Measures to quantify the accuracy and the erosion of genomic predicted breeding values

B.C.D. Cuyabano[1*], D. Boichard[1] and C. Gondro[2]

[1] Université Paris Saclay, INRAE, AgroParisTech, GABI, Domaine de Vilvert, 78350 Jouy-en-Josas, France; [2] Department of Animal Science, Michigan State University, 474 S Shaw Ln, East Lansing, MI 48824, USA; [*]beatriz.castro-dias-cuyabano@inrae.fr.

Abstract

Genomic predicted breeding values (GPBV) to select animals in breeding programs are nowadays routinely adopted by most commercial livestock production systems. Accuracy of the GPBV (defined here as their correlation with the true phenotypes) has a theoretical limit of $\sqrt{h^2}$ that is achieved when, among other factors, available SNPs are in sufficient LD with the QTL. However, even under such assumptions, realized accuracies of GPBV remain generally below $\sqrt{h^2}$, because allele frequencies and LD patterns differ between reference and target populations (particularly noticeable across generations), resulting in the so-called erosion of SNP effects and consequently, erosion of the accuracy of the GPBV. We present here a measure to quantify the erosion of the GPBV's accuracy through the genomic correlation between reference and target populations and validate this measure through simulations.

Introduction

Nowadays, GPBV are routinely used for the selection of animals in many commercial livestock breeding programs. The accuracy of GPBV is therefore a very important factor for the success of a breeding program. However, realized accuracies of GPBV remain below the theoretical maximum ($\sqrt{h^2}$), even when the reference population is sufficiently large, and SNPs included in the model are in sufficient LD with the QTL. That is particularly noticeable over generations, as we observe the so-called erosion of SNP effects accompanied by the erosion of the GPBV's accuracy. Erosion occurs mostly due to differences in LD patterns and allele frequencies between reference and target populations; for example, if in the reference population a SNP is in strong LD with the QTL, a large effect will be assigned to it. However, if due to segregation over the generations, the LD between this SNP and the QTL becomes weaker in the target population, an effect closer to zero should be assigned to this SNP. The decay in prediction accuracy due to differences in allele frequencies and LD patterns, specially over generations, is a topic widely known and discussed by animal breeders and quantitative geneticists (Daetwyler *et al.*, 2008; Habier *et al.* 2013; Wientjes *et al.*, 2015, 2016; Pszczola and Calus, 2016; van den Berg *et al.*, 2019; Dekkers *et al.*, 2021). Quantifying the erosion at the individual SNP level is in fact, a difficult and unresolved task. It is, however, more tractable to quantify the erosion of the accuracy of the GPBV through a metric based on the genomic correlation between reference and target populations, which we present in this paper and validate through simulations.

Materials & Methods

Consider the genomic model $\boldsymbol{y}_1 = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{M}_1\boldsymbol{\alpha} + \boldsymbol{\varepsilon}_1$, in which $\boldsymbol{y}_1$ are the phenotypes measured in the reference population, $\boldsymbol{\beta}$ are the fixed effects and $\boldsymbol{X}_1$ their design matrix in the reference population, $\boldsymbol{M}_1$ is the (centred) SNP-genotypes matrix of the reference population, $\boldsymbol{\alpha} \sim N(\boldsymbol{0}, \boldsymbol{I}_m\sigma_\alpha^2)$ are the SNP effects, and $\boldsymbol{\varepsilon}_1 \sim N(\boldsymbol{0}, \boldsymbol{I}_n\sigma_\varepsilon^2)$ are the random residuals. Solving Henderson's mixed model equations (HMME) (Henderson *et al.*, 1959), we have the estimates $\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}_1'(\boldsymbol{M}_1\boldsymbol{M}_1'\sigma_\alpha^2 + \boldsymbol{I}_n\sigma_\varepsilon^2)^{-1}\boldsymbol{X}_1]^{-1}\boldsymbol{X}_1'(\boldsymbol{M}_1\boldsymbol{M}_1'\sigma_\alpha^2 + \boldsymbol{I}_n\sigma_\varepsilon^2)^{-1}\boldsymbol{y}_1$ and $\widehat{\boldsymbol{\alpha}} = \boldsymbol{M}_1'(\boldsymbol{M}_1\boldsymbol{M}_1'\sigma_\alpha^2 +$

