Effect of subdivision of the Lacaune dairy sheep breed on the accuracy of genomic prediction

M. Wicki, J. Raoul, A. Legarra

J. Dairy Sci. 106:5570–5581
https://doi.org/10.3168/jds.2022-23114

Effect of subdivision of the Lacaune dairy sheep breed on the accuracy of genomic prediction

M. Wicki,[1,2]* J. Raoul,[1,2] and A. Legarra[1]†
[1]INRAE, INP, UMR 1388 GenPhySE, F-31326 Castanet-Tolosan, France
[2]Institut de l'Elevage, Castanet-Tolosan 31321, France

ABSTRACT

Genomic selection was deployed in Lacaune dairy breed in 2015. Lacaune population split in 1972 into 2 breeding companies with associated flocks, and there have been very few exchanges of animals between the subpopulations, leading to divergence of the 2 subpopulations. In spite of that, there is a joint genomic prediction. The objective of this study is to understand how this structuring affects prediction accuracy. We analyzed all the data available from Lacaune breeding program for milk yield: around 6 million phenotypes, 2 million animals in the pedigree and more than 29,000 genotyped animals, including 3,434 and 2,868 AI rams for each company. To consider missing pedigree, we set up genetic groups using the theory of metafounders. First, we studied the pedigree and genomic structures of the 2 subpopulations calculating F_{st}, evolution of average pedigree relationships across time and principal components analysis of genomic relationships. In a second part, we compared the reliability between different scenarios: an evaluation with a single reference population (Alone), an evaluation with a joint reference population (Together) and an evaluation of one subpopulation based on the reference population of the other group (Indirect). The low F_{st} value (0.02) reveals that the 2 subpopulations are still genetically close. Nevertheless, a low and constant average relationship between the animals of the 2 subpopulations confirms the absence of recent connections between them. We can see with principal component analysis results that even if they are close, they diverge over time. Finally, we observe small gains in accuracy of Together versus Alone, in spite of whereas doubling the reference population size in Together. These gains vary across years and subpopulations: less than 0.08 (0.46 to 0.54; ratio of accuracy for the partial and whole evaluations—corresponding to the greatest change in this ratio for breeding company 1, observed for the cohort 2016) for one subpopulation and between 0.03 (0.55 to 0.58) and 0.17 (0.48 to 0.65) for the other. To conclude, the 2 subpopulations remain close enough genetically so that their combined evaluation is advantageous, even if only slightly.

Key words: genomic prediction, accuracy, population structure, reference population

INTRODUCTION

Genomic evaluation aims at predicting the genetic value of an individual from phenotypic data, pedigree, and genotypes at SNP markers. These genomic predictions are used in animal breeding to classify the animals and manage matings to create genetic gain, under the concept of genomic selection.

Genomic selection in French Lacaune dairy sheep started in 2015. Each year, 250 young AI rams are selected among ~2,000 genotyped prospective rams, based on their genomic EBV (**GEBV**) and used to inseminate females. Compared with pedigree-based selection, the increase in genetic gains from genomic selection has been estimated to be 57% (Astruc et al., 2022). The accuracy of milk yield (**MY**) breeding values of young genotyped rams (AI candidates) increased from 0.32 to 0.47 (i.e., a relative increase of 47%), when transitioned from pedigree-based to genomic-based selection (Baloche et al., 2014; Macedo et al., 2022). However, this genomic accuracy is lower than the genomic accuracy observed in some large dairy cattle breeds as Holstein (VanRaden et al., 2009), which indicates (perhaps, as Holstein and Lacaune are very different populations in numbers and structure) room for improvement. Improving the GEBV accuracy would result in an increase in the genetic gain provided by genomic selection.

Several factors have already been identified as influencing the reliability of the genomic prediction: the trait architecture [heritability of the trait, number and

Received December 5, 2022.
Accepted February 16, 2023.
*Corresponding author: marine.wicki@inrae.fr
†Current address: Council on Dairy Cattle Breeding, Bowie, MD 20716.

minor allele frequency of the QTL (Daetwyler et al., 2008), linkage disequilibrium between SNP markers and QTL (Habier et al., 2007), effective population size (Goddard, 2009), marker density, minor allele frequency for markers, number of independent chromosome segments]. In addition, the properties of the reference population, and in particular its size, influence the GEBV accuracy. It has been demonstrated that the GEBV accuracy is dependent on the number of phenotyped individuals (Liu et al., 2011). The relationships of animals within the reference population and the relationships between candidates to selection and training population also influence the GEBV accuracy. It has been shown that the accuracy decreases when the average relationship within the reference population increases (Pszczola et al., 2012). However, the more the selection candidates are related to the animals of the reference population, the more accurate their GEBV are (Habier et al., 2013).

Questions arise regarding the set up and the optimization of the reference populations. Some studies have focused on the interest of setting up exchanges between populations to increase the reference population size. They have shown that combining genetically close populations (same breed or related breeds) in the reference population increases the reliability (Brøndum et al., 2011; Lund et al., 2011) and the gain in accuracy increases with the relationship between populations (Zhou et al., 2014). However, other studies have shown no improvement in accuracy of the evaluation by combining multiple populations (Legarra et al., 2014). Combining data from distant populations, for example including crossbred animals, may even have a negative effect on accuracy (Moghaddar et al., 2014).

In French Lacaune dairy sheep, the breeding program is organized into 2 tiers: a number (roughly 450) of breeding flocks that create genetic progress (nucleus flocks) through performance recording and selection, and a larger number (roughly 1,500) of commercial flocks that only receive genetic material (mostly semen) from the nucleus flocks, but do not contribute genetic material to them. In turn, in 1972, nucleus flocks split into 2 nuclei, each managed by a breeding company (**BC**). The 2 BC are Confederation and Ovitest; their names are anonymized in the following. In fact, each BC manages its AI center, which collects rams from nucleus flocks and redistributes semen to "their" nucleus flocks and commercial flocks. Assignation of nucleus flocks to either of the BC is rigid. Thus, for the last decades, nucleus flocks have been contributing rams to a single BC and receiving semen from a single BC. Since this date, there were very few genetical exchanges between the 2 nuclei, although all animals benefit from a common genetic evaluation with a single reference population. In the following, we will use the wording "subpopulation" to indicate the set of animals belonging to nucleus flocks attached to a given BC.

Therefore, the 2 subpopulations, one per BC, have diverged over time, but the extent of this divergence and its effects on genomic predictions are not known. The objectives of our study are therefore (1) to assess the genetic and genomic structure of the Lacaune population based on pedigree and genomic data and (2) to study the accuracy and bias of GEBV according to different reference population compositions including one, the other, or both breeding subpopulations, focusing on the trait MY. Previous studies by Baloche et al. (2014) and Macedo et al. (2022) did not address these points.

MATERIALS AND METHODS

Records, Pedigree, and Genotypes

Milk recording was performed by the breeding scheme according to the International Committee for Animal Recording Rules.

We studied all the data available in Lacaune dairy sheep until 2021 for MY. Regular performance recording for MY started in 1978 with pedigree recording starting in the 1960s. The number of animals in pedigree, the number of records, the number of animals with records, and the percentage of animals with unknown parents are summarized in Table 1 considering each or both subpopulations. Considering both subpopulations, there are 1,974,901 animals in the pedigree with ~11% missing pedigree and around 1 million animals for each subpopulation. Note that there is overlap of pedigrees of each subpopulation, because of the common ancestors before the time of split which explains that the number of animals in the pedigree considering both subpopulations is less than the sum of the animals in pedigree for each subpopulation. The number of animals with records was 1,782,445 with 6,010,370 phenotypes for the complete data set and, again, is roughly split into 2 halves, one per subpopulation.

