Expected values for the accuracy of predicted breeding values accounting for genetic differences between reference and target populations

Beatriz C. D. Cuyabano, Didier Boichard, Cedric Gondro

Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Expected values for the accuracy of predicted breeding values accounting for genetic differences between reference and target populations

Beatriz C. D. Cuyabano[1*], Didier Boichard[1] and Cedric Gondro[2]

Abstract

Background Genetic merit, or breeding values as referred to in livestock and crop breeding programs, is one of the keys to the successful selection of animals in commercial farming systems. The developments in statistical methods during the twentieth century and single nucleotide polymorphism (SNP) chip technologies in the twenty-first century have revolutionized agricultural production, by allowing highly accurate predictions of breeding values for selection candidates at a very early age. Nonetheless, for many breeding populations, realized accuracies of predicted breeding values (PBV) remain below the theoretical maximum, even when the reference population is sufficiently large, and SNPs included in the model are in sufficient linkage disequilibrium (LD) with the quantitative trait locus (QTL). This is particularly noticeable over generations, as we observe the so-called erosion of the effects of SNPs due to recombinations, accompanied by the erosion of the accuracy of prediction. While accurately quantifying the erosion at the individual SNP level is a difficult and unresolved task, quantifying the erosion of the accuracy of prediction is a more tractable problem. In this paper, we describe a method that uses the relationship between reference and target populations to calculate expected values for the accuracies of predicted breeding values for non-phenotyped individuals accounting for erosion. The accuracy of the expected values was evaluated through simulations, and a further evaluation was performed on real data.

Results Using simulations, we empirically confirmed that our expected values for the accuracy of PBV accounting for erosion were able to correctly determine the prediction accuracy of breeding values for non-phenotyped individuals. When comparing the expected to the realized accuracies of PBV with real data, only one out of the four traits evaluated presented accuracies that were significantly higher than the expected, approaching $\sqrt{h^2}$.

Conclusions We defined an index of genetic correlation between reference and target populations, which summarizes the expected overall erosion due to differences in allele frequencies and LD patterns between populations. We used this correlation along with a trait's heritability to derive expected values for the accuracy (R) of PBV accounting for the erosion, and demonstrated that our derived E[R|erosion] is a reliable metric.

*Correspondence:
Beatriz C. D. Cuyabano
beatriz.castro-dias-cuyabano@inrae.fr
Full list of author information is available at the end of the article

Background

Today, many commercial livestock breeding programs use genetic merit for the selection of individuals within their programs. Genetic merits, which are often referred to as breeding values (BV), comprise the individual's additive genetic effects that are directly transmitted to its offspring [1]. Estimation of BV relies on the relationships between individuals, and while such an estimation depends on having recorded phenotypes for the traits of interest, it is possible to predict BV for individuals without phenotypic records through their relationships with phenotyped individuals. Henderson's mixed model equations (MME) [2–4] provided a method that yields the so-called best linear unbiased predictors (BLUP) of the individuals' BV, a method which in its original conception used pedigree-based relationships. Combined with the rapid computational advancements during the second half of the twentieth century, Henderson's MME (HMME) revolutionized livestock production systems, enabling large-scale genetic evaluations (i.e. the estimation of BV).

Thanks to the rapid development of molecular technologies, genotype information in the form of single nucleotide polymorphisms (SNPs) is now available at a relatively low cost for the agricultural industry. The first decade of the twenty-first century was marked by significant developments in statistical methods to perform genetic evaluation including either exclusively genomic information [5, 6], or by combining both genomic and pedigree information, either to perform the single-step genetic evaluation [7, 8], or to enhance genetic relationships even when all individuals are genotyped. These developments resulted in dramatic rates of improvement in agricultural production [9], and today BV can be obtained through either pedigree relationships, a genomic relationship matrix (GRM) [6], or the single-step relationship matrix [7, 8], by implementing Henderson's BLUP or a variety of Bayesian methods [5, 10–14], with the reproducing kernel Hilbert spaces (RKHS) being among the most popular of the latter methods when using relationship matrices [11].

In breeding programs, obtaining predicted BV (PBV) for young candidates prior to observing their phenotypes allows the selection at a very early age, thus enabling a reduction of the generation interval, a benefit of particular relevance for example in cattle breeding populations. Thus, the accuracy of PBV is a very important factor for the success of a breeding program. However, realized accuracies of PBV remain below the theoretical maximum even when the reference population is sufficiently large, and SNPs included in the model are in sufficient linkage disequilibrium (LD) with the quantitative trait loci (QTL) within the reference population. This is particularly noticeable over generations, as we observe the so-called erosion of SNP effects [15] accompanied by the erosion of the accuracy of PBV. Erosion occurs mostly because of 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 a QTL, a large effect will be assigned to it. However, if due to segregation over 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. In this paper, accuracy of the PBV will be defined as the correlation with own performance in a validation procedure, with a maximum theoretical value of $\sqrt{h^2}$ (where h^2 is the trait's heritability).

The decay in prediction accuracy due to differences in allele frequencies and LD patterns, especially across generations, is a topic widely known and discussed by animal breeders and quantitative geneticists, with a number of different deterministic equations proposed [16–22]. Dekkers et al. [15] proposed a deterministic method to predict the accuracy of PBV based on selection index theory and on Fisher's information theory, a method that depends on the effective number of chromosome segments (M_e), which in turn relies on quantifying the erosion at the individual SNP level. While their method was successful with simulated data, for which the recombination at the individual SNP level, i.e. the erosion factor, is known, it may lead to wrong predictions of the accuracy of PBV with real datasets, for which the erosion factor is unknown and has to be estimated. In order to address this challenge in real datasets, a factor that accounts for long-distance LD was added to the deterministic formula of the predicted accuracy of PBV, however the values for this factor are quite arbitrary. Accurately 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 PBV through a metric based on the relationships between reference and target populations.

In this work, we propose a statistical method that accounts for erosion to derive the expected accuracy of the PBV through an index of genetic correlation (IGC) between reference and target populations. By considering the accuracy of the PBV as a population parameter measured on the target population, we evaluated our proposed approach using simulated and real data. Accurate expectations for the accuracy of PBV, accounting for erosion, will improve our understanding of the gap between the theoretical maximum, i.e. $\sqrt{h^2}$, and the observed prediction accuracy. Moreover, defining expectations for the accuracy of PBV based on the correlations between reference and

target populations allows us to determine whether accuracies lower than $\sqrt{h^2}$ can be further increased by enhancing the model, or if a low accuracy is only a feature of a target population that is poorly represented by the reference population.

Methods

Model for the prediction of breeding values

Consider the animal model for a genetic evaluation $\mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is a vector of the phenotypes measured in the reference population and pre-corrected for the fixed effects, $\mathbf{g} \sim N(\mathbf{0}, \mathbf{W}\sigma_g^2)$ is the vector of the additive genetic effects, referred to as the BV in animal and plant breeding, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}_n\sigma_\varepsilon^2)$ is the vector of the random residuals, and σ_g^2 and σ_ε^2 are the additive genetic variance and the residual variance, respectively. $\mathbf{W}$ is the matrix of the relationship coefficients between the individuals, and we assume that this relationship can be any of the following three: (i) the pedigree-based relationship matrix, i.e. $\mathbf{W} = \mathbf{A}$; (ii) the genomic relationship matrix [6], i.e. $\mathbf{W} = \mathbf{G}$; or (iii) the single-step relationship matrix [7, 8], i.e. $\mathbf{W} = \mathbf{H}$. We emphasize here that the type of relationship (pedigree, genomic, or single-step) will not affect the future derivations and results with respect to the expected values for the accuracy of PBV.