$\boldsymbol{I}_n\sigma_\varepsilon^2)^{-1}\boldsymbol{y}_1^*\sigma_\alpha^2$, such that $\boldsymbol{y}_1^* = \boldsymbol{y}_1 - \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}$. Finally, the fitted breeding values for the n_1 animals in the reference population are $\widehat{\boldsymbol{g}}_1 = \boldsymbol{M}_1\widehat{\boldsymbol{\alpha}}$, and let $\boldsymbol{M}_2$ be the (centred) SNP-genotypes matrix of the target population, the GPBV for the n_2 animals in the target population are $\widetilde{\boldsymbol{g}}_2 = \boldsymbol{M}_2\widehat{\boldsymbol{\alpha}}$.

Theoretical limit of GPBV's accuracy without accounting for erosion. Our interest lies on the accuracy of the GPBV, *i.e.* on $R = \widehat{cor}(\widetilde{\boldsymbol{g}}_2, \boldsymbol{y}_2^*)$, such that $\boldsymbol{y}_2^* = \boldsymbol{y}_2 - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}$. Using Fisher's z-transformation on R, we have that $Z = log\left(\frac{1+R}{1-R}\right) \sim N\left(log\left(\frac{1+\rho}{1-\rho}\right), \frac{4}{n_2-3}\right)$, with ρ being the true accuracy, *i.e.* $\rho = \sqrt{h^2}$. By reversing Fisher's z-transformation $R = \varphi(Z) = \frac{e^Z-1}{e^Z+1}$ is a function of Z, and using Jensen's inequality we show that $E[R|\text{no erosion}] = E[\varphi(Z)|\text{no erosion}] \leq \varphi(E[Z]|\text{no erosion}) = \frac{e^{E[Z]}-1}{e^{E[Z]}+1} = \left[e^{log\left(\frac{1+\rho}{1-\rho}\right)} - 1\right] \Big/ \left[e^{log\left(\frac{1+\rho}{1-\rho}\right)} + 1\right] = \rho = \sqrt{h^2}$.

Theoretical limit of GPBV's accuracy accounting for erosion. We hypothesized that the genomic correlation between reference and target populations (r) affects linearly the mean of the z-transformed correlations, resulting in $Z = log\left(\frac{1+R}{1-R}\right) \sim N\left(r\ log\left(\frac{1+\rho}{1-\rho}\right), \frac{4}{n_2-3}\right)$. Using Jensen's inequality again we show that, under our hypothesis, $E[R|\text{erosion}] \leq \frac{\left(1+\sqrt{h^2}\right)^r - \left(1-\sqrt{h^2}\right)^r}{\left(1+\sqrt{h^2}\right)^r + \left(1-\sqrt{h^2}\right)^r} \underset{r\to 1}{\longrightarrow} \sqrt{h^2}$.

Erosion of the GPBV. We consider the erosion of the GPBV (δ_{GPBV}) as the difference between $\sqrt{h^2}$ and the theoretical limit of GPBV's accuracy accounting for erosion, scaled by $\sqrt{h^2}$,

