# GenTORE

## *Genomic management Tools to Optimise Resilience and Efficiency*

Grant agreement n°: 727213

**H2020 - Research and Innovation Action**

# D4.4
# Paper on genomic prediction of efficiency traits in a major crossbreed

**Due date:** M57 (February 2022)

**Actual submission date:** M57 (February 2022)

**Project start date:** 1st June 2017 **Duration:** 60 months

**Workpackage concerned:** WP4

**Concerned workpackage leader:** Mogens Lund (AU)

**Lead Beneficiary:** INRAE

**Dissemination level:**

**X PU:** Public (must be available on the website)

☐ **CO:** Confidential, only for members of the consortium (including the Commission Services)

☐ **CI:** Classified, as referred to in Commission Decision 2001/844/EC

## Table of content

Summary ..... 3
1. Introduction ..... 3
2. Results ..... 3
3. Partners involved in the work ..... 4
4. Annex ..... 5
4.1. Annex 1 ..... 5

## Summary

This paper corresponds to deliverable D4.4 of GenTORE workpackage 4 (*Genomic management tools to optimise resilience and efficiency).* This contribution describes a prototype of genomic evaluation in a real dairy population in three-way rotational crossbreeding (Montbéliarde x Holstein x Red Danish) and provides an estimate of its accuracy. The methodology used is a SNP-BLUP considering SNP effects defined according to the breed of origin of the alleles. Two software were developed: (a) to estimate breed of origin of each allele of all SNP of crossbreds; (b) to estimate SNP effect from purebred and crossbred data jointly. Prediction reliability in a validation population of the most recent crossbred animals varied from 0.38 to 0.70 according to traits, showing that practical implementation is possible.

## 1. Introduction

Crossbreeding strategy in dairy cattle breeding schemes has significantly grown over the past years, however only few countries routinely use genetic evaluation considering non-terminal crossbred animals in their breeding program. Due to limited linkage disequilibrium conservation across breeds, the SNP effects are not expected to be the same in all breeds. Therefore, genomic evaluation approaches based on the Breed of Origin of the Alleles (BOA) were identified as a promising tool to answer this question. These approaches are based on three steps: (a) impute and phase genomic information of crossbred animals, (b) accurately determine the breed of origin of the alleles of each markers, (c) run the genomic evaluation estimating an effect for each SNP in each breed.

## 2. Results

The imputation and phasing process relied on the FImpute program (Sargolzaei et al, 2014). This program is already used in pure breed in many situations and was found to be very accurate in bovine populations with a strong familial structure. To obtain a high imputation accuracy, a large reference population of purebred animals was included. Purebred animals included direct ancestors of crossbred animals and a set of major ancestors contributing to their breeds. With >20,000 purebred animals included, the collection of haplotypes was fairly complete and imputation was found to be very accurate. This work was performed at the very beginning of GenTORE.

Breeds of origins were determined from phased genotypes with the "*BreedOrigin*" fortran software developed in house. This software was designed to be very fast and with reasonable requirements in memory. For each haplotype of 16 markers, it determines the frequency in each pure breed and, when found in crossbreds, allocates it to the breed with the largest frequency. Possible gaps are filled when flanking segments are from the same origin. Undetermined segments are resolved by

considering smaller haplotypes. This approach was found to be very accurate and easy to use. The *BreedOrigin* software is available to GenTORE partners and was provided to Aarhus University for a comparison study. After the project, it will be made available under free licence.

Finally, genomic evaluation was carried out with an in-house fortran software which is an extension of the SNP-BLUP model but where a SNP effect is estimated for each breed.

This method was applied on a set of 5238 crossbred animals in addition of 20,000, 22,265 and 6,866 pure Montbéliard, Holstein and Red Danish animals respectively. Five production traits that are major determinants of efficiency were studied (Milk yield, Fat yield, Protein yield, Fat content and Protein content) through the use of Yield Deviation (YD) adjusted or not for Heterosis effect.

Regression coefficients between YD and genomic breeding values and associated slope of regression were studied. On the training population, all the correlations were around 0.80 for lactation milk yield, fat yield, and protein yield and around 0.90 for fat content and protein content. Adjusting for Heterosis had only a weak effect on these correlations. The associated slopes of regression were slightly higher than 1 for all the traits.

In the validation population, without integration of the Heterosis effect, correlations ranged from 0.34 for protein yield to 0.63 for fat content trait, corresponding to reliability estimates between 0.38 and 0.70. The gain of correlation obtained after adjusting for Heterosis was marginal (+1 point for lactation milk yield, fat yield, and protein yield). Initial regression slope estimates were higher than 1 both with and without adjustment of the performances for Heterosis (from 1.47 to 1.69). After the submission of the paper, it was found that calculating frequencies exclusively within purebred animals (see equation 2 in Annex 1) led to a strong improvement of the slope of regression which were all very close to 1.

