# ESTIMATION OF SNP EFFECTS IN LACAUNE DAIRY SHEEP DEPENDING ON THE REFERENCE POPULATION COMPOSITION

**M. Wicki[1,2], J. Raoul[1,2] and A. Legarra[1,3]**

[1]INRAE, INP, UMR 1388 GenPhySE, F-31326 Castanet-Tolosan, France
[2]Institut de l'Elevage, Catstanet-Tolosan 31321, France
[3]Current address: Council on Dairy Cattle Breeding, Bowie, MD 20716, USA

## SUMMARY

The Lacaune dairy sheep breed split in 1972 into two subpopulations with no exchange of genetic material but a single genetic evaluation and same selection objectives. Previous work has shown that this led to the creation of two disconnected but genetically close subpopulations. Previous work also demonstrated that the currently performed combined genomic evaluation of both subpopulations is slightly advantageous, in terms of accuracy, as opposed to within-subpopulation genomic evaluations. This paper focuses on the study of the estimated SNPs effects related to the three training populations: composed of one, the other or both subpopulations. The estimated SNP effects are strongly correlated across years within the training population. When subpopulations are predicted separately, there is low correlation between estimated SNP effects, but when they are predicted jointly, there is a strong correlation of the joint estimate with subpopulation estimates. The regression of "early" (only based on genomic information) on "late" (including progeny information) SNP predictions is lower than one for one of the subpopulations but not for the other, and close to one for the joint prediction. This shows some bias in this particular subpopulation whose origin is not understood.

## INTRODUCTION

Selection in French Lacaune dairy sheep started in the 70's with Genomic selection starting in 2015. Each year, young AI rams are selected, among genotyped prospective rams, based on their Genomic Estimated Breeding Values (GEBVs) and used to inseminate females. The accuracy of Milk Yield BV 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). However, it is of interest to understand if this genomic accuracy can be enhanced further by increasing the size and optimizing the setting up of the reference population.

In 1972, the structure of genetic improvement split, with each flock participating in the AI programs of only one of two existing ram AI studs (breeding companies BC) 1 or 2), exclusively, i.e. a flock only sends rams and receives semen to and from the chosen BC. This created in fact two different subpopulations (1 and 2), subpopulations which do not exchange as breeders rarely exchange sheep and the flux of males and semen is handled by the BC within their participant flocks. Moreover, flocks respect the initial assignation of flocks to BC. Thus, for the last 5 decades, flocks have been contributing rams to a single BC and receiving semen from a single BC. In the following, we will use the wording "subpopulation" to indicate the set of animals belonging to flocks attached to each BC.

A first study (Wicki *et al*. 2023) revealed a low genetic differentiation between the two subpopulations observable, on the one hand, by a low Fst value (0.02), and on the other hand by the results of a Principal Component Analysis (PCA) of the genomic relationship matrix. Indeed, this PCA shows two distinct groups corresponding to each BC, separated on the second component. However, the percentage of variance explained (1.6%) implies that most variation is within-subpopulation, not across. Pedigree analyses showed a low and constant average pedigree relatedness between BC which confirms the very low genetic exchanges between companies.

Finally, Wicki *et al*. (2023) observed a small gain in GEBVs accuracy from the evaluations with training populations of a single BC to the evaluation based on combined reference population.

In this paper, we focus on the study of estimated SNPs effects obtained from genomic evaluations based on reference populations using one company (BC1), the other (BC2) or both of them together (T). We compare SNP effects across years, and across the three possible reference populations.

## MATERIALS AND METHODS

This study used all the pedigree, genotypes (50K Illumina chip OvineSNP50) and phenotypic data obtained from regular performance recording of Milk Yield from 1972 to 2021 available in Lacaune dairy sheep (Table 1). The correlation between allele frequencies of each subpopulation is 0.905.

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

| Population | Animals in the pedigree | Animals with unknown parent(s) (%) | Number of records | Animals with records | Animals genotyped |
|---|---|---|---|---|---|
| BC 1 | 1,087,161 | 11.5% | 2,968,758 | 908,116 | 16,792 |
| BC 2 | 1,060,862 | 13.5% | 3,041,612 | 874,329 | 12,225 |
| T (BC1+2) | 1,974,901 | 10.8% | 6,010,370 | 1,782,445 | 29,017 |

**Genomic prediction based on different reference populations.** We performed genomic evaluations according to several scenarios in which the subpopulations were studied together or separately (Table1). In two scenarios, only the reference population of one subpopulation (BC1 and BC2) was included in the prediction model. In the scenario Together (T), information of both subpopulations was included.

