Methods to Estimate Erosion Factors of Genomic Breeding Values of Candidates due to Long-Distance Linkage Disequilibrium

D Boichard, S Fritz, P Croiseau, V Ducrocq, T Tribout, M Barbat, B C D Cuyabano

Methods to Estimate Erosion Factors of Genomic Breeding Values of Candidates due to Long-Distance Linkage Disequilibrium

D. Boichard[1], S. Fritz[1,2], P. Croiseau[1], V. Ducrocq[1], T. Tribout[1], M. Barbat[3], and B.C.D. Cuyabano[1]

[1]Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France

[2]Eliance, 75012 Paris, France

[3]Geneval, 78350 Jouy-en-Josas, France

Abstract

Most validation studies of genomic evaluation observe inflation, *i.e.* regression coefficients of the later phenotypes on early predictions smaller than one. This pattern does not reflect a bias in the evaluation model, it rather reflects long distance associations between markers and quantitative trait loci (QTLs). Due to linkage disequilibrium (LD), SNP effects estimated from a reference data capture non-zero contributions from distant QTLs located not only in the same, but also in the other chromosomes, and we show that some across-chromosome LD does exist in different French dairy cattle breeds. This LD results from limited effective population size and, more importantly, from the relationship within the reference population. Long distance associations are partly broken and rebuilt at random at each generation. Therefore, corresponding SNP effects are partly lost in the next generations and we shall refer to this effect loss as erosion. This erosion can be predicted by different methods based on the following equations applied to simulated QTLs. If the breeding values are **Pq** with **P** the QTL genotypes and **q** their effects, the expected contribution of QTL j to the estimated SNP effect i is $\mathbf{c}_i \mathbf{M'} \mathbf{P}_j q_j$, where **M** is the matrix of SNP genotypes and $\mathbf{c}_i$ is line i (corresponding to SNP i) of $\mathbf{C} = (\mathbf{M'M} + \lambda \mathbf{I})^{-1}$. Two methods based on simulations are proposed to estimate the erosion factor ρ. In Method 1, the direct genomic value (DGV) of the progeny based on SNP effects estimated in this new simulated generation are regressed on the DGV of the same progeny based on SNP effects estimated in the reference population. In Method 2 all the QTL contributions to SNP effects are regressed based on SNP-QTL recombination rates and summed to predict the breeding value at the next generation. The regression coefficient of the DGV based on eroded contributions on the raw DGV is also an estimate of erosion. An illustration is given with the French Normande female reference population in 2021. Method 1 is simpler to implement on a routine basis, and yields good estimates of erosion over one generation. Erosion is also dependent on the distance between the young candidates and their reference population and formulae are proposed to apply erosion. We recommend accounting for erosion in genetic evaluations to provide unbiased predictions for the young candidates. Accordingly, erosion has been accounted for in the French Single Step bovine evaluation since March 2022.

Key words: Genomic evaluation, inflation, erosion of genomic values, validation methods

Introduction

In genomic evaluation, single nucleotide polymorphism (SNP) effects are estimated in a reference population and applied to selection candidates. This method is extensively used to select candidates at an early stage of their life or not yet with phenotypic information. The standard interpretation is that SNPs are in close LD with causal mutations (or QTL) and therefore, good proxies for these QTL. Implicitly, this assumes that estimated SNP effects reflect those of the neighbouring causal mutations. Under this assumption, SNP effects observed in the reference sample should be very similar in the next generation as short-distance LD erodes slowly due to recombination. It is, however, well known that genomic evaluation

efficiency is highly dependent on the close relationship of the candidates to the reference sample (Habier et al, 2007, 2013; Legarra et al, 2008; Pszczola et al, 2012). Many studies have shown the limited gain in accuracy in multi-breed evaluation (Erbe et al, 2012; Hozé et al, 2014), illustrating that distant reference data are not informative. Other studies have shown a decrease in accuracy over generations when the reference population is not updated (Soneson et al, 2009; Solberg et al, 2009). Moreover, it has been observed that the absence of parents in the reference population directly influences the prediction accuracy of the selection candidates. All these results suggest that SNP effects erode as the distance between candidates and reference sample increases.

