cattle breeds based on microsatellite markers. *Asian-Australasian Journal of Animal Sciences, 32*(4), 467–476. https://doi.org/10.5713/AJAS.18.0283

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics, 12*(1), Article 246. https://doi.org/10.1186/1471-2105-12-246

Anderson, E. C., & Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics, 160*(3), 1217–1229. https://doi.org/10.1093/GENETICS/160.3.1217

Brasil, B. S. A. F., Coelho, E. G. A., Drummond, M. G., & Oliveira, D. A. A. (2013). Genetic diversity and differentiation of exotic and American commercial cattle breeds raised in Brazil. *Genetics and Molecular Research, 12*(4), 5516–5526. https://doi.org/10.4238/2013.NOVEMBER.18.2

Carneiro Vieira, M. L., Santini, L., Lima Diniz, A., & De Freitas Munhoz, C. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and Molecular Biology, 39*(3), 312–328. https://doi.org/10.1590/1678-4685-GMB-2016-0027

Coates, B. S., Sumerford, D. V., Miller, N. J., Kim, K. S., Sappington, T. W., Siegfried, B. D., & Lewis, L. C. (2009). Comparative performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis. *Journal of Heredity, 100*(5), 556–564. https://doi.org/10.1093/JHERED/ESP028

Cordero-Solórzano, J. M., Vargas-Leitón, B., León-Rodríguez, B., Chacón-González, I., & Martínez-Pichardo, M. (2015). Diversidad genética en bovinos de ocho regiones en Costa Rica. *Agronomía Mesoamericana, 26*(2), 191–202. https://doi.org/10.15517/am.v26i2.19275

Dufresnes, C., Dutoit, L., Brelsford, A., Goldstein-Witsenburg, F., Clément, L., López-Baucells, A., Palmeirim, J., Pavlinić, I., Scaravelli, D., Ševčík, M., Christe, P., & Goudet, J. (2023). Inferring genetic structure when there is little: population genetics versus genomics of the threatened bat *Miniopterus schreibersii* across Europe. *Scientific Reports, 13*, Article 1523. https://doi.org/10.1038/s41598-023-27988-4

Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources, 4*, 359–361. https://doi.org/10.1007/S12686-011-9548-7

Edea, Z., Bhuiyan, M. S. A., Dessie, T., Rothschild, M. F., Dadi, H., & Kim, K. S. (2015). Genome-wide genetic diversity, population structure and admixture analysis in African and Asian cattle breeds. *Animal, 9*(2), 218–226. https://doi.org/10.1017/S1751731114002560

Egito, A. A., Paiva, S. R., Albuquerque, M. do S. M., Mariante, A. S., Almeida, L. D., Castro, S. R., & Grattapaglia, D. (2007). Microsatellite based genetic diversity and relationships among ten Creole and commercial cattle breeds raised in Brazil. *BMC Genetics, 8*, Article 83. https://doi.org/10.1186/1471-2156-8-83

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology, 14*(8), 2611–2620. https://doi.org/10.1111/J.1365-294X.2005.02553.X

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics, 164*(4), 1567–1587. https://doi.org/10.1093/GENETICS/164.4.1567

Food and Agriculture Organization of the United Nations. (2011). *Molecular genetic characterization of animal genetic resources*. https://www.fao.org/4/i2413e/i2413e00.htm

Gebrehiwot, N. Z., Strucken, E. M., Aliloo, H., Marshall, K., & Gibson, J. P. (2020). The patterns of admixture, divergence, and ancestry of African cattle populations determined from genome-wide SNP data. *BMC Genomics, 21*(1), Article 869. https://doi.org/10.1186/S12864-020-07270-X

Instituto Nacional de Estadística y Censos. (2021). *Encuesta Nacional Agropecuaria 2021 Resultados generales de la actividad ganadera vacuna y porcina*. https://admin.inec.cr/sites/default/files/2022-10/reagropecENAPECUARIO2021-01.pdf

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics, 23*(14), 1801–1806. https://doi.org/10.1093/BIOINFORMATICS/BTM233

Kumar, P., Freeman, A. R., Loftus, R. T., Gaillard, C., Fuller, D. Q., & Bradley, D. G. (2003). Admixture analysis of South Asian cattle. *Heredity, 91*(1), 43–50. https://doi.org/10.1038/sj.hdy.6800277

Lawson, D. J., Van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications, 9*(1), Article 3258. https://doi.org/10.1038/s41467-018-05257-7

Ma, H., Yu, D., Li, J., Qin, Y., Zhang, Y., & Yu, Z. (2022). Construction of first genetic linkage map based on microsatellite markers and characterization of di- and tri-nucleotide microsatellite markers for *Crassostrea hongkongesis*. *Aquaculture, 556*, Article 738272. https://doi.org/10.1016/J.AQUACULTURE.2022.738272