To account for the missing pedigree, we used the theory of metafounders (Legarra et al., 2015), which can be seen as a generalization of unknown-parent groups. The missing parents were assigned to metafounders according to the year of birth of the animal whose parent is unknown. At the beginning (beginning of the pedigree recording, before 1978) many animals have both parents unknown, but later (after 1978) pedigree recording was consolidated and most animals have only the sire unknown. The same metafounder was assigned to all unknown parents of animals born before 1978, moreover because performance recording for MY started in 1978. Then, from 1978 to 2018, the metafounders

Table 1. Number of animals in pedigree, number of records, and animals in records

Population[1]	Animals in the pedigree	Animals with unknown parent(s) (%)	No. of records	Animals with records
BC1	1,087,161	11.5	2,968,758	908,116
BC2	1,060,862	13.5	3,041,612	874,329
BC1+2	1,974,901	10.8	6,010,370	1,782,445

[1]BC = breeding company associated with each subpopulation.

changed every 2 yr. Finally, the last metafounder was assigned to the unknown parents of animals born after 2018. According to the scenario, the attribution of metafounders was separated per subpopulation, except for the pre-1978 metafounder, which was the same.

This study examined 29,138 genotyped (50K Illumina chip OvineSNP50) animals including 6,302 AI males with offspring (3,434 for BC1 and 2,868 for BC2) born from 1996 to 2019, whereas the remaining 22,836 genotyped individuals were 18,541 genotyped but not selected young males (thus with no progeny) and 4,295 females (mostly genotyped in research projects). We included only autosomal SNPs. The quality control criteria applied included keeping animals and markers with call-rate higher than 0.90, minor allele frequency higher than 0.05, removal of Mendelian conflicts, and removal of loci with deviation higher than 15% from Hardy-Weinberg equilibrium. We obtained 37,312 effective SNPs after quality control.

Population Structure and Genetic Diversity

To infer the degree of divergence of the 2 subpopulations over time, we estimated different statistics based on pedigree and genomic analysis. The pedigree-based statistic that we computed was the change of average pedigree relationship across time cohorts (defined every 4 yr), within and between subpopulations.

The genomics-based statistics we computed were, first, the Hudson's fixation index F_{st} (Hudson et al., 1992), which measures divergence of 2 populations b and b′ based on allele frequencies p_i across n loci. We used the 2 estimators: the average of ratios and the ratio of averages (Bhatia et al., 2013), such as

$$\hat{F}_{st} = \frac{\frac{1}{n}\sum_i \left(\left(p_i^{b'} - p_i^{b}\right)^2\right)}{\frac{1}{n}\sum_i \left(p_i^{b'}\left(1-p_i^{b}\right) + p_i^{b}\left(1-p_i^{b'}\right)\right)} \quad \text{or}$$

$$\hat{F}_{st} = \frac{1}{n}\sum_i \frac{\left(p_i^{b'} - p_i^{b}\right)^2}{\left[p_i^{b'}\left(1-p_i^{b}\right) + p_i^{b}\left(1-p_i^{b'}\right)\right]}.$$

A further genomic analysis was by principal component analysis (**PCA**) of the genomic relationship matrix.

Scenarios

The 2 subpopulations (1 and 2) were studied together or separately according to 6 scenarios. The information considered for each scenario is summarized in Table 2. The same genomic prediction model was applied for each of these scenarios. In scenarios Alone1 and Alone2, only the pedigree, phenotypes, and the reference population of the subpopulation 1 or 2, respectively, was included in the statistic model.

In the scenarios TogetherSameMF, information of both subpopulations was included, and the attribution of the metafounders did not depend on the subpopulation but only on the year of birth. A total of 22 metafounders were used in this scenario. The TogetherDifferentMF is a similar scenario, but we assigned different metafounders according to the subpopulation

Table 2. Pedigree, phenotypes, genotypes, and criteria for the allocation of metafounders (MF) and number of AI males in the pedigree for each scenario[1]

Scenario	Pedigree of subpopulation	Phenotypes of subpopulation	Genotypes of subpopulation	MF attribution (n)	Size of reference population[2]
Alone1	1	1	1	YOB (22)	3,434
Alone2	2	2	2	YOB (22)	2,868
TogetherSameMF	1+2	1+2	1+2	YOB (22)	6,302
TogetherDifferentMF	1+2	1+2	1+2	YOB + subpopulation (43)	6,302
Indirect1	2	2	2 for ssGBLUP and 1 for indirect prediction	YOB (22)	2,868
Indirect2	1	1	1 for ssGBLUP and 2 for indirect prediction	YOB (22)	3,434

[1]YOB = year of birth; ssGBLUP = single-step genomic BLUP.

[2]Size of reference population = number of genotyped AI males with offspring.

after 1978, resulting in a total of 43 metafounders. The same metafounder was assigned for the 2 subpopulations before 1978 because we considered that, even if the 2 subpopulations were created in 1972, there is no real differentiation for a few years.

In the scenario Indirect1, we back-solved SNPs effects from model used in Alone2 ("only subpopulation 2 information") and then we computed the direct genomic value of animals of subpopulation 1. The scenario Indirect2 was the reverse of Indirect1 (use of "only subpopulation 1 information" to predict "subpopulation 2 direct genomic values").

Model

For all the genetic evaluations the following linear model was fitted:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{W_u u} + \mathbf{W_p p} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of records for MY (on average 160 DIM), $\boldsymbol{\beta}$ is the vector of fixed effects, $\mathbf{u} \sim N\left(\mathbf{0}, \mathbf{H}\sigma_u^2\right)$ is the vector of animal breeding values, $\mathbf{p} \sim N\left(\mathbf{0}, \mathbf{I_p}\sigma_p^2\right)$ is the vector of permanent environmental effects, and $\mathbf{e} \sim N\left(\mathbf{0}, \mathbf{I_e}\sigma_e^2\right)$ is the vector of residuals, where σ_u^2, σ_p^2, and σ_e^2 are the genetic, permanent environmental, and residual variances, respectively. We estimated these variances at 133,334, 88,889, and 222,223, respectively. $\mathbf{H} = \mathbf{A}_{(\boldsymbol{\Gamma})}^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{G}_{05} - \mathbf{A}_{(\boldsymbol{\Gamma})22}^{-1} \end{pmatrix}$ is the relationship matrix based on genomic and pedigree information with $\mathbf{A}_{(\boldsymbol{\Gamma})}$ including metafounders and genomic relationships obtained as $\mathbf{G}_{05} = \frac{2}{m}\mathbf{MM}'$, where m is the number of markers and $\mathbf{M}$ contains $\{-1,0,1\}$ for the 3 different genotypes. The variance component σ_u^2 was scaled as in Legarra et al. (2015). $\mathbf{I_p}$ is the identity matrix of order equal to the number of animals with records and $\mathbf{I_e}$ is the identity matrix of order equal to the number of records. $\mathbf{X}$, $\mathbf{W_u}$, and $\mathbf{W_p}$ are the incidence matrices for fixed effects, breeding values, and permanent environmental effects, respectively. The fixed effects were the interactions of herd × year × parity, age-parity × year × parity, month-parity × year × parity, and interval parity-first recording × year × parity.

