

Genetic kinship and divergence analysis among Alberes cattle breed from Spain and France assessed by SNP chip

Departament de Ciència Animal i dels Aliments || Universitat Autònoma de Barcelona

Period of the internship: July 2021 – September 2021

Tutor: Jesús Piedrafita Arilla

Delivery date: 12nd October 2021

Emiliano Navarro Garre

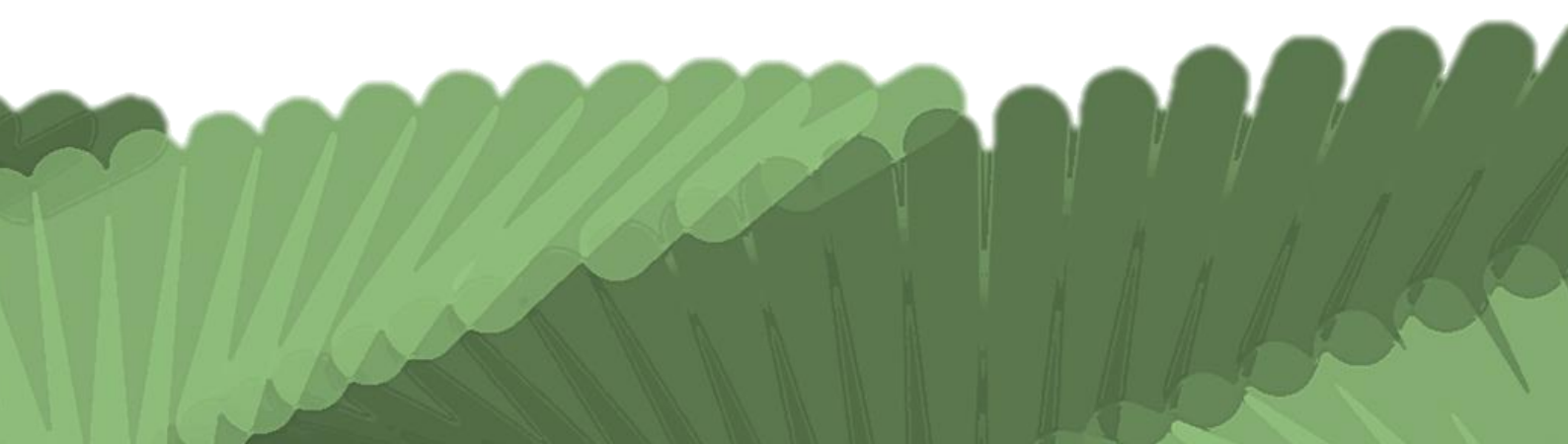


Table of contents

1 Department Description	1
2 Introduction and Objectives	1
3 Methodology	2
Project Timeline	2
3.1 Analysis tutorials	2
3.2 Quality Control	3
3.3 File format transformation	3
3.4 Principal Component Analysis	4
3.5 Structure Analysis	4
3.6 Genetic Relationship Matrix	5
4 Results	6
4.1 Principal Component Analysis	6
4.2 Structure Analysis	7
4.3 Genetic Relationship Matrix	8
5 Discussion	10
6 Supplementary Information	11
6.1 Initial Files	11
6.2 RStudio scripts	11
7 References	11
8 Annex I: Valuations	13

1 Department Description

This internship project has been carried out at the **Conservation and Improvement of Animal Resources group** of the Animal and Food Science Department of the Autonomous University of Barcelona (UAB), within their current project: *“Valorisation of bovine and silvopastoral resources of the across border Mediterranean Pyrenean massif”*.

The aim of the group is to study the structure and genetic diversity of populations through morphological characters and molecular markers to establish conservation and improvement programs for information management, for evaluation of breeders or for the detection of QTLs in order to define optimal selection strategies [Department, 2021].

2 Introduction and Objectives

As I have previously commented, this internship is within the group’s current project, which consists of the realization of a plan for the conservation of the Alberes bovine breed through the incorporation of animals of the two slopes of the eastern Pyrenees evaluated morphologically and genetically.

The Alberes population constitute an autochthonous and endangered breed (Fig. 1, b), according to the FAO classification, from northern Catalonia (Fig. 1, a), that lives free all the year between Catalonia and France territories and it is essential for the control of vegetation and the reduction of fire risk [Fina et al., 2008].

