Which individuals to choose to update the reference population? Minimizing the loss of genetic diversity in animal Genomic Selection programs

SE Eynard *,+,(sonia.eynard@wur.nl)
P Croiseau *(pascal.croiseau@inra.fr)
D Laloë *(denis.laloe@inra.fr)
S Fritz *,+,(sebastien.fritz@inra.fr)
MPL Calus †(mario.calus@wur.nl)
G Restoux *(gwendal.restoux@agroparistech.fr)

* GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy en Josas, France
† Wageningen University & Research Animal Breeding and Genomics, 6700 AH
Wageningen, The Netherlands
‡ Wageningen University & Research, Centre for Genetic Resources the Netherlands, 6700
AA Wageningen, The Netherlands
§ Allice, 75595 Paris Cedex12, France

© The Author(s) 2013. Published by the Genetics Society of America.
Running title: Genetic diversity in Genomic Selection

Key words: genomic selection, genetic diversity, reference population, optimal contribution

Corresponding author:

Sonia E. Eynard
Wageningen University & Research
ABGC
P.O. Box 338, 6700 AH Wageningen, NL
Radix building nr. 107
Droevendaalsesteeg 1, 6708 PB Wageningen, NL
sonia.eynard@wur.nl
Genomic selection is commonly used in livestock and increasingly in plant breeding. Relying on phenotypes and genotypes of a reference population, genomic selection allows performance prediction for young individuals having only genotypes. This is expected to achieve fast high genetic gain but with a potential loss of genetic diversity. Existing methods to conserve genetic diversity depend mostly on the choice of the breeding individuals. In this study we propose a modification of the reference population composition to mitigate diversity loss. Since the high cost of phenotyping is the limiting factor for genomic selection our findings are of major economic interest. This study aims to answer the following questions:

How would decisions on the reference population affect the breeding population? How to best select individuals to update the reference population and balance maximizing genetic gain and minimizing loss of genetic diversity? We investigated three updating strategies for the reference population: random, truncation and optimal contribution strategies. Optimal contribution maximizes genetic merit for a fixed loss of genetic diversity. A French Montbéliarde dairy cattle population with 50K SNP chip genotypes and simulations over ten generations were used to compare these different strategies using milk production as the trait of interest. Candidates were selected to update the reference population. Prediction bias and both genetic merit and diversity were measured. Changes in the reference population composition slightly affected the breeding population. Optimal contribution strategy appeared to be an acceptable compromise to maintain both genetic gain and diversity in the reference and the breeding populations.
INTRODUCTION

The development of genomic selection (GS), as described by Meuwissen et al. (2001), is the most important recent innovation in animal breeding. In livestock breeding, GS comprises the estimation of genomic estimated breeding values (GEBVs) and the actual selection of individuals with only genotypes available, e.g. young individuals that are candidates for selection, based on these GEBVs (Figure S1). A reference population, composed of individuals with known phenotypes and genotypes based on many markers across the genome, is used to set up prediction equations and infer GEBVs of selection candidates. The main advantages of GS compared to the traditional methods based on phenotype and pedigree, are that generation intervals are reduced since phenotypes of mature progenies are no longer needed to perform genetic evaluation. Secondly, selection can still be performed with the same accuracy as classical selection and lastly, it allows selection for new traits that are difficult and costly to record (Meuwissen et al. 2001; Calus and Veerkamp 2011).