Using this model, we realized single-step genomic BLUP evaluations for all the scenarios described using the software blup90iod2. We used, postGSf90 to compute SNP effects and then predf90 to compute direct genomic values of the scenarios Indirect (Tsuruta et al., 2001; Aguilar et al., 2010). We used the $\boldsymbol{\Gamma}$ matrix containing the relationship between metafounders (Legarra et al., 2015) in a similar manner (but with some differences) to the "Trend" method in Macedo et al. (2022), as follows. For a closed population, with an assumed linear trend of increasing inbreeding we hypothesize the following form for $\boldsymbol{\Gamma}$ (Sorensen and Kennedy, 1983):

$$\boldsymbol{\Gamma} = \begin{bmatrix} \Gamma_0 & \Gamma_0 & \Gamma_0 & \Gamma_0 & \dots \\ \Gamma_0 & \Gamma_0 + 2\Delta F_{(\gamma)} & \Gamma_0 + 2\Delta F_{(\gamma)} & \Gamma_0 + 2\Delta F_{(\gamma)} & \dots \\ \Gamma_0 & \Gamma_0 + 2\Delta F_{(\gamma)} & \Gamma_0 + 4\Delta F_{(\gamma)} & \Gamma_0 + 4\Delta F_{(\gamma)} & \dots \\ \Gamma_0 & \Gamma_0 + 2\Delta F_{(\gamma)} & \Gamma_0 + 4\Delta F_{(\gamma)} & \Gamma_0 + 6\Delta F_{(\gamma)} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}.$$

To consider 2 populations drifting from a common base, we considered the following structure:

$$\boldsymbol{\Gamma} = \begin{bmatrix} \Gamma_0 & \Gamma_0 & \Gamma_0 & \dots & \Gamma_0 & \Gamma_0 & \dots \\ \Gamma_0 & \Gamma_0 + 2\Delta F_{(\gamma)}^{(1)} & \Gamma_0 + 2\Delta F_{(\gamma)}^{(1)} & \dots & \Gamma_0 & \Gamma_0 & \dots \\ \Gamma_0 & \Gamma_0 + 2\Delta F_{(\gamma)}^{(1)} & \Gamma_0 + 4\Delta F_{(\gamma)}^{(1)} & \dots & \Gamma_0 & \Gamma_0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \Gamma_0 & \Gamma_0 & \Gamma_0 & \dots & \Gamma_0 + 2\Delta F_{(\gamma)}^{(2)} & \Gamma_0 + 2\Delta F_{(\gamma)}^{(2)} & \dots \\ \Gamma_0 & \Gamma_0 & \Gamma_0 & \dots & \Gamma_0 + 2\Delta F_{(\gamma)}^{(2)} & \Gamma_0 + 4\Delta F_{(\gamma)}^{(2)} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix}.$$

This matrix has size $(1 + n1 + n2)$, with 1 for the population between split, and subsequent $n1$ ($n2$) 2-yr interval metafounders within subpopulation 1 (2). In this case $n1 = n2$. This structure considers that the relationship between the 2 subpopulations (in the block $n1$, $n2$) is identical to the relationship before split (block 1,1), and differential increases of relationships within blocks ($n1$,$n1$) and ($n2$,$n2$).

Using the GLS method of Garcia-Baccino et al. (2017) we estimated the first value Γ_0 of the matrix $\boldsymbol{\Gamma}$ corresponding to the metafounder assigned before year of birth 1978. This value is estimated with accuracy because the pedigree of rams is very complete. Next, we computed the increase of inbreeding per year as a regression of inbreeding on year of birth, either within the whole breed or per subpopulation. Finally, we obtained $\Delta F_{(\gamma)}$, that is, the increase of inbreeding from one metafounder to the next, as $\Delta F_{(\gamma)} = \Delta F\left(1 - \Gamma_0 / 2\right)$ (Legarra et al., 2015), again, within subpopulation or across the whole breed. The value of ΔF was estimated as the regression of individual pedigree-based inbreeding coefficients on year of birth.

Depending on whether we considered Lacaune as a single population or as 2 populations, we obtained 4 $\boldsymbol{\Gamma}$ matrices:

- A matrix calculated based on the animals of BC1 (dimensions 22 × 22).
- A matrix for the animals of BC2 (dimensions 22 × 22).
- A matrix (dimensions 22 × 22) for the case of metafounders were assigned independently of the BC as in TogetherSameMF.
- A matrix combining the information calculated for each subpopulation (dimensions 43 × 43). This matrix was used in scenario TogetherDifferentMF where the animals of both subpopulations were considered and metafounders varied according to the subpopulation.

Validation

The scenarios were compared using the linear regression method (Legarra and Reverter, 2018). This method focuses on a particular group of individuals called focal individuals. Focal individuals correspond to genotyped individuals of interest, for whom we want to calculate the predictivity of genomic predictions (i.e. their evaluation when they were candidates for selection for the first time and when they did not have phenotypes or offspring) by comparison with predictions after they have been evaluated based on progeny records. In our case, these individuals correspond to the rams selected each year for AI and that have offspring with phenotypes. Groups constituted of focal individuals were selected according to their year of birth; for inclusion, rams must be genotyped and have at least 10 daughters with performances in the whole data set. For all the studied cohorts, the number of focal individuals (AI rams, in our case) was between 142 and 151 for BC1 and between 96 and 127 for BC2. For these groups of animals, we compared 2 values: the GEBV calculated with the "whole" data (i.e., the daughter's performances; $\mathbf{GEBV_w}$) available at the end of the studied period that is in 2021 $\left(\hat{u}_w\right)$; and GEBV calculated with "partial" data ($\mathbf{GEBV_p}$), that is by eliminating phenotypes recorded after the year of birth of the Focal Individuals $\left(\hat{u}_p\right)$. This "partial" evaluation mimics the genetic evaluation done when focal individuals were candidates to selection for the first time. We considered 5 groups of focal individuals, from 2015 to 2019.

For each scenario we compared the partial value $\hat{u}_p$ to a reference value (the benchmark), that was the value $\hat{u}_w$ for TogetherDifferentMF scenario, which we consider to be the most accurate scenario. This comparison was made using 3 estimators: bias, dispersion, and ratio of accuracies (Legarra and Reverter, 2018). The bias was estimated by the difference of GEBV means between partial and whole evaluations: $\hat{\Delta}_p = \left(\overline{\hat{u}_p}\right) - \left(\overline{\hat{u}_w}\right)$. We expect a value close to 0 for this estimator. The dispersion was studied using the slope of regression of "whole" on "partial" GEBV: $\hat{b}_p = \frac{Cov\left(\hat{u}_p, \hat{u}_w\right)}{Var\left(\hat{u}_p\right)}$. This estimator is supposed to be close to 1. Finally, the estimator for ratio of accuracies $\frac{acc_p}{acc_w}$ was the correlation between partial and whole values: $\hat{\rho}_{p,w} = \frac{Cov\left(\hat{u}_p, \hat{u}_w\right)}{\sqrt{Var\left(\hat{u}_p\right) Var\left(\hat{u}_w\right)}}$.