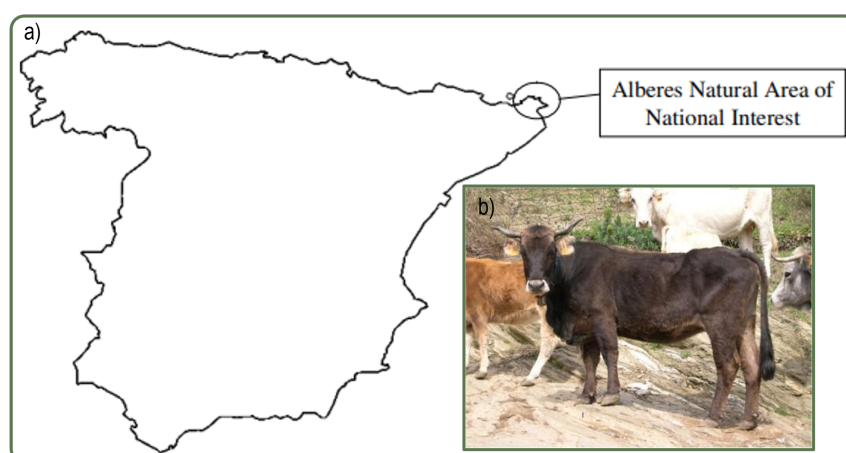


Figure 1 – a) The Alberes cattle is located in the Natural Park of the Alberes Massif, in the eastern extreme of the Pyrenees and b) a typical Alberes Black cow. Modified from Fina et al. [2008]

That is why the project's **objectives** have been to study the genetic diversity and divergence between the cows from both countries from genotyping carried out using a SNP marker chip, based on the work made by Cañas-Álvarez et al. [2015] to establish the bases for the joint genetic management of the aforementioned populations with a view to the conservation of this important breed.

Therefore, these months, I have carried out bioinformatics work that allows me to become familiar with the principal components analysis (**PCA**), to study the population structure of a set of animals, as well as to calculate the genomic kinship between them.

3 Methodology

The methodology steps that were used to carry out this internship project can be divided into five kinds of analysis, presented in Fig. 2 in chronological order that were done.

Project Timeline

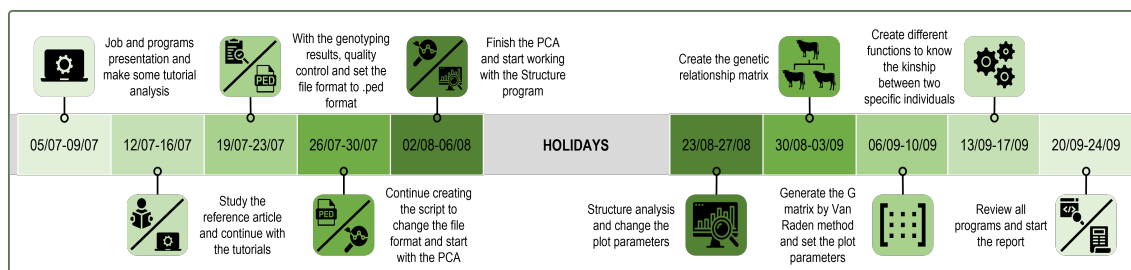


Figure 2 – Internship timeline.

All the scripts that are explained below were created in RStudio [RStudioTeam, 2021] and were carried out in the Linux operating system.

Before starting the project, a bibliographic search was carried out focused on works related to the objective of the study, in addition to the reference paper.

3.1 Analysis tutorials

First of all, a tutorial [Mészáros, 2020] was made to understand the operation of both PLINK [Purcell et al., 2007] and RStudio, as well as the file formats required

and generated with PLINK software version 1.09 and how to do a quality control on genotypic data and a principal component analysis.

The results of the tutorial are not presented in this report, because they have been carried out in order to gain some practice and knowledge about the programmes and the analyses that would be carried out later.

3.2 Quality Control

To analyze the genetic diversity of Alberes cattle, the first step was to perform a quality control (**QC**) on the initial files obtained from the sequencing of ~63,000 SNPs from the 781 cow samples, distributed in 4 populations that we have assigned as Catalan Alberes (**Cat-Albera**), Catalan Not Alberes (**Cat-notAlbera**), French Alberes (**Fr-Albera**) and French Not Alberes (**Fr-notAlbera**)(Supplementary Information, Initial Files).