Despite the confirmed advantages, most of the knowledge on the long-term impact of GS is based on simulation studies (for example: Colleau et al. (2009), Jannink (2010), Bastiaansen et al. (2012) and Clark et al. (2013)) and many questions remain concerning its use. In particular, the following questions remain about the design of the reference population: how many individuals are needed (Pszczola et al. 2011; Khatkar et al. 2012; Pryce and Daetwyler 2012), how often marker effects should be re-estimated (Calus 2010; Heslot et al. 2013), how closely related individuals in reference population should be to the selection candidates (Pszczola et al. 2012a; Meuwissen et al. 2013), and which individuals should be used to update the reference population (Rincent et al. 2012; Isidro et al. 2015).
Many livestock breeds have high inbreeding rates and low genetic diversity as a result of intensive selection (LEROY et al. 2011). Limited genetic diversity restricts the potential long-term genetic gain of the populations (LI et al. 2008; GODDARD 2009; JANNINK 2010; ENGELSMA et al. 2012; LIU 2013; HENRYON et al. 2014) and reduces their ability to respond to new challenges (TORO et al. 2009; ALLENDORF et al. 2010; STOCK and REENTS 2013; BRUFORD et al. 2015). To allow for long-term maintenance, individuals representing the overall population diversity need to be used for breeding (RINCENT et al. 2012; HESLOT et al. 2013; ISIDRO et al. 2015). Different strategies have been previously suggested: 1) limiting the number of offspring per male to avoid the sire ‘star system’ (DANCHIN-BURGE et al. 2012; BOICHARD et al. 2015), 2) distinguishing individuals according to the marker variation they carry and giving extra weights to the low-frequency favorable markers (JANNINK 2010) or 3) choose individuals to represent the highest overall population diversity (MEUWISSEN 1997; RINCENT et al. 2012; HESLOT et al. 2013). One of the available methods developed for such a goal is the optimal contribution (OC) strategy as defined by Meuwissen (1997). The OC strategy can be used to simultaneously conserve genetic diversity and achieve genetic gain by minimizing the relationships between the individuals (ENGELSMA et al. 2011; SONesson et al. 2012; CLARK et al. 2013; DE CARA et al. 2013; EYNard et al. 2016). The effectiveness of these methods relies on the final choice of the breeding individuals. In the case of dairy cattle, such strategies to conserve overall population genetic diversity may be insufficiently used in the context of competitive economical markets promoting the use of elite reproducers. Methods implicitly driving selection toward both genetic gain and the maintenance of genetic diversity may be the alternative. With the design of the reference population there is the potential to modify the breeding population by changing the genetic evaluation.
In this study we addressed the following question; how does one choose individuals to update the reference population of a GS scheme in order to balance genetic gain and genetic diversity? We anticipate that changes in the composition of the reference population will be associated with changes in the breeding population due to adjustments of the prediction equations for GS. To test this hypothesis we compared three different strategies (Random, Truncation and OC strategy) to select individuals for the update of the reference population. Using a real dataset of French dairy cattle (Montbéliarde), we focused on the effect of updating strategies on the population of selected candidates. Using simulations, we inferred the long-term effect of these updating strategies on the breeding population. For both real and simulated datasets, updating strategies were evaluated in terms of genetic merit, genetic diversity and performances of GS.

MATERIALS AND METHODS

Real dataset

A population of 14,052 individuals from the French Montbéliarde dairy cattle breed, 2,459 males and 11,593 females, born between 1969 and 2011 was available for the analysis. The complete pedigree record contained 50,852 individuals born from the 1940s until 2011. All individuals had, at the very least, complete pedigree records for their parental generation with a maximum of seven complete generations. The generation equivalents (sum of the proportion of known ancestors in all available generations (MAIGNEL et al. 1996)) ranged from two to nine. For all individuals 50K SNP genotypes were available. Males were genotyped using the BovineSNP50 v2 BeadChip (Illumina ®) and females were genotyped using the 10K SNP chip (Illumina ®) and subsequently imputed, by Hoze et al. (2013), to the BovineSNP50 v2
BeadChip using the BEAGLE software (Browning and Browning 2007). The software DAGPHASE (Druet and Georges 2010) was used for phasing. Subsequent quality control steps were required for each SNP: i) a minimum call rate higher than 90%, ii) non-departure from Hardy-Weinberg equilibrium (p-values <10\(^{-4}\)) and iii) MAF > 1%, to minimize potential genotyping errors. The final genotype data comprised 43,801 markers genotyped on the 29 autosomes. In this study we focused on milk yield having heritability of 0.3, a genetic variance of 423,390 kg\(^2\) and a residual variance of 987,910 kg\(^2\). Milk yield was measured as the corrected milk yield for the females with, on average, 1.66 records per female. For the progeny-tested males, milk yield was measured as daughter yield deviation (DYD), reflecting the average milk yield of their daughters adjusted for fixed and non-genetic random effects and the additive genetic value of their dam (Mrode and Swanson 2004). Weights used for male records were defined as effective daughters’ contribution (EDC) (Fikse and Banos 2001) and were on average 26.21. The dataset was divided into three groups according to individuals’ birth years. The first group included 5,969 individuals (2,325 males and 3,644 females) born between 1969 and 2007 and was used as the initial reference population for GS (A\(_1\)). The second group included 3,791 individuals (134 males and 3,657 females) born in 2008 and 2009, and those individuals were considered to be available to add in the updated reference population (A\(_2\)). The third group included 4,292 individuals (all females) born in 2010 and 2011, and was used for validation of the GS (V) (Figure S2).