This method is found to be efficient and relatively easy to implement when sufficient purebred data are available.

## 3. Partners involved in the work

INRAE, Allice, AU.

# 4. Annex

These results are presented in a communication by Croiseau et al. (2022) submitted to the World Congress of Genetics Applied to Livestock Production (Rotterdam, NL). The paper is in Annex 1.

## 4.1. Annex 1

**Accuracy of prediction for a genomic evaluation in rotational crossbreeding scheme (Montbéliarde x Holstein x Red Danish).**

**P. Croiseau[1*], R. Saintilan[1,2], A. Baur[1,2], I. Croue[1,2], V. Ducrocq[1], J.R. Thomasen[3], E. Karaman[4], D. Boichard[1], H. Leclerc[1,2] and B.C.D. Cuyabano[1]**

[1] Université Paris Saclay, INRAE, AgroParisTech, GABI, Domaine de Vilvert, 78350 Jouy-en-Josas, France; [2] ALLICE, Paris, France; [3] Viking Genetics, Ebeltoftvej 16, 8960 Randers SO, Denmark ; [4] Center for Quantitative Genetics and Genomics, Faculty of Technical Sciences, Aarhus University, 8830 Tjele, Denmark. [*]pascal.croiseau@inrae.fr

**Abstract**

In recent years, crossbreeding strategy has grown in dairy cattle farms at an international level. Breeders are interested in keeping crossbred cows in their herd both to combine the strengths of the pure breeds, compensate for their weaknesses and benefit from heterosis. However genetic tools are still lacking to manage these crossbred animals. In this study, we evaluate the performances of a genomic evaluation adapted for rotational crossbreeding schemes with real data. This genomic evaluation was applied to a population that includes pure-breed animals from Holstein, Montbéliarde, and Red Danish breeds, as well as crossbreds between these three breeds. The genomic evaluation approach was based on the estimation of breed specific SNP specific according to the Breed of Origin of the Alleles.

**Introduction**

Crossbreeding strategy in dairy cattle breeding schemes has significantly grown over the past years, and this trend may be reinforced in the future to face challenges linked to agroecology or to new constraints induced by climate change. Indeed, this strategy is particularly suitable to combine the strengths of the pure-breeds and to compensate for their weaknesses, being an efficient way for breeders to obtain more adaptable and robust animals, resulting in a more sustainable breeding system. Crossbreeding is also relevant to decrease inbreeding or, equivalently, generate heterosis.

Genomic evaluation in pure breed is common, it is developing in terminal crossbreeding, but only a few countries routinely evaluate crossbred animals in continuous crossbreeding programs. For most countries, breeders engaged in a crossbreeding process have information limited to purebred bulls and the raw performances of their crossbred females to manage their selection and mating. In this context, the literature dedicated to continuous crossbreeding relates, for the most part, to simulated data (Van Raden et al., 2021 ; Eiriksson et al., 2021; Karaman et al, 2021) and rarely to applications on real data (Sevillano et al., 2017).

In this study, we propose an application of genomic evaluation for continuous crossbreeding programs, through a SNP-BLUP based on Breed Origin of Alleles (BOA). This approach estimates specific SNP effect depending on its breed of origin. The rationale for this approach is that linkage disequilibrium between SNP and the causal mutation differs according to the breed; additionally, QTL effects may be breed dependent.

The data used in this analysis consisted of the first steps of a three-way crossbreeding scheme including Montbéliarde (Mo), Holstein (Ho) and Red Danish (RD) breeds. This data set was augmented by purebred data to ensure accurate estimation of within breed SNP effects. The procedure included several steps: imputation and phasing at the 50K level of all the animals; identification of BOA of each allele of crossbred animals; genomic evaluation on five production traits, including a cross-validation procedure. Accuracy of prediction and slope of regression were calculated.

**Materials & Methods**

***Genotypes.*** The data used in this study originated from France and Nordic Countries (NAV). They consisted of 5,238 genotypes of crossbred animals, and 20,000, 22,265, and 6,866 genotypes of pure Mo, Ho, and RD animals, respectively. 53,498 autosomal SNP markers were retained from the Illumina 50K chips used routinely in France for genomic selection. Imputation was carried out into two steps. In the first step, genotypes of purebreds were imputed with FImpute (Sargolzaei et al., 2014) with the pipeline used in the routine French national evaluation system and in the second step, these imputed genotypes were treated as known genotypes. In the second step, the crossbred genotypes were imputed with FImpute software using the purebred Ho, Mo, and RD as a reference, and without using any pedigree information.