Validation studies of genomic evaluations are generally based on the regression of later performances on the early predictions. These studies frequently observe an inflation pattern, *i.e.*, the regression coefficient is systematically lower than 1, meaning that later performances of the best candidates are below those initially predicted (and later performances of the worst candidates, if any, are above those initially predicted).

One plausible interpretation is the existence of long-range LD, even across different chromosomes. Consequently, many markers may capture partial effects of supposedly unlinked QTL. Although long-distance LD is notably lower than short-distance LD, the number of long-distant variants is considerably higher and their combined effects can account for a substantial proportion of the genetic variance in a genomic prediction.

In the first part of this study, we demonstrate that markers do capture part of the effects of distant QTL due to the long-distance LD. Because this long-range LD gradually decays over generations, it is imperative to account for the erosion of marker effects to predict the genomic values for the candidates. In the second part, we propose two methods to estimate the specific erosion factor of a reference population and suggest how to use it in practice to adjust DGV.

Materials & Methods

Evidence for linkage disequilibrium across chromosomes

LD across chromosomes was assessed using data from the 2021 female reference populations of six French dairy cattle breeds (Holstein, Montbéliarde, Normande, Abondance, Tarentaise, Vosgienne). The reference populations exhibited varying sizes, ranging from 2617 to 362,363 animals. It is worth noting that Vosgienne, Tarentaise, and Abondance are local mountain breeds, whereas Montbéliarde and Normande are national populations, comprising 18% and 7% of the French dairy herd, respectively. On the other hand, the international Holstein breed accounts for 70% of the French dairy cattle population. In our analysis, we selected one every 20 SNP of the Illumina EuroGMD BeadChip on the 29 autosomes, resulting in a sample of ~3 million r^2 values for each breed (vs ~1.2 billion in total for all SNPs).

Table 1 presents various LD statistics across chromosomes in the female reference populations of six French dairy cattle breeds. While the average r^2 values appear to be small, suggesting limited across-chromosome disequilibrium, it is important to note that the focus here is on the parameter r, as the impact of a QTL on a SNP effect is directly proportional to the correlation between them. These correlation values decreased when the size of the breed (reference population, number of females in the breed, or effective size) increased. The proportion of SNP pairs with |r| exceeding 5% fluctuated notably, ranging from 1.5% to 33%, depending on the breed. Notably, as the reference population size increased, this proportion also tended to decrease. Nevertheless, it is worth highlighting that even with a small percentage, we still observe non-null correlations between 600 to several thousand SNP with a QTL located on a different chromosome (assuming that these r distributions between SNP and QTL are the same as between SNP).

Table 1. Statistics of |r| and r^2 values across the 29 chromosomes in female reference populations of six French dairy cattle breeds (selection of one every 20 SNP within chromosome).

Breeds	# cows	Mean (|r|) *	% |r| > 0.05	Mean(r^2)
Vosgienne	2617	0.0420	33	0.0029
Tarentaise	3788	0.0225	18	0.0015
Abon-dance	7115	0.0268	15	0.0012
Normande	69,220	0.0206	7	0.00073
Montbe-liarde	185,053	0.0173	4	0.00053
Holstein	362,363	0.0148	1.5	0.00038

*statistics based on 2,812,741 to 3,231,800 SNP pairs per breed

Impact of long distance LD on genomic predictions

To investigate the impact of long-distance LD on SNP effects, we focused on the Normande population, which consisted of 69,220 cows (N) with genotypes and phenotypes. In this study, we considered the first five chromosomes comprising 13,608 SNP. Two hundred additive causal mutations (nq=200) were randomly sampled among SNPs with a minor allele frequency (MAF) higher than 0.02. The additive effects of these mutations were independently drawn from a normal distribution, assuming a heritability of 0.3. Two scenarios were tested: (1) the SNP-BLUP model accounted for the ns=13,408 SNPs excluding the QTL; (2) in addition to these SNPs, the SNP-BLUP model also accounted for an additional residual polygenic effect explaining 20% of the genetic variance. Note that in a previous study (Boichard et al., 2022), we have shown with a similar approach that erosion was minimal when the causal variants were included in the analysis, and this scenario is not replicated here. This also agrees with de los Campos et al (2015) who showed that missing heritability does not exist when causal variants are in the model.