Simulation process

We simulated a population with characteristics similar to a domestic cattle population and a trait similar to milk yield. An ancestral population of 1000 males and 1000 females that had
undergone selection based on estimated breeding values (EBVs) estimated from a best linear unbiased prediction (BLUP) method was used as the starting point of our simulations. Ten more generations of selection and breeding were simulated. In every generation the 150 males and 500 females from the previous generations with the highest GEBVs were selected to produce the next generation \( n+1 \) (a selection rate of 0.6 for the males, of 1 for the females from the generation \( n \) and of 0.5 for the females from the generation \( n-1 \)). Males could reproduce for one generation while females could produce offspring in multiple generations assuming that their GEBVs were high enough. We assumed that selection excluded them from the population after two years. Each female produced one offspring per generation and the sex ratio in the offspring generation was 0.5 (Figure 1). The simulated design is simpler than what occurs in a real breeding scheme. Simulations were performed using QMSim (SARGOLZAEI and SCHENKEL 2009). Details of the simulation process are provided in supplementary material (Supplementary1).

**Genomic best linear unbiased prediction**

To investigate the impact of an update to the reference population on GS in terms of subsequent predicted GEBVs we used both real and simulated datasets. The real dataset allowed us to study the impact of reference population updating strategies on the choice of breeding individuals for the next generation only. Simulations were used to study the impact on the breeding population over multiple generations. GEBVs were predicted by a genomic best linear unbiased prediction (GBLUP) model fitted with the GS3 software (LEGARRA et al. 2011). For the GBLUP model (CROISEAU et al. 2011): i) the estimated relationship matrix was calculated according to the VanRaden (2008) equation:

\[
G = \frac{ZZ'}{2\sum_{i=1}^{m} p_i (1-p_i)}
\]

where \( Z \) is the
genotype matrix and $p_i$ the allele frequency of marker $i$, ii) the variance components for this
trait were the ones used in the routine evaluation in France and were fixed in the model, and
iii) only random effects were fitted as the phenotypes used were already corrected for fixed
effects and non-genetic random effects.

Reference population update

Three updating strategies were compared: 1) selection at random (Random) repeated 100
independent times, 2) truncation selection based on the highest GEBVs (Sel) and 3) selection
to simultaneously maximize the genetic diversity and the genetic merit of the group of
selected individuals (SelDiv) using OC strategy and the Gencont program (Meuwissen 1997).
Genetic merit of a set of selected individuals is the average breeding value (BV) of the
selected individuals. The rate of inbreeding ($\Delta F$) between the current and next generation is
estimated from the average genomic relationships of selected individuals. The OC method
identifies a set of individuals with maximum genetic merit with the restriction that the
expected rate of inbreeding is no more than 1%, as recommended by the FAO (1998). If the
given constraint of 1% cannot be met because of population structure, then the choice of
individuals is made to minimize the rate of inbreeding and genetic merit is effectively not
considered. The SelDiv strategy used genomic relationships, computed as similarities that
count the number of identical alleles, averaged across loci between two individuals (Nejati-
Javaremi et al. 1997; Eding and Meuwissen 2001):

$$G_{jk} = \frac{2}{N} \sum_i (x_{ij} - 1)(x_{ik} - 1)$$
Where $N$ is the number of markers and $G_{jk}$ is the estimated relationship between individual $j$ and $k$ across all markers. At each marker, $i$ and $x_i$, the individual variant genotype is coded as 0, 1 or 2. Note that computing these relationships using the methods described by VanRaden (2008) and Yang et al. (2010), assuming allele frequencies of 0.5 for all loci, yields exactly the same result. This relationship matrix has been shown to reduce the loss of overall genetic diversity better than other relationship matrices when applying OC strategy (Eynard et al. 2016).

**Update of reference population in real datasets:** The initial reference population ($A_1$) was used to predict GEBVs of the individuals in the candidates’ population ($A_2$). Using these GEBVs and the relationships between individuals in $A_1$ and $A_2$ we selected subgroups of individuals to build updated reference populations ($A_{1+2}$). For all strategies (Random, Sel and SelDiv) the initial reference population ($A_1$) of 5,969 individuals was updated with 100, 200, 500, 1000 or 2000 new individuals, which represented adding approximately 1.5, 3, 8, 15 and 30% to the initial reference population, respectively. The updated reference populations ($A_{1+2}$) were used to predict GEBVs of the candidates’ group $V$. Based on their GEBVs the top 100 individuals from $V$ were selected as breeding populations, $V_{sel}$. A detailed review of all results is available in the supplementary Table S1.