When the goal is to predict BV for young candidates prior to observing their phenotypes, our model can be re-written as $\mathbf{y}_1 = [\mathbf{I}_{n_1}\mathbf{0}_{n_1 \times n_2}]\mathbf{g} + \boldsymbol{\varepsilon}_1$, such that $\mathbf{g} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} \sim N\left(\mathbf{0}, \mathbf{W}\sigma_g^2\right) \overset{\text{def}}{=} N\left(\mathbf{0}, \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{bmatrix}\sigma_g^2\right)$, such that sub-index 1 indicates the reference population of phenotyped individuals, and sub-index 2 indicates the target population without phenotypes (young candidates). From Henderson's MME [2–4], the analytical solutions for the breeding values $\hat{\mathbf{g}}_1$ for the n_1 animals in the reference population, and the PBV $\tilde{\mathbf{g}}_2$ for the n_2 animals in the target population are:

$$\hat{\mathbf{g}}_1 = \mathbf{W}_{11}\left(\mathbf{W}_{11}\sigma_g^2 + \mathbf{I}_{n_1}\sigma_\varepsilon^2\right)^{-1}\mathbf{y}_1\sigma_g^2, \quad (1)$$

$$\tilde{\mathbf{g}}_2 = \mathbf{W}_{21}\left(\mathbf{W}_{11}\sigma_g^2 + \mathbf{I}_{n_1}\sigma_\varepsilon^2\right)^{-1}\mathbf{y}_1\sigma_g^2. \quad (2)$$

Theoretical limit for the accuracy of predicted breeding values

Our interest lies on the accuracy of the PBV, i.e. on $R = \widehat{cor}\left(\tilde{\mathbf{g}}_2, \mathbf{y}_2\right) = \frac{\sum_{i=1}^{n_2}(\tilde{g}_{2i} - \bar{\tilde{g}}_2)(y_{2i} - \bar{y}_2)}{\sqrt{\sum_{i=1}^{n_2}(\tilde{g}_{2i} - \bar{\tilde{g}}_2)^2 \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2}}$. More specifically, our interest lies on the expected value E(R). While the distribution of R, a realized correlation, is not straightforward, Fisher demonstrated that [23]:

$$Z = \log\left(\frac{1+R}{1-R}\right) \sim N\left(\log\left(\frac{1+\rho}{1-\rho}\right), \frac{4}{n_2 - 3}\right), \quad (3)$$

such that ρ is the true correlation. When predicting BV for young candidates without phenotypes, we can consider the true correlation as $\rho = cor\left(\hat{\mathbf{g}}_1, \mathbf{y}_1\right)$, as it is intuitive that the expected prediction accuracy in the target population, *i.e.* the young candidates, is the same accuracy obtained on the reference population. If we assume that the training dataset is sufficiently large, and that the available SNPs are representative of the QTL, such that BV are accurately obtained for the reference population, we may consider the true correlation as $\rho = cor\left(\hat{\mathbf{g}}_1, \mathbf{y}_1\right) \approx \sqrt{h^2}$.

We acknowledge that the distributions in Eq. (3) rely on the assumption that observations are independent, an assumption that does not hold when individuals are related. However, this lack of independence among the elements in $\mathbf{g}_2$ and $\mathbf{y}_2$ has an impact only on Var(Z). Since the interest of our present study is restricted to E(Z), and ultimately E(R), we will disregard potential changes to the variance of the distribution defined in Eq. (3).

The distribution of Z will allow us to comprehend E(R), although the exact distribution of R, a realized correlation, is not straightforward. To do so, we will study R as a function of Z. Since $Z = \log\left(\frac{1+R}{1-R}\right)$, then $R = f(Z) = \frac{e^Z - 1}{e^Z + 1}$, a function that is concave for $Z > 0$ and convex for $Z < 0$, as illustrated in Fig. 1. Thus, Jensen's inequality [24] allows us to conclude that:

$$E(R) = E[f(Z)] \leq f(E[Z]) = \frac{e^{E(Z)} - 1}{e^{E(Z)} + 1}, \quad \text{for } Z > 0, \quad (4)$$

$$E(R) = E[f(Z)] \geq f(E[Z]) = \frac{e^{E(Z)} - 1}{e^{E(Z)} + 1}, \quad \text{for } Z < 0. \quad (5)$$

Finally, since $E(Z) = \log\left(\frac{1+\rho}{1-\rho}\right)$, as per the distribution in Eq. (3), the inequalities in Eqs. (4) and (5) can be re-written as:

$$\begin{aligned} E(R) \leq f(E[Z]) &= \frac{e^{E(Z)} - 1}{e^{E(Z)} + 1} = \frac{e^{\log\left(\frac{1+\rho}{1-\rho}\right)} - 1}{e^{\log\left(\frac{1+\rho}{1-\rho}\right)} + 1} \\ &= \frac{\left(\frac{1+\rho}{1-\rho}\right) - 1}{\left(\frac{1+\rho}{1-\rho}\right) + 1} = \frac{(1+\rho) - (1-\rho)}{(1+\rho) + (1-\rho)} \\ &= \frac{2\rho}{2} = \rho, \quad \text{for } Z > 0, \end{aligned} \quad (6)$$

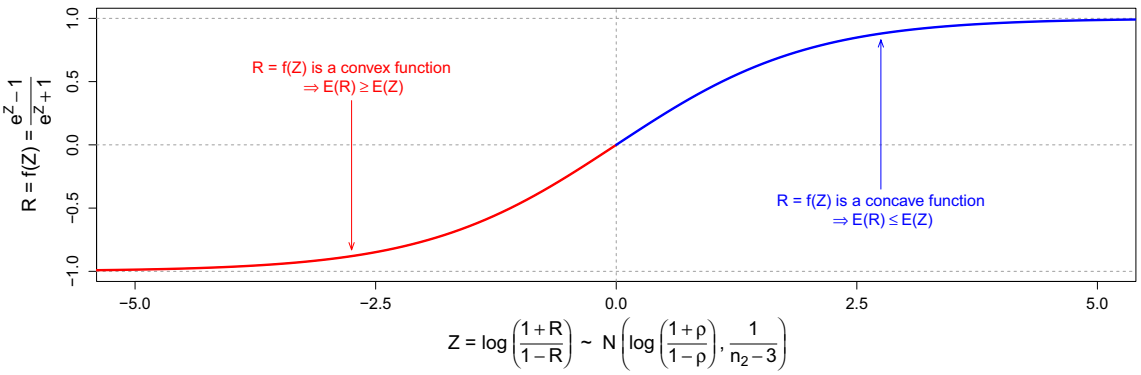


Fig. 1 Description of the relationship between $Z = \log\left(\frac{1+R}{1-R}\right)$ and $R = \widehat{cor}\left(\widetilde{\mathbf{g}}_2, \mathbf{y}_2\right) = \frac{e^Z-1}{e^Z+1}$, highlighting the regions in which R = f(Z) is a convex (in red) or concave (in blue) function of $Z \sim N\left(\log\left(\frac{1+\rho}{1-\rho}\right), \frac{4}{n_2-3}\right)$

$$E(R) \geq \rho, \quad \text{for } Z < 0. \tag{7}$$

Because the animal model is defined as $\mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon}$, it is straightforward that both $\rho = cor\left(\widehat{\mathbf{g}}_1, \mathbf{y}_1\right)$ and $R = \widehat{cor}\left(\widetilde{\mathbf{g}}_2, \mathbf{y}_2\right)$ should be greater than zero, and therefore we focus only on the inequality described in Eq. (6), in which all elements Z, R, and ρ assume positive values.

Some remarks must be made to conclude this section. Previous works that evaluated the accuracy of PBV traditionally assumed that $E(R) \leq \sqrt{h^2}$, while here we assume that $E(R) \leq \rho = cor\left(\widehat{\mathbf{g}}_1, \mathbf{y}_1\right)$. In fact, as previously stated, if the training dataset is sufficiently large and the QTL are correctly represented by the SNPs, the BV for the reference population are accurate enough to ensure, $\rho \approx \sqrt{h^2}$. However, if the BV of the reference population are not accurate, $\rho < \sqrt{h^2}$. This may occur either because of an insufficient number of samples, or because the SNPs are unable to correctly capture the QTL effects (or both). Finally, we emphasize that, in this work, we do not intend to address the drivers of inaccurate BV estimation in the reference population. Instead, we address how the genetic connections between reference and target populations impact the accuracy of the PBV for the target population of individuals without phenotypes. Therefore, $E(R) \leq \rho = cor\left(\widehat{\mathbf{g}}_1, \mathbf{y}_1\right) \leq \sqrt{h^2}$. If the target population is well represented by the reference population, meaning that the relationships in $\mathbf{W}_{21}$ are strong, then E(R) should be in the upper boundary of the inequality in Eq. (6), reaching the equality $E(R) = \rho$. A final remark is that, if the number of records is not sufficiently large to adequately estimate $\widehat{\mathbf{g}}_1$, then $\widehat{\mathbf{g}}_1$ and $\widehat{\mathbf{e}}_1$ may present a level of correlation because of the model's inability to adequately separate the random effects from the residual effects, resulting in an accuracy at the training population of $\rho = cor\left(\widehat{\mathbf{g}}_1, \mathbf{y}_1\right) > \sqrt{h^2}$. In this situation, the limit for $R = \widehat{cor}\left(\widetilde{\mathbf{g}}_2, \mathbf{y}_2\right)$ will be $\sqrt{h^2}$, and thus $E(R) \leq \rho = \min\left\{cor\left(\widehat{\mathbf{g}}_1, \mathbf{y}_1\right), \sqrt{h^2}\right\} \leq \sqrt{h^2}$.