***BOA of alleles.*** The frequency (f) of each haplotype of n consecutive SNPs was estimated by counting within each pure breed. BOA i was selected if $\frac{f_i}{\sum_{j=1}^{nbreed} f_j}$ was higher than a given threshold (0.9). The whole genome was scanned by a sliding window moving one SNP at a time. The initial value of n was set to 16 (leading to 65,536 theoretical haplotypic combinations). When a haplotype origin remained undetermined, the process was iterated after n was divided by 2. Undetermined origins surrounded by identical breed origins were assigned to this breed. When n=1, the small proportion of finally remaining undetermined origins were allocated according to allelic frequencies. This procedure was implemented in the in-house BreedOrigin fortran software and applied to a target population of 5,238 crossbred animals. The results were compared to expected breed proportions based on pedigree.

**Crossbreeding parameters.** Fractions of heterosis (H) and recombination losses (R) were calculated for each crossbred animal from proportions of breed origins of their parents based on pedigree. Values of the H and R coefficients were calculated as in Dechow et al. (2007) using the following formulae $H = 1 - \sum_{i=1}^{nb_{breed}} s_i d_i$ and $R = 1 - \sum_{i=1}^{nb\_breed} \frac{(s_i^2 + d_i^2)}{2}$ where $s_i$ and $d_i$ are the proportions of sire and dam genes from breed i, respectively. Heterosis and recombination losses were then estimated by regressing on H and R.

***Phenotypes.*** The genomic evaluation was a two-step procedure. In a first step, a polygenic model was used to estimate all non-genetic effects (herd-year, age-year and year-month of calving), heterosis and recombination losses (as regression coefficients) and heterogeneous

variances, as in Dezetter et al (2015). Unknown parent groups accounted for breed of origin, and breed composition of crossbreds was accounted for by the pedigree. Yield deviations (YD) produced by this model correspond to performances adjusted for fixed and non-genetic random effects. Five production traits were analysed: lactation milk yield, protein yield, protein content, fat yield, and fat content. Two batches of evaluations were performed, the first one accounting for heterosis and recombination losses, and the second one ignoring them.

***Crossbreed genomic evaluation accounting for BOA.*** To account for crossbred animals in a genomic evaluation model, we proposed an extension of the SNP-BLUP model where a SNP effect $\beta_i$ is estimated for each breed as described in the following model:

$$y_i = \sum_{b=1}^{nb\_breed} (p_{i,b}\mu_b) + \sum_{b=1}^{nb\_breed} \left( \sum_{j=1}^{nb\_SNP} (\beta_{i,j,b} X_{i,j,b}) \right) + e_i \quad (1)$$

where $y_i$ is the YD of the animal i, μ is a vector of means defined within each breed, $p_{i,b}$ is the proportion of breed b in the genome of individual i, estimated with the BOA approach. $X_{i,j,b}$ is the allele content of SNP j that originates from the breed b for animal i, centered for the allelic frequency of the SNP in breed b :

$$X_{i,j,b} = (k_{i,j,b} - n_{i,j,b} f_{j,b}) \quad (2)$$

where $k_{i,j,b}$ and $n_{i,j,b}$ are the number of "2" alleles and the total number of alleles of the SNP j that originates from breed b for the animal i, respectively; $f_{j,b}$ is the frequency of allele "2" of the SNP j in breed b.
To assess the performance of the genomic evaluation, the data set was divided into a training dataset with both genotypes and phenotypic records used to estimate SNP effects according to their BOA, and a validation data set for which predicted breeding values were computed using the effects obtained with the training dataset and then compared to the observed phenotypes. The validation population consisted of 2000 crossbred cows without progeny.

## Results

***Breed of origin of alleles.*** Results for breed composition based on pedigree information and BOA methodology are presented in Table 1. For genotyped crossbred animals, 94% of the alleles were assigned to a breed. 48% of markers were from Ho origin, 34% from RD origin and 13.6% from Mo origin. In comparison, based on pedigree information, the corresponding origins were 56.3%, 29.8% and 11.8% respectively. Correlation between breed compositions for both methodologies ranged from 0.95 for the RD breed to 0.99 for the Mo breed.

