Genetic correlations: a parameter or a latent phenotype in genetic evaluations?

B.C.D. Cuyabano[1], P Croiseau[1], F Shokor[1,2], MR Motta[3], S Aguerre[1,4], and S Mattalia[1,4]

[1]INRAE – GABI, Université Paris Saclay, AgroParisTech, France
[2]Eliance, UMT eBIS, France
[3]University of Campinas, SP, Brazil
[4]Idele, UMT eBIS, France
Corresponding author: beatriz.castro-dias-cuyabano@inrae.fr

Abstract

Genetic correlations are relevant parameters in genetic evaluations, particularly when a breeding program aims to achieve genetic progress for multiple traits altogether. These correlations are usually estimated from a base population as one of the many parameters that define the distribution used to predict breeding values for the selection candidates. In such a fashion, genetic correlations are assumed to be identical for all selection candidates. However, with a preliminary study on the output predicted breeding values of sires with more than 500 daughters from the French Montbéliarde population, we observed that the genetic correlation among daughters from different sires may differ substantially, *i.e.*, different sires expressed different genetic correlations between traits through their daughters. Thus, if genetic correlations are specific values inherent to each individual, they could be considered as a phenotype; in other words, genetic correlations may be the observable consequence of a concealed regulatory trait guiding the relationship between observable traits. For antagonistic traits (*e.g.* production and fertility in dairy cattle), it is reasonable to believe that individuals on the extremes of the trade-off distribution are likely to present a low breeding value for this concealed regulatory trait. However, due to our inability to directly measure this potential regulatory trait, it can be considered a latent phenotype. Although a method to consider such hypothesis that genetic correlations may be a latent phenotype is yet undefined, there is no doubt that such hypothesis has an impact on the medium to long-term perspectives of a breeding program, given its breeding goals. Hypothesizing that genetic correlations are latent phenotypes, simulations can then be used to assess the genetic progress for multiple traits of interest in a breeding program over many generations, as well as to assess the trajectory of genetic correlations between traits and the genetic progress of the latent regulatory phenotype driving such correlations. Such comprehension of the genetic progress for the latent phenotype is of particular relevance, since a regulatory trait is likely to impact more than only two antagonistic traits, but many of the traits selected for in a breeding program.

Key words: correlated traits; multi-trait evaluation; physiological trait regulation; non-linear genetic correlation; genetic progress

Introduction

Genetic correlations (GC) are relevant parameters in genetic evaluations, particularly when a breeding program aims to achieve genetic progress for multiple traits altogether. These correlations are typically estimated from a base population as parameters from the joint distribution of the breeding values of the traits of interest, a distribution that is then used to predict the breeding values (BV) for the selection candidates (Patterson and Thompson, 1971; Henderson et al., 1959; Henderson, 1975). In such a fashion, GC are assumed to be identical for all evaluated individuals. In terms of statistical modelling, the assumption that

GC are population parameters, thus identical to all selection candidates, enables the implementation of the best linear unbiased prediction (BLUP) (Henderson, 1975) and Bayesian methods (Meuwissen et al., 2001; Gianola and van Kaam, 2008), widely used to predict BV in genetic evaluations.

While the assumption that GC is a population parameter is of great value to describe the underlying genetic architecture that drive the relationship between traits, such assumption ignores potential physiological genetic effects in the regulation of multiple traits (Berry et al., 2016). A preliminary study on the output predicted BV of sires with more than 500 daughters from a dairy cattle population, showed that the GC among daughters from different sires may differ substantially, *i.e.*, different sires may express different GC between traits through their daughters.

Physiological traits may impact both positive and negatively correlated traits. Our study focused on the latter case, particularly on the classic antagonism between production and fertility traits in dairy cattle (Boichard and Manfredi, 1994; Hoekstra et al., 1994; Veerkamp et al., 2001), traits of great commercial interest for this production system. Our hypothesis is that, rather than a modelling parameter, GC may be the observable consequence of an underlying physiological trait, responsible to regulate the trade-off between production and fertility, and such regulatory trait is not directly measurable.