**Update of reference population in simulated datasets:** The initial reference population ($A_1$) consisted of 1000 males from the ancestral individuals and was updated every generation by adding 150 individuals, males and/or females, selected based on one of the proposed strategies (Random, Sel and SelDiv). The size of the reference population, therefore, rose from 1000 in the first generation to 2,350 individuals in the tenth generation. In each generation the reference population was updated based on GEBVs from the candidates’ population, and
subsequently used for prediction of GEBVs of the simulated offspring. Therefore, individuals in the reference population could be included as part of the breeding population provided that they have been selected for breeding based on their GEBVs. The whole simulation and updating process was replicated 50 times for each strategy.

**Evaluation of updating strategies**

To compare the different updating strategies several parameters were evaluated for the selected candidates’ population ($V_{sel}$, top 100 individuals) in the real dataset and for the breeding population in the simulated dataset. Those parameters included: i) the response to selection, ii) the genetic diversity, iii) prediction bias and iv) the effective population size of the reference population. Response to selection was measured as the change in average BV. Genetic diversity was measured as: i) observed heterozygosity and ii) the inbreeding coefficient obtained from pedigree following the Sargolzaei *et al.* (2005) algorithm. The bias of GEBV was measured by the absolute standardized prediction errors for the BV as follows:

$$\text{Bias}_k = \frac{|GEBV_k - BV_k|}{\sigma_i},$$

Where $GEBV_k$ is the GEBV of the individual $k$, $BV_k$ is the Breeding Value (based on multiple records in the real dataset or given by the simulations in the form of a True Breeding Value) of the individual $k$ and $\sigma_G$ is the True Breeding Value standard deviation of the population under scrutiny $i$. The effective population size of the reference population, $N_e$, was also estimated following the classical formula derived from the inbreeding coefficient definition (Falconer and Mackay 1996):

$$N_e = \frac{1}{2 \times f_t}$$

With $f_t$ representing the mean inbreeding coefficient of the population in the $t^{th}$ generation.
The effects of the different updating strategies on BV, heterozygosity, inbreeding and prediction bias were tested using linear models implemented in R and the *lme4* package (Bates et al. 2015; R Core Team 2016) considering the Random strategy as the null hypothesis distribution. When dealing with heterozygosity or inbreeding, an arcsine-square root transformation was applied to ensure the applicability of linear models. The effects of strategy and the size of the update were tested using a type II ANOVA (R package *car* (Fox and Weisberg 2011)). Coefficients of change throughout generations were compared using least square means for qualitative variables and least square trends to compare regression slopes for quantitative variables (R package *lsmeans* (Lenth 2016)).

For the real dataset, linear models were applied on the candidates’ populations as follows,

$$Y_{ijk} = \mu + \text{strategy}_i + \text{update size}_j + (\text{strategy} \times \text{update size})_{ij} + \beta_1 \left(\frac{N_e}{N}\right)_{ij} + \varepsilon_{ijk},$$

Where $Y_{ijk}$ is the variable measured on individual $k$, for strategy $i$ (Random, Sel or SelDiv), when adding update size, number of individuals added to the reference population, fitted here as a qualitative effect (100, 200, 500, 1000 or 2000). $\beta_1$ is the regression coefficient on the ratio $\frac{N_e}{N}$ of the reference population (with $N$ the census population size) and $\varepsilon_{ijk}$ is the gaussian residual. For simulated datasets, we focused on the breeding and offspring populations using the following mixed effects models,

$$Y_{ilk} = \mu + \text{strategy}_i + \beta_2,i(generation_l) + \alpha_i(generation_l) * (\text{strategy}_i) + \beta_1 \left(\frac{N_e}{N}\right)_{il} + Sim_l + \varepsilon_{ilk},$$

$$Sim_l \sim N(\mu = 0, \sigma^2_{sim}).$$

Where $Y_{ilk}$ is the variable measured on individual $k$, for the strategy $i$, in generation $l$ of simulation, $\beta_2,i$ the regression coefficient on the generation number for strategy $i$, $\alpha_i$ is the
interaction effect of method with generation, and $\text{Sim}_t$ was the random effect of the simulation where $\sigma^2_{\text{sim}}$ represented the data variability among simulation replicates and $\varepsilon_{ijk}$ the gaussian residuals. The ratio $\frac{N_e}{N}$ of the reference population was used in the model to account for the effect of the change in reference population size through time while accounting for a parallel growth of census population size. This allows one to distinguish between the increases in size over time from the cumulative effect due to consecutive population changes over the ten generations.