The strategy used to compute erosion relied on the determination of the contribution of each QTL to each SNP, as follows. Omitting fixed effects, the SNP-BLUP equations can be written as

$$[\, \boldsymbol{M'M} + \lambda \boldsymbol{I} \,]\, \hat{\boldsymbol{s}} = \boldsymbol{M'y}$$

with **M** the (*N* x *ns*) matrix of cantered and scaled genotypes, **s** the vector of SNP effects, **y** the vector of phenotypes adjusted for the fixed effects, and $\lambda = \sigma_e^2 / \sigma_s^2$ with σ_e^2 and σ_s^2 the residual and the SNP variances, respectively. According to the simulation, the phenotype can be written as **y = Pq + e**, *i.e.*, the sum of *nq* QTL effects and an error term, with **P** the (*N* x *nq*) matrix of genotypes at the QTL level and **q** the vector of true QTL effects. Therefore, the equations can be rewritten as

$$\hat{\boldsymbol{s}} = [\, \boldsymbol{M'M} + \lambda \boldsymbol{I} \,]^{-1} \quad \boldsymbol{M'(Pq + e)} \qquad [1]$$

Let us denote $\mathbf{C} = [\, \mathbf{M'M} + \lambda \mathbf{I} \,]^{-1}$ the inverse of the coefficients matrix. If $\mathbf{c_i}$ is line *i* (corresponding to SNP *i*) of **C,** the contribution of QTL *j* to SNP effect *i* is

$$f_{ij} = \boldsymbol{c_i M' P_j} q_j. \qquad [2]$$

There were *nq* x *ns* = 200 x 13,408 = 2,681,600 such contributions, distributed in the 4 following categories based on the distance (*d*) between the QTL and the SNP: (1) $d < 5$ Mb; (2) $5 < d < 20$ Mb; (3) $d > 20$ Mb with both the QTL and the SNP located on the same chromosome; (4) the QTL and the SNP are located on different chromosomes. Within each of these categories, we computed a partial DGV for each cow within the reference population. Summary statistics were then calculated over 30 replicates quantifying their relative contributions to the total DGV and their correlations.

The same strategy can also be applied in a model including a residual polygenic effect, denoted as **u**. The equations corresponding to **s** and **u** are as follows:

$$\begin{bmatrix} \boldsymbol{Z'Z} + \kappa \boldsymbol{A^{-1}} & \boldsymbol{Z'M} \\ \boldsymbol{M'Z} & \boldsymbol{M'M} + \lambda \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{u}} \\ \hat{\boldsymbol{s}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z'y} \\ \boldsymbol{M'y} \end{bmatrix} \qquad [3]$$

with **Z** being the incidence matrix linking the records of **y** to **u** and $\kappa = \sigma_e^2 / \sigma_u^2$ the corresponding variance ratio.

The **u** equations can be absorbed into **s** equations, resulting in the following formula:

$$\left[\boldsymbol{M}'\left(\boldsymbol{I} - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z} + \kappa\boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{Z}'\right)\boldsymbol{M} + \lambda\,\boldsymbol{I}\right]\hat{\boldsymbol{s}} = \mathbf{M}'\left(\mathbf{I} - \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + \kappa\mathbf{A}^{-1}\right)^{-1}\mathbf{Z}'\right)(\mathbf{Pq} + \mathbf{e})$$

Let us denote **C*** the inverse of the coefficient matrix after absorption

$$\mathbf{C}^{*} = \left[\mathbf{M}'\left(\mathbf{I} - \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + \kappa\mathbf{A}^{-1}\right)^{-1}\mathbf{Z}'\right)\mathbf{M} + \lambda\,\mathbf{I}\right]^{-1}$$

and **M*** the adjusted genotype matrix after absorption

$$\boldsymbol{M}^{*\prime} = \boldsymbol{M}'\left(\boldsymbol{I} - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z} + \kappa\boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{Z}'\right)$$

Then the contribution of QTL *j* to each SNP effect *i* is:

$$f_{ij} = \mathbf{c}^{*}_{i}\ \boldsymbol{M}^{*\prime}\ \boldsymbol{P}_j\ q_j \qquad [4]$$

where $\mathbf{c}^{*}_{i}$ is line i of **C***

Table 2 presents the relative contribution of the 4 categories based on QTL-SNP distance to the total DGV variance.

