SMARTER

SMAll RuminanTs breeding for Efficiency and Resilience

Research and Innovation action: H2020 – 772787
Call: H2020-SFS-2017-2
Type of action: Research and Innovation Action (RIA)
Work programme topic: SFS-15-2016-2017
Duration of the project: 01 November 2018 – 30 June 2023

Report on demography and genetic diversity of underutilised breeds

Arianna Manunza[1], Filippo biscarini[1], Paolo Cozzi[1], Johanna Ramirez-Diaz[1], Valentina Tsartsianidou[2], Konstantinos Gkagkavouzis[2], Nikoletta Karaiskou[2], Alexandros Triantafyllidis[2], Bertrand Servin[3], Alessandra Stella[1]

[1] Partner CNR
[2] Partner AUTH
[3] Partner INRAE

[1]Deliverable leader – Contact: alessandra.stella@ibba.cnr.it

DELIVERABLE D4.2

Workpackage N°4

Due date: M48

Actual date: 03/07/2023

Dissemination level: Public

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 772787

Table of contents

1 Summary

The challenge facing small ruminant populations worldwide is to increase productivity while maintaining genetic diversity and their ability to adapt to climate change. The objective of the WP4 on Smarter project was to analyse the genetic diversity and variability levels of underutilised small ruminant breeds as well as the local adaptation.

Research on the genetic diversity of the local worldwide underutilised sheep and goat provide an understanding of their population structure, admixture and inbreeding levels, and the relationships among them. Likewise, it is important to better understand which forces guided the formation of each breed starting from their origin. Demographic events occurred in the past (also recent), natural and artificial selection are the factors that play a key role in shaping their genome.

In addition, local adaptation studies can be used to identify causal factors underlying breed adaptation to specific environmental conditions and have a potential to predict future loss of adaptive genomic variation under climate change. In particular, the adaptation to environmental conditions lead to the selection of alleles that are maintained over the time and that allow the populations to thrive in their habitat and in challenging environmental conditions. The underutilised breeds constitute a reservoir of genetic variation in comparison with the commercial breeds and a useful source of specific alleles and variants that can be relevant in specific breeding objectives.

To do so, we used different approaches:

1. We explored the population structure and the demography in the sheep and goat datasets by using:

- Principal Components Analysis, highlighting the contribution of the SMARTER project and to fill the gap in the analysis of worldwide local breeds
- A distance-based method to explore the relationship among breeds
- ADMIXTURE and individual ancestry components for the foreground data
- Runs of Homozygosity and inbreeding levels

2. We explored the local adaptation using a LFFM algorithm

3. We addressed the problem of the missingness (when missing SNP-genotype data) through imputation

4. We analysed a case-study (commercial vs local breeds) for the detection of Heterozygosity-rich regions, and we performed several tests for improving the parameters setting for both ROH and HRR.

The relationships between breeds represented by a Neighbour-Net graph for sheep and goat highlighted that both species are clustering according to the geographical origin even though some groups present an inner separation. The analysis of individual ancestry in the new genotyped goat breeds revealed an introgression from Alpine breed in the genome of Fossée, a Nordic origin of the Swedish breed and the influence of the Eghoria to Skopelos population that supports previous studies. For the new genotyped sheep breeds, the same analysis supports previous relationships among groups already detected by the Network graph. By investigating the relationship between genotype and environment in some underutilised local breeds from South America and Africa countries, we identifies genes that play a role in adaptation to altitude and temperature, and we inferred that this approach could help to provide recommendations on favourable genotypes for specific climatic and environmental conditions.

These results provide useful information about the genetic status of underutilised european breeds and on the mechanisms that regulate the evolution of genes and traits of interest. This contributes to the planning of effective breeding programs and to more sustainable use of the AnGR.

2 Introduction

Centuries of intensive breeding selection programs aimed to satisfy the increasing demand of livestock products (milk, meat, wool) at global level contributed to the formation of new breeds (often called “synthetic breeds”) by crossbreeding local ecotypes with the more productive ones. In the last decades, several studies on local adaptation were made to better understand the mechanisms of action of the genes involved in this process. Small ruminant traditional breeds are valuable animal resources for small-scale farmers above all in marginal areas and in developing countries, sustaining the local production systems. These breeds have evolved to adapt to production needs and agro-ecological conditions, maintaining adaptive traits shaped by natural selection over the time. In addition, local breeds have an important value as a source of genetic variation for commercial breeds. Despite their importance, most sheep and goat breeds are still uncharacterized and their response to natural selection is still unresolved. Genetic characterization is key for the conservation of livestock and to analyse the genetic diversity provides an understanding of the relationships that exist among breeds as well as the within-breed differentiation. Uncovering their genetic makeup shaped by selection pressures and demographic events can help to identify genes and markers, to determine their functions and to assist in identifying traits of economic importance. The objective of this task was to analyse the genetic diversity and variability levels of underutilised small ruminant breeds (sheep and goat). The findings revealed by these results can also provide a basis for future studies focused on the associations between phenotype and genotype in response to the environment. Finally, to assess their contribution for long-term

food security in changing environments is crucial for making more sustainable, productive and resilient livestock systems.

3 Genetic diversity and demographic analyses

The datasets included 285 breeds for sheep and 145 breeds for goats with a worldwide distribution. Both include transboundary/selected breeds (background data) and underutilised breeds that are the target of this task (foreground data). The SNP data were remapped against OAR3 and ARS1.2 assemblies and the quality control was performed using PLINK v1.9 (Chang et al 2015), following the FAO guidelines for the genomic characterisation of animal genetic resources (Ajmone-Marsan et al 2023). In order to balance the number of animals per breed/population, we used BITE (Milanesi et al., 2017) with a threshold of 30 individuals and excluding the breeds that have less than 5 individuals. PGDSpider v2.0.4.0 (Lischer et al., 2012) was used to convert the file in a specific format suitable for the different programs. Some of the results that follow consist of analyses of foreground dataset for both species. Results shown here are relative to subset including foreground data (representing mainly underutilised breeds) and other background breeds (mostly commercial breeds) as a base of comparison. In details, the goat subset included the following as foreground data (5 breeds): Eghoria (EGH) and Skopelos (SKO) - Greece, Fossés (FSS) and Provencale (PVC) – France, Landrace, (LNR_SE) - Sweden. For comparison we included: French Alpine (ALP_FR), Italian Alpine (ALP_IT), Swiss Alpine (ALP_CH), French Saanen (SAA_FR), Italian Saanen (SAA_IT), Swiss Saanen (SAA_CH), French Angora (ANG_FR), Swiss Boer BOE_CH). Total of (13 breeds/populations). For the sheep foreground data (29 breeds): Boutsko (BOU), Chios (CHI), Frizarta (FRZ), Mytilini (MYT), Pelagonia (PEL) – Greece; Manech Tête Noire (MTN), Bizet (BIZ), Rouge du Roussillon (RDR), Solognote (SOL) – France; Castellana (CAS), Assaf (ASF), Churra (CHU), Ojalada (OJA) – Spain; Bábolna Tetra (BAT), Tsigai (TSI), Dorper (DRP), Hortobágy Racka (HRR), White Dorper (WDR) – Hungary; Rusty Tsigai (RST), Turcana (TRC) – Romania; Creole (CRL), Corriedale (CRR) – Uruguay. Île de France (IDF) - France and Hungary; Suffolk (SUF) - France and Hungary; Merino (MER) – Hungary and Uruguay; Texel (TEX) – Uruguay and France. The breed codes that we used in the analyses are indicated in brackets.

The distance matrices were calculated with Arlequin 3.5 (Excoffier and Lischer 2010), the population structure (Principal Components Analysis – PCA) and the individual ancestry with SNPRelate 4.1.2 (Zheng, et al 2012) and ADMIXTURE 1.3 (Alexander et al 2009), the demography (Runs of Homozygosity detection) with DetectRUNS 0.9.6 (Biscarini et al 2018) using the consecutive method. The Neighbour-Net were calculated and drawn with SplitsTree v4.14.2 (Huson 1998).

3.1. Population structure of the whole datasets

The PCA analyses were performed on the whole dataset, in order to highlight the contribution of the partners of Smarter and the achievement of the objective.

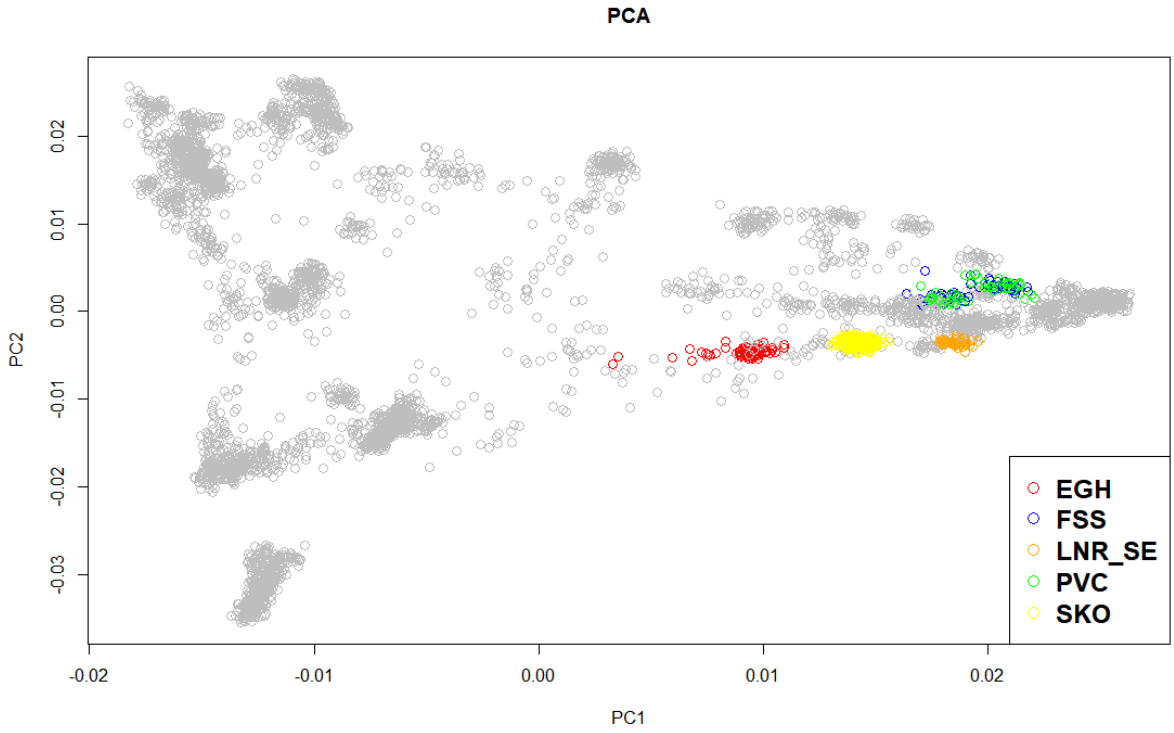

Figure 1 Principal Component Analysis for the goat dataset. The foreground data are labelled in different colours and the background data are in grey.