Under this hypothesis that GC is a consequence of an underlying physiological trait, it is reasonable to believe that individuals on the extremes of a trade-off distribution (*i.e.* individuals who present a very high breeding value for production and a very low breeding value for fertility, or *vice-versa*) are likely to present a low breeding value for this concealed regulatory trait. Conversely, individuals on the center of a trade-off distribution, with average breeding values for both commercial traits, are likely to present a good regulatory capacity. Therefore, GC would be a measure inherent to each individual, representing their genetic capacity to regulate a trade-off. However, due to our inability to directly measure this potential regulatory trait, it can be considered as a latent phenotype, making it difficult to be evaluated and included in the unified index for the selection candidates.

Rather than aiming on how to include the hypothesis that GC are latent phenotypes in a genetic evaluation, the objective of the present work was to compare the genetic progress of production and fertility traits in a simulated breeding program, with data simulated under the assumptions that GC was either a parameter or a latent phenotype. Simulations were performed for different scenarios of selection, and the consequences of these scenarios on the regulatory trait and on the observed GC between the measurable traits was also studied.

Although the objective of our study was the discussion of the hypothesis that GC are latent phenotypes, without neither developing novel methods to evaluate antagonistic traits, nor proposing a manner to consider the possibility that GC are latent phenotypes in the unified index for the selection candidates, the discussion of this hypothesis is still relevant, since it sheds a light on the medium to long-term consequences of breeding decisions on the genetic progress of traits of interest.

Materials and Methods

Preliminary study on real data

The data set analyzed consisted of records from production (PROD) and fertility (FERT) traits from the French Montbéliarde population. PROD consisted of milk yield on first lactation corrected for 305 days, and FERT consisted of the cow conception rate at the first insemination after the beginning of the first lactation. Records on both traits were available for 806,159 cows, for which pedigree data with ~ 4 million animals were available. The phenotypes analyzed were recorded for

cows that began their first lactation between the years of 2002 and 2021.

The model used for both the variance component estimation and the genetic evaluation was a two-trait model (PROD and FERT), with an overall mean, age, and herd-year-season included as fixed effects for both traits; lactation length was included as a fixed effect only for PROD, and calving-insemination interval, sexed semen, artificial insemination (AI) operator, and day of the week as fixed effects for FERT only; for both traits, the random additive genetic effect was included assuming a normal distribution with mean zero and variances $\boldsymbol{A}\sigma^2_{PROD}$ and $\boldsymbol{A}\sigma^2_{FERT}$ for PROD and FERT respectively, and a covariance $\boldsymbol{A}\sigma_{P,F}$ between the two traits, such that $\boldsymbol{A}$ was the nominal relationship matrix, σ^2_{PROD} and σ^2_{FERT} were the total additive genetic variances of PROD and FERT respectively, and $\sigma_{P,F}$ was the additive genetic covariance between the two traits evaluated; the random effect of the AI bull was included into the model for FERT only, assuming independence between the bulls and a normal distribution with mean zero and variance $\sigma^2_{bull\times year}$; finally, for both traits the random residuals were considered to be normally distributed with mean zero and a heterogeneous variance per herd-year group.

Variance components were estimated using the residual maximum likelihood (REML), and the genetic evaluation was performed using the BLUP, to obtain the estimated breeding values (EBV) for all animals in the pedigree. After the evaluation model was performed, from the pedigree we subset 247 sires with more than 500 daughters evaluated among those 806,159 cows with records on both traits. For each of these sires, we calculated the mean EBV of their daughters for both PROD and FERT per year of their first lactation, and the GC between their daughters' EBVs for PROD and FERT, over all the years, and separated by year of the daughters' first lactation. This descriptive study per sire was performed to confirm the genetic progress for both traits (thus, selection for both PROD and FERT), and to verify whether different sires expressed different GC between the traits of interest, through their daughters.