The data pruning was performed using the pertinent PLINK commands thanks to the script [QC.r](#) and following the criteria detailed below:

- **Missingness per SNP:** 0.05.
- **Missingness per individual:** 0.05.
- **Minor allele frequency:** 0.05.
- **Hardy-Weinberg threshold:** 0.0001.
- **Mendel error rates:** 0.05.

This quality control is quite lax, unlike the one carried out in Cañas-Álvarez et al. [2015], since otherwise a large amount of genotypic information is lost, especially from French populations, which could be due to differences in the quality of the samples.

After the pruning process, the resulting files have a total of 41,418 variants and 781 bovines, due to the elimination of 50 variants due to the exact Hardy-Weinberg test.

3.3 File format transformation

The files obtained after completing the quality control, do not present the desired format to carry out the analyzes, so, with the script [addIndividualData.r](#), the individual information was added to the original data by PLINK, according to the PED file format,

which is a white-space (space or tab) delimited file, where the first six columns are mandatory in the order that follows:

- Family ID
- Individual ID
- Paternal ID
- Maternal ID
- Sex
- Phenotype

The pruned file has the same family and individual ID (e.g. 81.cel) and there is no information for the rest of the columns, since 0 means "no information" like -9 in the phenotype column. With the script previously commented, the family ID becomes to Cat/Fr-Albera/notAlbera, representing the breed; the individual ID is the same without the ending ".cel" and the only additional information that was added was the sex (1=male, 2=female, other=unknown), since it was the only one present in the file [InfoSamplesAlbera_BP_Group.xlsx](#).

3.4 Principal Component Analysis

To achieve an approach to characterize divergence, a PCA was applied to the relationship matrix, built up from the pairwise identity by state (IBS) between all animals using PLINK . Each entry of this relationship matrix relates any 2 animals genotyped and is computed as 1 minus the pairwise IBS, averaged across markers [Cañas-Álvarez et al., 2015]. All estimates and plots were performed using [PCAgraph_4.r](#) script, developed under an R environment.

From the data pruned files, SNPs in strong linkage disequilibrium (**LD**) were excluded for this analysis, because they can affect the PCA as indicated in Cañas-Álvarez et al. [2015] and in the work done by Visser et al. [2016], so 19,299 variants were pruned in a window of 50 SNP, sliding the window by 5 SNP and $r^2 = 0.3$ (where r^2 is the pairwise genotypic correlation) at a time as in Moorjani et al. [2013] and 22,119 remains in the file used for the PCA.

3.5 Structure Analysis

Additionally, to characterize the genetic structure across the 4 cattle populations, a cluster analysis using the ADMIXTURE software version 1.3 [Alexander et al., 2009] was performed in an Anaconda environment [AnacondaDocumentation, 2020]. The program implements a maximum likelihood method to infer the genetic ancestry of

each animal from a mixture of K predefined ancestral groups [Cañas-Álvarez et al., 2015]. The number of clusters (K) tested ranged from 2 to 12. A preferable value of K will exhibit a low crossvalidation error compared with other K values.

The code used to make this analysis is described below:

```
1 #From the directory with the PED files
2 > plink --file alberaDef --cow --recode12 --out alberaDef
3 #Admixture analysis
4 > for K in 2 3 4 5 6 7(...); do admixture --cv alberaDef.ped $K | tee log${K}.out; done
5 #CV error
6 > grep -h CV log*.out
```

The second line of the previous code recode the genotypic information as 1,2 way instead of the allelic way, due to Admixture works with 1,2 codification. Also, the admixture plot was performed using the self-written [plotspophelper.r](#) script, using the pophelper package [Francis, 2017].

3.6 Genetic Relationship Matrix

Finally, the genetic relationship matrix was calculated from the initial files, without having passed the QC (Supplementary Information, Initial Files), because the package calculating the matrix does its own QC. The self-written script that generates the matrix and performs other analysis and statistics is [G_VR_Albera.r](#).

To begin with the matrix, the genotypic information was recoded as 0,1,2 way with the "--recode A" command in PLINK and the resulting file was read and manipulated by different R packages ("data.table", "stringr" and "dplyr").

Immediately after, the genetic relationship matrix was calculated by the VanRaden method [VanRaden, 2008] using the AGHmatrix package [Amadeu et al., 2016] with a minor allele frequency (**MAF**) equal to 0.01.

The next step was the creation of a function that returns the relationship between two given IDs of the samples we have and the elaboration of the heatmap diagram, as well as some summary statistics that are discussed in the results.