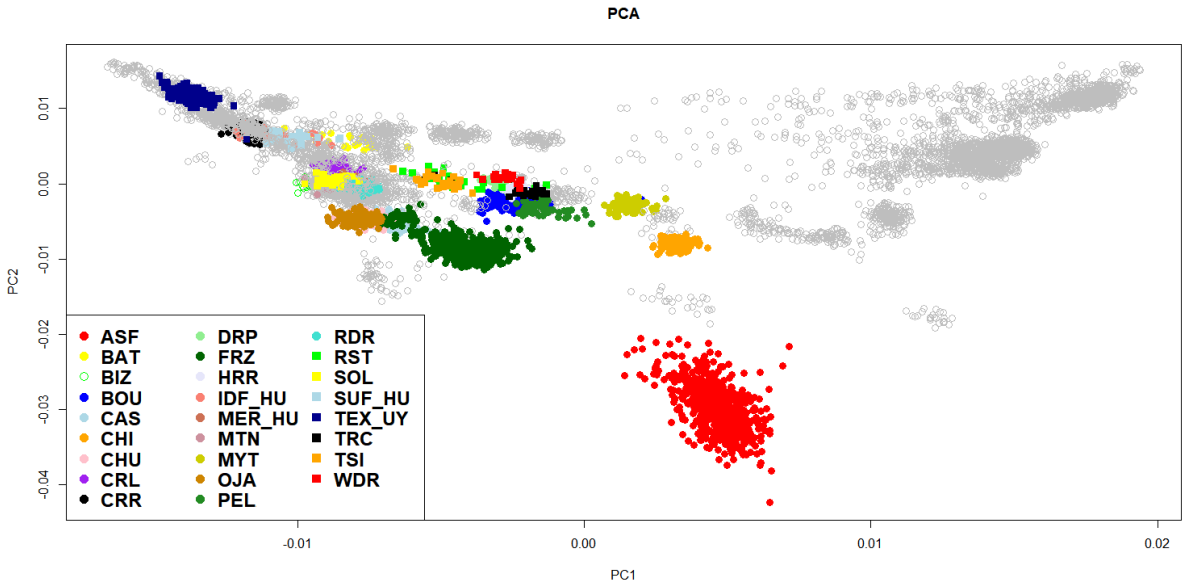

Figure 2 Principal Component Analysis for the sheep dataset. The foreground data are labelled in different colours and the background data are in grey.

3.2. Relationship between breeds using the whole datasets

The relationships between all the breeds were analysed using the Reynolds' distance matrix and a Neighbour-Net method as shown in Figure 3 (goat) and 4 (sheep) below.

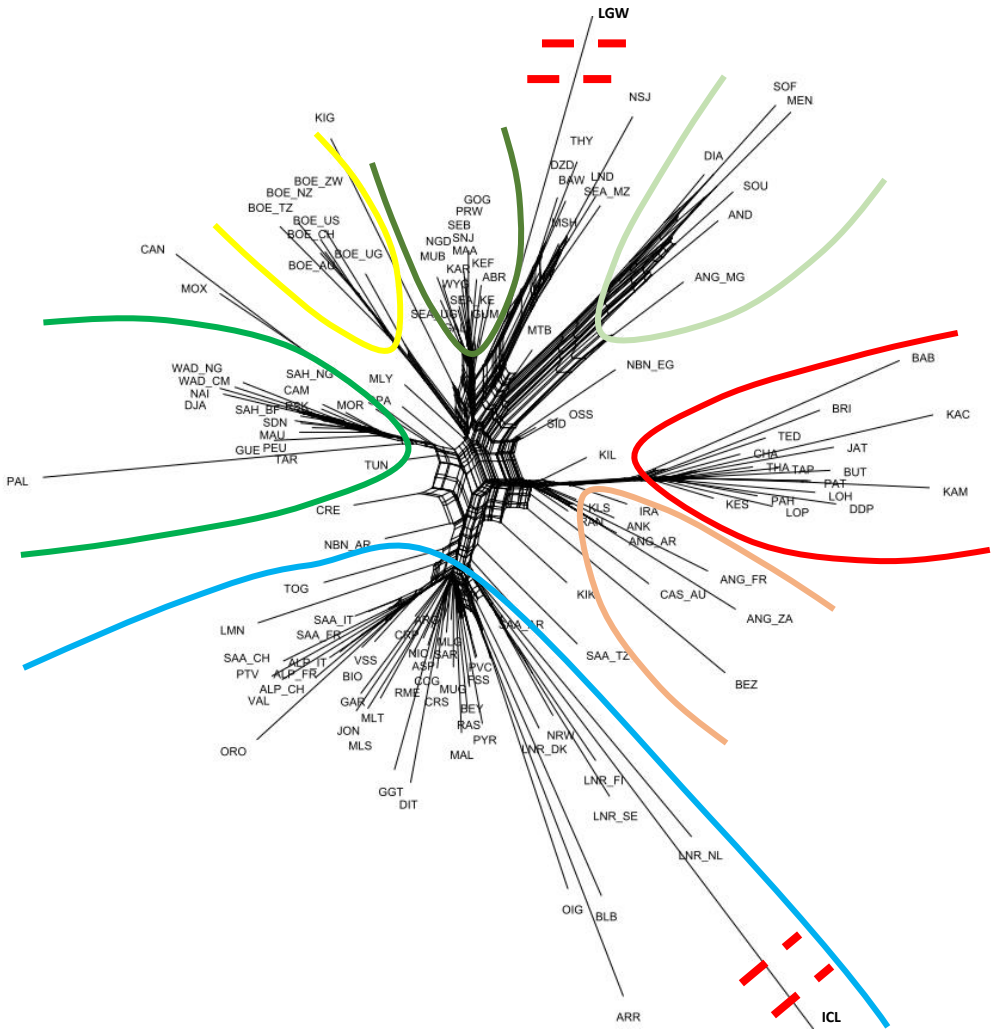

Figure 3 NeighbourNet analysis displaying the relationship and distances among the goat breeds. The red dotted lines indicate that the branch is much longer, and it has been cut for a better visualisation.

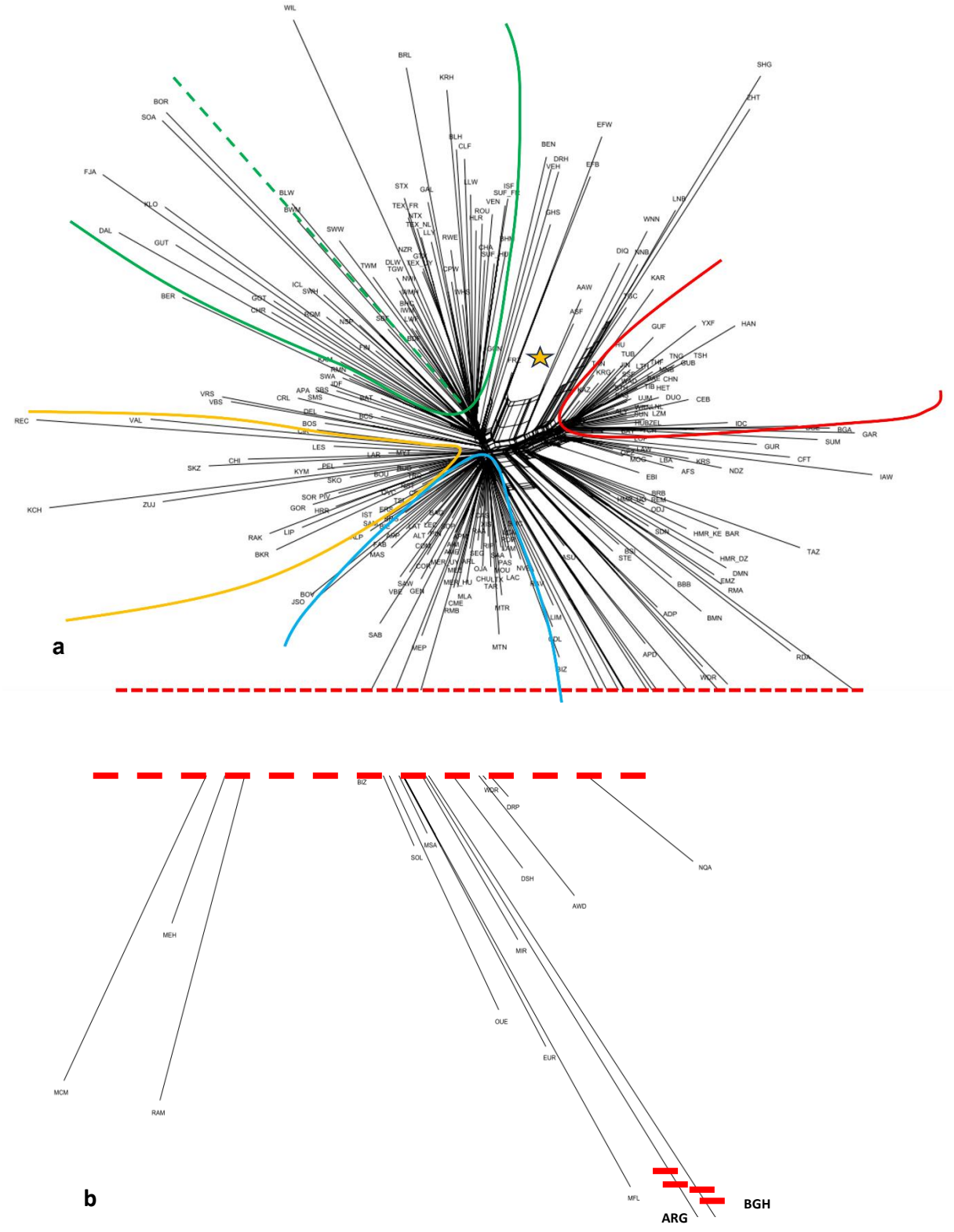

Figure 4**a**. NeighbourNet analysis displaying the relationship and distances among the goat breeds. The red dotted lines indicate that the branch is much longer, and it has been cut for a better visualisation. In **b,** the zoomed bottom part of the graph.

In the goat network graph (Figure 3), we can observe a clear subdivision of the breeds according to the geographical origin. The transboundary breeds such as Angora (light orange) and Boer (yellow) are also clustering all together. More in detail, for the African continent (green curve, different shades) there are three groups: north-west (green), south-east (dark green) and Madagascar (light green). The European counterpart is included in the blue curve, with a more internal partition per country (Italian, French breeds closer to the Alpine and Saanen, followed by the Spanish one. The Scandinavian breeds (Northern Europe) form a unique branch very close to the Irish breeds (Western Europe). The Pakistani ones are grouping inside the red curve.

The NeighbourNet for sheep is more complex and a less clear division is evident. The few remarkable observations are that i) the Chinese breed are well distinct and separate group -in red; ii) the Western European group (Ireland and the British Isles- mainly Wales) are close to the Scandinavian ones – green curve, Scandinavian to the left and British Isles to the right of the dotted green line; iii) the Greek breeds are all together with some breed from The Eastern Europe -in yellow- (Hungry, Romania, Albania, etc.) except the synthetic breed Frizarta (indicated by a star) that is included in the same branch of the source populations (East Friesian white and brown) and a bit more distant from Assaf and Awassi (the last one a fat-tail-type). Assaf is a crossbreed between Awassi and East Friesian). The European counterpart (light blue curve) is also forming a unique big group with some internal subdivision per country.