Erosion in the accuracy of predicted breeding values

In the previous section, we have established that $E(R) \leq \rho = \min\left\{cor\left(\widehat{\mathbf{g}}_1, \mathbf{y}_1\right), \sqrt{h^2}\right\} \leq \sqrt{h^2}$, and that if the target population is well represented by the reference population, meaning that the relationships in $\mathbf{W}_{21}$ are strong, then we can expect that $E(R) = \rho$. However, if the target population is poorly represented by the reference population, the relationships in $\mathbf{W}_{21}$ are weak. Finally, since $\mathbf{W}_{21}$ is the key to obtaining the PBV ($\widetilde{\mathbf{g}}_2$), as per Eq. (2), the weaker the relationships in $\mathbf{W}_{21}$, the more inaccurate the PBV, and we can expect that $E(R) < \rho$. The question that we address in this work is how much smaller than ρ is E(R).

In other words, we shall say that when $E(R) < \rho$ there is an erosion in the accuracy of PBV, and we use again Fisher's Z-transformation [23] to quantify the eroded E(R). First, we must define a population parameter $r \in [0, 1]$, a single value which summarizes the relationships in $\mathbf{W}_{21}$, resembling a correlation. Hereafter, we will refer to r as the index of genetic correlation (IGC) between reference and target populations, and in the next section we describe in detail how this parameter can be calculated. Note that $r \in [0, 1]$ is not simply equal to the average relationships in $\mathbf{W}_{21}$, and in fact obtaining r directly from the relationship matrix $\mathbf{W}$ poses some challenges, which we address in "Index of genetic correlation between populations" section, along with two suggested methods to estimate r.

Returning to the Z-transformed accuracy of PBV, we have that $E(Z) = \log\left(\frac{1+\rho}{1-\rho}\right)$, as per the distribution in Eq. (3), and we will say that this expected value holds

when there is no erosion effect, i.e. $E(Z|\text{no erosion}) = \log\left(\frac{1+\rho}{1-\rho}\right) = \mu_Z$. We hypothesized that the eroded E(Z) is linearly affected by the IGC (r), thus:

$$E(Z|\text{erosion}) = r\log\left(\frac{1+\rho}{1-\rho}\right) = r\mu_Z, \tag{8}$$

such that $E(Z|\text{erosion}) \underset{r\to 1}{\to} E(Z|\text{no erosion})$, and $E(Z|\text{erosion}) \underset{r\to 0}{\to} 0$, this last scenario being equivalent to a reference population, thus distinct from the target population, this BV cannot be predicted at all. Finally, operating with $R = f(Z) = \frac{e^Z-1}{e^Z+1}$, we have:

$$\begin{aligned} E(R|\text{erosion}) &= f(E[Z|\text{erosion}]) = \frac{e^{E(Z|\text{erosion})}-1}{e^{E(Z|\text{erosion})}+1} \\ &= \frac{e^{r\log\left(\frac{1+\rho}{1-\rho}\right)}-1}{e^{r\log\left(\frac{1+\rho}{1-\rho}\right)}+1} = \frac{e^{\log\left(\frac{1+\rho}{1-\rho}\right)^r}-1}{e^{\log\left(\frac{1+\rho}{1-\rho}\right)^r}+1} \\ &= \frac{\left(\frac{1+\rho}{1-\rho}\right)^r-1}{\left(\frac{1+\rho}{1-\rho}\right)^r+1} = \frac{(1+\rho)^r-(1-\rho)^r}{(1+\rho)^r+(1-\rho)^r}, \end{aligned} \tag{9}$$

such that $E(R|\text{erosion}) \underset{r\to 1}{\to} \rho = E(R|\text{no erosion})$, and $E(R|\text{erosion}) \underset{r\to 0}{\to} 0$. Figure 2 describes the behaviour of $E(R|\text{erosion}) = \frac{(1+\rho)^r-(1-\rho)^r}{(1+\rho)^r+(1-\rho)^r}$ as a function of $E(Z|\text{no erosion}) = \mu_Z$, ρ, and ρ^2, and shows that E(R|erosion) is an increasing function on both ρ and r. The description of E(R|erosion) as a function of ρ^2, although redundant with the description of E(R|erosion) as a function of ρ, has relevance to interpret E(R|erosion) on the same scale as a function of the trait's heritability (h^2).

Index of genetic correlation between populations

In order to quantify $E(R|\text{erosion}) = \frac{(1+\rho)^r-(1-\rho)^r}{(1+\rho)^r+(1-\rho)^r}$, the IGC represented by r must be calculated. As mentioned in the previous section, obtaining r directly from the relationship matrix $\mathbf{W}$ can be challenging. In the next paragraph, we show how to calculate r when $\mathbf{W} = \mathbf{G}$, the genomic relationship matrix. This method is computationally heavy and may be unfeasible for very large datasets. We also propose a simpler method which uses empirical results on simulated phenotypes and can be a practical alternative that may be applied to any type of relationship matrix $\mathbf{W}$ (pedigree, genomic or single-step). A sample code for implementing both methods to calculate the IGC in R is provided in Additional file 1.

IGC calculated from genomic data

Let $\mathbf{M}$ be a n × m centered and scaled matrix of SNP-genotypes, where the scaling factor is $\left(\sum_{j=1}^{m} 2p_j(1-p_j)\right)^{-1/2}$ and p_j's are the alleles frequencies, and its singular-value decomposition (SVD) is $\mathbf{M} = \mathbf{UDV}'$. In this SVD, $\mathbf{D} = \mathbf{diag}(d_1, d_2, \ldots, d_{n-1}, d_n)$ is a diagonal matrix of the n singular-values, such that $d_1 \geq \cdots \geq d_n \geq 0$ with $d_i = 0$ for every $i > \text{rank}(\mathbf{M})$; $\mathbf{U}_{n\times n} = [\mathbf{U}_1 \ldots \mathbf{U}_n]$ and $\mathbf{V}_{m\times n} = [\mathbf{V}_1 \ldots \mathbf{V}_n]$ are matrices of unitary eigen-vectors, such that $\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}_n$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}_n$.

Each of the components $d_1^2, \ldots, d_n^2$ explains a portion of the variation from the whole system $\mathbf{M}$; each element $U_{ik}(i = 1, \ldots, n)$ in $\mathbf{U}_k = [U_{1k} \ldots U_{nk}]'$ represents the contribution of individual i to the variation explained by component k; each element $V_{jk}(j = 1, \ldots, m)$ in $\mathbf{V}_k = [V_{1k} \ldots V_{mk}]'$ represents the contribution of SNP j to the variation explained by component k.

To obtain the IGC between reference and target populations (r), we need to compare the different contributions of the SNPs to the system's variation in the