Simulation study

Datasets were simulated to contemplate the two hypotheses we intended to discuss with this present study: (1) GC are statistical parameters modulating the genetic relationship between two traits; (2) GC are observable consequences from a latent physiological trait (RGLT) that regulates the genetic relationship between two traits. Under both hypotheses, PROD and FERT were simulated with heritabilities $h^2_{PROD} = 0.3$ and $h^2_{FERT} = 0.04$ respectively, and with a GC $\rho_{P,F} = -0.2$ between them. The total phenotypic variances were 50 and 75 for PROD and FERT respectively.

For the simulated datasets under each of the two hypotheses, a base population with 2,000 individuals was simulated, with 50k SNPs allocated in 29 chromosomes, such that the number of SNPs per chromosome and the linkage disequilibrium (LD) pattern were adjusted to resemble the cattle genome.

Genomic data simulations were performed in R language (R core team, 2018), using routines from the GenEval package (https://github.com/bcuyabano/GenEval), and correlated traits were simulated using self-coded routines. All evaluations were also performed in R language.

Selection scenarios with different weights for the two traits in the breeding goals were defined to evolve the population over many generations. For every simulated scenario, selection was performed on sires only, by selecting the top 20% bulls in agreement with the scenario's breeding goal. The choice of selection of sires only was made so that the simulation resembled a dairy cattle breeding program.

Genetic correlation as parameter

When GC was simulated as a parameter, its origin was purely quantitative. This means that part of the correlation was due to pleiotropic quantitative trait loci (QTL) for both PROD and FERT, and part of this correlation was due to different QTL for each trait, but that were in close proximity, such that these QTL were in a sufficient level of LD for GC to arise. For each trait, 3,000 out of the 50k simulated SNPs were assigned as QTL; 1,000 of these QTL were shared by both traits (pleiotropic); 1,000 QTL were trait-specific, but in close proximity to those trait-specific from the other trait so that these QTL were in LD; 1,000 QTL were trait-specific and far enough from any QTL from the other trait, so that their between-trait effects were completely independent. Figure 1 illustrates the described scheme of the QTL display on the simulated genome.

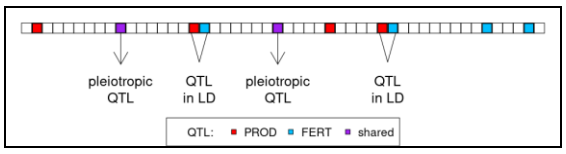

Figure 1. Scheme of the QTL display on the simulated genome, indicating the QTL responsible for creating genetic correlation between PROD and FERT (pleiotropic QTL, and QTL in LD), and the QTL that created independent effects on each trait.

For the simulation study under the hypothesis that GC was a parameter, the population was evolved over 40 generations under five different scenarios, one scenario in completely random mating, and four scenarios with selection of the top 20% bulls, with different weights for %PROD-%FERT in the breeding goal: (1) 100-0; (2) 90-10; (3) 80-20; and (4) 50-50. Each scenario was replicated 100 times.

Genetic correlation as a latent phenotype

When GC was simulated as a latent phenotype, we initially simulated RGLT was with heritability $h^2_{RGLT} = 0.1$, and then both PROD and FERT were simulated to have a concave parabolic relationship with RGLT, following the simulation method in Shokor et al. 2024. Figures 2 and 3 illustrate the relationship between the simulated BV for the three traits.

For the simulation study under the hypothesis that GC was a latent phenotype, the population was evolved over 50 generations under three selection scenarios of the top 20% bulls, with different weights for %PROD-%FERT-%RGLT in the breeding goal: (1) 100-0-0; (2) 80-20-0; and (3) 80-10-10. Each scenario was replicated 1,000 times.