$$\delta_{GPBV} = 1 - \frac{E[R|\text{erosion}]}{\sqrt{h^2}}. \quad (1)$$

Genomic correlation between populations. Let $\boldsymbol{M}$ be a (centred) SNP-genotypes matrix ($n \times m$), with singular-value decomposition (SVD) $\left(\sum_{j=1}^m 2p_j(1-p_j)\right)^{-1/2}\boldsymbol{M} = \boldsymbol{UDV}'$, such that $p_j's$ are the alleles frequencies. In this SVD, $\boldsymbol{D}$ is a diagonal matrix of the $k = rank(\boldsymbol{M})$ singular-values, and $\boldsymbol{U}_{n\times k}$ and $\boldsymbol{V}_{m\times k}$ are matrices of unitary eigen-vectors relative to the individuals and to the SNPs, respectively. Knowing that components in $\boldsymbol{D}^2$ explain the variation of the whole system $\boldsymbol{M}$, each U_{ik} explains the contribution of the $i-th$ individual to the $k-th$ component, and each V_{jk} explains the contribution of the $j-th$ SNP to the $k-th$ component. To obtain the genomic correlation between reference and target populations (r), we need to quantify the different contribution of the SNPs to the system's variation in the two populations. To do so, we perform the aforementioned SVD on $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$, then build a matrix $\boldsymbol{T} = \sqrt{(n_2/n_1)}\boldsymbol{V}_2'\boldsymbol{V}_1\boldsymbol{D}_1$ which correlates the contributions of the SNPs in both populations, while correcting for the different population sizes, weighting these correlations by the singular-values on the reference population. Next, we obtain the SVD $\boldsymbol{T} = \boldsymbol{U}_T\boldsymbol{D}_T\boldsymbol{V}_T$, and perform the linear regression $\boldsymbol{D}_T \sim \boldsymbol{D}_2$ with a quadratic term, *i.e.*, we fit $d_{Ti} = a + bd_{2i} + cd_{2i}^2$. Finally, we obtain the genomic correlation between reference and target populations as $r = a + b + c$.

Simulation study. We simulated 50K SNPs and additive phenotypes for $h^2 = 0.05, 0.15, \ldots, 0.9, 0.95$, using 2K SNPs as QTL. A base reference population of 5,000 individuals was used to estimate variance components using REML (Patterson and Thompson, 1971) and SNP effects $\widehat{\boldsymbol{\alpha}}$ (solving HMME). Using $\widehat{\boldsymbol{\alpha}}$ we obtained the GPBV for three different target populations (1,000 individuals each) with increasing generation distances (one, five, and ten) from the

reference population. Then we compared our theoretical measures $E[R|\text{erosion}]$ and δ_{GPBV} with the realized prediction accuracies ($R = \widehat{cor}(\tilde{\boldsymbol{g}}_2, \boldsymbol{y}_2^*)$) and erosion ($\hat{\delta} = 1 - R/\sqrt{h^2}$).

Results

In Figure 1 we observe that as the generation distance between reference and target populations increases, the relationship $d_{2i} \times d_{Ti}$ moves further from the identity line and becomes more quadratic, resulting in decreasing genomic correlations as the generation distances increase: $r = 0.49$, $r = 0.42$ and $r = 0.33$ for target populations respectively one, five and ten generations apart from the reference population.

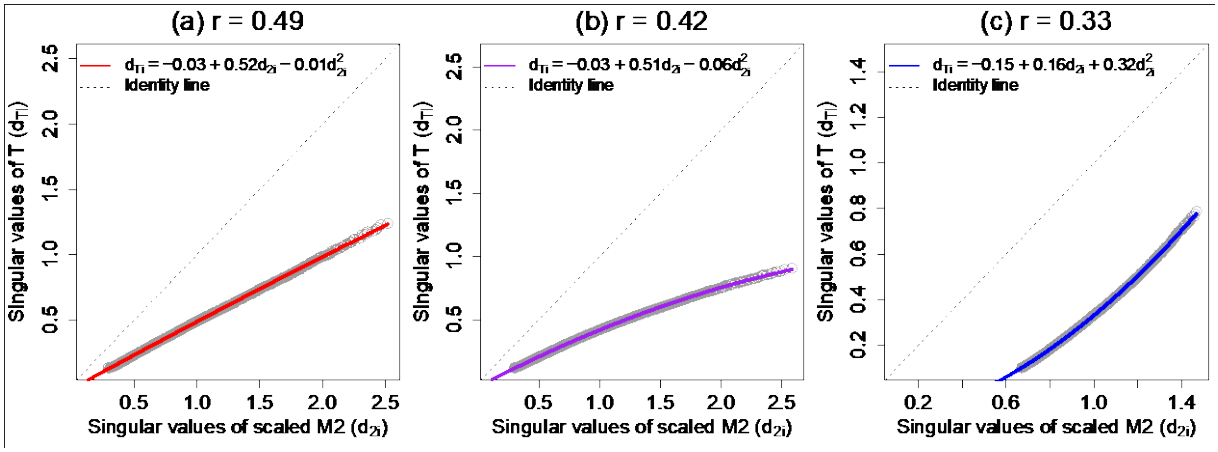

Figure 1. Singular values of the scaled M_2 (d_{2i}) vs. singular values of $T = \sqrt{(n_2/n_1)}V_2'V_1D_1$ (d_{Ti}), and estimated coefficients fitting the model $d_{Ti} = a + bd_{2i} + cd_{2i}^2$ for the target populations (a) one, (b) five, and (c) ten generations apart from the reference population.