Results and Discussion

In the model without a polygenic effect (scenario 1), the partial DGV derived from contributions of the QTL close to markers explained approximatively three-quarters of the DGV variance. Notably, more distant markers (d>20 Mb) and markers located on other chromosomes together accounted for ~13% of the total DGV variance. Markers located on other chromosomes explained more variance than markers at more than 20 Mb on the same chromosome. This result can be attributed to the larger number of marker-QTL pairs when markers are located on different chromosomes. The likely underlying reason of this pattern lies in the strong shrinkage of the effects of markers situated close to the QTL due to the influence of the prior information. Indeed, all markers receive the same prior variance, and the parameter λ had a relatively high value compared to the diagonal elements of the matrix **M'M**. Consequently, the estimated effects of markers in proximity to the QTL experience substantial shrinkage and are much smaller, even altogether, than the true QTL effect. As a result, the unexplained part of the true QTL effect becomes available for distant markers, potentially leading to their contribution to total DGV. This effect would be probably maximized when the number of true QTL is much lower than the number of SNPs ($\sigma_q^2 >> \sigma_s^2$), and when the reference population is relatively small (diagonal (**M'M**) does not dominate λ). It is important to note that when the size of the reference population is very large, the influence of prior information decreases, and this observed pattern is likely to gradually decrease.

Table 2. – Relative contributions (%) of each of the 4 classes of QTL-SNP pairs defined according to their distance (d). The contribution of a class is the percentage of DGV variance explained by each partial DGV in the reference population, over 30 replicates.

Classes of partial DGV defined according to QTL-SNP distance (*d*)	Scenario 1 Model without polygenic effect	Scenario 2 Model with polygenic effect
1: *d* < 5 Mb	73.5	68.9
2: 5 Mb < *d* < 20 Mb	13.7	14.3
3: *d* > 20 Mb	4.9	5.2
4: QTL and SNP located on different chromosomes	8.0	11.5

The inclusion of a polygenic effect in the model (scenario 2) resulted in a higher proportion of variance being explained by distant markers and by markers located on different chromosomes. At first glance this result may seem counterintuitive as one could

expect that the polygenic effect would help account for these long-distance effects since it captures the genetic relationships between individuals. Distant markers and markers on another chromosome altogether explained around 17% of the DGV variance whereas the share due to close markers (d<5Mb) decreased to 69%. As in scenario 1, a possible interpretation is the shrinkage of estimated SNP effects. Indeed, in the presence of a polygenic effect with variance σ_u^2, the total variance due to SNPs is reduced to σ_g^2 - σ_u^2 and results in an increased variance ratio λ (σ_e^2 / σ_s^2). It can then be concluded that inclusion of a polygenic effect into the model should primarily be motivated by the need to account for the genetic variance not captured by the SNPs, rather than as a means to reduce inflation.

Table 3a. Average correlations between partial DGV in Scenario 1, without polygenic effect. Results over 30 replicates

Partial DGV class	<5 Mb	5-20 Mb	>20 Mb
5-20 Mb	0.30		
>20 Mb	0.17	0.15	
Other Chromosomes	0.11	0.05	0.12

Table 3b. Average correlations between partial DGV in Scenario 2, with polygenic effect. Results over 30 replicates

Partial DGV class	<5 Mb	5-20 Mb	>20 Mb
5-20 Mb	0.54		
>20 Mb	0.26	0.29	
Other Chromosomes	0.25	0.22	0.27

Tables 3a and 3b present the correlations between the partial DGVs derived from different categories of QTL-SNP distances in scenarios 1 and 2, respectively. Both scenarios presented low to moderate positive correlations, illustrating that distant QTL contribute to the effects of many markers. Inclusion of a polygenic effect in the model (scenario 2, table 3b) increases these correlations showing that long distance effects are reinforced.

Methods to estimate erosion factor of SNP effects

The concept of erosion of the genomic breeding values has two distinct components: (a) a component that is characteristic of the reference population itself, and (b) a component that is specific to each candidate and its genetic distance from the reference population.