3.3 Individual ancestry components

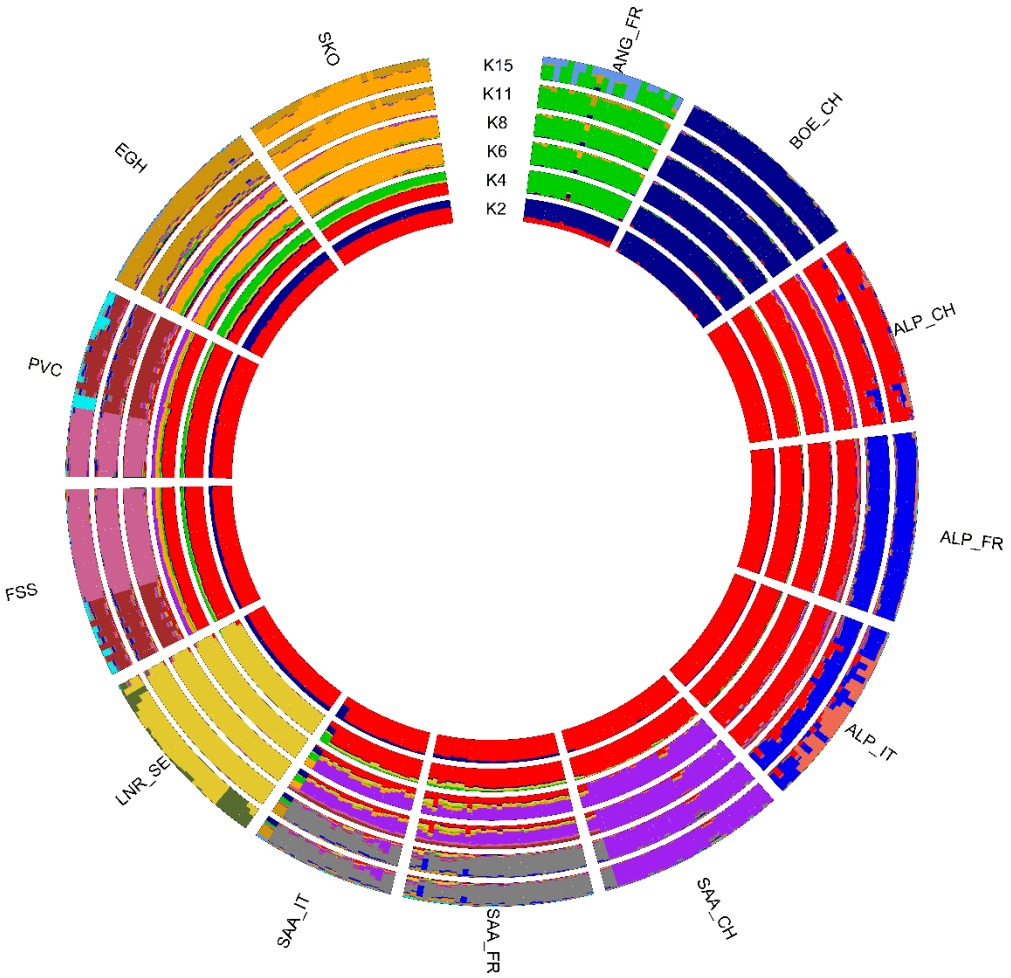

Figure 5 ADMIXTURE analysis for the goat subset. The value of K ranged from 2 to 15, with the most probable K = 11 (as detected by cross-validation method).

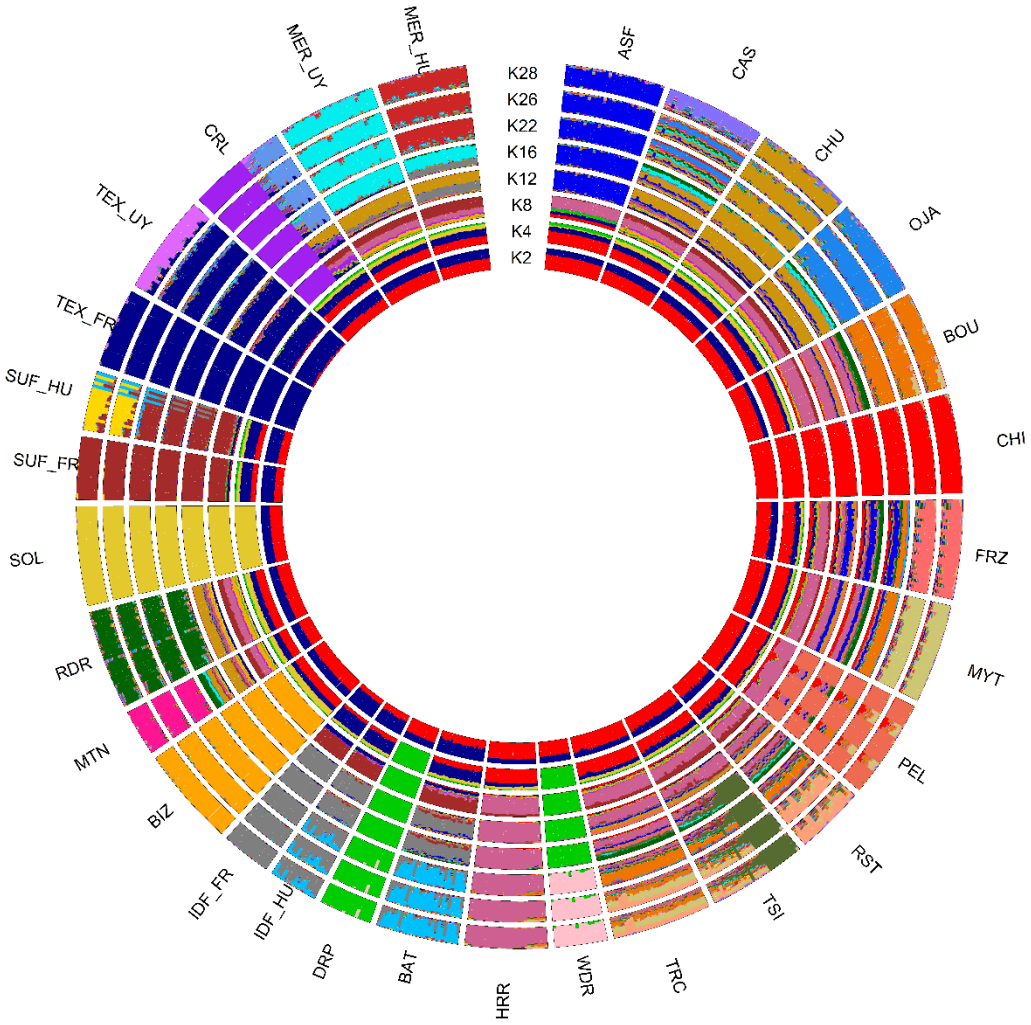

Figure 6 ADMIXTURE analysis for the sheep subset. The value ok K ranged from 2 to 31, with the most probable K = 28 (as detected by cross-validation method).

For the analysis of genetic components in the goat subgroup, we can observe a greater contribution from Alpine breed to the French Fossée, while the genetic background of the Swedish landrace suggests a different origin. The Greek breeds are distinct from each other but with some introgression from Eghoria to Skopelos population, as already detailed in Michailidou et al (2019). For the sheep subgroup, the ADMIXTURE analysis supports the previous relationship among groups already detected by the Network graph. In particular, we observe a common genetic component between the Spanish breeds and the Greek one together with some Hungarian breeds until K=12. In Frizarta we retrieve a background similar to the other Greek breeds but with a contribution from Assaf (linked to the origin of this crossbred population). Some French breeds look well genetically characterised (*e.g.* Solognote and Bizet) and the Uruguayan Creole clearly divided in two subpopulations, of which the smallest one showed an introgression with the Spanish Ojalada.

3.4. Analysis of demography through the Runs Of Homozygosity (ROH)

A general overview on the demography can be done through the detection, distribution and study of homozygous long stretches. As we can see from the distribution of ROH per class of length (Figure 7), the goat breeds Provencale and the Greek Eghoria have the longest segments (>16Mb), which suggests more likely to be the result of recent inbreeding. This last finding is supported by previous results (Michailidou et al 2019), where the Eghoria breed is more inbred in comparison with Skopelos.

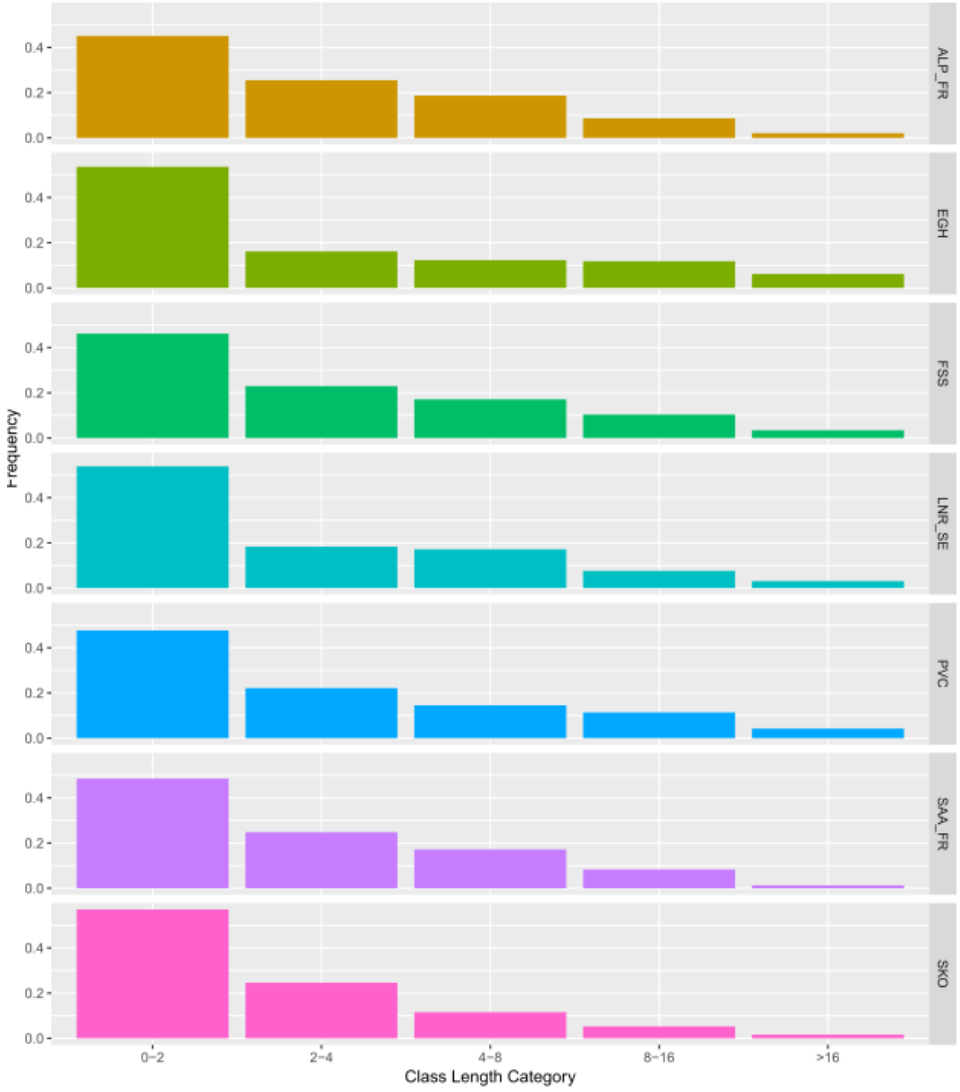

Figure 7 Whole genome distribution of Runs of Homozygosity for the goat subset (foreground data). We included here the French Alpine and Saanen as comparison.