For all the genetic evaluations we used an animal model ssGBLUP with metafounders as detailed in Wicki *et al*. (2023) using blup90iod2 (Tsuruta *et al*. 2001). We used postGSf90 to compute SNPs effects (Tsuruta *et al*. 2001; Aguilar *et al*. 2010), *i.e.* SNP effects are backsolved from GEBVs of genotyped individuals.

**Validation**. The scenarios were compared using the LR method (Legarra and Reverter 2018) but applied to SNP effects. We defined as "whole" the evaluation including all the phenotypes available until 2021. We compared the SNP effects estimated from this evaluation with SNP effects estimated from "partial" evaluations in which the phenotypes were truncated, i.e. phenotypes after a cut-off date were deleted, with cut-off dates ranging from 2015 to 2019. The correlation shows stability of SNP effects whereas the regression of SNP estimates on "whole" on SNP estimates on "partial" is expected to have a value of 1 for unbiased predictions.

## RESULTS AND DISCUSSION

We observe very high correlations of estimated SNPs effects (Figure 1) across years within each reference population (above 0.77, 0.87 and 0.77 respectively for reference subpopulation 1, subpopulation 2 and both), which is reassuring in regards to the correctness of the model and the stability of the genomic predictions, especially for the combined reference population. The correlation is slightly higher for subpopulation 2 across years although we don't have an explanation. The low correlations between subpopulations 1 and 2 (below 0.28) are on line with previous studies investigating combined genomic evaluations where differences in SNPs effects are observed according to the reference population design. Indeed, in our previous study (Wicki *et al*. 2023) we observed that "indirect" genomic predictions using SNP estimates from one subpopulation to obtain

GEBVs in the other subpopulation had very low accuracy of 0.10 on average. In addition, these results show that, when analysing both subpopulations together, the model forces the SNP effects to be "portable" across breeds, whereas the analysis of populations alone does not impose this. The correlation between "Together" with each subpopulation is lower than 1 and lower than correlations within each subpopulation, yet the "Together" evaluation increases accuracy of GEBVs (Wicki *et al*. 2023) from 0.56 to 0.60 for subpopulation 1 and from 0.45 to 0.55 for subpopulation 2 on average (ratios of accuracies). We believe that the increase in accuracy from separate subpopulation analyses comes from the increase in the reference population size.