The extent of long-distance LD in the reference population is influenced by the effective population size (Ne). Notably, when Ne is small, a non-zero LD baseline persists. More importantly, the level of long-distance LD is also strongly dependent on the genetic relatedness within the reference population, which can be different from the relatedness in the overall population. A higher average relationship between individuals within the reference population results in more long-distance LD. It can be argued that, on average, the long-distance LD appears relatively stable across generations, but this stability does not hold for individual pair of markers. At a given generation, existing LD is halved in the subsequent generation due to recombination, but new LD can emerge from different marker pairs as a consequence of the random processes associated with the sampling of parents and genetic drift, making average LD stable.

The theoretical derivation of the erosion factor ρ requires additional investigation. Nevertheless, practical and efficient solutions can be obtained through simulation. In this paper, we present two simulation-based approaches which offer practical and effective means to address the erosion phenomenon and to estimate ρ.

Method 1: by simulating a new generation

The real reference population G_r of the breed for a given trait is considered with its SNP genotypes **M**. As previously, *nq* QTL are simulated in this reference population by sampling SNP which are thereafter excluded from the analysis. Expectations of SNP effects are estimated by $\widehat{\mathbf{s_r}} = (\mathbf{M'M} + \lambda\mathbf{I})^{-1}\,\mathbf{M'}\,\mathbf{P}\,\mathbf{q}$, assuming the same previous notations.

A new generation, G_n, is then simulated, by sampling parents (at random or following a predefined design) in the reference population and performing matings. The expected DGV of this new generation is obtained from the genotypes M_n and from the SNP effects estimated in the reference population: $\mathbf{DGV_r} = \mathbf{M_n}\,\widehat{\mathbf{s_r}}$

Assuming phenotypes are known in this new generation, new SNP effect estimates can be obtained from generation *n* only $\widehat{\mathbf{s_n}} = (\mathbf{M_n'}\,\mathbf{M_n} + \lambda\mathbf{I})^{-1}\,\mathbf{M_n'}\mathbf{P_n}\mathbf{q}$, and a new set of DGV is obtained from these new SNP estimates $\mathbf{DGV_n} = \mathbf{M_n}\,\widehat{\mathbf{s_n}}$

These new SNP effects are different from the previous ones if the covariances between markers and QTL $\mathbf{M'P}$ and $\mathbf{M_n'}\,\mathbf{P_n}$ differ. A large change in the covariances between markers $\mathbf{M'M}$ and $\mathbf{M_n'}\,\mathbf{M_n}$ (i.e., in LD between SNP) may also affect the results but probably to a lesser extent.

From these two sets of DGV, an estimate of the erosion ρ between the two generations is obtained through a regression analysis, where:

$$\boldsymbol{DGV_n = 1\mu + \rho\, DGV_r + e} \qquad [5]$$

Method 2: by regressing contributions of QTL to marker effects.

As above, the real reference population G_r of the breed for a given trait is considered with its genotypes **M** and *nq* QTL are simulated in this reference population. All contributions f_{ij} of the QTL *j* to the effect of SNP *i* are computed as shown in equation [2]:

$$f_{ij} = \boldsymbol{c_i\, M'\, p_j\, q_j}.$$

DGV_r in the reference population is the sum of all contributions:

$$\boldsymbol{DGV_r = M\, f\, 1_q}$$

with $\mathbf{1_q}$ being a vector of 1 of size q.

Note that the DGV can also be obtained as $\mathbf{DGV_r} = \mathbf{M}\,\widehat{\mathbf{s_r}}$, as in Method 1.

Then, all f_{ij} are regressed according to the genetic map, with coefficients $(1\text{-}r_{ij})$ varying from 1 to 0.5, r_{ij} being the recombination rate between the loci *i* and *j*:

$$h_{ij} = r_{ij}\, f_{ij}. \qquad [6]$$

New eroded DGV ($\mathbf{DGV_e}$) are the sum of all regressed contributions

$$\boldsymbol{DGV_e = M\, h\, 1_q}$$

An estimate of $\sqrt{\rho}$ is obtained through a regression analysis, where:

$$\boldsymbol{DGV_e = 1\mu + \sqrt{\rho}\;\; DGV_r + e} \qquad [7]$$

Comparison of both methods

Both methods are based on QTL simulation, and their results are influenced by assumptions, particularly regarding the number of QTLs. However, as far as the number of QTLs is smaller than the number of markers and long-distance LD is present, we can anticipate that erosion exists.