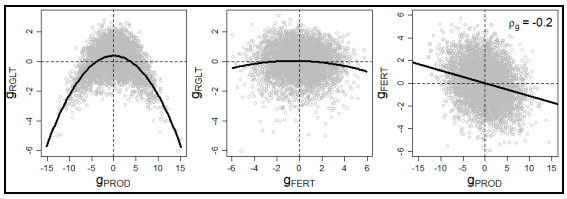

Figure 2. Scatterplots of the BV, denoted as g, simulated for the three traits when the genetic correlation between PROD and FERT was the consequence of a latent phenotype. The full black line in all the panels indicate the mean relationship between the pairs of traits.

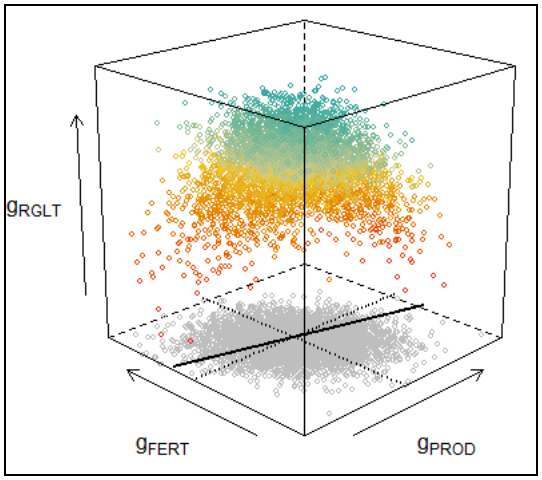

Figure 3. 3D scatterplot of the BV (colored dots, with the gradient red-yellow-blue representing negative-zero-positive values for RGLT), denoted as g, simulated for the three traits when the genetic correlation between PROD and FERT was the consequence of a latent phenotype. The gray dots are the projection of the simulated BV for PROD and FERT only, which are the observable traits. The full black line indicates the mean relationship between PROD and FERT, perceived as a linear correlation.

Results & Discussion

Descriptive results on real data

On the group of 247 subset sires, we could observe a clear pattern of genetic progress from 2002 for PROD, and from 2009 for FERT, as shown in Figure 4. Although FERT has been included in the breeding goals for the French Montbéliarde population in 2001, at this moment this breeding goal was defined mostly for the AI sires. Therefore, the genetic progress is not immediately perceived. Given the low heritability of FERT and considering the generation interval needed for a change in breeding goals to take effect, it not surprising that the clear pattern of genetic progress arises from 2009. Moreover, to further explain the trajectory of the genetic progress for FERT, it is only from 2006 that females began to be more systematically selected for fertility traits.

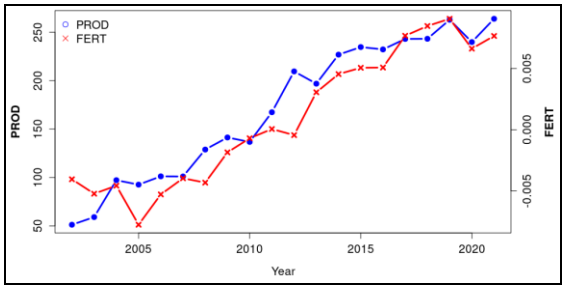

Figure 4. Yearly mean breeding values of PROD and FERT from 2002 until 2021, for the daughters of 247 sires with more than 500 daughters evaluated, all with both traits recorded.

With respect to the sire-specific GC between PROD and FERT, from Figure 5 we can observe that their values range from -0.3 to 0.3, a great dispersion around the GC of 0.051 estimated by REML. This dispersion is observed both on the sire-specific GC disregarding their daughters' year of birth and taking the year into account. When observing the distributions of the sire-specific GC per their daughters' year of birth, we observed that this distribution changes most visibly from 2009. From the year 2002 until 2008, sire-specific GC were on average negative, with a mean of -0.055. From 2009 on, these mean shifts to approach zero, and finally become mildly positive, with a mean of 0.063.