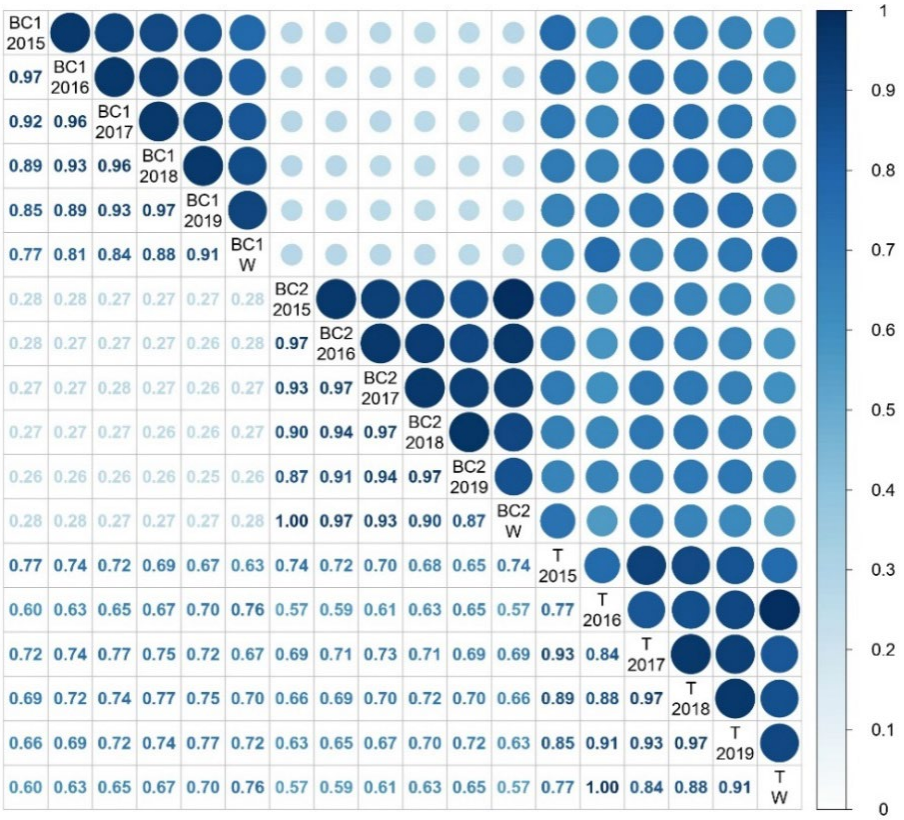


**Figure 1. Correlation of estimated SNPs effects between all the studied reference populations ("BC1" = reference population based on subpopulation 1 only, "BC2" = reference population based on subpopulation 2 only, "T" = reference population based on both subpopulations, "W" = evaluation including all phenotypic information until 2021, "2015" to "2019" = evaluation with phenotypic information truncated after year 2015 to 2019)**

We expected regression slopes close to 1 between SNPs effects whole and partial in each reference population. Similarly, we expected slopes slightly different from 1 between reference populations BC1 and T, BC2 and T; but far from 1 between BC1 and BC2. We indeed observed low slopes (below 0.31) when estimated SNP effects from one subpopulation were regressed on estimates from the other subpopulation. Within training population BC1, the slope increases over cohorts from 0.58 to 0.83, whereas within training population BC2 the slopes are very close to 1.

This would suggest some bias in BC1 but not in BC2 – the reasons for that are unknown. Slopes between single and combined populations are also variable across cohorts and BC but not too far from 1. Technically, they don't need to be 1 because the "partial" Together contains information that it is not in the "whole" subpopulation.

**Table 2. Slopes of regression between estimated SNPs effects "whole" on "partial"**

| | | Cohort | | | | |
|---|---|---|---|---|---|---|
| Partial | Whole | 2015 | 2016 | 2017 | 2018 | 2019 |
| BC1 | BC1 | 0.58 | 0.64 | 0.71 | 0.77 | 0.83 |
| | BC2 | 0.25 | 0.28 | 0.30 | 0.30 | 0.31 |
| | T | 0.36 | 0.39 | 0.43 | 0.47 | 0.50 |
| BC2 | BC1 | 0.22 | 0.22 | 0.23 | 0.23 | 0.24 |
| | BC2 | 1.00 | 1.01 | 1.02 | 1.01 | 1.00 |
| | T | 0.35 | 0.38 | 0.41 | 0.43 | 0.46 |
| T | BC1 | 0.62 | 0.96 | 0.73 | 0.78 | 0.84 |
| | BC2 | 0.94 | 0.92 | 0.96 | 0.95 | 0.94 |
| | T | 0.60 | 1.00 | 0.72 | 0.78 | 0.84 |

## CONCLUSIONS

Although the evaluations within each subpopulation alone or combined lead to very similar results, this study showed that the estimation of SNP effects was different depending on whether each of the two Lacaune subpopulations was considered separately or together. However, the estimation of SNP effects across subpopulations were too different to be portable, leading to very poor-quality cross-subpopulations evaluations.

## ACKNOWLEDGEMENTS

This study was funded by Institut de l'Elevage, INRAE and Apis-gène. We gratefully acknowledge the funding. Furthermore, we want to thank Jean-Michel Astruc for the preparation and 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 access to the database.

## REFERENCES

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S. and Lawlor T.J. (2010) *J. Dairy Sci.* **93**: 743.

Baloche G., Legarra A., Sallé G., Larroque H., Astruc JM, Robert-Granié C. and Barillet F. (2014) *J.Dairy Sci.* **97**: 1107.

Legarra A. and Reverter A. (2018) *Genet. Sel. Evol.* **50**: 53.

Tsuruta S, Misztal I. and Strandén I. (2001) *J. Anim. Sci.* **79**: 1166.

Wicki M., Raoul J. and Legarra A. (2023) *J. Dairy Sci.* In revision.