RESULTS

Genetic Diversity and Population Structure

We obtained very close values for the 2 F_{st} estimators, respectively 0.024 and 0.023 for the ratio of averages and the average of ratios. These low values reveal that the 2 Lacaune subpopulations are genetically close to each other. The results of the principal component analysis (Figure 1) show 2 distinct groups corresponding to the 2 BC, which confirm very low genetical exchanges between them. The percentage of variance explained is 35.6% for the first component (which clearly separates a few very old individuals from the rest) and 1.6% for the second component (which may be attributed to across-breed separation). When the old individuals were removed (not shown) the percentage of variance explained did not change. This small value of 1.6% implies that most variation is within subpopulation, not across-subpopulations (as also shown by the F_{st} coefficient), and thus the 2 subpopulations are indeed very similar. In addition, and even if the subpopulations are not very distant from each other, they seem to diverge over time, as we can see, for the second principal component, the oldest animals (darkest blue) in the middle of the graph and the youngest (lightest blue) toward the extreme. Some young individuals are located between the 2 clouds. We do not have a clear explanation for this, but to the best of our knowledge, it is neither DNA sample mislabeling nor genetic exchange. A possible hypothesis (very hard to verify) is that because both populations are selected toward the same objectives, moreover in a genomic manner, there is de facto convergence in their genetic background.

Average Pedigree Relationship

Figure 2 represents the average pedigree relationship according to the BC and the year of birth (by groups of 4 yr) for all the AI males (genotyped or not). The bottom left-hand and upper right-hand parts of the

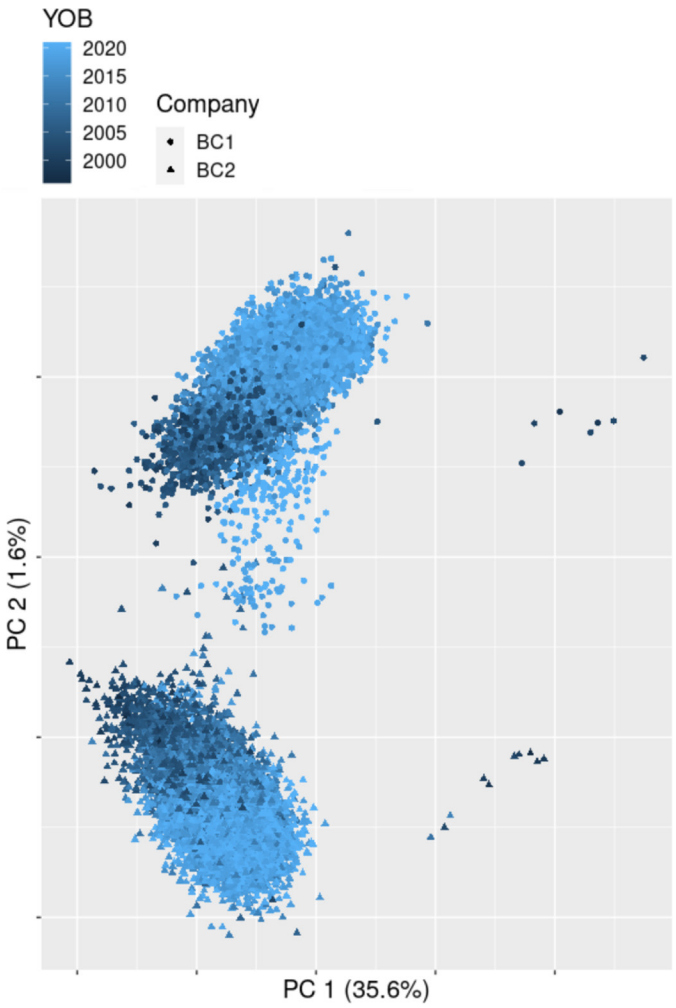

Figure 1. Principal component analysis for the 2 Lacaune subpopulations. YOB = year of birth; BC = breeding company associated with the subpopulation; PC = principal component.

graph contains average relationships, respectively for subpopulations 1 and 2. The 2 other parts show the relationship between the subpopulations. We observe a progressive increase of the relationship within the subpopulations over time from 0.0116 to 0.0714 (0.15% per year) for the subpopulation 1 and from 0.035 to 0.0651 (0.15% per year) for the subpopulation 2. On the other parts of the graph, we see a constant relationship, close to the initial value for the animals born before 1978. This result confirms the disconnection between the subpopulations since their creation. We obtain similar results for the females (not shown).

Estimates of Metafounder Relationships

The ancestral metafounder relationship Γ_0 obtained was 0.4638. This is not in the usual scale which assumes pedigree founders as unrelated; this is the average relationship relative to a (nonexisting) conceptual population with all markers at $p = 0.5$ (i.e., of maximum heterozygosity). For instance, if the genetic variation in the conceptual population with maximum heterozygosity would be 100, the genetic variation in the founders of Lacaune is $100\left(1-\frac{\Gamma_0}{2}\right) = 76.81$. This value was used for all the scenarios since we assume that the subpopulations were not yet disconnected before 1978. The average increase of inbreeding from one metafounder to the next in the metafounder scale, applying $\Delta F_{(\gamma)} = \Delta F\left(1-\frac{\Gamma_0}{2}\right)$, $\Delta F_{(\gamma)}$ were, respectively, 0.00185, 0.00145, and 0.00085 for subpopulation 1 (1), subpopulation 2 (2), and the animals of both subpopulations considered together. Finally, the matrix used in the scenario TogetherDifferentMF was a combination of the matrices (1) and (2). These values are shown in the Supplemental Material (Wicki et al., 2023; https://data.mendeley.com/datasets/smx65dvkgp/draft?a=df6774bd-5129-4f76-bca7-a337ce5012da). Note that the diagonals of the Metafounders relationships contain functions of homozygosity, e.g., $\frac{\Gamma_0}{4} = 0.5 - 2\overline{pq}$ (Garcia-Baccino et al., 2017), in which 0.5 is the maximum possible heterozygosity (all alleles are at $p = 0.05$) and $2\overline{pq}$ is the average heterozygosity of the founders of Lacaune. Thus, we do not expect diagonals of $\mathbf{\Gamma}$ to be 1.

Bias and Accuracy When Transitioning from Single to Combined Evaluation

Table 3 shows the bias $\hat{\Delta}_p$ of early (at selection time) predictions $\hat{u}_p$ ($GEBV_p$), expressed in genetic standard deviations, according to the scenarios. We did not obtain the expected value of 0, but this bias is small and corresponds to 1 to 2 yr of genetic progress. Bias appears to be higher in the scenarios with combined evaluation of the 2 subpopulations (TogetherSameMF and TogetherDifferentMF) than in the Alone scenarios. It also seems that Alone1 is less biased than Alone2. For the scenario Indirect, and in the particular case of metafounders, the software predf90 does not correctly include the shift (a constant identical for all animals) between pedigree and genomic relationships; this is a current limitation of the software. Thus, indirect predictions are shifted by an unknown constant compared with the Alone and Together evaluations, that is why we cannot estimate bias for Indirect scenarios.