**Data availability**

Genetic information (in the form of a $G$-matrix), pedigree (for the individuals under scrutiny) and BV for the trait of interest are available for the real dataset, as well as the script allowing the production of the simulated datasets and documents describing each files for real and simulated datasets on the following depository: doi.org/10.5281/zenodo.1000534. Programs and scripts used for the GS, reference population update and for the post-hoc analysis are available upon request.

**RESULTS**

**Effect of updating strategy on selected candidates (real dataset)**

**Genetic merit of the selected candidates:** Individual BVs in $V_{sel}$ exhibited large variability and ranged from 461 to 5674. Average BV of $V_{sel}$ populations, across all combinations of strategies and the size of updates, ranged from 3153.56 to 3185.63 ($\pm$ 5.21), thus revealing limited variation in genetic gain between different strategies to update the reference
population. Even though none of these differences were significant, genetic merit tended to increase when increasing the size of the group used to update the reference population.

**Genetic diversity of the selected candidates:** Individuals’ inbreeding ranged from 0.02 to 0.11. Over all combinations of strategies and size of updates, per $V_{sel}$, the inbreeding coefficients were all on average 0.05 ($\pm 1.14 \times 10^{-4}$) and not significantly different from each other. Individuals’ heterozygosity ranged from 0.28 to 0.33 and average populations’ heterozygosities were all close to the mean value of 0.31 ($\pm 5.65 \times 10^{-5}$), and not significantly different across scenarios.

**Precision of GEBV prediction procedure:** The prediction bias of GEBVs of the full candidates’ population, $V$, ranged from 0.00 to 7.73, indicating substantial disparity in how well individuals’ GEBVs are predicted. Across all combinations of strategies and size of updates, average absolute bias of GEBV ranged from 1.05 to 1.08 ($\pm 0.01$) without any significant difference among them (Table 1).

Overall, no significant differences could be observed between the three tested strategies when considering the top 100 candidates for selection.

**Long-term effect of updating strategy on breeding population (simulated datasets)**

**Genetic merit of the breeding population:** The average BV of the breeding population always increased from one generation to the next. Despite the fact that strategy significantly affected the realized genetic merit (all p-values <10^{-5} Supplementary Table S3), the actual differences between the $Sel$, $SelDiv$ and $Random$ strategies were very modest (Table 2, Figure 2, Supplementary Table S2).

**Genetic diversity of the breeding population:** Whatever the strategy, the inbreeding coefficient increased from one generation to the next. Despite large standard errors (Figure 2)
the increase in inbreeding coefficients throughout the ten generations appeared to be
significantly slower for SelDiv than for the two other strategies (Table 2). Inbreeding level
was significantly associated with both generation number and \( \frac{N_e}{N} \) (p-values <10^{-16},
Supplementary Table S3). Both an increase in generation number and a decrease in \( \frac{N_e}{N} \) was
associated with an increase of the average population inbreeding. After the fourth generation,
the SelDiv strategy resulted in higher heterozygosity than the Sel or Random strategies (Figure
2) due to a slower decrease over generations (Table 2). All the parameters, strategy (p-
value=1.12*10^{-2}), \( \frac{N_e}{N} \) (p-value=1.26*10^{-6}), generation number and the interaction between
strategy and generation (both with p-values <10^{-16}), significantly affected the heterozygosity
(Supplementary Table S3). The effect of \( \frac{N_e}{N} \) was positive; an increase in \( \frac{N_e}{N} \) caused an increase
in average heterozygosity of the population. Average heterozygosity decreased from one
generation to the next faster for the Random and Sel strategies than for SelDiv.

**Precision of GEBV prediction procedure:** For all generations, on average the Sel strategy
and even more so the SelDiv strategy, resulted in lower prediction bias of the offspring’s
GEBVs than the Random strategy (Supplementary Table S2). The parameters strategy,
generation number, interaction between strategy and generation, and \( \frac{N_e}{N} \) significantly affected
prediction bias, with p-values <10^{-10} (Supplementary Table S3). The Random, Sel and SelDiv
strategies were significantly different from each other (Table 2). A shift was observed at the
fourth generation, with the Random strategy having the largest bias whereas the SelDiv
strategy had the lowest bias (Figure 2). Despite the apparently chaotic behavior of this
variable, prediction bias tended to increase over time faster for the Random and Sel strategies
than SelDiv. The small effect of \( \frac{N_e}{N} \) on the prediction bias is presumably due to the decline in
relationships between reference and candidate populations through time as a result of the constant addition of new individuals without the removal of older ones.