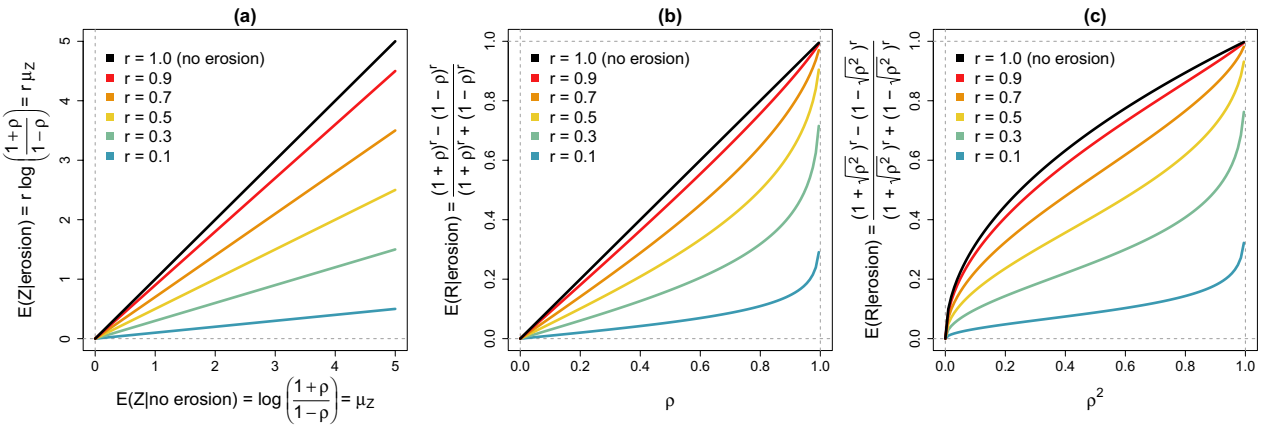


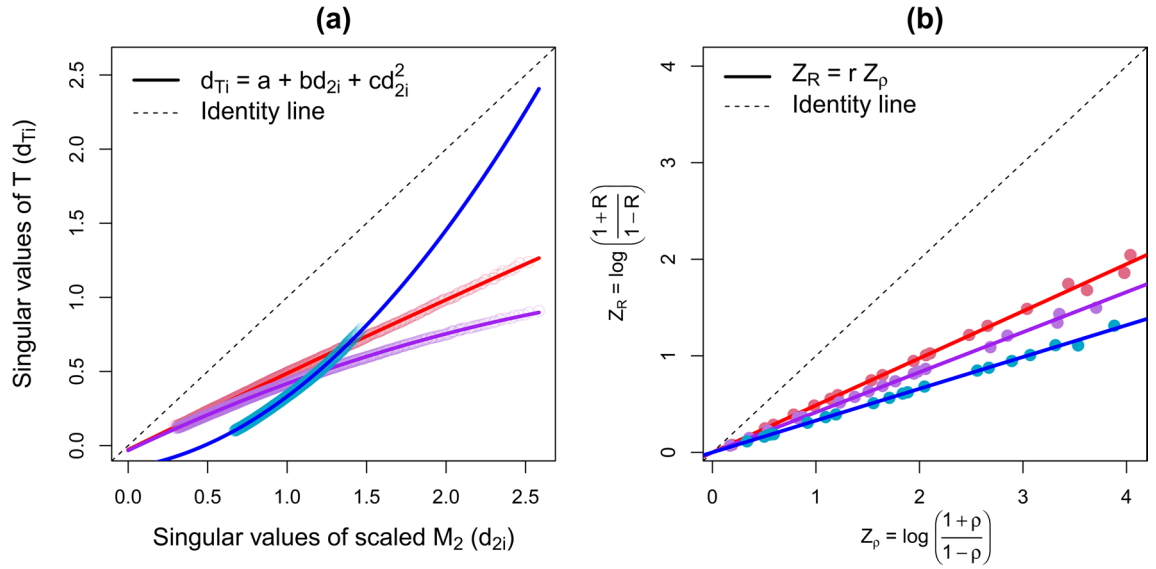
Fig. 2 Description of $E(R|\text{erosion}) = \frac{(1+\rho)^r-(1-\rho)^r}{(1+\rho)^r+(1-\rho)^r}$ as a function of (**a**) $E(Z|\text{no erosion}) = \log\left(\frac{1+\rho}{1-\rho}\right) = \mu_Z$; (**b**) $\rho = \text{cor}(\hat{\mathbf{g}}_1, \mathbf{y}_1)$; (**c**) $\rho^2 = \text{cor}^2(\hat{\mathbf{g}}_1, \mathbf{y}_1)$

two populations. To do so, we perform the aforementioned SVD on $\mathbf{M}_1$ and $\mathbf{M}_2$ (bearing in mind that sub-index 1 refers to the reference population and sub-index 2 refers to the target population), then build a matrix $\mathbf{T} = \sqrt{(n_2/n_1)}\mathbf{V}_2'\mathbf{V}_1\mathbf{D}_1$ which correlates the contributions of the SNPs in both populations, while correcting for the different population sizes, and weighting these correlations by the singular-values of the reference population. Note that the term $\mathbf{V}_2'\mathbf{V}_1\mathbf{D}_1$ that is used to define matrix $\mathbf{T}$ is not arbitrary; this term is the kernel of the solution to $\widetilde{\mathbf{g}}_2$ in Eq. (2), when $\mathbf{W}_{21} = \mathbf{M}_2'\mathbf{M}_1$ and $\mathbf{W}_{11} = \mathbf{M}_1\mathbf{M}_1'$, and we replace $\mathbf{M}_2$ and $\mathbf{M}_1$ by their SVD. By saying that $\mathbf{V}_2'\mathbf{V}_1\mathbf{D}_1$ is the kernel of the solution to $\widetilde{\mathbf{g}}_2$ we mean that for any trait with the same reference and target populations, $\mathbf{V}_2'\mathbf{V}_1\mathbf{D}_1$ is a systematic linear transformation on the observed phenotypes that dictates the projected solutions $\widetilde{\mathbf{g}}_2$ for any set of observed performances, and for any heritability. Next, we obtain the SVD $\mathbf{T} = \mathbf{U}_T\mathbf{D}_T\mathbf{V}_T'$, and perform the linear regression $\mathbf{D}_T \sim \mathbf{D}_2$ with a quadratic term, i.e., we fit $d_{Ti} = a + bd_{2i} + cd_{2i}^2$. Finally, based on extensive observational testing on empirical results, the IGC between reference and target populations can be calculated as $r = a + b + c$. Figure 3a presents an example of the $\mathbf{D}_T \sim \mathbf{D}_2$ obtained for different scenarios of relationships between reference and target populations, which were consistently repeating the pattern of a linear or quadratic relationship between $d_{Ti} \sim d_{2i}$. Throughout all the simulated replicates, we observed that the sum of the coefficients $a + b + c$ was always between 0 and 1, which led us to attempt setting $r = a + b + c$, and finally observing that this value empirically satisfied $E(R|\text{erosion}) = \frac{(1+\rho)^r - (1-\rho)^r}{(1+\rho)^r + (1-\rho)^r}$ for the replicates.

IGC calculated from simulated phenotypes

A less computationally demanding alternative to the method proposed to obtain the IGC r is to use simulated phenotypes. There are two advantages from simulating phenotypes: (1) there is no computational burden from performing SVD on the genotype matrices; and (2) it enables obtaining r irrespective of the genotypes being available.

The first step is to define an arbitrary phenotypic variance σ_y^2, and then simulate $\mathbf{g} \sim N\left(\mathbf{0}, \mathbf{W}h^2\sigma_y^2\right)$ and $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \mathbf{I}_{\mathbf{n}_1+\mathbf{n}_2}(1-h^2)\sigma_y^2\right)$ for a sequence of h^2 that covers a range from low to high heritabilities. If genotypes are available for all animals, $\mathbf{g}$ can be simulated from the genotypes instead, by simulating the vector of SNP effects



Generation	a	b	c	r = a + b + c	r : $Z_R = r\,Z_\rho$
1	−0.028	0.525	−0.009	0.487	0.487
5	−0.032	0.508	−0.057	0.418	0.415
10	−0.151	0.161	0.321	0.331	0.329

Fig. 3 Index of genetic correlation between the simulated reference and three different target populations (one, five, and ten generations after the base reference population) using (**a**) singular-value decompositions on the matrices of SNP-genotypes as proposed in section IGC calculated from genomic data ($r = a + b + c$); (**b**) simulated phenotypes and their breeding values solutions as proposed in section IGC calculated from simulated phenotypes ($\log\left(\frac{1+R}{1-R}\right) = r\log\left(\frac{1+\rho}{1-\rho}\right)$)

$\boldsymbol{\alpha} \sim N\left(\mathbf{0}, \mathbf{I}_m \frac{h^2\sigma_y^2}{\sum_{j=1}^{m} 2p_j(1-p_j)}\right)$ and setting $\mathbf{g} = \mathbf{M}\boldsymbol{\alpha}$, such that $\mathbf{M}$ is the (centred) matrix of SNP-genotypes (n × m). The simulated phenotypes are then $\mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon}$.

The next step is to obtain, for each h^2, the BV for the reference and the PBV for the target populations as established in Eqs. (1) and (2), setting $\sigma_g^2 = h^2\sigma_y^2$, i.e. $\widehat{\mathbf{g}}_1 = \mathbf{W}_{11}\left(\mathbf{W}_{11}\sigma_g^2 + \mathbf{I}_n\sigma_\varepsilon^2\right)^{-1}\mathbf{y}_1 h^2\sigma_y^2$ and $\widetilde{\mathbf{g}}_2 = \mathbf{W}_{21}\left(\mathbf{W}_{11}\sigma_g^2 + \mathbf{I}_n\sigma_\varepsilon^2\right)^{-1}\mathbf{y}_1 h^2\sigma_y^2$, and to obtain $\hat{\rho} = \widehat{cor}\left(\widehat{\mathbf{g}}_1, \mathbf{y}_1\right)$ and $R = \widehat{cor}\left(\widetilde{\mathbf{g}}_2, \mathbf{y}_2\right)$. By doing this procedure for each h^2, a sample of ρ's and R's is generated. Now the Z-transformed correlations are calculated: $Z_\rho = \log\left(\frac{1+\rho}{1-\rho}\right)$ and $Z_R = \log\left(\frac{1+R}{1-R}\right)$.