Method 1 is relatively straightforward to implement, as it involves simulating one additional generation and estimating expected SNP effects in both the reference population and in the new generation using standard software. In contrast, Method 2 requires a specific program to compute all the contributions and erode them. Nevertheless, it provides an explicit biological basis to understand and interpret erosion across generations.

Method 1 generates progeny from pairs of parents, leading to erosion on both sire-progeny and dam-progeny pathways. Method 2, on the

other hand, simulates erosion through recombination at only one meiosis, resulting in the erosion factor estimated by method 2 being the square root of that estimated by method 1. This scale difference should be considered when interpreting results.

Additionally, method 2 considers the whole reference population whereas method 1 generates a new generation based on assumptions about the number and the choice of parents sampled. Therefore, results between the two methods can exhibit slight variations.

Numerical example

In our numerical example, we applied both method 1 and method 2 to the same 2021 Normande female reference population. As before, we focused on only 5 chromosomes and simulated 200 QTLs. In method 1, a new generation was created by sampling 1,000 sires and 50,000 dams. Each sire had 50 progeny, while each dam had one progeny, resulting in a new generation of 50,000 animals. The results obtained with both methods are presented in Table 4. Recombination rates were based on ARS-UCD1.2 bovine genome assembly, assuming 1 cM for 1 Mb.

Table 4. Estimation of erosion factors by Method 1 and Method 2 (30 replicates)

	$\hat{\rho}$	SD($\hat{\rho}$)
Method 1	0.87	0.015
Method 2	0.84	0.010

The number of replicates was set to 30. Variability across replicates was small (at least with such a large reference population) and this number of replicates was sufficient to obtain reliable estimates of ρ.

Discussion on how to apply erosion in practice

Here, we consider that, by definition, individuals in the reference population are assumed to possess non-eroded SNP effects. It is worth noting that this assumption may warrant discussion due to potential heterogeneity within the reference population. The estimated SNP effect represents an expectation and may not align with individual situations. However, this point goes beyond the scope of this initial study.

Erosion primarily concerns selection candidates, *i.e.* genotyped individuals without phenotype data and, therefore, out of the reference population. When their parents, referred to as *s* and *d*, are part of the reference population, we assume that the parent's average (PA) DGV, denoted as

$$PA = 0.5\,(DGV_s + DGV_d),$$

remains unaffected. Indeed, their DGVs are based on performances; and this is especially the case for sires with progeny evaluation, and therefore with very reliable DGVs. Erosion influences the deviation from PA, *i.e.*, the predicted Mendelian sampling term. We propose applying the following formula:

$$DGV_{eroded} = PA + \rho\,(DGV\text{-}PA) \qquad [8]$$

When the parents of a candidate are not in the reference population, erosion applies at each generation between the reference population and the candidate. Following a similar approach as described by Dekkers et al (2021), the number of generations between the candidate and its closest relatives within the reference population is determined on both the sire and dam pathways and k is their sum. Erosion is applied following equation [9]:

$$DGV_{eroded} = PA + \rho^{k/2}\,(DGV\text{-}PA) \qquad [9]$$

When the parents themselves are candidates, *i.e.,* genotyped and not in the reference population, erosion also applies to them. This erosion affects the PA of their progeny in the following way:

$$DGV_{eroded} = PA_{eroded} + \rho^{k/2}\,(DGV\text{-}PA)$$

This formula should be applied recursively, processing parents before progeny. This recursive approach highlights that the DGV of candidates born from very young parents experience significant erosion, which aligns

well with practical observations. Therefore, breeding schemes with accelerated generations without updating reference data tend to accumulate more erosion than initially anticipated. These schemes may be less appealing due to the rapid erosion effect on genomic values.

Furthermore, as shown by Dekkers et al (2021), erosion also affects reliability, but with a coefficient equal to ρ^2 instead of ρ. The loss in genomic accuracy is therefore very fast. Theoretical accuracies calculated based on the inverse of the coefficient matrix tend to overestimate the reliabilities for candidates and must be adjusted accordingly. In the French evaluation system, reliabilities are computed by combining effective record contributions (ERC) associated to polygenic information and genomic information, the latter (and the latter only) being eroded in candidates.