The estimators of dispersion $\hat{b}_p$ (the slope of the regression of $GEBV_w$ in TogetherDifferentMF on GEB-

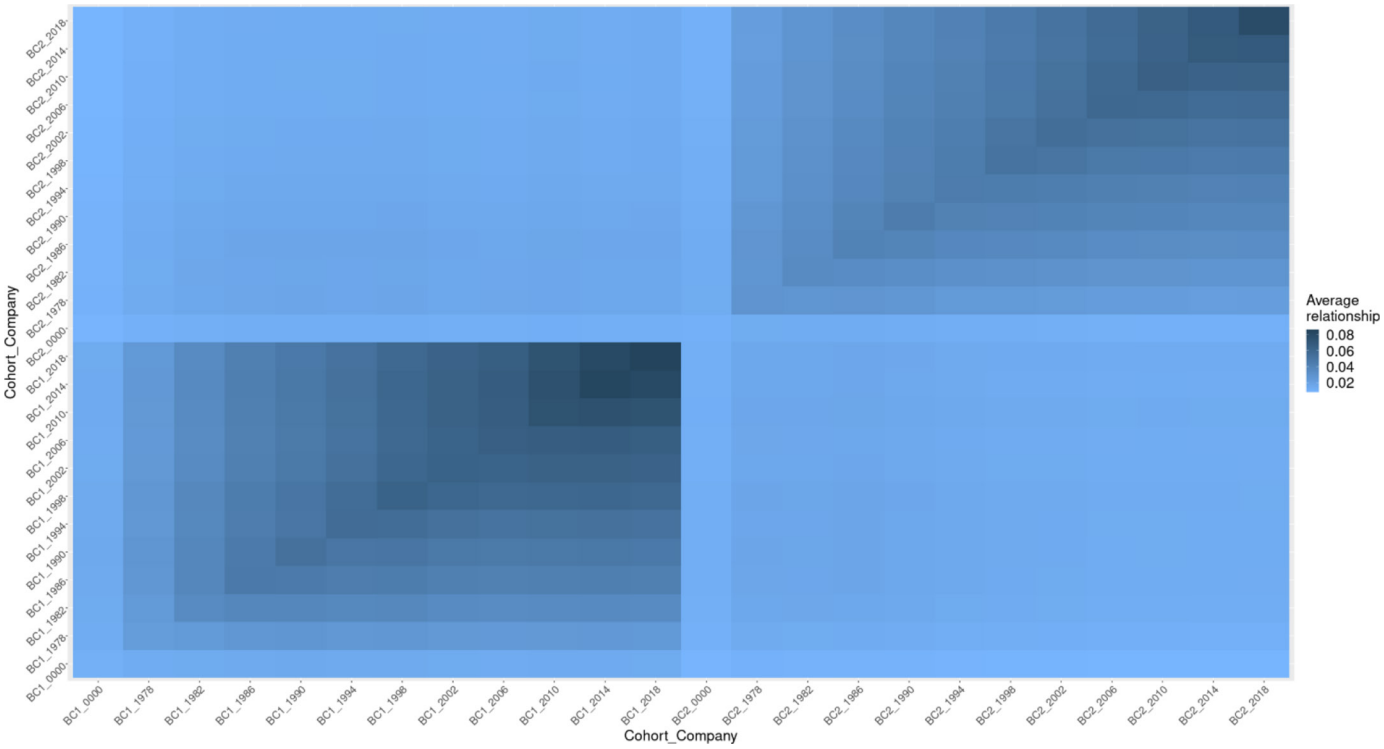

Figure 2. Evolution of average relationship by cohort (defined by time, animals born every 4 yr) and subpopulation in males. BC = breeding company associated with the subpopulation.

V_p), are presented in Table 4. This value is expected to be close to 1. We observe in most scenarios and for the different focal individuals slopes lower than 1 so an overdispersion. The slopes are closer to 1 for the 2 scenarios with records from both BC (TogetherSameMF and TogetherDifferentMF). We also observe very similar slopes between these 2 scenarios. Therefore, the modeling of metafounders according to the BC has no influence on the dispersion. Also, both "Together" scenarios slightly improve the slope $\hat{b}_p$ and slightly deteriorate the bias $\hat{\Delta}_p$. Finally, indirect predictions show rather high overdispersions.

Figure 3 shows the ratio of accuracies $\hat{\rho}_{p,w}$ for each BC, according to the focal individuals' groups and to the scenarios Alone, Indirect, and Together, (which includes TogetherSameMF and TogetherDifferentMF). The closer the ratio is to 1, the closer is the predictive capacity from early genomic information to progeny information (i.e., the more accurate the genomic prediction). For this statistic we obtained exactly the same values for the scenarios TogetherSameMF and TogetherDifferentMF. We can see that the scenario Together is always the most accurate (with all ratios above 0.45). Considering the scenarios Alone, it seems that BC1 is better predicted than BC2 by at least 0.04. Indirect predictions based on the "other" population generally have low accuracy, sometimes with values close to 0.

To investigate the contribution of one subpopulation to the genomic prediction of the other, the most important result in Figure 3 is the difference of accuracies between the scenarios Alone and Together, for each

Table 3. Bias $\left(\hat{\Delta}_p\right)$ between whole and partial genomic EBV for the different scenarios for the focal individuals 2015 to 2019[1]

	Focal individuals									
	2015		2016		2017		2018		2019	
Subpopulation	BC1	BC2	BC1	BC2	BC1	BC2	BC1	BC2	BC1	BC2
Alone	−0.04	0.34	−0.10	0.26	−0.03	0.29	0.02	0.26	−0.06	0.26
TogetherSameMF	0.31	0.43	0.23	0.42	0.33	0.44	0.41	0.44	0.29	0.38
TogetherDifferentMF	0.32	0.33	0.24	0.33	0.35	0.34	0.43	0.32	0.34	0.25

[1]MF = metafounders; BC = breeding company associated with each subpopulation.

Table 4. Slope $\left(\hat{b}_p\right)$ of regression of the genomic EBV calculated with "whole" data ($GEBV_w$) on the genomic EBV calculated with "partial" data ($GEBV_p$) for the different scenarios for the cohorts 2015 and 2019[1]

	Focal individuals									
	2015		2016		2017		2018		2019	
Subpopulation	BC1	BC2	BC1	BC2	BC1	BC2	BC1	BC2	BC1	BC2
Alone	0.82	0.98	0.77	0.52	0.90	0.87	0.66	0.90	0.81	0.56
TogetherSameMF	0.90	1.03	0.87	0.69	0.93	0.96	0.83	0.90	0.89	0.82
TogetherDifferentMF	0.89	1.04	0.87	0.67	0.93	0.97	0.82	0.90	0.89	0.82
Indirect	0.08	0.4	0.28	0.03	0.09	1.60	0.06	0.21	0.09	0.12

[1]MF = metafounders; BC = breeding company associated with each subpopulation.

subpopulation, across years. This difference is globally low (less than 0.08) for BC1 in spite the fact that the reference population size doubles from Alone to Together. For BC2, the differences are more variable from one year to the other: from 0.03 for focal individuals 2015 to 0.17 for focal individuals 2018.

Correlations of Estimated SNPs Effects

The correlations between estimated SNPs effects across different reference populations and truncation years are presented in Figure 4. In this figure, the reference population differs according to the subpopulation considered (BC1, BC2, or both of them) and the phenotypic records included in the evaluation (i.e., including all available phenotypes or deleting phenotypes after the years 2015 to 2019). We observe very high correlations across years within each reference population, especially when the reference population is only one subpopulation. The correlations are greater than 0.77 for BC1, 0.87 for BC2, and 0.77 for reference population Together composed of BC1 and BC2.