Finally, the last step is to perform the linear regression $Z_R \sim Z_\rho$ without an intercept, and according to Eq. (8), the slope of this regression is the IGC, i.e. $r = \frac{\widehat{cov}(Z_R, Z_\rho)}{\widehat{Var}(Z_\rho)}$.

Data for the empirical study

Simulated data

We used the R package GenEval (https://github.com/bcuyabano/GenEval) to simulate 50 k SNPs and additive phenotypes ($\sigma_y^2 = 100$) for a wide range of heritabilities $h^2 = 0.05, 0.15, \ldots, 0.9, 0.95$, using a random subset of 2 k SNPs as QTL. SNP-genotypes were simulated in LD, as per the function simGeno() from the R package GenEval, with the LD structure set to resemble that of a cattle population. A base reference population of 5000 individuals was used to estimate variance components using the residual maximum likelihood (REML) [25, 26] and then to obtain the PBV as in Eq. (2), for three different target populations (1000 individuals each) with an increasing number of generations (one, five, and ten) from the base reference population. All the generations were simulated with a 50/50 ratio of males and females, and random mating was performed with no selection. For each $h^2 = 0.05, 0.15, \ldots, 0.9, 0.95$, we simulated 500 independent replicates of phenotypes for the entire population (base and generations one, five, and ten), thus creating a large sample of $\widehat{h^2}$, $\hat{\rho} = \min\left\{\widehat{cor}\left(\widehat{\mathbf{g}}_1, \mathbf{y}_1\right), \sqrt{\widehat{h^2}}\right\}$, and $R = \widehat{cor}\left(\widetilde{\mathbf{g}}_2, \mathbf{y}_2\right)$ at each heritability level.

We performed this study for $\mathbf{W} = \mathbf{G}$, which allowed us to calculate the IGC r with both proposed methods, and to compare the values obtained. We wanted to ensure $\hat{\rho} = \min\left\{\widehat{cor}\left(\widehat{\mathbf{g}}_1, \mathbf{y}_1\right), \sqrt{\widehat{h^2}}\right\} \approx \sqrt{\widehat{h^2}}$, thus the QTL were kept among the genotypes used for analysis. Then, we compared the realized prediction accuracies $R = \widehat{cor}\left(\widetilde{\mathbf{g}}_2, \mathbf{y}_2\right) = \frac{\sum_{i=1}^{n_2}(\widetilde{g}_{2i} - \bar{\widetilde{g}}_2)(y_{2i} - \bar{y}_2)}{\sqrt{\sum_{i=1}^{n_2}(\widetilde{g}_{2i} - \bar{\widetilde{g}}_2)^2 \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2}}$ with the theoretical curve $E(R|\text{erosion}) = \frac{(1+\rho)^r - (1-\rho)^r}{(1+\rho)^r + (1-\rho)^r}$, to evaluate how accurately this equation quantifies the erosion in the accuracy of PBG. The comparison of R to the theoretical curve $\frac{(1+\rho)^r - (1-\rho)^r}{(1+\rho)^r + (1-\rho)^r}$ was performed for both $\rho = \sqrt{\widehat{h^2}}$ and $\rho = \min\left\{\widehat{cor}\left(\widehat{\mathbf{g}}_1, \mathbf{y}_1\right), \sqrt{\widehat{h^2}}\right\}$, to test our hypothesis that the latter is a more suitable measure to be considered, even in a scenario in which $\widehat{cor}\left(\widehat{\mathbf{g}}_1, \mathbf{y}_1\right) \approx \sqrt{\widehat{h^2}}$.

Real data

We compared the accuracies of PBV to their derived expected values using real data from a dairy cattle population within a breeding program. In total, 9636 cows were used as the reference population, and the target population comprised 2130 cows. The pedigree relationship matrix $\mathbf{A}$ was built by tracing back three generations on the pedigree for each of the 11,766 cows used for the analysis. These 11,766 animals were a subset of an original dataset comprising records from ~100,000 cows covering six generations within the breeding program, i.e. this is a population under selection. Our subset ensured that all cows in the target population had their dams in the reference population, and that all dams could be fully traced back by three generations in the pedigree. The subset of animals belonged to the last three generations of the breeding program.

The phenotypes were available for the cows in the form of yield deviations (YD), and four traits were evaluated: one fertility (FERT), one health (HEALTH), and two production (PROD1, PROD2) traits. A preliminary study indicated that the heritabilities of these traits were $h^2_{FERT} = 0.02$, $h^2_{HEALTH} = 0.188$, $h^2_{PROD1} = 0.375$, and $h^2_{PROD2} = 0.625$.

Both pedigree information and 53,469 autosomal SNP-genotypes (EuroGMD v1, a customized ILLUMINA genotyping microarray that contains approximately 70,000 SNPs) were available for all animals, allowing us to perform the study with the three possible genetic relationship matrices: $\mathbf{W} = \mathbf{A}$ (pedigree), $\mathbf{W} = \mathbf{G}$ (genomic), and $\mathbf{W} = \mathbf{H}$ (single-step, using genotypes for all animals in the target population and for 25% of the animals in the reference population). For this study on real data, we applied only the method that uses simulated phenotypes to obtain the IGC (r), and these additive phenotypes ($\sigma_y^2 = 100$) were simulated for heritabilities $h^2 = 0.1, 0.2, \ldots, 0.9$, using a random subset of 2 k SNPs as QTL. Finally, the expected prediction accuracies accounting for

erosion were calculated for each studied trait, and their values compared to the accuracies obtained with real phenotypes (in the form of the yield deviations, as mentioned previously), *i.e.* $R = \widehat{\mathrm{cor}}\left(\widetilde{\mathbf{g}}_2, \mathbf{y}_2\right)$.

Genomic BLUP reference for relationships between populations and reliability of prediction

We compared our derived expected accuracy of prediction $\mathrm{E(R|erosion)} = \frac{(1+\rho)^r-(1-\rho)^r}{(1+\rho)^r+(1-\rho)^r}$ to the theoretical mean reliability obtained from Henderson's MME [2–4], assuming the same animal model that was considered to compute E(R|erosion), i.e. $\mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon}$, with $\mathbf{y}$ being the vector of the phenotypes measured in the reference population and pre-corrected for the fixed effects. The theoretical mean reliability of the genomic (G)BLUP can be considered as the following average reliability in the target population:

$$\mathrm{E(R|GBLUP)} = \frac{\sqrt{\hat{h}^2}}{n_2}\sum_{i=1}^{n_2}\sqrt{\left[\mathbf{W}_{21}\mathbf{W}_{11}^{-1}\left(\mathbf{W}_{11} - \left(\mathbf{LHS}\times\hat{\sigma}_g^2\right)^{-1}\right)\mathbf{W}_{11}^{-1}\mathbf{W}_{12}\right]_{ii}}, \quad (10)$$

such that **LHS** is the left-hand-side of Henderson's MME for BLUP, and the sub-index *ii* indicates the diagonal elements of the matrix inside the square brackets.

Moreover, in order to assess our proposed IGC (r), we compared its value to the average weighted relationships between reference and target populations:

$$r_{GBLUP} = \frac{1}{n_2}\sum_{i=1}^{n_2}\sqrt{\left[\mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}\right]_{ii}}. \quad (11)$$

These comparisons were performed on the simulated data only, as their purpose was to verify if our derived expected value E(R|erosion) and IGC (r) were a better fit for making inferences on the expected accuracy of prediction.