4 Results

4.1 Principal Component Analysis

The first approach to characterise genetic diversity among our four populations was a PCA, which was applied to a distance matrix constructed from the IBS marker relationship matrix (Fig. 3) as in Cañas-Álvarez et al. [2015]. This type of analysis allows us to represent each animal based on the PCA coordinates, where the first and second principal components explain 34.6% and 19.07% of the total variance. As shown in Fig. 3, the PCA clustering suggest heterogeneity among the four populations.

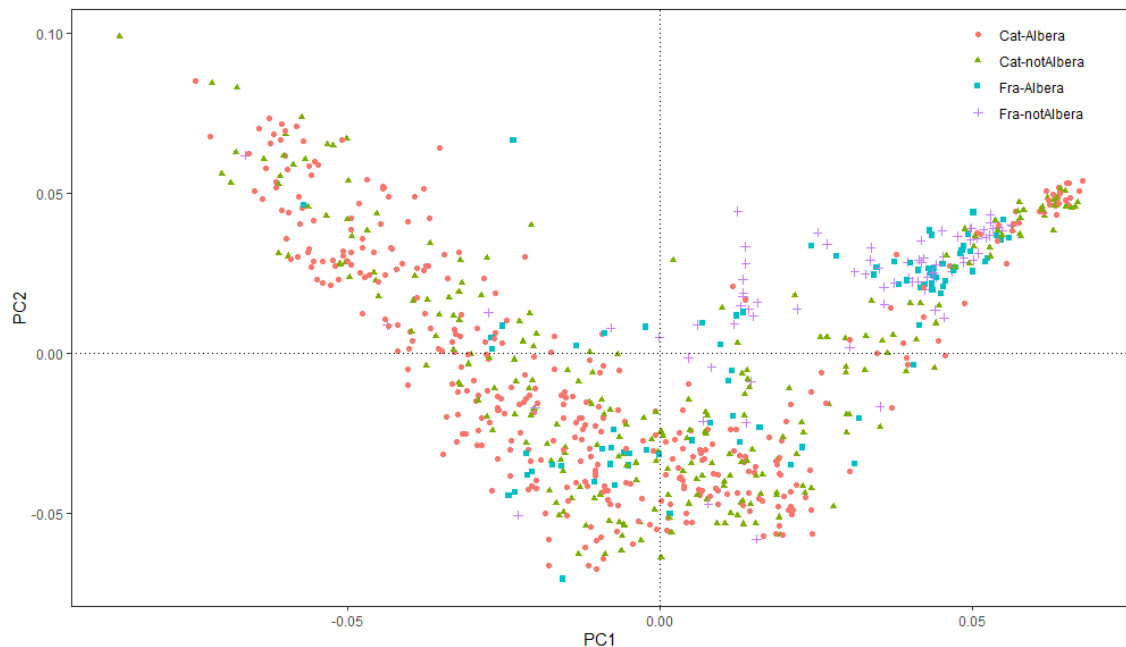


Figure 3 – Population groups defined by PCA, where PC1 and PC2 correspond to principal components 1 and 2, respectively.

Firstly, it can be observed that most of the animals on the French side of the border are grouped together, as well as a small group of cows on the Catalan side of the border, without discriminating between those belonging to the Alberes breed and those not belonging to it. In any case, the mixing of most of the animals analysed and the centrality of their distribution suggests a certain genetic flow between all the cattle populations due to the geographical proximity throughout their reproductive life and the non-existence of physical barriers throughout the year, except in the winter months, when the populations are separated in each country.

4.2 Structure Analysis

To further characterise the divergence between Alberes cows, a cluster analysis was performed using a maximum likelihood method that infers the genetic ancestry of each animal from a mixture of predefined ancestral groups as in Cañas-Álvarez et al. [2015]. This clustering by ADMIXTURE assumes the absence of LD and related animals, although in our case, as we have seen in the previous section, animals seem to have a high gene flow between them. However, all the SNPs analysed in our samples have a mean LD of ~ 0.433 , which is excessively high. However, this analysis has been performed assuming that this assumption is not met and without pruning the data, since doing so would mean the average LD would be approximately 0.242, but we lose a large amount of genotypic information, so this error is assumed.