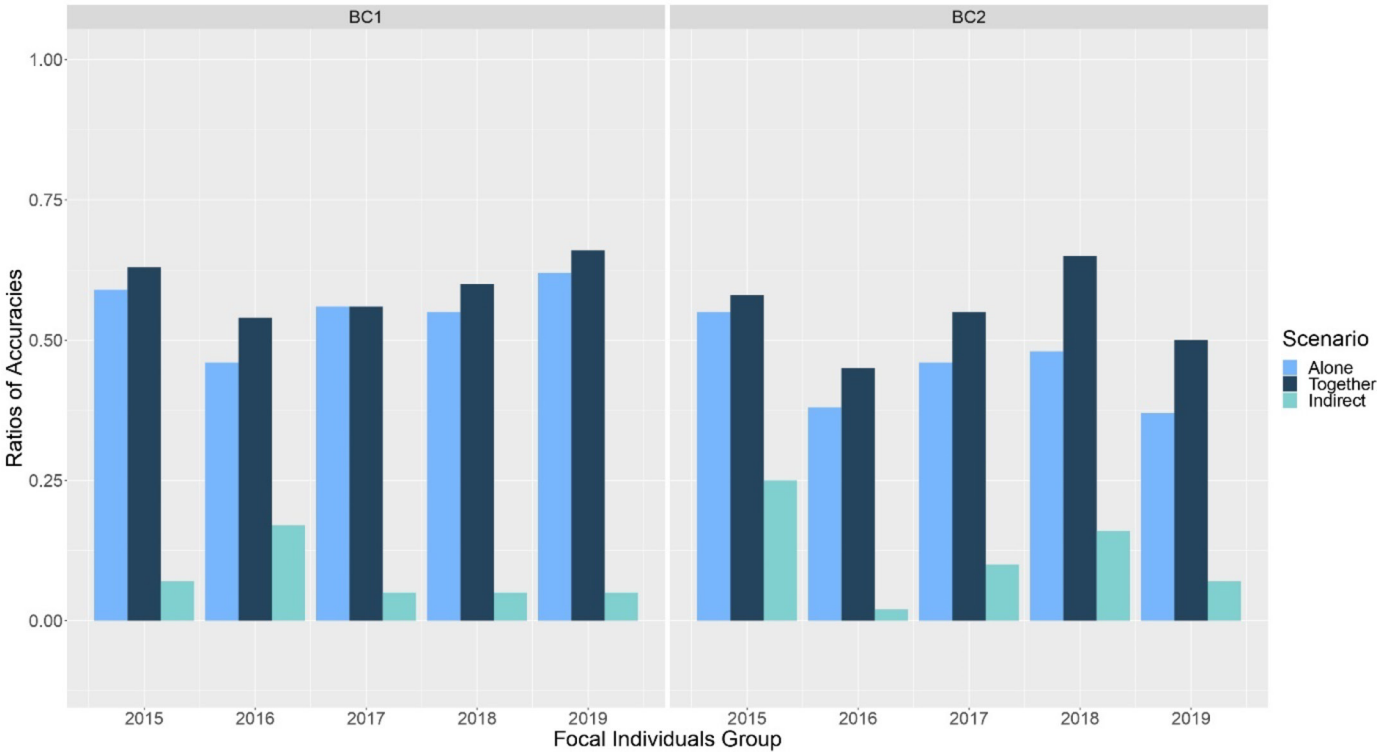

Figure 3. Ratios of accuracies according to focal individuals' groups and subpopulation (BC = breeding company associated with the subpopulation). Alone scenario = genetic evaluation based on single reference population (BC1 or BC2); Together scenario = genetic evaluation based on joined reference population (BC1 and BC2); Indirect scenario = indirect genetic evaluation of training population of one subpopulation based on the reference of the other one.

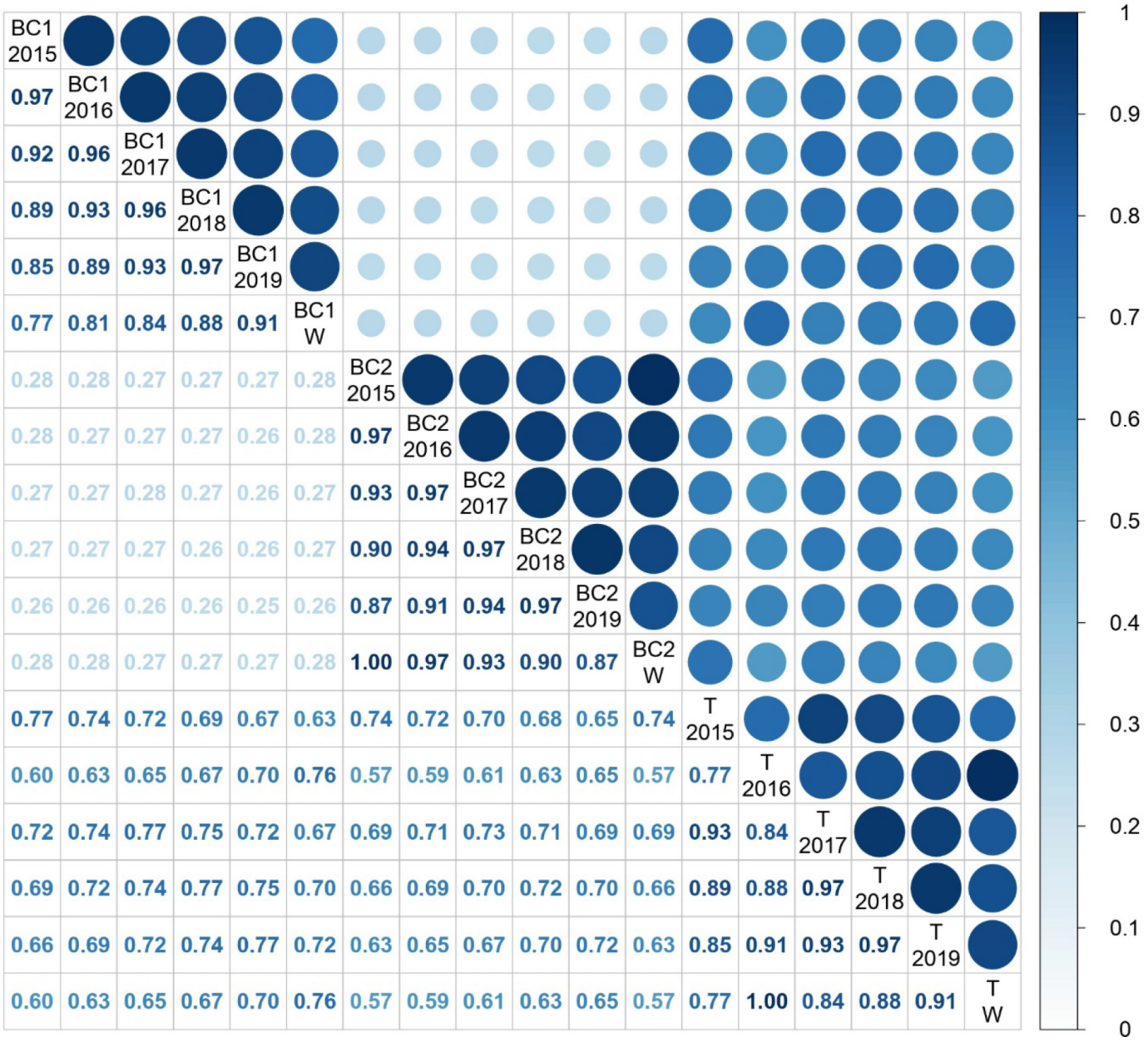

Figure 4. Correlation of estimated SNPs effects between all the studied reference populations. BC1 = breeding company 1 in reference population; BC2 = breeding company 2 in reference population; T = companies 1 and 2 in reference populations; W = all the phenotypic records available considered; 2015 to 2019 = phenotypic records after this year deleted.

The correlations are slightly lower between reference population Alone and Together but still high: from 0.60 to 0.77 between Together and BC1, and from 0.57 to 0.74 between Together and BC2. However, we observe very low correlations between the 2 reference populations (BC1 and BC2) in scenarios Alone; all correlations are below 0.28 between BC1 and BC2.