Looking at the sheep subset, Boutsko (Greece), Sufflok (Hungary) and Tsigai (Hungary) are the breeds with a slightly highest number of ROH segments that fall in the last class of length (Figure 8). All the breeds showed an increased number of homozygous stretches that are included in the 2-4 Mb-class.

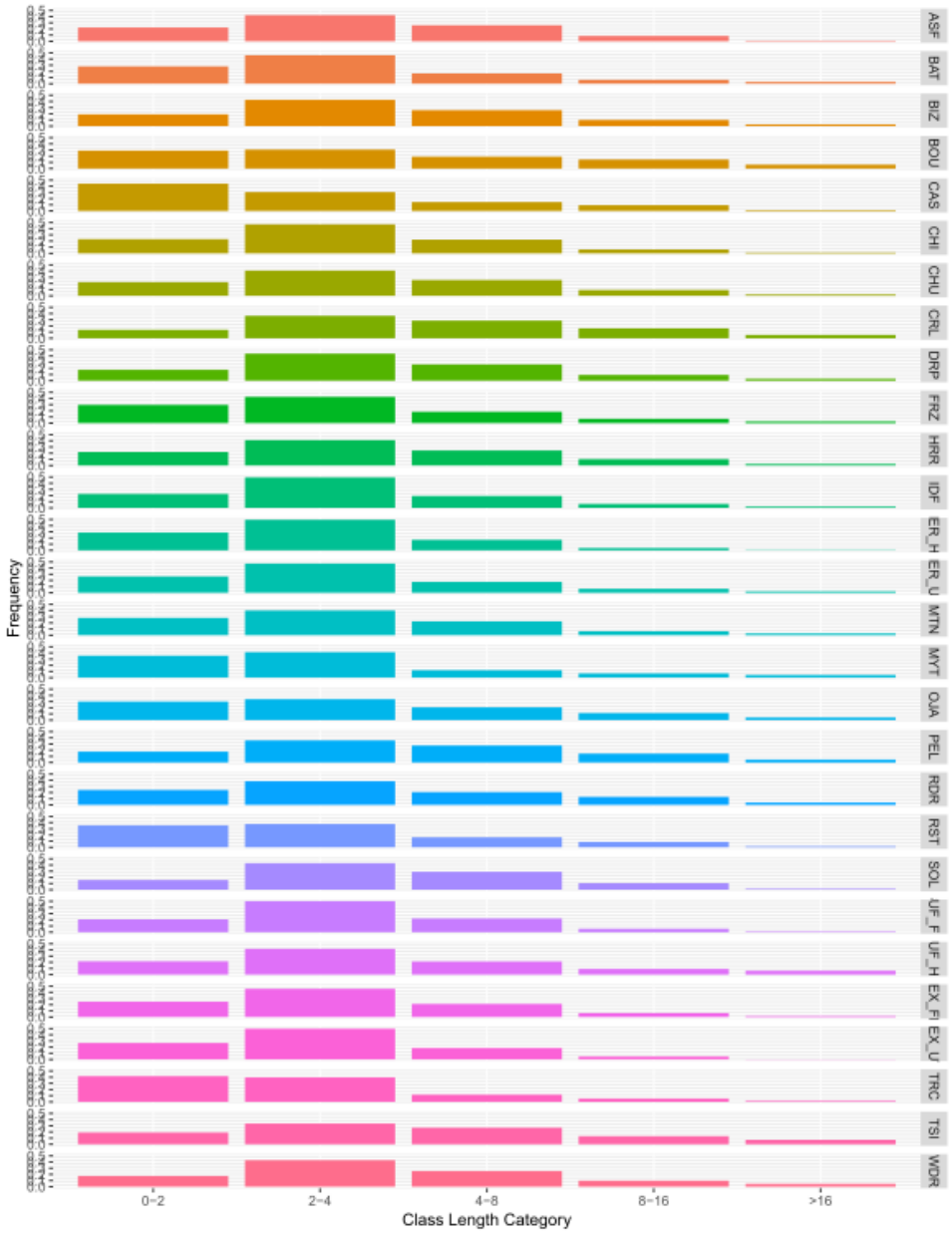

Figure 8 Whole genome distribution of Runs of Homozygosity for the sheep subset (foreground data).

4 Landscape genomics for local adaptation in underutilised sheep breeds

Small ruminants are an important source of livelihood for thousands of rural communities worldwide, playing a key role in local economies of both developing countries and the Western world. Local sheep breeds are ecoregional and genetically diverse (Wanjala et al., 2023) because, over time, natural and artificial selection has shaped their genome with morphological (Whannou et al 2021), behavioural (Dwyer and Lawrence, 2005; McManus et al., 2020) and physiological (Collier et al., 2019) changes to suit different breeding objectives. Therefore, the small ruminants provide a valuable model for identifying the genetic pathways and mechanisms that drive adaptations. Landscape genomics can help to understand the relationship between genetic architecture and environmental variables, as well as provide information on the evolutionary history of a species at different spatial scales. Thus, we aimed to investigate regions of the genome that may be associated with environmental adaptation.

- Genotype Dataset

The dataset included samples from sheep creole populations from Europe (n= 111), Asian (n=58), Africa (n=12) and America (n=8)) and belonging to the SMARTER project (Table 1), and populations from Brazil (n=1) and Colombia (n=8) from a private dataset. A total of 6519 animals from 189 populations were included in the analysis. The populations were genotyped with the OvineSNP50, Ovine HD BeadChip (Illumina Inc., USA) or Affymetrix Axiom. A common map was created using the Ovine SNP50 BeadChip coordinates of SNPs on the OAR v3.1 reference genome assembly. After the merge with the SNP50 data, we obtained 40,455 genotypes. Genotype quality control was performed using PLINK v1.9 and pruned for Linkage Disequilibrium (indep-pairwise: 50, 5, 0.2) and following filtering thresholds: (i) SNP call rate ≤ 95%; (ii) SNP minor allele frequency (MAF) ≤ 1%; (iii) animals displaying ≥ 5% of missing genotypes.

Continent	**Country**	**Breeds**	**Animals**
Africa	Algeria	Barbarine, Hamra, Tazegzawth	17
	Central African Republic	Sidaoun	14
	Congo	Sidaoun	6
	Ethiopia	EthiopianMenz	34
	Kenya	Hamra, Red Maasai	45
	South Africa	NamaquaAfrikaner, RonderibAfrikaner	29
	South Sudan	Sidaoun	13

	Uganda	Hamra	10
America	Barbados	Barbados BlackBelly	24
	Brazil	Morada Nova, Santa Ines	69
	Colombia	Brazilian Creole	23
	Jamaica	StElizabeth	10
	United States	GulfCoastNative	94
	Uruguay	Creole	203
Asia	Bangladesh	BangladeshiBGE, BangladeshiGarole	48
	China	Argali, Lop, Kirghiz, Celei black, Diqing, Guangling fat-tail, Guide Black Fur, Hanzhong, Hulun Buir, Jingzhong, Lanping Black-bone, Lanzhou Large-tailed, Luzhong Mountain, Minxian Black Fur, Ninglang Black, Shiping Gray, Taihang Fur, Tan, Tashkurgan, Tengchong, Tong, Turfan Black, Ujimqin, Wuranke, Yuxi Fat-tailed, Zhaotong, Bashbay, Sunite, Changthangi, Bayinbuluke, Altay, Duolang, Kazakh, Sishui Fur, Small Tailed Han, Large Tailed Han, Hu, Wadi, Tibetan	2093
	India	Deccani, indianGarole	50
	Indonesia	Garut, Sumatra	46
	Iran	Iranian sheep, Iranian mouflon, Lori-Bakhtiari, Zel, Moghani, Afshari	141
	Kazakhstan	Karakul	6

Europe	Albania	Recka, Lara, Shkodrane, Ruda	37
	Bosnia and Herzegovina	Dubska, Privorska	6
	Croatia	Dalmatian, Lika, Croatian Isles, Istrian	49
	Cyprus	CyprusFatTail	30
	Czechia	Valachian	10
	Finland	Finnsheep	154
	France	Romanov, Tarasconnaise, Corse, Mourerous, Pr.alpes du Sud, Mourerous, Rouge de l'Ouest, Limousine, Merinos d'Arles, Ouessant, Berrichon du Cher, Noire du Velay, Blanc du Massif Central, Causses du Lot, Rava, Roussin de la Hague, Mouton Venden, Ãle de France, Mouton Charollaise, Manech Tete Rouge, Merinos de Rambouillet, Charmoise	408
	Germany	Bentheimer, EastFriesianWhite, EastFriesianBrown	53
	Greece	Kymi, Lesvos, Pelagonia, Mytilini, Chios, Boutsko, Frizarta	1040
	Hungary	Racka, Bbolna Tetra, Hortobegy Racka, Tsigai	122
	Iceland	Icelandic	54
	Ireland	Galway	49

	Italy	Sardinian Ancestral Black, Biellese, Sambucana, Bagnolese, Fabrianese, Alpagota, Altamurana, Appenninica, Bergamasca, Comisana, Delle Langhe, Gentile di Puglia, Laticauda, Leccese, Massese, Pinzirita, Sardinian White, Sopravissana, Valle del Belice	445
	Montenegro	Zuja, Oivska, Sora	17
	Netherlands	Schoonebeker, Drenthe Heath, Veluwe Heath	14
	North Macedonia	Karakachanska, Ovchepolean	16
	Norway	Old Norwegian spaelsau, Norwegian White, Spael-white	71
	Poland	Kamieniec, Polish Mountain	11
	Romania	Rusty Tsigai, Tsigai, Turcana	65
	Russian Federation	Romanov	80
	Serbia	Lipska	7
	Slovenia	Jezersko-SolÄava	5
	Spain	Segurena, Merino Estremadura, Ripollesa, Aragonesa, Castellana, Latxa, Ojalada, Sasi-Ardi, Xisqueta, Churra	307
	Switzerland	Bundner Oberlander Sheep, Engadine Red Sheep, Swiss Mirror Sheep, Valais Blacknose Sheep, Valais Red Sheep	120

	Turkey	Karakas, Norduz, Sakiz, Qezel	95
	United Kingdom	Boreray, Wiltshire, BorderLeicester, IrishSuffolk, Soay	253

- Bioclimatic information

Bioclimatic variables were obtained from the WordClim database with a spatial resolution of 30 seconds using the GPS coordinates for each population and smarterapi R package version 0.1.2 (Cozzi 2022). Bioclimatic variables represent trends data, seasonality and extreme or limiting environmental factors (Table 2).