In Figures 2 and 3 we observe that our theoretical curve $E[R|\text{erosion}]$ accurately fits the realized $\hat{h}^2$ and $R = \widehat{cor}(\tilde{\boldsymbol{g}}_2, \boldsymbol{y}_2^*)$, and that δ_{GPBV} is quite accurate to assess the erosion. For extreme h^2 (<0.1 and >0.85) however, there is still a small over expectation for R, which results in δ_{GPBV} being lower than the observed $\hat{\delta}$. Low heritability traits ($h^2 < 0.2$) also present a large $Var(\hat{\delta})$, indicating that precise assessment of erosion may be difficult for such traits.

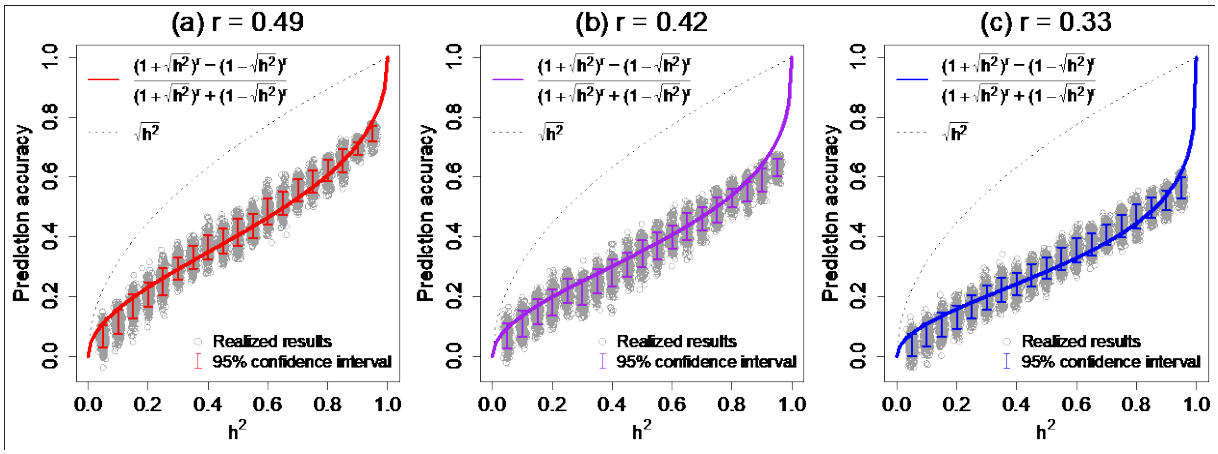

Figure 2. Theoretical $E[R|\text{erosion}]$ (full coloured line) and $E[R|\text{no erosion}]$ (dashed line), and the realized results for the target populations (a) one, (b) five, and (c) ten generations apart from the reference population.