Finally, we do not observe any difference in correlations from one truncation year to the other.

DISCUSSION

Genetic Diversity and Population Structure

For the 2 F_{st} estimators computed and considering all the available genotypes, we obtained a value of 0.02. Kijas et al. (2009) studied the F_{st} between different sheep breeds. They found highly variable values ranging from 0.053 between the breeds Merino and Sarda (both originally from Europe) to 0.349 between the

breeds Namaqua Afrikaners and Soay (respectively African and European origins). Concerning Pyrenean dairy sheep breeds who are supposed to be genetically close populations, Garcia-Baccino et al. (2022) found a value of 0.053 between the Manech Tête Noire and Laxa Cara Negra and a very low value of 0.014 between the breeds Manech Tête Rousse (**MTR**) and Laxa Cara Rubia (**LCR**). It is known that MTR started as an importation of LCR animals into France in the 1960s and 1970s, and there is use of MTR in the LCR breeding program. All in all, the 2 Lacaune subpopulations thus seem genetically close.

As for our PCA results, we can see quite distinct groups for each subpopulation. In Garcia-Baccino et al. (2022) the groups for LCR and MTR overlap, as explained by the use of LCR AI rams that are offspring of MTR AI rams. In our PCA we can see that there are very few connections between the 2 Lacaune subpopulations compared with the LCR and MTR breeds.

We can also see in our PCA plot, on which the color of the individuals changes according to their year of birth, that the divergence between the 2 subpopulations tends to increase over time. However, this observation must be tempered in view of other work considering the 2 Lacaune subpopulations as well as other breeds; for instance the 2 Lacaune populations are very close when shown together with the Pyrenean dairy breeds (Legarra et al., 2014).

Average Pedigree Relationship

The progressive increase in the average relatedness observed within each subpopulation (in Figure 2) was expected. Indeed, because these subpopulations are closed, relatedness can only increase over generations. On the opposite, the relatedness between subpopulations remains constant and equal to the original relatedness observed before the split between BC1 and BC2. This confirms that the subpopulations have not used reproducers from the other subpopulation (or very few), so there were very few exchanges between them. This was our expectation based on current knowledge of how flocks work and it has been confirmed with actual data.

Bias, Dispersion, and Accuracy

The biases obtained in our study (Table 3) are slightly higher than those obtained by Macedo et al. (2022) for the Lacaune breed, who performed the genetic evaluations and the linear regression method without considering a structuring of the breed in 2 subpopulations.

The regression slopes (Table 4) and correlations (Figure 3) are broadly similar to those obtained in Macedo et al. (2022) except for the Indirect scenarios for which the indices are more dispersed and the predictions less accurate. Many studies have shown that prediction accuracy is strongly dependent on genetic relationship. Indeed, we observed very few relationships between subpopulations (Figures 1 and 2), so it seems logical that the indirect prediction of one subpopulation by the other is little accurate. In our case, this indirect prediction seems extremely inaccurate for some years where we obtain correlations very close to 0 (Figure 3). Studies in dairy cattle have shown that evaluations of one breed based on the reference of another breed were less accurate than the evaluation of candidates of the same breed, but accuracies remained above 0.4 (Wientjes et al., 2015). However, in Wientjes et al. (2015), the accuracy of within-breed evaluations were higher (>0.9) than those observed in our study for the Alone and Together scenarios.

Another result that may seem surprising is the fact that, in some cases, we obtain an improvement in accuracy from Alone to Together, whereas Indirect is very inaccurate (e.g., focal individuals of BC2 born in 2018). This result can be explained by the results concerning correlations of SNPs effects as we discuss below.

Regarding the attribution of metafounders, we obtained very close (Tables 3 and 4) or even identical results (Figure 3) between the scenarios TogetherSameMF and TogetherDifferentMF for the 3 estimators and all the cohorts. The assignment of metafounders independently of the subpopulation did not affect the results. This makes sense because the 2 subpopulations have a similar genetic trend.

The pooling of subpopulations in the evaluation showed a clear, although variable across years, interest in terms of accuracy gain because we observe higher accuracies moving from scenarios Alone to Together (Figure 3). These accuracies are lower than expected as the reference population size in fact doubles between these scenarios. The increase in the ratios $\hat{\rho}_{p,w}$ appears to be greater for subpopulation 2 than subpopulation 1 (Figure 3); therefore, it seems more interesting to join subpopulations 1 and 2 for the prediction of 2, than for prediction of 1. Zhou et al. (2014) explain the observed nonreciprocal contribution between the Danish Red, Finnish Ayrshire, and Swedish Red composite breed and the Norwegian Red breed by the nonequivalent reference populations sizes: 3,300 and 2,353 bulls, respectively. In our case, we also study reference populations with sizes that are not exactly equivalent: 3,434 AI rams for BC1 and 2,868 for BC2.

We can wonder what would be the gain in accuracy observed between the scenarios Alone and Together if the population was not structured (i.e., if it was a

single population with no subpopulations). In their study, Lourenco et al. (2017) simulated the genomic evaluations of a dairy trait with a heritability of 0.3 that could correspond to MY. They obtain an increase in accuracy approximatively from 0.65 to 0.80 between a training population with 2,000 and 5,000 animals. We can think that a higher gain would be obtained with our reference population size but we would perhaps observe differences in dairy sheep for example due to the much lower number of offspring per AI sires.

Finally, we can say that today the 2 subpopulations remain close enough genetically so that their combined evaluation is slightly advantageous, because we do see an increase in accuracy from the scenarios Alone to Together, unlike the case observed between distant breeds where accuracy does not increase (Brøndum et al., 2011).

Correlations of Estimated SNPs Effects

For the 3 reference populations studied (BC1, BC2, and Together), the high correlations of estimated SNPs effects obtained within a subpopulation, from one year to the other, are reassuring about the correctness of the model and the stability of the genomic predictions, especially for the combined reference population.

The low correlations between BC1 and BC2 are on line with previous studies dealing with combined genomic evaluations where differences in SNPs effects are observed according to the reference population design (Pryce et al., 2011). Note that these correlations underestimate the genetic correlation as SNP effects are shrunken. In fact, the model imposes that SNP effect are defined within subpopulation, so this explains the inability to predict one subpopulation from the other.

In particular, these results show that when analyzing both subpopulations together, the model forces the SNP effects to be "portable" across breeds, whereas the analysis of populations alone does not impose this. In addition, the Together model increases in the reference population size. That explains our gain in accuracy from the scenario Alone to Together. There must be a threshold beyond which too much distance implies that "portable" SNP effects are in fact not that portable and yield less accurate GEBVs, but this is not the case here.

Thus, there seems to be an interest in pursuing a joint evaluation for both subpopulations because this allows a larger reference population size, even if the corresponding gain in accuracy is small. In addition, we hypothesize that a greater genetic connectedness between the 2 subpopulations, for example by setting up systematic exchanges between the 2 subpopulations, would be a way to improve the reliability and increase the value of the joint evaluation currently performed. Furthermore, in the current situation with no exchange, the 2 subpopulations will continue to diverge, and we wonder if the joint evaluation would eventually become disadvantageous.