Table 2. Nineteen bioclimatic variables derived from WordClim database at http://www.worldclim.org/bioclim.

Variable	**Definition**
Bio1	Annual mean temperature
Bio2	Mean Diurnal Range (Mean of monthly – max temp - min temp)
Bio3	Isothermality (bio2/bio7) x 100
Bio4	Temperature Seasonality (SD x 100)
Bio5	Max Temperature of Warmest Month
Bio6	Min Temperature of Coldest Month
Bio7	Temperature Annual Range (bio5-bio6)
Bio8	Mean temperature of Wettest Quarter
Bio9	Mean temperature of Driest Quarter
Bio10	Mean temperature of Warmest Quarter
Bio11	Mean temperature of Coldest Quarter
Bio12	Annual Precipitation
Bio13	Precipitation of Wettest Month
Bio14	Precipitation of Driest Month
Bio15	Precipitation seasonality (Coefficient of Variation)
Bio16	Precipitation of Wettest Quarter
Bio17	Precipitation of Driest Quarter
Bio18	Precipitation of Warmest Quarter
Bio19	Precipitation of Coldest Quarter

- Principal components analysis and LFFM algorithm

Principal Component Analysis was conducted to evaluate the genetic structure and reduce the risk of false positive detection. The numbers of PCA that adequately describe the dataset are included as population structure predictors for the association analysis. The PCA analysis was

divided in two subsets: 1) On the entire dataset, selecting 10-20 randomly animals by population in order to limit the size of bigger groups and prevent biassed estimations and 2) data subsets separated by Continent. As a first approach, we separated a dataset that included some populations of Colombia, Brazil, Spain and Africa and applied the Latent Factor Mixed Models (LFFM) to determine SNPs significantly associated with the geographic and environmental variables. The FDR - q value was calculated for each locus based on the p-values in R. We identified genes related to oxidative stress (CAT, BLF), thermotolerance (FGF2, GNAI3, PLCB1) and altitude (PPP1R12A, RELN, PARP2), showing that some environmental and geographic selection pressures drive evolution and local adaptation.

5. Imputation of missing SNP genotypes

Every time SNP genotype data are obtained, be it from sequencing (e.g. whole-genome resequencing, RAD-sequencing/GBS etc.) or array-based genotyping technologies, a fraction of SNP remains uncalled, generating missing SNP-genotype data. Many types of analysis in statistical genetics can not handle missing data points, and require a complete dataset: therefore, the imputation of missing SNP data is necessary. Also, in many circumstances SNP genotype data with different densities are available for different animals (e.g. low- and high-density SNP array data), and a common approach is to impute the low-density data to the high-density array. Given the pervasive presence of the imputation of missing SNP data, it is relevant to measure the accuracy of such imputation. As a matter of fact, imputation accuracy can vary widely depending on the breed/population, the data size and the imputation scenario.

We used 50k SNP array data from 6-7 goat breeds (Alpine, Angora, Boer, Barki, Creole, Landrace, Saanen) to simulate two imputation scenarios:

1. gap-filling: impute residual missing SNP genotypes in the 50k SNP array data; to measure the accuracy of imputation, artificially missing SNP genotypes (SNP genotypes that we know, but pretend we don't and set them to missing) were injected in the dataset: 1%, 5%, 10%. Missing SNP genotypes were then imputed with different sample sizes: 100, 80, 60, 40, 20
2. low-to-high density imputation scenario: from 50k SNP data, we subsampled 10, 20, 30 and 40 animals and pretended they had only low density SNP data. We then imputed back to 50k data (higher density) and measured the accuracy of imputation. Different total sample sizes were tested: 100, 80, 60. The low-density SNP array contained 15k SNPs and was subsampled randomly from the 50k SNP array.

In all cases, the accuracy of imputation was measured with Cohen's kappa coefficient of concordance between imputed and known SNP genotypes. Figure 9 and Figure 10 show the results of the imputation experiments. Each experiment (combination of scenario, sample size, missing rate) was run 10 times, and averages are presented (solid lines) plus boxplots of distributions of all accuracy values. As for the gap-filling scenario, we see that imputation accuracy varies greatly in the different goat breeds. In general, imputation accuracy decreases with decreasing sample sizes, while the missing rate does not seem to have too much influence (at least in the interval tested: 1% - 10%). However, while in the Boer, Landrace, Angora and Alpine breeds the accuracy was high for most sample sizes tested, degrading only where 40 samples or fewer were used; for Creole and especially for the Barki breed, imputation accuracy was always very low.

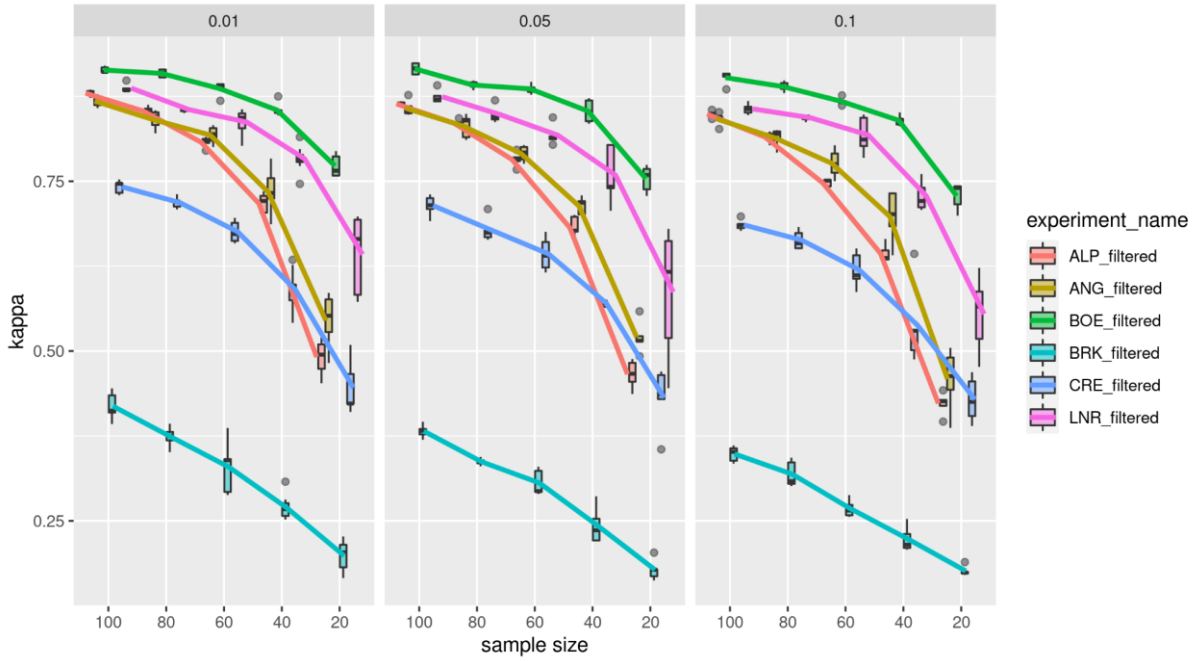

Figure 9. Gap-imputation scenario. Imputation accuracy with decreasing sample size (x axis) and increasing missing rate (panels with 1%, 5% and 10% missing SNP data), in different goat breeds. Imputation accuracy is measured as the Cohen kappa coefficient of concordance (y axis) between known and imputed SNP genotypes. ALP: Alpine; ANG: Angora; BOE: Boer; BRK: Barki; CRE: Creole; LNR: Landrace.

Similar results were observed for the low-to-high imputation scenario, although with generally lower imputation accuracies. This came as no surprise, since imputing from low to high density is a much harder exercise compared to imputing residual missing SNP genotypes ("filling the gaps"). The general trend showed reduced imputation accuracy when a larger fraction of animals with low density SNP array was used (e.g. 40 low density samples out of 100/80/60). The overall sample size also had an impact, with lower average imputation accuracy when 60 samples were used rather than 100 or 80. Again, and more markedly, we observed huge variation between goat breeds: The accuracy of imputing from low to high density was reasonably high in Landrace, Boer, Angora and Alpine goats, much worse in Saanen and Creole goats. In Barki goats the accuracy of imputation was practically zero in all tested scenarios.

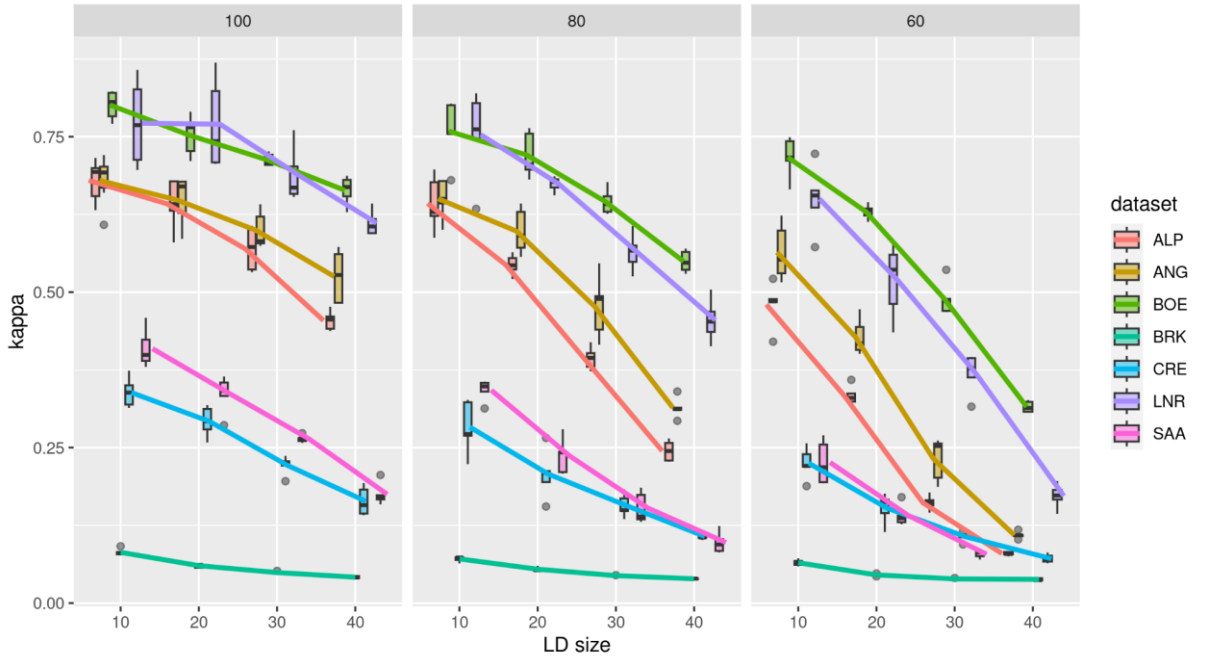

Figure 10. Low-to-high density imputation scenario. Imputation accuracy with increasing fractions of samples with low-density SNP data (x axis) to be imputed to high(er) density SNP data. Different total sample sizes were tested (panels with 100, 80 and 60 samples). Imputation accuracy is measured as the Cohen kappa coefficient of concordance (y axis) between known and imputed SNP genotypes. ALP: Alpine; ANG: Angora; BOE: Boer; BRK: Barki; CRE: Creole; LNR: Landrace; SAA: Saanen.