To summarize, the results above show that different strategies to update the reference population have a significant, but small, impact on the breeding population. The SelDiv strategy resulted in slightly higher genetic diversity in the breeding population accompanied by a minor impact on the genetic gain and lower long-term prediction bias.

DISCUSSION

In this study we compared the impact of different strategies to update the reference population in a GS framework on the genetic merit and diversity of the resulting breeding population. Optimizing the updating strategy is especially important in artificial selection based on the genotypes of individuals at an early age. This is because phenotyping is the limiting factor due to the time and money investment for the rearing of the individuals (COLLEAU et al. 2009; KONIG et al. 2009). It is also relevant when both phenotypes and genotypes are available but only a fraction can be included in the reference population, for example, when designing a core collection in plant breeding (RINCENT et al. 2012; ISIDRO et al. 2015). In GS, reference population design and breeding decisions are linked through EBVs of selection candidates. Our hypothesis was that the choice of individuals in building the reference population might impact the EBVs of selection candidates and consequently the breeding population, both in terms of genetic gain and diversity.
Analysis based on a single generation in the real dataset did not show significant differences between the three proposed updating strategies, however, analysis based on a simulated dataset over ten generations did show significant effects of the updating strategy on the breeding populations’ over time. A small beneficial response of the truncation strategy was observed for genetic merit whilst the OC strategy performed best at conserving genetic diversity.

A recent study by De Beukelaer et al. (2017) focused on the similar question of how to balance genetic gain and genetic diversity conservation in populations under selection. The authors used simulations to compare established selection strategies: GS including OC (GOCS) and GS weighting for rare alleles (GSW) for long-term genetic diversity conservation in plant breeding. Even though both GOCS and GSW outperformed GS for long-term genetic gain, they were not successful in controlling inbreeding rate and loss of rare variants in the breeding population. These authors proposed two new strategies combining index-based method and expected heterozygosity (IND-HE) or rare allele frequencies (IND-RA) as alternatives outperforming GS, GOCS and GSW in balancing genetic gain and diversity. These methods require further investigation to confirm their benefit in practice.

Approaches proposed in plant breeding to design reference populations representing the population structure and diversity (LALOE 1993; RINCENT et al. 2012; ISIDRO et al. 2015; BARTHOLOME et al. 2016) could also be alternatives in the context of animal breeding. In fact, the current concerns of how to best design reference population by targeting only relevant individuals has also become of interest for animal breeding due to the increasing availability of individual information both for phenotypes and genotypes. The data on livestock reference
population are now far more comprehensive and should enable choices of which individuals should be present in the reference population to take place. Therefore, methods used in plant breeding, mostly to design core collections, may be of interest for animal breeders.

**Potential implication for animal breeding**

Breeding decisions in practice are mainly based on the genetic merit of individuals. This is because breeders’ incomes come from production. This phenomenon is putting small breeds in a difficult situation, in a market mostly dominated by mainstream breeds, because of their limited population size, high inbreeding rates and lower fitness potential (TORO *et al.* 2009; ALLENDORF *et al.* 2010; PRYCE and DAETWYLER 2012). Livestock breeding has to balance the conservation of genetic diversity against genetic gain. Within GS, the adoption of alternative selection strategies, such as OC, are not common in practice. Acting on the reference population to directly mitigate the loss of genetic diversity of the breeding population while only marginally affecting the genetic gain over generations is a promising way to incorporate genetic diversity into breeding programs. Indeed current methods to cope with the loss of genetic diversity are mainly dealing with the choice of which individuals to keep in the breeding population according to their estimated performances. On the one hand, direct selection of breeding individuals has the advantage of having a strong impact on both the level of genetic diversity and genetic gain, depending on the method used. On the other hand, it relies on the choice of the breeders and is thus not systematic. Here we propose an integrated method to cope with genetic diversity at the genetic evaluation level, making it systematically incorporated. Thus, even if its impact on the conservation of genetic diversity is weaker than direct choices in the short term, it has a potentially more consistent impact on a
long-term basis. We expect that in the ideal case of operating on both the reference and breeding population, the effect observed would be further amplified and thus have an important impact on genetic diversity conservation.