Martínez, A. M., Gama, L. T., Cañón, J., Ginja, C., Delgado, J. V., Dunner, S., Landi, V., Martín-Burriel, I., Penedo, M. C. T., Rodellar, C., Vega-Pla, J. L., Acosta, A., Álvarez, L. A., Camacho, E., Cortés, O., Marques, J. R., Martínez, R., Martínez, R. D., Melucci, L., … Zaragoza, P. (2012). Genetic footprints of iberian cattle in America 500 years after the arrival of Columbus. *PLoS ONE, 7*(11), Article e49066. https://doi.org/10.1371/JOURNAL.PONE.0049066

Martínez, M., Vargas, B., Cordero, J., Chacón, I., & León, B. (2015). Diversidad genética entre subpoblaciones raciales bovinas de Costa Rica. *Agronomía Costarricense, 39*(2), 33–45. https://doi.org/10.15517/RAC.V39I2.21772

Martinez, R., Bejarano, D., Ramírez, J., Ocampo, R., Polanco, N., Perez, J. E., Onofre, H. G., & Rocha, J. F. (2023). Genomic variability and population structure of six Colombian cattle breeds. *Tropical Animal Health and Production, 55*, Article 185. https://doi.org/10.1007/S11250-023-03574-8

Ocampo, R. J., Martínez, J. F., & Martínez, R. (2021). Assessment of genetic diversity and population structure of Colombian Creole cattle using microsatellites. *Tropical Animal Health and Production, 53*, Article 122. https://doi.org/10.1007/S11250-021-02563-Z

Pérez-González, J., Carranza, J., Anaya, G., Broggini, C., Vedel, G., De la Peña, E., & Membrillo, A. (2023). Comparative Analysis of Microsatellite and SNP markers for genetic management of red deer. *Animals, 13*(21), Article 3374. https://doi.org/10.3390/ANI13213374

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., & Lareu, M. V. (2013). An overview of STRUCTURE: Applications, parameter settings, and supporting software. *Frontiers in Genetics, 4*, Article 98. https://doi.org/10.3389/FGENE.2013.00098

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945–959. https://doi.org/10.1093/GENETICS/155.2.945

Puckett, E. E. (2017). Variability in total project and per sample genotyping costs under varying study designs including with microsatellites or SNPs to answer conservation genetic questions. *Conservation Genetics Resources, 9*, 289–304. https://doi.org/10.1007/S12686-016-0643-7

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics, 195*(3), 693–702. https://doi.org/10.1534/GENETICS.113.154138

United Nations. (2024). *World Population Prospects*. Department of Economic and Social Affairs. Retrieved January 20, 2025, from https://population.un.org/wpp/

Upadhyay, M., Bortoluzzi, C., Barbato, M., Ajmone-Marsan, P., Colli, L., Ginja, C., Sonstegard, T. S., Bosse, M., Lenstra, J. A., Groenen, M. A. M., & Crooijmans, R. P. M. A. (2019). Deciphering the patterns of genetic admixture and diversity in southern European cattle using genome-wide SNPs. *Evolutionary Applications, 12*(5), 951–963. https://doi.org/10.1111/EVA.12770

Utsunomiya, Y. T., Milanesi, M., Fortes, M. R. S., Porto-Neto, L. R., Utsunomiya, A. T. H., Silva, M. V. G. B., Garcia, J. F., & Ajmone-Marsan, P. (2019). Genomic clues of the evolutionary history of *Bos indicus* cattle. *Animal Genetics, 50*(6), 557–568. https://doi.org/10.1111/AGE.12836

Vargas, B., & Van Arendonk, J. A. M. (2004). Genetic comparison of breeding schemes based on semen importation and local breeding schemes: framework and application to Costa Rica. *Journal of Dairy Science, 87*(5), 1496–1505. https://doi.org/10.3168/JDS.S0022-0302(04)73301-9

Vásquez-Loaiza, M., & Molina-Coto, R. (2020). Caracterización de la población bovina cebú con certificado de registro genealógico en Costa Rica. *Agronomía Mesoamericana, 31*(3), 679–694. https://doi.org/10.15517/AM.V31I3.39059

Vaughan, L. K., Divers, J., Padilla, M. A., Redden, D. T., Tiwari, H. K., Pomp, D., & Allison, D. B. (2009). The use of plasmodes as a supplement to simulations: a simple example evaluating individual admixture estimation methodologies. *Computational Statistics & Data Analysis, 53*(5), 1755–1766. https://doi.org/10.1016/J.CSDA.2008.02.032

Webster, M. S., & Reichart, L. (2005). Use of microsatellites for parentage and kinship analyses in animals. *Methods in Enzymology, 395*, 222–238. https://doi.org/10.1016/S0076-6879(05)95014-3

Wickenheiser, R. A. (2002). Trace DNA: a review, discussion of theory, and application of the transfer of trace quantities of DNA through skin contact. *Journal of Forensic Sciences, 47*(3), 442–450. https://doi.org/10.1520/JFS15284J

Zimmerman, S. J., Aldridge, C. L., & Oyler-Mccance, S. J. (2020). An empirical comparison of population genetic analyses using microsatellite and SNP data for a species of conservation concern. *BMC Genomics, 21*, Article 382. https://doi.org/10.1186/S12864-020-06783-9