6. Runs Of Homozygosity and Heterozygosity-Rich Regions: two case study for their detection

6.1 Distribution of heterozygosity-rich regions (HRR) in the genome of local vs commercial goat breeds

Heterozygosity-rich regions (HRRs) are regions of unusually high heterozygosity in the genome of diploid organisms (Williams et al. 2016) and are still largely uncharacterized in animals. In the present study we looked for HRRs in the goat genome, specifically looking for common HRRs across commercial and local breeds. We used SNP genotype data from widely distributed commercial breeds ("commercial") and locally adapted breeds ("local"): the Barki (BRK), Creole (CRE) and Landrace (LNR) (local), and the Alpine (ALP), Boer (BOE) and Saanen (SAA) (commercial) breeds, for a total of 1072 goats. SNPs were obtained from the GoatSNP50 BeadChip (Illumina Inc., San Diego, CA) which includes 53,347 SNPs (mapped on ARS1.2). Only SNP loci on the goat autosomes (chromosomes 1 - 29) were used, filtered for MAF > 0.05 and call-rate (> 95% locus-wise, > 80% sample-wise). After filtering, 1051 goats and 47,689 SNPs were left for the analysis. Plink v.1.9 was used for data management and filtering (Chang et al. 2015). HRR were detected by scanning the ordered sequence of SNP

loci with the window-free method described by Marras et al. (2015). The following parameters were used for the detection of HRR: minimum 15 SNPs and minimum length of 250 kbps to define a HRR; maximum gap 1e+03 kbps between adjacent SNPs in a HRR; maximum 3 homozygous SNPs and 2 missing SNPs in a HRR. The R package detectRUNS was used for the detection of HRR (https://cran.r-project.org/web/packages/detectRUNS/). HRR islands have been defined as HRR found (identical) in at least 20% of the samples (within breed).

A total of 73,173 HRRs were detected in the 1,051 goats remaining after filtering (69.6 HRRs per goat, on average). More HRRs have been detected in commercial breeds than in local breeds, both overall (50,477 vs 22696) and per goat (83.3 vs 51). The average length of detected HRRs was 805.7 kbps in commercial breeds and 794.5 kbps in local breeds. Looking at HRRs shared by individual goats, 79 HRR islands (HRRs in >= 20% of the samples) have been identified. Figure 11 shows HRR islands in chromosomes harbouring more than one island. Common HRR islands are found on chromosomes 1, 10 and 14 (across 3 breeds), chromosomes 3, 13 and 18 (across 4 breeds), and on chromosome 12 (across 5 breeds). The HRR islands most shared across breeds is the one found on chromosome 12 at 49.884 - 51.611 Mbps. Figure 12 shows a visualisation of this common HRR island across samples (B, right), and the corresponding peak of SNP loci found to be lying inside it (A, left).

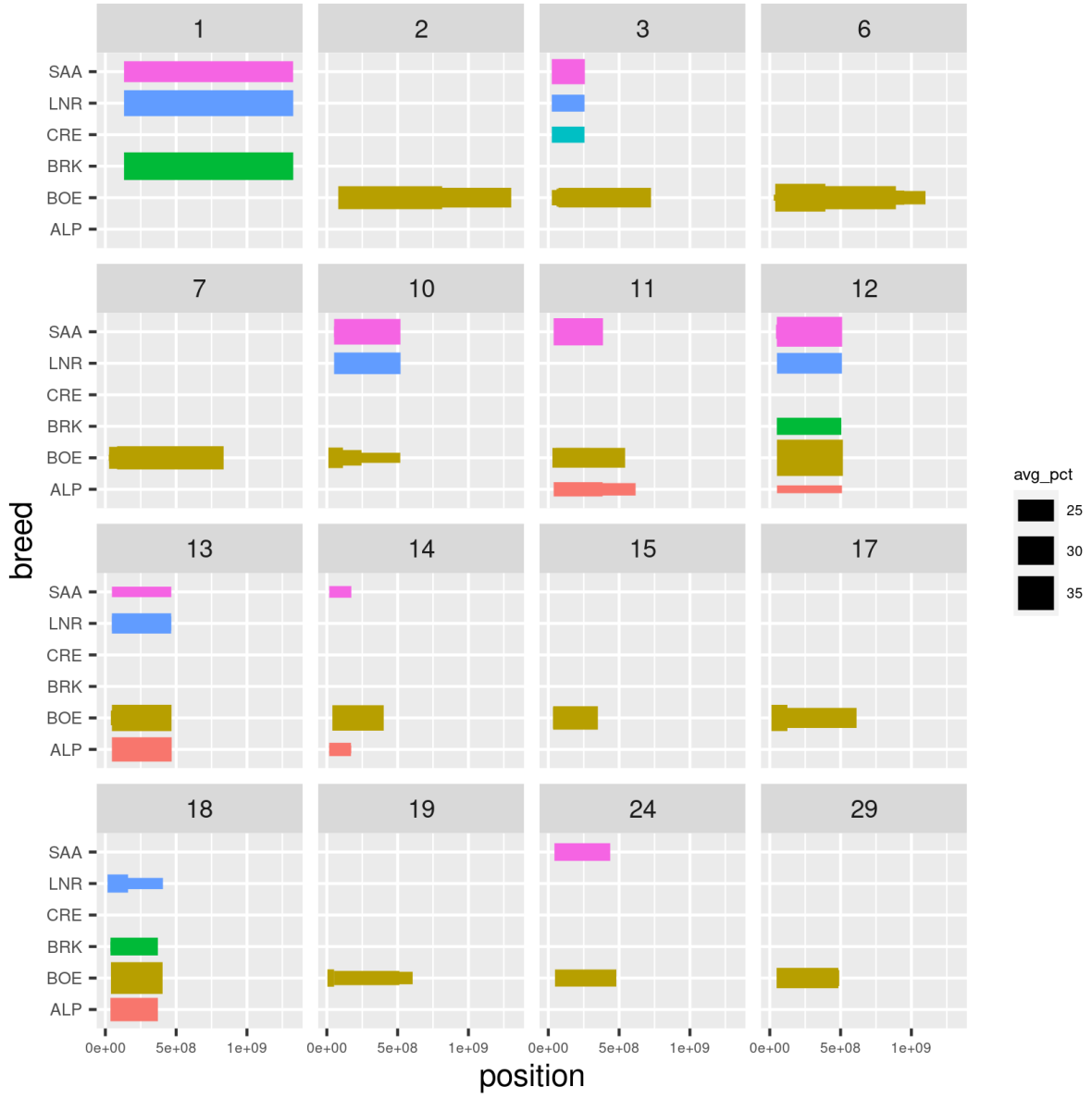

Figure 11. Visualisation of HRR islands detected in the genome of commercial (ALP, BOE, SAA) and local (BRK, CRE, LNR) goat breeds. Chromosomes with only one HRR have been excluded. The thickness of the HRR is proportional to the percentage of samples in which it has been detected. The length of the HRRs has been artificially extended (multiplied by 10) for the sake of visualisation.

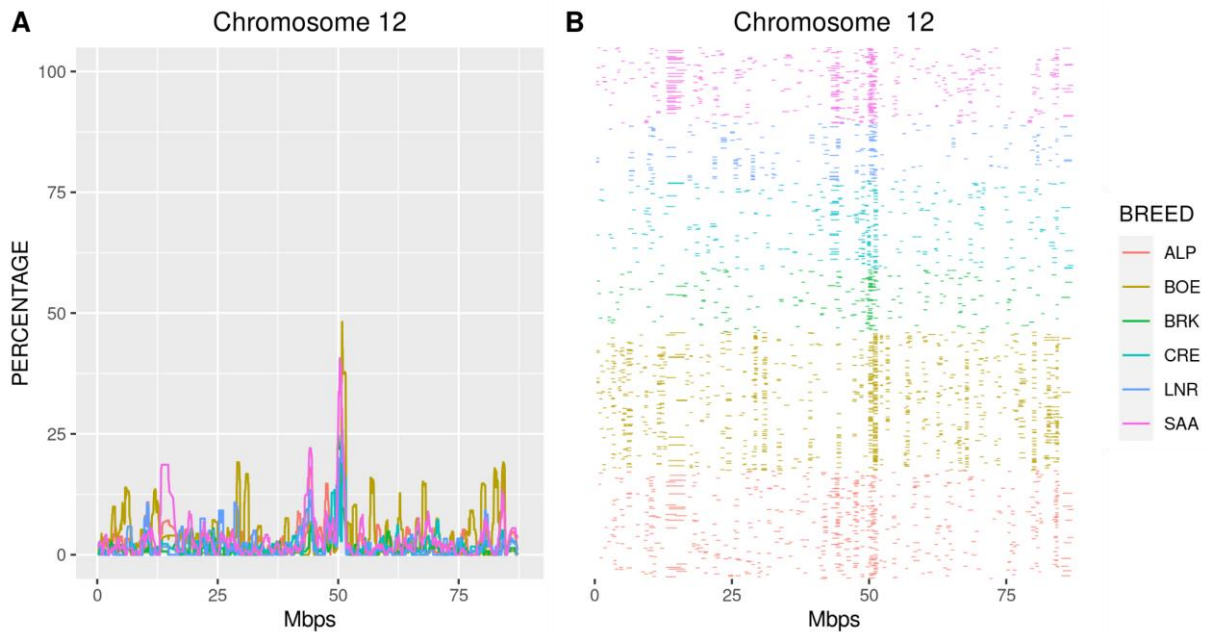

Figure 12. HRR island on goat chromosome 12. A) proportion of times SNP loci fall within HRRs in the analysed goat samples (across breeds); B) distribution of HRRs on chromosome 12 in goat samples (a clear signature of heterozygosity is visible at 49.9-51.6 Mbps).

6.2 Choosing parameters for the detection of ROH and HRR in the genomes of sheep and goats

To detect heterozygosity-rich regions (HRRs) in the genome, several parameters are used, e.g. the minimum length of the HRR, the minimum number of SNPs inside the HRR, the maximum gap between adjacent SNP loci, the maximum number of missing and homozygous SNP loci allowed within the HRR. It is however unclear how sensitive results are to the specific values of the detection parameters. We used 50k SNP array data from sheep (Lacaune) and goats (Saanen) recruited for the project, and from publicly available cow (Holstein) data (Gautier et al. 2012). We selected only autosomes, and filtered SNP data for MAF > 0.05, call-rate > 95% (locus) and > 90% (sample). For the detection of HRR we used the window-free method (Marras et al. 2015) implemented in detectRUNS (Biscarini et al. 2018). Around a base scenario (min. 15 SNP, min. length 250 kb, max gap 10^3 kb, max 3 homozygous SNP, max 2 missing SNP) we tweaked parameters and looked at the average n. of HRR detected, their average length, minimum n. of SNP within HRR, n. of samples with at least one HRR).