Results

Figure 3 presents the IGC between the reference and the three different target populations (one, five, and ten generations after the base reference population) using the two proposed methods. Figure 3a presents the relationship between the singular values of $\mathbf{T} = \sqrt{(n_2/n_1)}\mathbf{V}_2'\mathbf{V}_1\mathbf{D}_1$ (d_{Ti}) and the singular values of the scaled $\mathbf{M}_2$ (d_{2i}), used to obtain the IGC calculated from genomic data in which $r = a + b + c$, such that $d_{Ti} = a + bd_{2i} + cd_{2i}^2$, as described in the Methods section. Figure 3b presents the linear relationship between Z_R and Z_ρ, used to obtain the IGC calculated from simulated phenotypes, in which r satisfies $Z_R = rZ_\rho$. We observed that the values of r obtained with the two proposed methods are very similar for the three target populations. We observed that $r < 1$ for all target populations, and as expected, it decreases when the number of generations of the target population from the base reference population increases. The results presented for r obtained using the simulated phenotypes in Fig. 3b were based

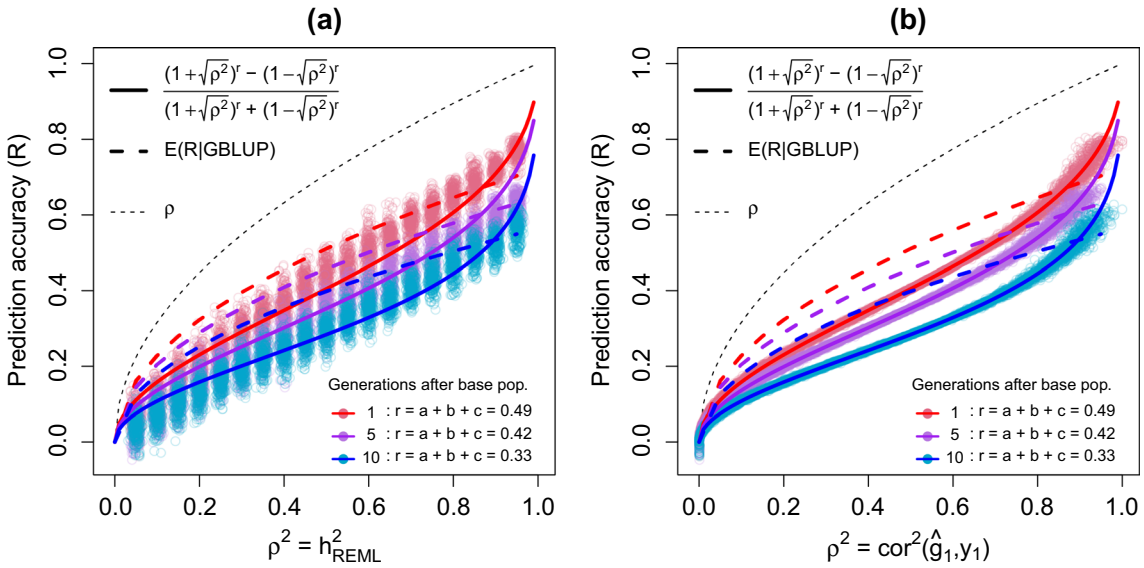


Fig. 4 Relationship between the realized prediction accuracy $R = \widehat{\mathrm{cor}}\left(\widetilde{\mathbf{g}}_2, \mathbf{y}_2\right)$ and (**a**) the REML heritability estimates $\hat{h}^2$; (**b**) the squared accuracy of the breeding values of the reference population, $\rho^2 = \widehat{\mathrm{cor}}^2\left(\widehat{\mathbf{g}}_1, \mathbf{y}_1\right)$. Results presented are of the 500 replicates of simulated phenotypes for each h^2

on one of the 500 replicates. However, when analysing r for each one of those replicates, the values are comparable with standard deviations of 0.01, 0.005, and 0.005, respectively, for the target populations one, five, and ten generations apart from the base reference population.

Figure 4 presents the relationship between the realized prediction accuracy $R = \widehat{cor}(\widetilde{\mathbf{g}}_2, \mathbf{y}_2)$ and both the REML heritability estimates $\widehat{h}^2$ and the squared accuracy of the breeding values of the reference population, $\widehat{cor}^2(\widehat{\mathbf{g}}_1, \mathbf{y}_1) \approx \widehat{h}^2$. We observed that there is a greater variation of R with respect to the heritability estimates, than with respect to the accuracy of the BV in the reference population. Since for all three target populations $r < 1$, we expected that all $R < \sqrt{\widehat{h}^2}$ and $R < \rho$, and this was confirmed by the results presented in Fig. 4a and b. Then, we evaluated how well the theoretical curve $\frac{(1+\rho)^r-(1-\rho)^r}{(1+\rho)^r+(1-\rho)^r}$ described the observed results, and compared it to the average GBLUP reliability in the target population, calculated as in Eq. (10). In Fig. 4a, we observed that the theoretical curve is a reasonable mean to describe the relationship between R and $\rho^2 = \widehat{h}^2$, however it overestimates R for the very low or very high heritabilities. This is not surprising, since in fact, the issue with a very lowly or very highly heritable trait is that, due to heritabilities being close to the boundaries of possible values, we expect a loss in the accuracy of BV for the reference population, i.e. $\widehat{cor}(\widehat{\mathbf{g}}_1, \mathbf{y}_1) < \sqrt{\widehat{h}^2}$. This result supports our hypothesis that the theoretical curve $\frac{(1+\rho)^r-(1-\rho)^r}{(1+\rho)^r+(1-\rho)^r}$ will better describe the relationship between R and $\rho^2 = \widehat{cor}^2(\widehat{\mathbf{g}}_1, \mathbf{y}_1)$, presented in Fig. 4b, in which we observed that the theoretical curves are very accurate to describe this relationship. In both panels (a) and (b) of Fig. 4, we observed that the average GBLUP reliability in the target population failed to correctly describe the average realized prediction accuracy $R = \widehat{cor}(\widetilde{\mathbf{g}}_2, \mathbf{y}_2)$ across the range of heritabilities. For the simulations with h^2 below 0.8, E(R|GBLUP) does serve as an upper boundary, closer to the realized results when R is compared to $\rho^2 = \sqrt{\widehat{h}^2}$ in Fig. 4a, than when R is compared to $\rho^2 = \widehat{cor}^2(\widehat{\mathbf{g}}_1, \mathbf{y}_1)$ in Fig. 4b. For the simulations with h^2 above 0.8, E(R|GBLUP) was much closer to the average realized prediction accuracy. With respect to the IGC, while our method to compute this index yielded values of 0.49, 0.42, and 0.33 for generations one, five, and ten, respectively, following the derivation based on GBLUP calculated as in Eq. (11), the values obtained were 0.96, 0.86, and 0.75 for generations one, five, and ten, respectively. Finally, these values, if used to compute the curve $\frac{(1+\rho)^r-(1-\rho)^r}{(1+\rho)^r+(1-\rho)^r}$ would result in a great overestimation for E(R|erosion).

Figure 5 and Table 1 present the results on IGC r and on the accuracies of BV and PBV obtained for the reference and target populations, respectively ($\hat{\rho} = \widehat{cor}(\widehat{\mathbf{g}}_1, \mathbf{y}_1)$ and $R = \widehat{cor}(\widetilde{\mathbf{g}}_2, \mathbf{y}_2)$, respectively), for the study on real data and for the four traits evaluated. Figure 5a, for r obtained using the simulated phenotypes, indicates that

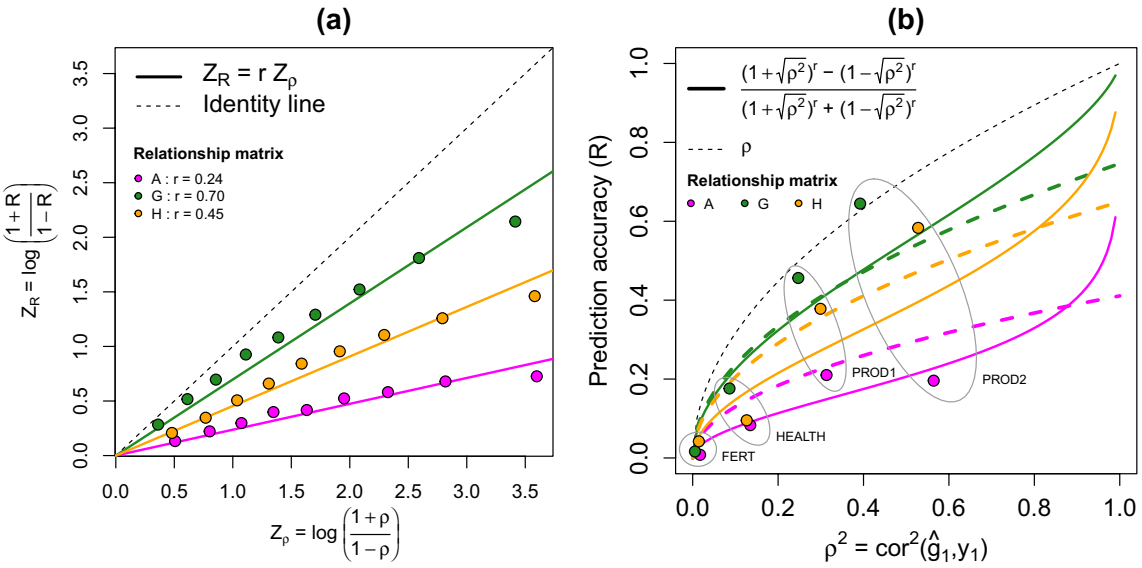


Fig. 5 Results of the study on real data: (**a**) Index of genetic correlation (r) between the reference and target populations using simulated phenotypes and their breeding values solutions as proposed in section IGC from simulated phenotypes. (**b**) Relationship between the realized prediction accuracy $R = \widehat{cor}(\widetilde{\mathbf{g}}_2, \mathbf{y}_2)$ and the squared accuracy of the breeding values of the reference population, $\hat{\rho}^2 = \widehat{cor}^2(\widehat{\mathbf{g}}_1, \mathbf{y}_1)$