Conclusions

The practical implications of erosion in genetic evaluations and breeding programs are important. When considering the overall prediction of selection candidates, contributions from short-distance LD tend to remain relatively stable because they are only mildly eroded by recombination (Dekkers et al, 2021). However, contributions from long-distance LD are halved at each generation. The extent of erosion varies with the relative weight of short and long-distance LD, but it should never be disregarded.

Further investigations are needed to theoretically determine the erosion factor ρ. Nonetheless it is clear this factor depends on baseline LD in the population, *i.e.,* effective population size (Ne) and genome length (L), as well as on the structure of the reference population. Additionally, it is likely influenced by the genetic architecture of the traits (such as the number of QTL and magnitude of their QTL effects) and the model used (SNP-BLUP/GBLUP vs Bayesian models, the latter being likely less affected by erosion).

It is also relevant to explore the impact of the structure of the reference population, such as its heterogeneity in terms of time span, selection, and relationship. For instance, the impact on erosion of old generation data in the reference population would be worth investigating. While the proposed erosion methods consider the smallest distance between the candidate and the reference population, alternative approaches using the barycentre of the reference population warrant investigation.

However, one can anticipate that:

(1) Inflation factors, frequently observed between 0.8 and 0.9, give the magnitude of the erosion phenomenon;

(2) Models including causal variants tend to be more persistent and less subject to erosion, as demonstrated by Boichard et al (2022);

(3) Models that incorporate a residual polygenic component may appear to have less inflated predictions for candidates because they combine two estimates of the MS term: the genomic estimate, which is inflated, and a polygenic estimate, which is equal to zero (i.e., 100% deflated). However, the polygenic effect does not capture long-distance LD effects and, therefore, does not improve predictions in terms of persistence. We believe that accounting for erosion is a more rigorous and accurate approach, even if it requires post-processing.

This methodology has been implemented in the French Single Step bovine evaluation since March 2022.

References

Boichard D., Fritz S., Croiseau P., Ducrocq V., Cuyabano B., and Tribout T. 2022. Long-distance associations generate erosion of genomic breeding values of candidates for selection. 12th World Congress on Genetics Applied to Livestock Production, Rotterdam, The Netherlands, July 3-8 2022,

288. https://doi.org/10.3920/978-90-8686-940-4_288.

Dekkers J.C.M., Su H.L., and Cheng J. 2021. Predicting the accuracy of genomic predictions. *Genet Sel Evol* 53, 81. https://doi.org/10.1186/s12711-021-00647-w.

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., et al. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J Dairy Sci* 95(7), 4114-4129. https://doi.org/10.3168/jds.2011-5019

Habier D., Fernando R.L., and Dekkers J.C.M. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389-2397. https://doi.org/10.1534/genetics.107.081190

Habier D., Fernando R.L., and Garrick D.J. 2013. Genomic BLUP decoded: A look into the black box of genomic prediction. *Genetics*, 194, 597–607. https://doi.org/10.1534/genetics.107.081190

Hozé C., Fritz S., Phocas F., Boichard D., Ducrocq V., et al. 2014. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population *J Dairy Sci* 97(6), 3918-3929. https://doi.org/10.3168/jds.2013-7761

Legarra A., Robert-Granie C., Manfredi E., and Elsen J.M. 2008. Performance of genomic selection in mice. *Genetics* 180(1), 611-618. https://doi.org/10.1534%2Fgenetics.108.088575.

Pszczola M., Strabel T., Mulder H.A., and Calus M.P.L. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J Dairy Sci* 95(1), 389-400. https://doi.org/10.3168/jds.2011-4338

Solberg T R., Sonesson A.K., Woolliams J.A., Odegard J., and Meuwissen T.H.E. 2009. Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. *Genet Sel Evol* 41, 53. https://doi.org/10.1186/1297-9686-41-53.

Sonesson A.K., Meuwissen T.H.E. 2009. Testing strategies for genomic selection in aquaculture breeding programs. *Genet Sel Evol* 41, 37. https://doi.org/10.1186/1297-9686-41-37