Figures 13-16 show the results. The parameters minimum number of SNP required and maximum number of homozygous SNP allowed (Figures 13 and 15) are those with the largest impact on the detection of HRR in all three species: in particular, when these parameters are too stringent (more than 20 SNP required and 0 homozygous SNP allowed) we observed a dramatic decline in the number of HRR detected. The other two parameters, minimum length of HRR required and maximum number of missing SNP allowed, appeared to have little effect on results, at least in the tested ranges. Interestingly, the only counterintuitive result was observed in sheep for the maximum number of homozygous SNP allowed in the HRR. In goats

and cows, as more homozygous SNP are allowed, more numerous and longer HRR are detected. In sheep, we do detect more HRR, as expected, but these are initially long, then get shorter with a minimum at 3 homozygous SNP, to then reprise slightly for 4 and 5 homozygous SNP. This may be linked to a different distribution of heterozygosity in Lacaune sheep compared to Saanen goats and Holstein cows.

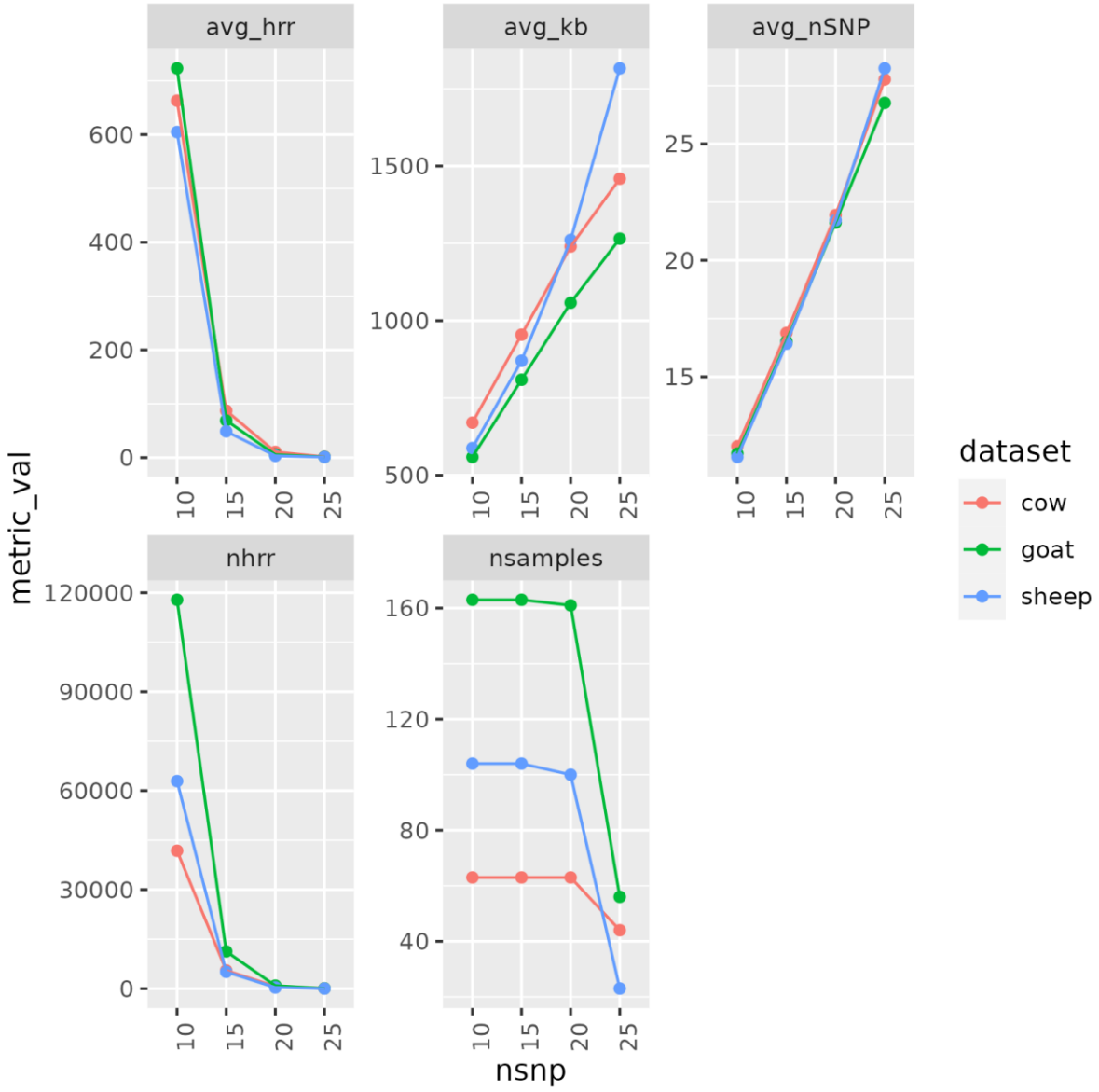

Figure 13. Detection of HRR for different values (10-25) for minimum number of SNP required. avg_hrr: average number of HRR/animal; avg_kb: average length of HRR; avg_nSNP: average number of SNP within HRR; nhrr: total number of HRR detected; samples: number of samples with at least one HRR.

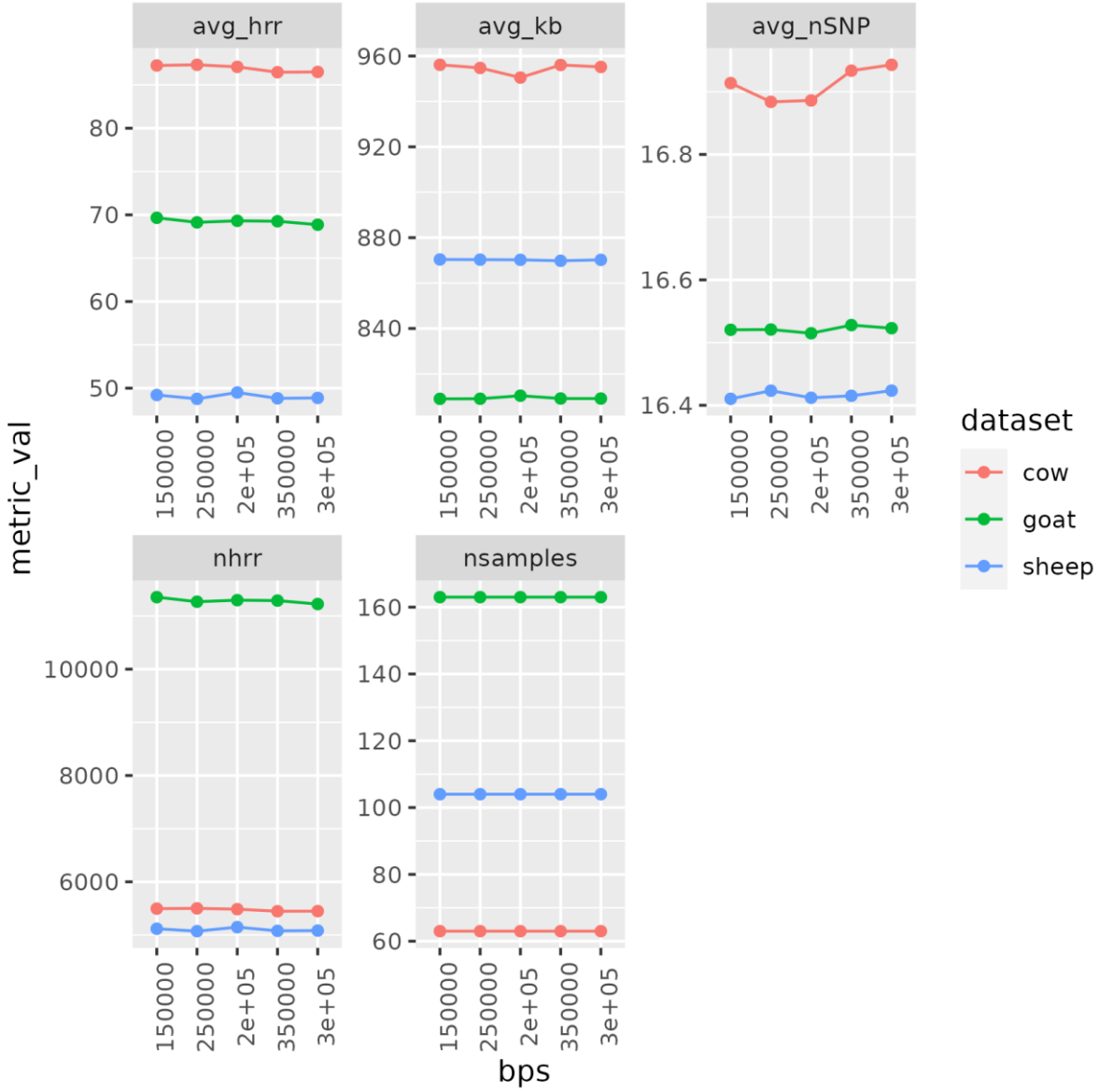

Figure 14. detection of HRR for different values (150-350 kb) for minimum length of HRR required. avg_hrr: average number of HRR/animal; avg_kb: average length of HRR; avg_nSNP: average number of SNP within HRR; nhrr: total number of HRR detected; samples: number of samples with at least one HRR.