Limitations and perspectives of the study

The 50K SNP chip is routinely used in GS because of its low cost and fair performance for genetic gain. Several studies cautioned that the accuracy of prediction in GS when using whole genome sequence (WGS) was at best marginally higher than of the SNP chips (VAN BINSBERGEN et al. 2015; CALUS et al. 2016; LUND et al. 2016; VAN DEN BERG et al. 2016; NI et al. 2017). Still, we can hypothesize that using WGS or genotypes of higher density could favor larger differences in genetic diversity conserved between the described scenarios. This may be especially be the case for rare variant sites since they are underrepresented in the SNP chip compared to WGS (EYNARD et al. 2015; EYNARD et al. 2016). Using WGS could enable the OC strategy during the update of the reference population to better conserve rare variants. Prediction bias appeared to be smaller in the case of the OC strategy compared to the other two strategies. Increasing the genetic diversity of the reference population increases our representation of the overall population diversity and seems to have led to a slightly more accurate overall prediction. This is potentially thanks to an improved prediction of ‘outsider’ variants. Additionally, particular attention should be paid to how many and which individuals should be removed. In fact, bias was first reduced by the addition of specifically selected individuals (PSZCZOLA et al. 2012b). However, after some generations, adding individuals elevated the prediction bias. This is probably due to a lack of relationship between the old individuals of the reference population and the candidates for selection. There is a need for
further investigations in order to give recommendations as to the total updating strategy of the reference population, accounting for the addition and removal of individuals. Finally, our study is based on milk production, a trait of major interest for the current livestock with a moderate heritability (0.3) similar to the ones for composite index traits representing the entire breeding goal. An important question is how results would change when the heritability is lower, because GS is especially appealing for low heritability traits. Using a lower heritability, while leaving the reference population size unchanged, would have yielded lower prediction accuracies and also smaller differences between scenarios. A lower accuracy means that more emphasis is put on information of relatives, such that EBV of relatives becomes more correlated and thus selected individuals are more likely to be related. This would result in conserving less genetic diversity and more inbreeding depression. Increasing the size of the reference population could counteract these effects of a low heritability trait, because it would increase the accuracy (Daetwyler et al. 2010). This is provided that increasing the reference population is possible given, for example, the size of the actual population.

CONCLUSIONS

The aim of this study was to investigate ways to reduce the loss of genetic diversity in GS breeding programs. The choice of individuals to be phenotyped and/or added to the reference population appeared to modestly impact the genetic gain and genetic diversity of the breeding population. The use of OC strategy, taking into account both relationships and performances of the individuals, to update the reference population: i) allowed for better conservation of genetic diversity in the breeding population, ii) predicted more accurate BV and iii) had only
minor repercussions on the genetic gain. The results of this study support the use of OC
strategy as a way to update the reference population, especially for breeds in need of diversity
conservation wanting to implement long-term GS programs. Making changes in the
composition of the reference population impacted the breeding population characteristics and
enabled the incorporation of genetic diversity in GS without revising farmers’ practices.

Authors’ contributions: GR, DL, PC and SEE designed the study. SF provided the data. PC
provided analytic tools. SEE and GR performed the statistical analysis and SEE, GR and
MPLC drafted the manuscript. GR, DL, PC, MPLC and SEE contributed to the interpretation
of results. GR, DL, PC, MPLC and SF contributed to the discussion and commented the
manuscript. All authors read and approved the manuscript.

Acknowledgements: The authors want to thank V. Ducrocq for the discussions and
suggestions given on the analysis, J. Vandenplas for his help with the programming and R.
Rincent for the discussions on the methodology. The authors want to thank the editor, J. B.
Holland, and the reviewers for their comments and contribution to the improvement of the
manuscript. The authors would also like to thank the projects: VALOGENE, CARTOFINE,
AMASGEN and LACTOSCAN, funded by the French National Agency for Research (ANR)
and APIS-GENE for producing the data. SE Eynard benefited from a grant from the European
Commission, within the framework of the Erasmus Mundus joint doctorate ‘EGS-ABG’, co-
funded by the Dutch Ministry of economic Affairs (KB-21-004-003).
REFERENCES


Liu, H., 2013 Application of dense marker genotypes for long-term genetic gain in animal breeding schemes, pp. 103 in Molecular biology and genetics. Aarhus, Aarhus, Danemark.