Based on our knowledge of the historical breeding goals for the French Montbéliarde population in the period from 2002 to 2021, and the observed response in genetic progress for PROD and FERT, it is then of no surprise that visible changes in the distribution of the sire-specific GC arise from 2009, as shown in Figure 5.

The great dispersion of the sire-specific GC between PROD and FERT suggests that considering these correlations as a static parameter may not reflect the true nature of what drives the relationship between PROD and FERT.

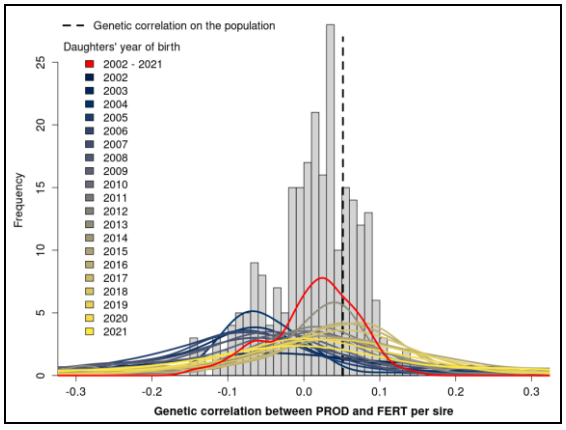

Figure 5. Histogram of the sire-specific genetic correlations (GC) between PROD and FERT, for 247 sires with more than 500 daughters evaluated, all with both traits recorded.

Genetic progress and genetic correlation on simulated data

The different assumptions of what causes GC between the antagonistic traits have a remarkably different impact on the trajectory of GC over generations for populations under selection.

When GC was simulated as a parameter, different selection scenarios generally presented significantly different outcomes. Under this hypothesis, a single-trait selection resulted in an attenuation of GC, *i.e.*, the negative GC evolved towards zero as generations progressed, as shown in Figure 6. Although we solely present the trajectory of

the GC for single-trait selection on PROD, the exact same trajectory was observed when selection was performed for FERT only.

Still under the hypothesis that GC is a parameter, with the exception of the scenario 90%PROD-10%FERT, for which GC remained stable around its original value, selection scenarios for both PROD and FERT inevitably lead to an intensification of GC, *i.e.*, the negative GC evolved to a farther more negative value as generations progressed, as shown in Figure 6. Moreover, the greater the equilibrium in the breeding goal between the two antagonistic traits, the faster this intensification of GC was observed.

Although the observed results in Figure 6 were initially surprising, these trends were statistically supported when we performed the calculus on the expected GC for the truncated bivariate normal distribution to select progressively increasing values for both means (calculus not shown), and can be explained in terms of loss of genetic diversity, assuming that the hypothesis that GC is a parameter is true, *i.e.*, that GC arises uniquely due to QTL effects. Nonetheless, questions remained about whether the observed trends in the simulated data were biologically sound, specially when compared to the results observed with the real data, as presented in Figures 4 and 5. This, combined with research in bovine physiology (Berry et al., 2016) lead us to hypothesize that genetic correlations may be a latent phenotype, or in other words, the observable consequence of a concealed physiological trait, responsible to regulate the trade-off between the antagonist traits.

When GC was simulated as the consequence of a latent phenotype, as shown in Figure 7, we observed that the different selection scenarios did not present the great differences as previously observed in Figure 6. In fact, in the short to medium term (up to approximately generation 15), the trajectory of GC was statistically the same for all selection scenarios. During the first 15 generations