Thus, the initial breed groupings were tested between K values ranging from K = 2 to K = 12, since, due to the low divergence between the populations analysed, the crossvalidation errors were very similar for all values of K and decreased as the value of K increased, with the smallest of those shown in Fig. 4 belonging to K = 6 (CError = 0.36248). However, if we continued to increase the value of K, the crossvalidation errors would continue to decrease, probably due to population heterogeneity or strong LD. Thus, the maximum likelihood estimation of the ancestries performed by ADMIXTURE assigned clusters belonging to different populations to each breed (Fig. 4), highlighting the high admixture between populations, both between populations on the same side of the border but different breeds, and between breeds on both sides of the border.

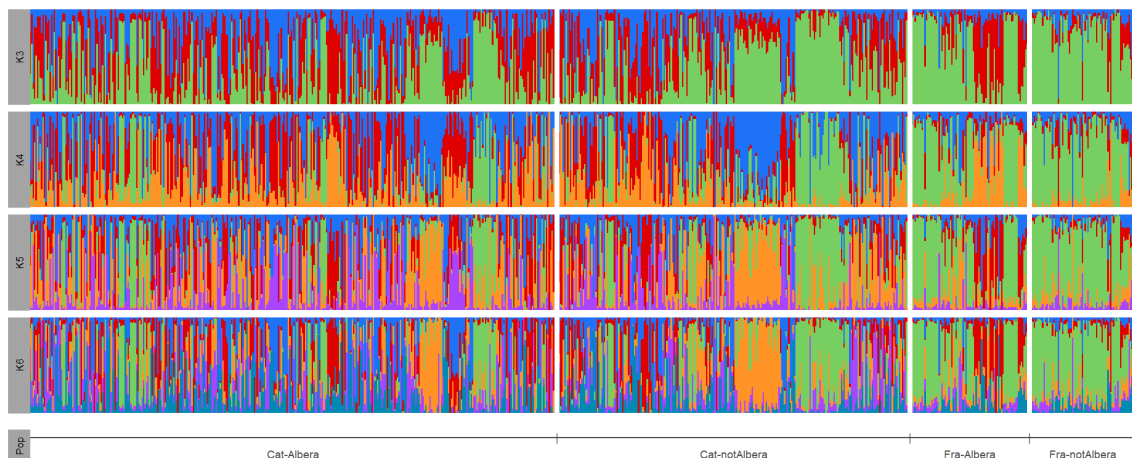


Figure 4 – Estimated membership coefficients for each animal for K = 3 through 6.

It should also be noted that, within the high heterogeneity of the populations, many

of the animals on the French side show a fairly high level of differentiation (green colour, Fig. 4), especially in those cows that do not belong to the Alberes species. This supports the results obtained in the PCA, where a large group of French cows showed a certain distance within the point cloud, as well as a few animals from the Catalan region, but not belonging to the Alberes breed, which present clusters similar to the French cows for the K values studied. This allows us to reaffirm the absence of a reproductive barrier due to belonging to one or the other side of the border.

4.3 Genetic Relationship Matrix

Finally, the relationship matrix of the animals analyzed was calculated in order to obtain different statistics from it and to confirm the results obtained in the previous analyses. Thus, a genomic kinship matrix with dimensions of 781x781 was obtained, where the self-relationship coefficients have average values of 0.852 ± 0.154 , which is to be expected for these coefficients. Nevertheless, the rest of the values in the matrix, i.e. the relationships between different animals, show an average value of 0.001 ± 0.072 , indicating a low degree of genetic relationship, which is in contrast with previous analyses. This may be due to the LD present between the markers studied, since for this analysis the quality control for this parameter is carried out by the AGHmatrix package itself. In any case, it is true that the cattle populations studied show a high degree of heterogeneity. This can be seen graphically in the Fig. 5, where the diagonal represents the self-pairs and where we can observe some clusters of animals in French populations (Alberes and NotAlberes) and in Catalan populations not belonging to the Alberes breed that present a kinship coefficient above the average of the population analysed.