**Table 1. Results of determination of Breed of Origin based on genomic or Pedigree information**

| Breed of Origin | Breed composition with BOA | | Breed composition based on Pedigree | | |
|---|---|---|---|---|---|
| | *Mean* | *Sd* | *Mean* | *Sd* | *Correlation* |
| Montbéliarde (Mo) | 13.60% | 20.00% | 11.80% | 19.20% | 0.99 |
| Holstein (Ho) | 48.00% | 19.40% | 56.30% | 20.20% | 0.97 |
| Red Danish (RD) | 34.00% | 17.80% | 29.80% | 17.80% | 0.95 |

***Across breed genomic evaluation accounting for BOA.*** Correlations between YD and genomic breeding values and associated slope of regression are presented on Table 2. On the training population, all the correlations are around 0.80 for lactation milk yield, fat yield, and protein yield and around 0.90 for fat content and protein content. Adjusting for Heterosis weakly affected these correlations. The associated slopes of regression were slightly higher than 1 for all the traits. In the validation population, without integration of the Heterosis effect, correlations ranged from 0.34 for protein yield to 0.63 for fat content trait. The gain of correlation obtained after adjusting for Heterosis was marginal (+1 point for lactation milk yield, fat yield, and protein yield). Regarding the slopes of regression, they were overestimated both with and without adjustment of the performances for Heterosis (between 1.47 and 1.69).

**Table 2: Correlations and slopes of regression of the YD, adjusted or not for Heterosis for 5 production traits, on the estimated genomic breeding values, in the training and validation population.**

| Traits | performances without Heterosis effect | | | | performances adjusted for Heterosis effect | | | |
|---|---|---|---|---|---|---|---|---|
| | **Training Population** | | **Validation Population** | | **Training Population** | | **Validation Population** | |
| | *Corr* | *Slope* | *Corr* | *Slope* | *Corr* | *Slope* | *Corr* | *Slope* |
| milk yield | 0.85 | 1.11 | 0.38 | 1.62 | 0.84 | 1.12 | 0.39 | 1.66 |
| fat yield | 0.81 | 1.16 | 0.37 | 1.58 | 0.80 | 1.19 | 0.38 | 1.64 |
| protein yield | 0.80 | 1.18 | 0.34 | 1.63 | 0.78 | 1.20 | 0.35 | 1.69 |
| fat content | 0.93 | 1.11 | 0.63 | 1.47 | 0.93 | 1.11 | 0.63 | 1.47 |
| protein content | 0.94 | 1.08 | 0.59 | 1.50 | 1.00 | 1.08 | 0.59 | 1.50 |

## Discussion

In this study, we propose an application on real data of a genomic evaluation in a rotational crossbreeding scheme based on a SNP-BLUP model accounting for BOA.
This approach required to impute and phase genotyping data of crossbred animals in order to predict BOAs for these animals. These steps were tested and an average error rate of about 1.5% for the predicted BOA was observed (data not shown). To complete this information, we compare in this study the breed composition of crossbred animals measured using the pedigree information of the animals and estimated from the genotyped animals. The very strong correlations obtained between these two approaches are completely consistent with previous tests (Table 1).
Finally, the genomic evaluations carried out have shown an honourable accuracy of prediction of around 0.30 for traits with a heritability of 0.3 and around 0.60 for traits with a heritability of 0.50 (Table 2). The adjustment of the performances for Heterosis did not impact the accuracy of the prediction of the breeding values.

## Acknowledgments

This project has received funding from APIS GENE and from the European Union's Horizon 2020 research and innovation program - GenTORE - under grant agreement No. 727213.

## References

Christensen O.F., Madsen P., Nielsen B., & Su G. (2014) GSE 46(1):1-9. https://doi.org/10.1186/1297-9686-46-23.

Dechow C.D., Rogers G.W., Cooper J.B., Phelps M.I., & Mosholder A.L. (2007) JDS 90(7):3542-3549. https://doi.org/10.3168/jds.2006-889
Dezetter C., Leclerc H., Mattalia S., Barbat A. Boichard D. et al. (2015) JDS 98(7):4904-4913. https://doi.org/10.3168/jds.2014-8386
Karaman E., Su G., Croue I., and Lund M.S. (2021) GSE 53(1):1-15. https://doi.org/10.1186/s12711-021-00637-y
Sargolzaei M., Chesnais J.P., and Schenkel, F.S. (2014). BMC Genomics 15(1):1-12. https://doi.org/10.1186/1471-2164-15-478
Sevillano C.A., Vandenplas J., Bastiaansen J.W., Bergsma R., and Calus M.P. (2017) GSE 49(1):1-14. https://doi.org/10.1186/s12711-017-0350-1
VanRaden P.M., Tooker M.E., Chud T.C.S., Norman H.D., Megonigal Jr et al. (2020) JDS 103(2):1620-1631. https://doi.org/10.3168/jds.2019-16634