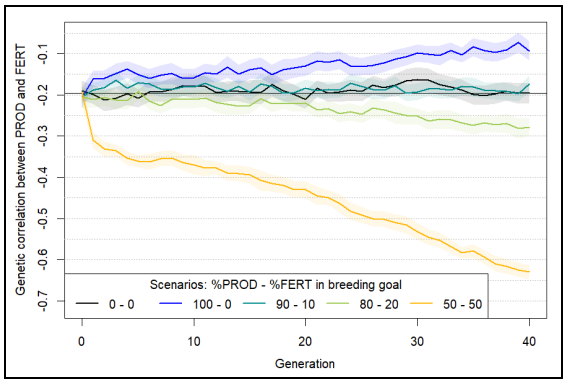

Figure 6. Trajectory of GC between PROD and FERT over 40 generations of populations under selection according to the five simulated scenarios of breeding goals (%PROD-%FERT). The full lines in the plot represent the mean GC observed with 100 replicates of each scenario, and the shaded area around the mean GC represent their 95% confidence interval.

under selection, the negative GC tended to be attenuated.

From generation 15 on, the overall trend of GC under selection for a single trait differs from that of selection for both traits. While under all simulated selection scenarios the GC reached a peak of attenuation, and then presented a trend of slow re-intensification, this trend seemed to be temporary when selection was performed for more than one trait, with GC reaching an apparent stabilization in its trend, from generation 30. When selection was performed for a single trait (in our simulations, PROD), the trend of re-intensification seemed constant throughout all generations after generation 15. These results presented in Figure 7 suggest that, although initially any breeding goal for a breeding program will lead to the attenuation of GC, in the long-term, single-trait selection will inevitably lead to stronger negative GC, compared to multi-trait selection.

To conclude the discussion with respect to the trends observed for the GC over generations in the simulated populations under different selection scenarios, the contrasting results from the two hypotheses considered to

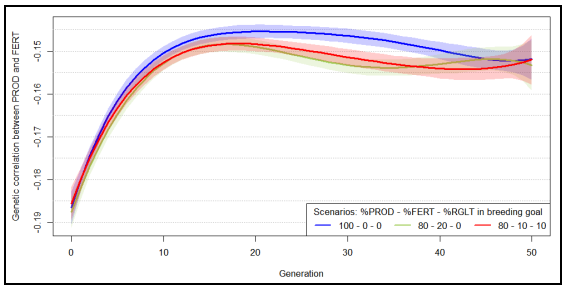

Figure 7. Trajectory of GC between PROD and FERT over 50 generations of populations under selection according to the three simulated scenarios of breeding goals (%PROD-%FERT-%RGLT). The full lines in the plot represent the mean GC observed with 1,000 replicates of each scenario, and the shaded area around the mean GC represent their 95% confidence interval.

simulate correlated traits presented in Figures 6 and 7, when compared to the results observed in real data, gives us information to support the hypothesis that GC is a consequence of a latent phenotype.

Finally, we evaluated the genetic progress achieved with the different selection scenarios, when GC was simulated under the hypothesis that they are a consequence of a latent phenotype. The results presented in Figure 8 show that, as expected, after 50 generations the average BV for production was mildly lower when multi-trait selection was in place. Since breeding goals for multi-trait selection kept a weight of 80% for production, although significant, the difference in PROD between the simulated scenarios was small. Therefore, the inclusion of FERT and RGLT did not largely decrease PROD. On the other hand, the inclusion of FERT alone, or FERT+RGLT to the breeding goal resulted in great changes to the average BV for these two traits, compared to the scenario in which selection was performed uniquely for PROD.

It was interesting to observe that, in the scenario for which selection was performed for PROD and FERT (without RGLT), the dual selection did impact positively the genetic progress of RGLT. Although not surprising, this result is reassuring that, if GC are the observed consequence of a latent physiological trait, selection for the observable traits is indirectly selecting for the latent trait.