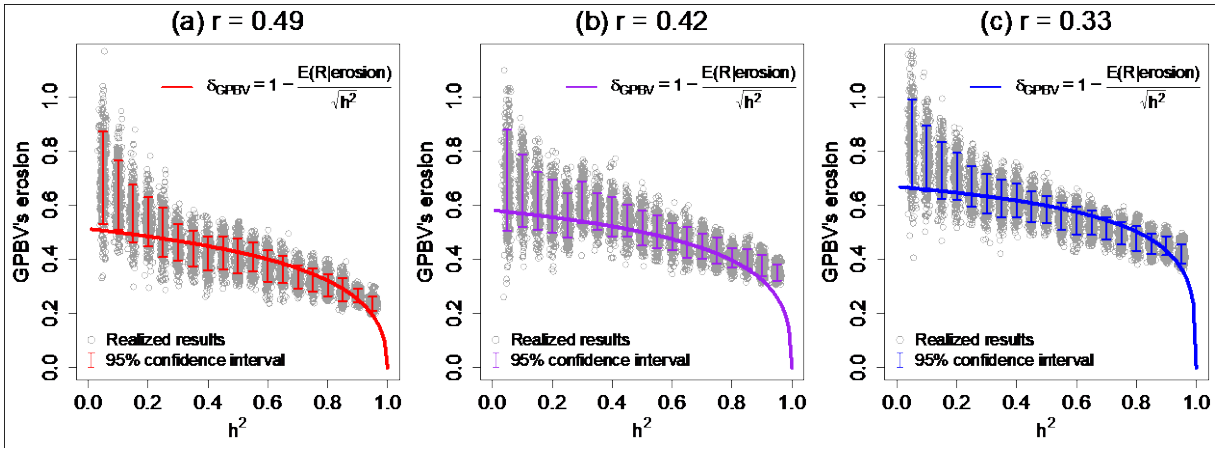

Figure 3. Theoretical erosion (δ_{GPBV}) and realized results for the target populations (a) one, (b) five, and (c) ten generations apart from the reference population.

Discussion

We hypothesized that once we know the genomic correlation between reference and target populations (r), we can define the maximum accuracy of the GPBV as $E[R|\text{erosion}] \leq \frac{(1+\sqrt{h^2})^r-(1-\sqrt{h^2})^r}{(1+\sqrt{h^2})^r+(1-\sqrt{h^2})^r} \underset{r\to 1}{\longrightarrow} \sqrt{h^2}$, and the results obtained with our simulations support this hypothesis. The measure we defined to quantify the erosion of the accuracy of the GPBV ($\delta_{GPBV} = 1 - E[R|\text{erosion}]/\sqrt{h^2}$) was also quite accurate, although it may be imprecise or underestimated for low heritability traits ($h^2 < 0.2$). One important element for calculating $E[R|\text{erosion}]$ and δ_{GPBV} is the genomic correlation r, a single value capable to summarize all the information about the differences in allele frequencies and LD patterns observed in both the reference and target populations. Although computationally costly for large populations and dense genotype data, our results indicate that the genomic correlation r obtained with the SVDs of genomic matrices is trustworthy. This work focused on the calculus of $E[R|\text{erosion}]$ and δ_{GPBV}, and did not explore the underlying meaning of the values of the coefficients a, b, and c that compose r, *i.e.* how allele frequencies, LD patterns, number of SNPs and population sizes affect each of these coefficients, but we do understand that such a study is relevant.

References

Daetwyler H.D., Villanueva B. and Wooliams J.A. (2008) PloS One 3(10):e3395. 10.1371/journal.pone.0003395.

Dekkers J., Su H. and Cheng J. (2021) Genet Sel Evol 53:55. 10.1186/s12711-021-00647-w.

Habier D., Fernando R.L. and Garrick D.J. (2013) Genetics 194(3):597-607. 10.1534/genetics.113.152207.

Henderson C.R., Kempthorne O., Searle S.R. and von Krosigk C.M. (1959) Biometrics 15(2):192. 10.2307/2527669.

Patterson H. and Thompson R. (1971) Biometrika 58(3):545. 10.2307/2334389.

Pszczola M. and Calus M.P.L. (2016) Animal 10(6):1018-1024. 10.1017/S1751731115002785.

Van den Berg I., Meuwissen T.H.E., MacLeod I.M. and Goddard M.E. (2019) J Dairy Sci. 102(4):3155-3174. 10.3168/jds.2018-15231.

Wientjes Y.C.J., Veerkamp R.F., Bijma P., Bovenhuis H., Schrooten, C. *et al.* (2015) Gen Sel Evol 47(1):5. 10.1186/s12711-014-0086-0.

Wientjes Y.C.J., Bijma P., Veerkamp R.F. and Calus M.P.L. (2016) Genetics 202(2):799-823. 10.1534/genetics.115.183269.