CONCLUSIONS

Our study has shown that the 2 Lacaune subpopulations are still genetically close but they diverge over time. The contribution in terms of accuracy of one subpopulation to the other is positive but fluctuating across years. We have also seen that the contribution between subpopulations is not always symmetrical. Globally, the increase in accuracy when information of both subpopulations was included is small given the doubling of the reference population, but the benefit is consistent and this implies that a joint evaluation is beneficial.

ACKNOWLEDGMENTS

This study was funded by Institut de l'Elevage (Paris, France), INRAE (Paris, France), and Apis-gène (Paris, France). We want to thank Jean-Michel Astruc for the preparation and all the helpful information provided about the database. We also are grateful to the Genotoul Bioinformatics Platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing storage and computing resources and to the organizations responsible for the Lacaune breeding program (Upra Lacaune, Ovitest, Confederation de Roquefort) for providing the database (pedigree, performances, and genotypes). The authors have not stated any conflicts of interest.

REFERENCES

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743–752. https://doi.org/10.3168/jds.2009-2730.

Astruc, J. M., G. Lagriffoul, A. Legarra, and D. Buisson. 2022. Six years of genomic selection have increased the genetic gain in French dairy sheep. Pages 2960–2963 in Proc. 12th World Congress on Genetics Applied to Livestock Production, Rotterdam, the Netherlands. Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-940-4_60.

Baloche, G., A. Legarra, G. Sallé, H. Larroque, J.-M. Astruc, C. Robert-Granié, and F. Barillet. 2014. Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. J. Dairy Sci. 97:1107–1116. https://doi.org/10.3168/jds.2013-7135.

Bhatia, G., N. Patterson, S. Sankararaman, and A. L. Price. 2013. Estimating and interpreting F_{ST}: The impact of rare variants. Genome Res. 23:1514–1521. https://doi.org/10.1101/gr.154831.113.

Brøndum, R. F., E. Rius-Vilarrasa, I. Strandén, G. Su, B. Guldbrandtsen, W. F. Fikse, and M. S. Lund. 2011. Reliabilities of genomic

prediction using combined reference data of the Nordic Red dairy cattle populations. J. Dairy Sci. 94:4700–4707. https://doi.org/10.3168/jds.2010-3765.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3:e3395. https://doi.org/10.1371/journal.pone.0003395.

Garcia-Baccino, C. A., A. Legarra, O. F. Christensen, I. Misztal, I. Pocrnic, Z. G. Vitezica, and R. J. C. Cantet. 2017. Metafounders are related to F_{st} fixation indices and reduce bias in single-step genomic evaluations. Genet. Sel. Evol. 49:34. https://doi.org/10.1186/s12711-017-0309-2.

Garcia-Baccino, C. A., C. Pineda-Quiroga, J. M. Astruc, E. Ugarte, and A. Legarra. 2022. High genetic correlation for milk yield across Manech and Latxa dairy sheep from France and Spain. JDS Commun. 3:260–264. https://doi.org/10.3168/jdsc.2021-0195.

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica 136:245–257. https://doi.org/10.1007/s10709-008-9308-0.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397. https://doi.org/10.1534/genetics.107.081190.

Habier, D., R. L. Fernando, and D. J. Garrick. 2013. Genomic BLUP decoded: A look into the black box of genomic prediction. Genetics 194:597–607. https://doi.org/10.1534/genetics.113.152207.

Hudson, R. R., M. Slatkin, and W. P. Maddison. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583–589. https://doi.org/10.1093/genetics/132.2.583.

Kijas, J. W. D., B. P. Townley, M. P. Dalrymple, J. F. Heaton, A. Maddox, P. McGrath, R. G. Wilson, R. Ingersoll, S. McCulloch, D. McWilliam, J. Tang, N. McEwan, V. H. Cockett, F. W. Oddy, H. Nicholas, and H. Raadsma. 2009. A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS One 4:e4668. https://doi.org/10.1371/journal.pone.0004668.

Legarra, A., G. Baloche, F. Barillet, J. M. Astruc, C. Soulas, X. Aguerre, F. Arrese, L. Mintegi, M. Lasarte, F. Maeztu, I. Beltrán de Heredia, and E. Ugarte. 2014. Within- and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. J. Dairy Sci. 97:3200–3212. https://doi.org/10.3168/jds.2013-7745.

Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal. 2015. Ancestral relationships using metafounders: Finite ancestral populations and across population relationships. Genetics 200:455–468. https://doi.org/10.1534/genetics.115.177014.

Legarra, A., and A. Reverter. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Genet. Sel. Evol. 50:53. https://doi.org/10.1186/s12711-018-0426-6.

Liu, Z., F. R. Seefried, F. Reinhardt, S. Rensing, G. Thaller, and R. Reents. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genet. Sel. Evol. 43:19. https://doi.org/10.1186/1297-9686-43-19.

Lourenco, D. A. L., B. O. Fragomeni, H. L. Bradford, I. R. Menezes, J. B. S. Ferraz, I. Aguilar, S. Tsuruta, and I. Misztal. 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. J. Anim. Breed. Genet. 134:463–471. https://doi.org/10.1111/jbg.12288.

Lund, M. S., A. P. de Roos, A. G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, F. Seefried, and G. Su. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet. Sel. Evol. 43:43. https://doi.org/10.1186/1297-9686-43-43.

Macedo, F. L., J. M. Astruc, T. H. E. Meuwissen, and A. Legarra. 2022. Removing data and using metafounders alleviates biases for all traits in Lacaune dairy sheep predictions. J. Dairy Sci. 105:2439–2452. https://doi.org/10.3168/jds.2021-20860.

Moghaddar, N., A. A. Swan, and J. H. van der Werf. 2014. Comparing genomic prediction accuracy from purebred, crossbred and combined purebred and crossbred reference populations in sheep. Genet. Sel. Evol. 46:58. https://doi.org/10.1186/s12711-014-0058-4.

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Sölkner, M. E. Goddard, and B. J. Hayes. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. J. Dairy Sci. 94:2625–2630. https://doi.org/10.3168/jds.2010-3719.

Pszczola, M., T. Strabel, H. A. Mulder, and M. P. L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95:389–400. https://doi.org/10.3168/jds.2011-4338.

Sorensen, D. A., and B. W. Kennedy. 1983. The use of the relationship matrix to account for genetic drift variance in the analysis of genetic experiments. Theor. Appl. Genet. 66:217–220. https://doi.org/10.1007/BF00251147.

Tsuruta, S., I. Misztal, and I. Strandén. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. J. Anim. Sci. 79:1166. https://doi.org/10.2527/2001.7951166x.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24. https://doi.org/10.3168/jds.2008-1514.

Wicki, M., J. Raoul, and A. Legarra. 2023. Supplemental material article "Effect of subdivision of the Lacaune dairy sheep breed on the accuracy of genomic prediction." Mendeley Data, V1. https://doi.org/10.17632/smx65dvkgp.1.

Wientjes, Y. C., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten, and M. Calus. 2015. Empirical and deterministic accuracies of across-population genomic prediction. Genet. Sel. Evol. 47:5. https://doi.org/10.1186/s12711-014-0086-0.

Zhou, L., B. Heringstad, G. Su, B. Guldbrandtsen, T. H. E. Meuwissen, M. Svendsen, H. Grove, U. S. Nielsen, and M. S. Lund. 2014. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. J. Dairy Sci. 97:4485–4496. https://doi.org/10.3168/jds.2013-7580.

ORCIDS

M. Wicki https://orcid.org/0000-0002-3585-1572
J. Raoul https://orcid.org/0000-0002-3140-4433
A. Legarra https://orcid.org/0000-0001-8893-7620