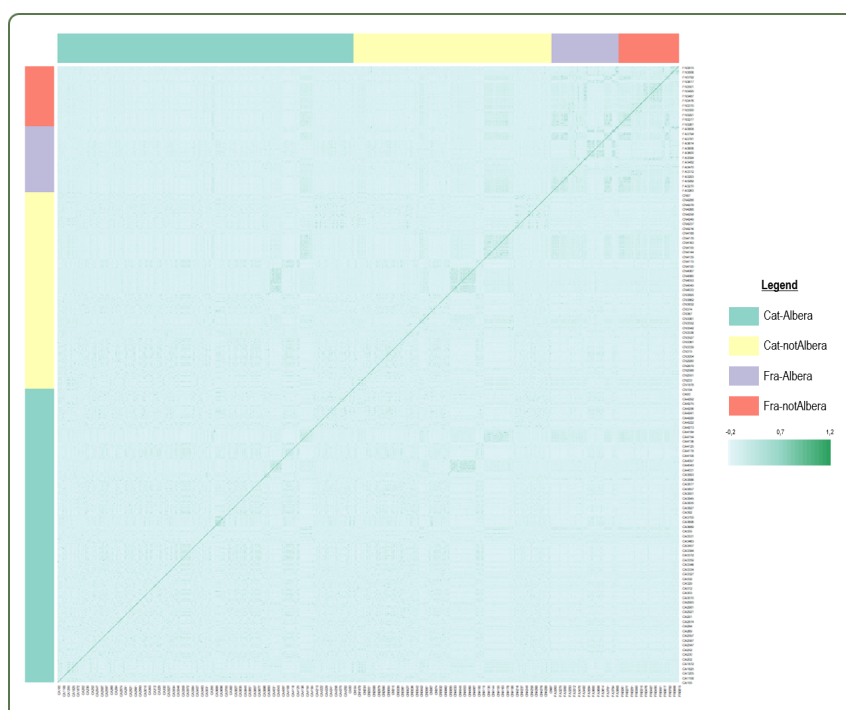


Figure 5 – Graphical representation of the genetic relationship matrix with the populations represented by colored bars outside the matrix. The colour scale for the genetic kinship within the matrix ranges from -0.2 to 1.2 as shown in the legend.

In addition, other statistics were calculated for each of the Alberes populations and for the breed as a whole (Table 1). It can be observed that the mean of the self-relationship coefficient for the Catalan population is slightly above the mean of the breed due to the high number of individuals analysed from this population, in the same way that, for the mean of the kinship coefficients of all the individuals, the mean of the Catalan population is more similar to the mean of the group, as it contributes a higher number of animals to the Alberes breed, being very close to 0, which can highlight the absence of relationship between crosses of animals of the same breed or with animals not belonging to the breed.

Table 1 – Some Alberes breed populations statistics

Breed population	Animal number	SR ¹ mean	UM ² animals	UM mean
Cat-Alberes	374	0.907±0.141	69,378	0.005±0.094
Fr-Alberes	84	0.759±0.098	3,570	0.035±0.133
Alberes	458	0.88±0.146	104,196	0.004±0.085

¹ Self-relationship

² Upper matrix

Finally, the previously created function allows any user to enter the genetic kinship matrix as calculated previously and two animal IDs in numerical format and returns a warning message with the possible relationship of the two queried animals according to the following parameters:

- **Parents/Full sibs:** 0.45 to 0.55.
- **Half sibs:** 0.2 to 0.4.
- **Same animal:** 0.8 to 1.
- **Two inbred animals:** Greater than 1.
- **Non related individuals:** Less than 0.2.

If two animals have a genetic kinship coefficient greater than 1, this means that these animals have a greater number of the loci studied that are identical by descent (**IBD**), indicating a greater degree of genetic relationship between these animals, and can be considered to be genetically related. Likewise, negative correlation means that detecting an allele in one individual makes it less likely that the allele will be detected in the other animal, indicating no genetic relationship between them.

5 Discussion

As we have been able to observe with the results obtained from the analyses carried out, the populations on both sides of the border show a high degree of admixture, which indicates the absence of any kind of reproductive barrier between the species in the region or between the cows on both sides of the border. Thus, the establishment of a joint genetic management plan for the animals of the Albera massif would be a good idea.

However, in order to carry it out, it would be advisable to repeat the analyses carried out taking into account a geographically close external group, such as the Bruna dels Pirineus breed, as was the objective of this internship report, as well as to ensure the absence of markers in excess of LD, in order to carry out the analyses without assuming any type of error and to be able to compare with an external group, differentiated from the breed studied. Furthermore, other types of analysis such as Runs of Homozygosity (**ROH**) to measure inbreeding, a molecular analysis of variance (**AMOVA**) or trying to reconstruct the phylogeny, and different statistics related to it, from the sequencing data, may help to clarify the results obtained here.

In any case, these analyses help us to have a first idea of the structure of the population and the level of divergence both between and within the established and studied populations.