To conclude the discussion with respect to the genetic progress, a final remark has to be done with respect to the selection including the latent physiological trait (RGLT) responsible to regulate the trade-off between PROD and FERT. If RGLT can be measured either directly or indirectly, and the included in the breeding goals, the genetic progress of this trait is relevant, counterbalancing the mild loss in genetic progress for the other traits of commercial interest (in the simulation PROD and FERT). The great genetic progress in RGLT due to its inclusion in the breeding goal has an importance, because such trait is very likely to have an influence in many other traits of commercial interest, beyond PROD and FERT. Thus, including RGLT in the breeding goals is expected to improve many of the traits considered in a real breeding program, which are far more traits than PROD and FERT.

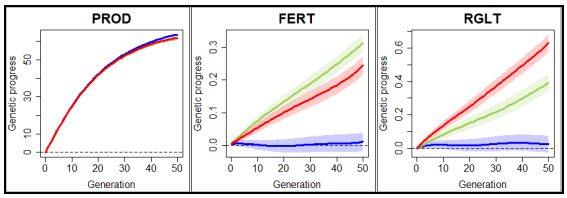

Figure 8. Genetic progress per generation, for PROD, FERT, and RGLT over 50 generations of populations under selection according to the three simulated scenarios of breeding goals (color-coded as in Figure 7, *i.e.* blue: 100%PROD-0%FERT-0%RGLT; green 80%PROD-20%FERT-0%RGLT; red 80%PROD-10%FERT-10%RGLT). The full lines in the plot represent the mean breeding values observed with 1,000 replicates of each scenario, and the shaded area around the mean GC represent their 95% confidence interval.

Conclusions

This work had the objective open a discussion about the nature of genetic correlations between traits. We evoked two hypothesis, one that assumes genetic correlations as a parameter driving the genetic architecture of correlated traits, and another that assumes that genetic correlations are the observable

consequence of a latent physiological trait responsible to balance the expression of measurable traits. Using simulations of breeding schemes considering different breeding goals under these two hypotheses, and comparing our simulated medium to long-term results with observations in real data, we believe that our study provides information to support the hypothesis that GC are a consequence of a latent phenotype. This hypothesis is relevant to define breeding objectives, since a regulatory trait may impact not two, but many traits altogether. Last but not least, although our study focused on the antagonism between production and fertility traits, the concept that genetic correlations may be the consequence of a latent phenotype can be extended to many other traits.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°101000226.

References

Berry DP, Friggens NC, Lucy M, and Roche JR. 2016. Milk production and fertility in cattle. *Annual Rev of Anim Biosci*. 4:269-90.

Boichard D and Manfredi E. 1994. Genetic analysis of conception rate in French Holstein cattle. *Acta Agric Scand A Anim Sci*. 44:138–45.

Gianola D and van Kaam JBCHM. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*. 178:2289-303.

Henderson CR, Kempthorne O, Searle SR, and von Krosigk CM. 1959. The estimation of environmental and genetic trends from records subject to culling. *Biometrics*.15:192-218.

Henderson CR. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics*.31:423-47.

Hoekstra J, van der Lugt AW, van der Werf JHJ, and Ouweltjes W. 1994. Genetic and phenotypic parameters for milk production and fertility traits in upgraded dairy cattle. *Livest Prod Sci*. 40:225–32.

Meuwissen TH, Hayes BJ, and Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*.157:1819-29.

Patterson HD and Thompson R. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika*.58:545-54.

R Core Team. 2018. R: A language and environment for statistical computing. Vienna. R foundation for statistical computing.

Shokor F, Croiseau P, Gangloff H, Saintilan R, Tribout T, Mary-Huard T, and Cuyabano BCD. 2024. Predicting nonlinear genetic relationships between traits in multi-trait evaluations by using a GBLUP-assisted Deep Learning model. *Biorxiv pre-print*. https://www.biorxiv.org/content/10.1101/2024.03.23.585208v1

Veerkamp RF, Koenen EPC, and De Jong G. 2001. Genetic correlations among body condition score, yield, and fertility in first-parity cows estimated by random regression models. *J Dairy Sci*. 84:2327–35.