Table 1 Results for single-trait evaluations performed on real data

Trait	h^2	$\sqrt{h^2}$	Matrix	$\rho = \widehat{cor}(\hat{g}_1, y_1)$		$R = \widehat{cor}(\tilde{g}_2, y_2)$		$\hat{E}$(R|erosion)	
FERT	0.020	0.141	**A**	0.132	[0.104;0.159]	0.008	[−0.052;0.068]	0.031	[−0.029;0.091]
			G	0.073	[0.045;0.101]	0.016	[−0.044;0.076]	0.051	[−0.009;0.111]
			H	0.120	[0.092;0.148]	0.042	[−0.018;0.102]	0.055	[−0.005;0.114]
HEALTH	0.188	0.434	**A**	0.367	[0.343;0.391]	0.083	[0.023;0.143]	0.091	[0.031;0.150]
			G	0.294	[0.268;0.319]	0.176	[0.117;0.234]	0.208	[0.150;0.265]
			H	0.356	[0.331;0.380]	0.096	[0.036;0.155]	0.167	[0.108;0.225]
PROD1	0.375	0.612	**A**	0.560	[0.540;0.597]	0.210	[0.152;0.267]	0.149	[0.090;0.207]
			G	0.497	[0.476;0.518]	0.456	[0.407;0.503]	0.363	[0.310;0.414]
			H	0.547	[0.527;0.567]	0.378	[0.325;0.428]	0.272	[0.215;0.327]
PROD2	0.625	0.791	**A**	0.751	[0.739;0.763]	0.196	[0.137;0.253]	0.227	[0.207;0.283]
			G	0.626	[0.609;0.643]	*0.654	[0.608;0.679]	0.472	[0.414;0.517]
			H	0.727	[0.713;0.740]	*0.583	[0.542;0.622]	0.396	[0.327;0.445]

* Observed R is significantly different from $\hat{E}$(R|erosion), at a significance level of 0.05

Heritabilities (h^2) were estimated in a preliminary study, and used to obtain the BV and PBV, which were then used to calculate the accuracies for the reference and target populations: $\hat{\rho} = \widehat{cor}(\hat{\mathbf{g}}_1, \mathbf{y}_1)$ and $R = \widehat{cor}(\tilde{\mathbf{g}}_2, \mathbf{y}_2)$, respectively

$\hat{E}$(R|erosion) were calculated according to the indexes of genetic correlation obtained for each of the relationship matrices used for the evaluation, their values being $r_A = 0.237, r_G = 0.697, r_H = 0.454$. Values between brackets are the 95% confidence intervals for the values presented

the pedigree relationship matrix (**A**) is the matrix that minimizes r, while the genomic relationship matrix (**G**) is the matrix that maximizes r. The values of the IGC were $r_A = 0.237$, $r_G = 0.697$, and $r_H = 0.454$, and please recall that the single-step relationship matrix (**H**) was built using genotypes for all animals in the target population and for 25% of the animals in the reference population. Therefore, the accuracies of the PBV are expected to be lowest when using **A**, and highest when using **G**. In fact, we observe in Fig. 5b that accuracies of the PBV are lowest when using **A** for all traits, and highest when using **G** for most of the traits, except for FERT. However, the 95% confidence intervals (CI) presented in Table 1, for the accuracies of the PBV obtained for FERT, the accuracy obtained when using **G** cannot be deemed as significantly different from those obtained when using **A** or **H**. When using **H**, the results for both r and the accuracies of the PBV are between those obtained when using **A** and **G**. This is not surprising, as the single-step combines the genotype information with the pedigree information from non-genotyped individuals. The values of R obtained when using **A** are all close to the curve of their expected values given erosion, i.e. E(R|erosion), when we observe Fig. 5b; the 95% CI presented in Table 1 for these two values confirm that the observed R are not significantly different from their expectations for all traits. In Fig. 5b, the values of R obtained when using **G** appear to be greater than E(R|erosion); however, the 95% CI presented in Table 1 for these two values indicate that the observed R is significantly different from the expectations only for PROD2. The same conclusions from R obtained when using **G** are drawn for the values of R obtained when using **H**.

Discussion

We hypothesized that once an IGC between reference and target populations (r) is calculated, we can define the maximum accuracy of the PBV as the expectation $E[R|\text{erosion}] \leq \frac{(1+\rho)^r-(1-\rho)^r}{(1+\rho)^r+(1-\rho)^r}$, such that ρ represents the true maximum prediction accuracy. We assumed this maximum prediction accuracy to be $\rho = \text{cor}(\hat{\mathbf{g}}_1, \mathbf{y}_1)$, the accuracy of the BV obtained for the reference population, as it is intuitive that the accuracy of prediction for the target population cannot be higher than that for the reference population. For ideal scenarios, in which the reference population is sufficiently large, and SNP-genotypes are available and in strong LD with the QTL, we can assume $\rho \approx \sqrt{h^2}$, i.e. the theoretical maximum prediction accuracy without erosion. The results obtained with extensive simulations supported our hypothesis, and indicated that indeed considering $\rho = \min\left\{\text{cor}(\hat{\mathbf{g}}_1, \mathbf{y}_1), \sqrt{\hat{h}^2}\right\}$ rather than $\rho = \sqrt{h^2}$ resulted in a more correct and consistent E[R|erosion].

One important element for calculating E[R|erosion] is the IGC, a single value capable of summarizing all the information about the genetic distance between reference and target populations. When working with genomic prediction, r summarizes the differences in allele frequencies and LD patterns observed in both the reference and target populations. We presented two methods

to calculate r, one method that simulates phenotypes and predicts them on the target population to infer r, and another method that can only be applied for genomic prediction by performing operations on the SVD of the genotype matrices from the reference and target populations. Our results show that both methods to obtain r are trustworthy.

Although computationally costly for large populations and dense genotype data, calculating r using the SVD of the genotype matrices may offer extra information about the genetic similarities or differences of the populations. Such decompositions are a very informative tool to evaluate the connections between the individuals studied, and the LD between the SNPs. Then, it is intuitive that allele frequencies, LD patterns, number of SNPs and population sizes affect the IGC by changing the coefficients a, b, and c that compose $r = a + b + c$ (a result obtained based on the extensive observations of empirical results). Because this work focused on calculating E[R|erosion], we did not explore the underlying meaning of the values of these coefficients (i.e. how they are affected by allele frequencies, LD patterns, number of SNPs and population sizes), but we do understand that such a study may be relevant, and should be conducted in the future.

When we compared the theoretical curves of E[R|erosion] obtained for the different relationship matrices (**A**, **G**, and **H**) to the realized prediction accuracies, we observed that the curve for **A** was the curve that best outlined expectations for the prediction accuracies using this relationship matrix. The realized prediction accuracies obtained using **G** and **H** were farther from the theoretical curves of E[R|erosion], however only the realized prediction accuracies for the PROD2 trait were significantly different from the expectations, as shown in Table 1.