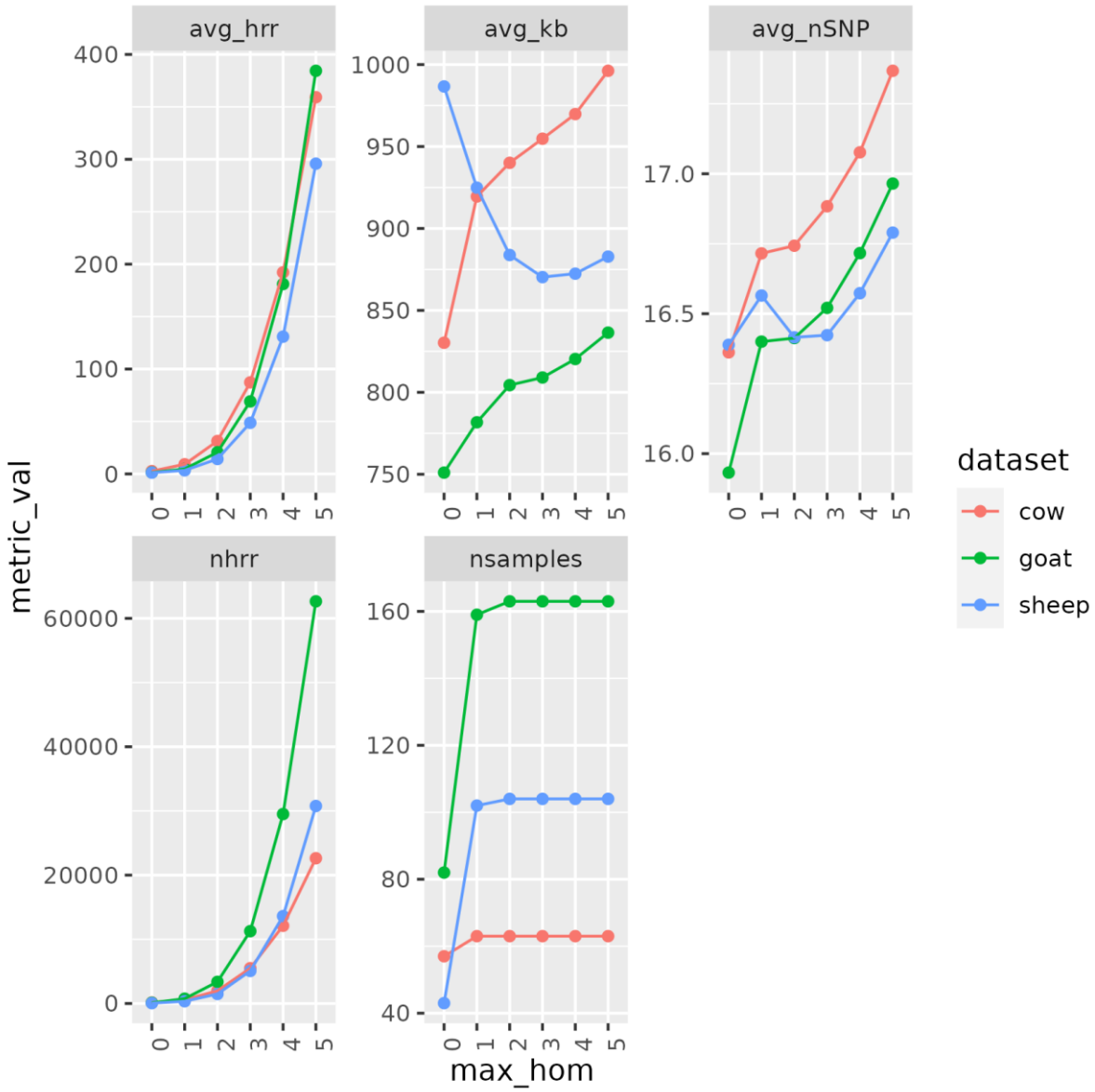

Figure 15. Detection of HRR for different values (0-5) for the maximum number of homozygous SNP allowed in HRRs. avg_hrr: average number of HRR/animal; avg_kb: average length of HRR; avg_nSNP: average number of SNP within HRR; nhrr: total number of HRR detected; samples: number of samples with at least one HRR.

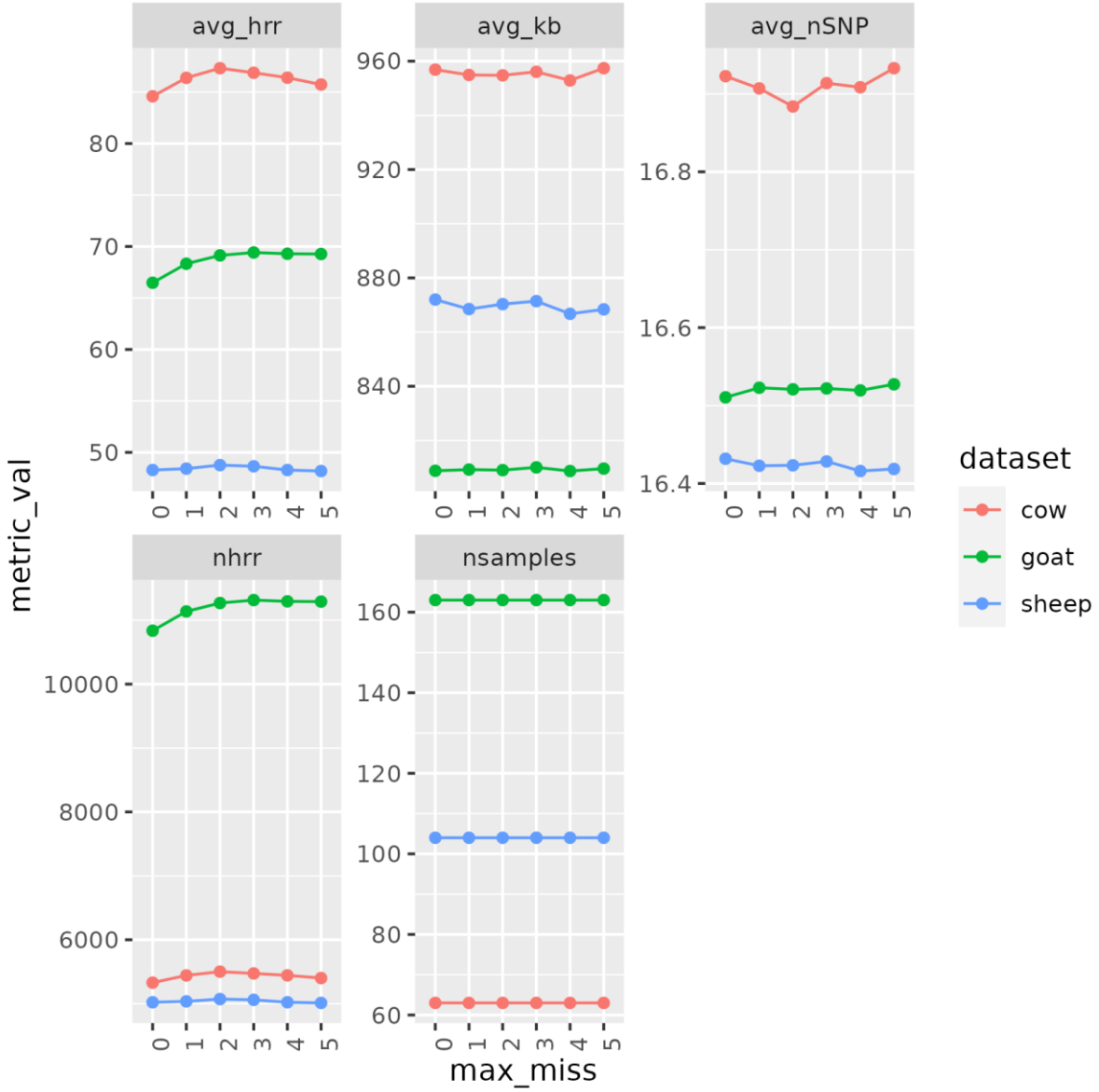

Figure 16: Detection of HRR for different values (0-5) for the maximum number of missing SNP allowed in HRRs. avg_hrr: average number of HRR/animal; avg_kb: average length of HRR; avg_nSNP: average number of SNP within HRR; nhrr: total number of HRR detected; samples: number of samples with at least one HRR.6. Conclusion

Exploring the demographic history of breeds is relevant to understand their current adaptation potential. So, investigating the demographic history, the genetic diversity and population structure is key, since they are keenly intertwined to the process of adaptation. We observed a common genetic component between the Spanish breeds and the Greek one together with some Hungarian sheep breeds. We retrieved a similar background in the Greek sheep breeds with some introgression in Frizarta from Assaf, supported by all the analyses and in agreement with the known origin of this breed. Some French breeds look well genetically distinct, and the Uruguayan Creole showed some introgression from the Spanish Ojalada. For the new genotyped local goat breeds we detected the influence of Alpine breed and other transboundary, even if much less in the Swedish goats. With these results we acquired more information about local and traditional breeds belonging to part of Europe

and the world still poorly unexplored, with a greater contribution for sheep. The results from the imputation experiments highlighted the importance of measuring (and being aware of) the accuracy of imputation, given that this varies wildly between scenarios (e.g. gap filling or low-to-high density SNP data) and between breeds. The analysis of HHRs showed the importance of this tool to estimate levels of heterozygosity and to identify genomic regions where genetic variability is conserved: choosing the right parameters for the detection of HHR is critical, since different values can sometimes lead to very different results.

7. Deviations or delays

A slight delay in the provision and uploading of the last data for foreground population was reflected in a few days delay in submission of the deliverable.

8. References

Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

Biscarini, F., Cozzi, P., Gaspa, G. & Marras, G. DetectRUNS: detect runs of homozygosity and runs of heterozygosity in diploid genomes. R package Version 0.9.5. (2018).

Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

Collier R.J., Baumgard L.H., Zimbelman R.B., Xiao Y. 2019. Heat stress: physiology of acclimation and adaptation. Anim. Front., 9-1: 12-19.

Cozzi P (2022). smarterapi: Fetch SMARTER data through REST API. R package version 0.1.2, https://cnr-ibba.github.io/r-smarter-api/.

Dwyer C.M., Lawrence A.B. 2005. A review of the behavioural and physiological adaptations of hill and lowland breeds of sheep that favour lamb survival. Appl. Anim. Behav. Sci., 92-3:235-260

Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).

Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).

Lischer, H. E. L. & Excoffier, L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299 (2012).

McManus C.M., Faria D.A., Lucci C.M., Louvandini H., Pereira S.A., Paiva S.R. 2020. Heat stress effects on sheep: Are hair sheep more heat resistant? Theriogenology, 155 (2020), pp. 157-167.

Michailidou S, Tsangaris GT, Tzora A, Skoufos I, Banos G, et al. (2019) Analysis of genome-wide DNA arrays reveals the genomic population structure and diversity in autochthonous Greek goat breeds. PLOS ONE 14(12): e0226179. https://doi.org/10.1371/journal.pone.0226179

Milanesi, M., Capomaccio, S., Vajana, E., Bomba, L., Garcia, J. F., Ajmone Marsan, P., & C. BITE: An R Package for Biodiversity Analyses. *bioRxiv*, doi:https://doi.org/10.1101/181610

Wanjala G., Kusuma A.P., Bagi Z., Kichamu N., Strausz P., Kusza S. 2022. A review on the potential effects of environmental and economic factors on sheep genetic diversity: Consequences of climate change. Saudi Journal of Biological Sciences, 30-1:1-10. DOI https://doi.org/10.1016/j.sjbs.2022.103505

Whannou H.R.V., Afatondji C.U., Ahozonlin M.C., Spanoghe M., Lanterbecq D., Demblon D., et al. 2021 Morphological variability within the indigenous sheep population of Benin. PLoS ONE 16-10: e0258761. https://doi.org/10.1371/journal.pone.0258761

Williams, J. L., Hall, S. J., Del Corvo, M., Ballingall, K. T., Colli, L. I.,Ajmone Marsan, P. A., & Biscarini, F. (2016). Inbreeding andpurging at the genomic level: The Chillingham cattle reveal extensive, non-random SNP heterozygosity. Animal Genetics, 47,19–27. https://doi.org/10.1111/age.12376

Zheng, X. *et al.* A High-Performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics* **28**, 3326–3328 (2012).