6 Supplementary Information

6.1 Initial Files

 || : [snpAlbera1c_r.ped](#)


 || : [snpAlbera1c_r.map](#)

 || : [snpAlbera2c_r_nBP.ped](#)

 || : [snpAlbera2c_r_nBP.map](#)

 || : [InfoSamplesAlbera_BP_Group.xlsx](#)

6.2 RStudio scripts

 || : [QC.r](#)

 || : [addIndividualData.r](#)

 || : [PCAgraph_4.r](#)

 || : [plotspophelper.r](#)

 || : [G_VR_Albera.r](#)

7 References

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19:1655. doi:10.1101/GR.094052.109.

Amadeu RR, Cellon C, Olmstead JW, Garcia AAF, Resende MFR, Muñoz PR. 2016. Aghmatrix: R package to construct relationship matrices for autotetraploid and diploid species: A blueberry example. *The plant genome* 9. doi:10.3835/PLANTGENOME2016.01.0009.

AnacondaDocumentation. 2020. Anaconda Software Distribution. <https://docs.anaconda.com/>.

Cañas-Álvarez J, González-Rodríguez A, Munilla S, Varona L, Díaz C, Baro J, Altarriba J, Molina A, Piedrafita J. 2015. Genetic diversity and divergence among spanish beef cattle breeds assessed by a bovine high-density snp chip. *Journal of animal science* 93:5164–5174. doi:10.2527/JAS.2015-9271.

Animal and food science department. 2021. <https://www.uab.cat/web/investigacion/genetica-y-genomica-1345750245558.html>.

Fina M, Casellas J, Tarrés J, Bartolomé J, Plaixats J, Such X, Jiménez N, Sánchez A, Piedrafita J. 2008. Characterisation and conservation programme of the alberes cattle breed in catalonia (spain). *Animal Genetic Resources/Resources génétiques animales/Recursos genéticos animales* 43:1–14. doi:10.1017/S1014233900002686.

Francis RM. 2017. pophelper: an r package and web app to analyse and visualize population structure. *Molecular ecology resources* 17:27–32. doi:10.1111/1755-0998.12509.

Moorjani P, Patterson N, Loh PR, Lipson M, Kisfali P, Melegh BI, Bonin M, Ludevít Kádaši, Rieß O, Berger B, Reich D, Melegh B. 2013. Reconstructing roma history from genome-wide data. *PLOS ONE* 8:e58633. doi:10.1371/JOURNAL.PONE.0058633.

Mészáros G. 2020. Genomics Boot Camp. <https://genomicsbootcamp.github.io/book/>.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, de Bakker P, Daly M, Sham P. 2007. Plink: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81:559. doi:10.1086/519795.

RStudioTeam. 2021. RStudio: Integrated Development Environment for R. <https://www.rstudio.com/>.

VanRaden PM. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91:4414–4423. doi:10.3168/JDS.2007-0980.

Visser C, Lashmar SF, Van Marle-Köster E, Poli MA, Allain D. 2016. Genetic diversity and population structure in south african, french and argentinian angora goats from genome-wide snp data. *PLOS ONE* 11:e0154353. doi:10.1371/JOURNAL.PONE.0154353.

8 Annex I: Valuations

As for the evaluation of the research group, I must say that only Dr. Jesús Piedrafita and I have worked on these analyses, so I can only evaluate him by supervising the project we have carried out and which has been presented here.

I must say that the involvement of the tutor at the beginning of the project was very good, but from the moment we realised that we could not obtain publishable results due to an error in the allocation between breeds and samples, I felt very lonely in carrying out all the analyses and received very little help in understanding the functioning of the programmes and the interpretation of the results. As a result, we were unable to carry out other analyses that we had planned with Arlequin or ROH and which would have completed my training in this area.

Anyway, on a personal level, I think I have learned how it works to perform a research project, where there are mistakes and you have to constantly look for solutions. Having done this project individually, I have learned to appreciate the importance of teamwork between members with different backgrounds who can help you change your point of view when something goes wrong, but it has also allowed me to learn autonomously and to learn how to solve problems that may arise during the analysis.

I think this experience has been very enriching for my future, as I have learned to handle a large number of programmes that can be applied to different fields of genetic research, to interpret the results obtained when carrying out the different analyses that have been discussed and to work and search for information autonomously, which does not detract from the great importance of teamwork.