PROD2 was the trait with the highest heritability ($h^2 = 0.625$). However, Fig. 4b, which presents the results on simulated data, indicated that the variance of prediction accuracies increases with the heritability of the trait. Thus, the variance of the prediction accuracy for PROD2 should be larger than the variances of the prediction accuracy for the other three traits. Although the realized prediction accuracies of PROD2 were significantly higher than E[R|erosion], we have to look at this result with caution. The CI presented in Table 1 are for correlations, and rely on the Z-transformed correlation, which are estimates, rather than exact CI. Moreover, it is relevant to consider that the real data evaluated comprised phenotypes measured on a breeding population, i.e. under selection, meaning that this may increase the accuracy of PBV obtained with genomic data, as selection can lead to an increase in the LD between the QTL and the most relevant SNPs, resulting in more accurate estimates for the SNP effects in the later populations. Thus, one possible explanation for the observed prediction accuracies of PROD2 being significantly higher than E[R|erosion] may be that, due to selection favouring a highly heritable trait, a stronger LD between the most relevant SNPs and the QTL is present, increasing the relationships between reference and target populations at the SNPs with a larger effect. Thus, genomic PBV for young candidates in a breeding program are expected to be quite accurate for highly heritable traits, even if the IGC between reference and target populations is low because the most significant SNPs have their effects quite accurately estimated in the reference population, and moreover, due to selection, reference and target populations should not present large differences for those SNPs with larger effects. Thus, if the SNPs that drive low values of r are those that are less significant, then r will have a smaller relevance for E[R|erosion]. We note that the observed prediction accuracies of PROD2 were also greater than those of E[R|GBLUP].

The accuracy of PBV has been previously studied from different perspectives [6, 17, 27, 28], and different deterministic equations have been proposed to calculate this accuracy [6, 15, 16, 18, 19, 29, 30]. The degrees of the genomic relationships [27, 28], as well as the LD and co-segregation of the QTL from pedigree [17], have already been evaluated as contributors to the accuracy of genomic prediction.

In order to predict the accuracy of genomic PBV, Goddard et al. [30] derived a method using the total genetic variance and the pairwise genomic correlations between individuals. Using a different approach, Wientjes et al. [18, 19] proposed a deterministic equation to predict the accuracy of PBV, accounting for differences between populations in across-breed predictions, and M_e, a function of the genotype data. Lee et al. [21] extended the proposed equation from Wientjes et al. [18, 19] by accounting for the effective population size (N_e) and studied how the degrees of the relationships between individuals, the size of the reference population and marker panel density impact the prediction accuracy. On the one hand, our study with simulated data, shows that the aforementioned deterministic methods yielded predictions for the accuracy of the genomic PBV that were very similar to the curves of E[R|GBLUP]. Also using a deterministic approach, Dekkers et al. [15] used selection index theory and Fisher's information theory to predict the accuracy of PBV, a method that ultimately relies on the information about the erosion at the individual SNP level. However, accurately quantifying the erosion at the individual SNP level is a difficult and unresolved task. On the other

hand, our work shows that quantifying the erosion of the accuracy of the PBV as a population parameter is a more tractable problem.

Taking a different approach from what was previously proposed, we defined a metric to quantify the IGC between reference and target populations. Then, we used this correlation to derive a statistical prediction for the accuracy of PBV, E[R|erosion], based on Fisher's Z-transformation [23] and treating the accuracy of the PBV as a population parameter, and demonstrated through simulated and real data that our derived E[R|erosion] is a reliable metric.

Conclusions

The accuracy of PBV is a very important factor for the success of breeding programs that make use of estimates of genetic merit to select individuals. While the advent of genomic prediction has greatly increased the accuracy of PBV, realized accuracies remain below $\sqrt{h^2}$, even when the reference population is sufficiently large, and SNPs included in the model are in sufficient LD with the QTL. This is particularly noticeable across generations, as we observe the so-called erosion of SNP effects [15] accompanied by the erosion of the accuracy of the PBV. We defined an IGC between reference and target populations, which summarizes the expected overall erosion due to differences in allele frequencies and LD patterns between reference and target populations. We used this correlation to derive a statistical prediction for the accuracy of the PBV accounting for erosion, E[R|erosion], an expectation based on Fisher's Z-transformation, and demonstrated that our derived E[R|erosion] is a reliable metric.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12711-024-00876-9.

Additional file 1. Example_IGC. Sample code to simulate genotypes/phenotypes, perform genomic prediction and compute the index of genetic correlation (IGC) between reference and target populations.

Acknowledgements
The pedigree, phenotypic and genotypic information used for the analysis on real data was extracted from the French National Bovine Selection database, hosted by INRAE-CTIG. The authors warmly thank Thierry Tribout who prepared the dataset and provided information on it.

Author contributions
BCDC performed the calculation of the theoretical results, performed the analyses on simulated and real data, and wrote the manuscript. DB and CG conceptualized the applications of the theoretical results, participated in the interpretation of the applied results, and assisted in the definition of the manuscript's final structure. All authors read and approved the manuscript.

Funding
This work was partially supported by the Rural Development Administration, Republic of Korea and by the National Institute of Food and Agriculture (AFRI Projects No. 2019-67015-29323 and 2021-67015-33411).

Availability of data and materials
Simulations of genotypes and phenotypes were performed using the R [31] package GenEval (https://github.com/bcuyabano/GenEval/). A sample code for implementing the method in R is provided in Additional file 1. The real data analysed are from a commercial source and not publicly available.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]INRAE, AgroParisTech, GABI, Université Paris Saclay, 78350 Jouy-en-Josas, France. [2]Department of Animal Science, Michigan State University, 474 S Shaw Ln, East Lansing, MI 48824, USA.

Received: 25 November 2022 Accepted: 8 January 2024
Published online: 29 February 2024

References

1. Falconer DS, Mackay TF. Introduction to quantitative geneticsm. 4th ed. Harlow: Pearson Education; 1996.
2. Henderson CR, Kempthorne O, Searle SR, von Krosigk CM. The estimation of environmental and genetic trends from records subject to culling. Biometrics. 1959;15:192–218.
3. Henderson CR. Use of relationships among sires to increase accuracy of sire evaluation. J Dairy Sci. 1975;58:1731–8.
4. Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975;31:423–47.
5. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
6. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
7. Misztal I, Legarra A, Aguilar I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J Dairy Sci. 2009;92:4648–55.
8. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 2010;42:2.
9. García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Van Tassell CP. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci USA. 2016;113:3995–4004.
10. Gianola D, Fernando RL, Stella A. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics. 2006;173:1761–76.
11. Gianola D, van Kaam JBCHM. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics. 2008;178:2289–303.
12. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12:186.

13. Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R. Additive genetic variability and the Bayesian alphabet. Genetics. 2009;183:347–63.
14. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 2015;11: e1004969.
15. Dekkers JCM, Su H, Cheng J. Predicting the accuracy of genomic predictions. Genet Sel Evol. 2021;53:55.
16. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One. 2008;3: e3395.
17. Habier D, Fernando RL, Garrick DJ. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics. 2013;194:597–607.
18. Wientjes YCJ, Veerkamp RF, Bijma P, Bovenhuis H, Schrooten C, Calus MPL. Empirical and deterministic accuracies of across-population genomic prediction. Genet Sel Evol. 2015;47:5.
19. Wientjes YC, Bijma P, Veerkamp RF, Calus MPL. An equation to predict the accuracy of genomic values by combining data from multiple traits, populations, or environments. Genetics. 2016;202:799–823.
20. Pszczola M, Calus MPL. Updating the reference population to achieve constant genomic prediction reliability across generations. Animal. 2016;10:1018–24.
21. Lee SH, Clark S, van der Werf JH. Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. PLoS One. 2017;12: e0189775.
22. van den Berg I, Meuwissen THE, MacLeod IM, Goddard ME. Predicting the effect of reference population on the accuracy of within, across, and multibreed genomic prediction. J Dairy Sci. 2019;102:3155–74.
23. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika. 1915;10:507–21.
24. Jensen J. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Math. 1906;30:175–93.
25. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. Biometrika. 1971;58:545–54.
26. Meyer K. An, "average information" restricted maximum likelihood algorithm for estimating reduced rank genetic covariance matrices or covariance functions for animal models with equal design matrices. Genet Sel Evol. 1997;29:97–116.
27. Hayes BJ, Bowman PJ, Chamberlain A, Verbyla K, Goddard ME. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol. 2009;41:51.
28. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res (Camb). 2009;91:47–60.
29. Goddard ME. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009;136:245–57.
30. Goddard ME, Hayes BJ, Meuwissen THE. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet. 2011;128:409–21.
31. R Core Team. R: a language and environment for statistical computing. Vienna: R foundation for statistical computing; 2018.

Publisher's Note