Maignel, L., D. Boichard and E. Verrier, 1996 Genetic variability of French dairy breeds estimated from pedigree information, pp. in Interbull meeting, Veldhoven, The Netherlands.


van den Berg, I., D. Boichard and M. S. Lund, 2016 Sequence variants selected from a multibreed GWAS can improve the reliability of genomic predictions in dairy cattle. Genetic Selection Evolution 48:83.


**Figure 1 – Simulation design.** This figure represents the scheme used for simulations. The highlighted boxes represent the population under consideration. The green arrows inform on the selection decision. P means phenotype, Pred. P means predicted phenotype and G means genotype.

**Figure 2 - Evolution of genetic merit, performance of genomic selection and genetic diversity over ten generations of simulations for different update strategies.** The four plots represent the average genetic merit of the breeding populations (top left), average prediction bias of genomic estimated breeding values of the offspring populations (top right), the average inbreeding (bottom left) and the average heterozygosity (bottom right) of the breeding populations over ten generations of selection. For the three update strategies *Random* (grey circle), *Sel* (magenta triangles) and *SelDiv* (blue squares) the average values and standard errors are represented.
Best GEBVs

update (based on Pred. P)

 phenotyping

A_i (1000)
P + G

X 10
(n)

Best GEBVs

selection

X 10
(n)

prediction

250

A_i+n
P + G

prediction

250

150

X 250+250

250

250

250

250

500 ♀ X 500 ♂
<table>
<thead>
<tr>
<th>Update size</th>
<th>Selection strategy</th>
<th>Breeding Value</th>
<th>Average</th>
<th>95% confidence interval</th>
<th>Absolute prediction bias</th>
<th>Average</th>
<th>95% confidence interval</th>
<th>Inbreeding</th>
<th>95% confidence interval</th>
<th>Observed Heterozygosity</th>
<th>Average</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Sel</td>
<td></td>
<td>3 182.63</td>
<td>-</td>
<td>1.08</td>
<td>-</td>
<td>5.06*10^2</td>
<td>-</td>
<td>-</td>
<td>3.07*10^1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>SelDiv</td>
<td></td>
<td>3 158.71</td>
<td>-</td>
<td>1.08</td>
<td>-</td>
<td>5.06*10^2</td>
<td>-</td>
<td>-</td>
<td>3.07*10^1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td></td>
<td>3 159.69</td>
<td>[3 159.30 ; 3 160.08]</td>
<td>1.08</td>
<td>[1.08 ; 1.08]</td>
<td>5.05*10^2</td>
<td>[5.04<em>10^2 ; 5.05</em>10^2]</td>
<td>3.07*10^1</td>
<td>[3.07<em>10^1 ; 3.07</em>10^1]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200</td>
<td>Sel</td>
<td></td>
<td>3 163.79</td>
<td>-</td>
<td>1.07</td>
<td>-</td>
<td>5.03*10^2</td>
<td>-</td>
<td>-</td>
<td>3.07*10^1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>SelDiv</td>
<td></td>
<td>3 159.21</td>
<td>-</td>
<td>1.08</td>
<td>-</td>
<td>5.03*10^2</td>
<td>-</td>
<td>-</td>
<td>3.07*10^1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td></td>
<td>3 161.05</td>
<td>[3 160.43 ; 3 161.67]</td>
<td>1.08</td>
<td>[1.08 ; 1.08]</td>
<td>5.04*10^2</td>
<td>[5.04<em>10^2 ; 5.04</em>10^2]</td>
<td>3.07*10^1</td>
<td>[3.07<em>10^1 ; 3.07</em>10^1]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>500</td>
<td>Sel</td>
<td></td>
<td>3 181.93</td>
<td>-</td>
<td>1.06</td>
<td>-</td>
<td>5.03*10^2</td>
<td>-</td>
<td>-</td>
<td>3.08*10^1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>SelDiv</td>
<td></td>
<td>3 165.91</td>
<td>-</td>
<td>1.07</td>
<td>-</td>
<td>5.04*10^2</td>
<td>-</td>
<td>-</td>
<td>3.07*10^1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td></td>
<td>3 162.64</td>
<td>[3 161.83 ; 3 163.45]</td>
<td>1.07</td>
<td>[1.07 ; 1.07]</td>
<td>5.04*10^2</td>
<td>[5.04<em>10^2 ; 5.05</em>10^2]</td>
<td>3.07*10^1</td>
<td>[3.07<em>10^1 ; 3.07</em>10^1]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1000</td>
<td>Sel</td>
<td></td>
<td>3 181.93</td>
<td>-</td>
<td>1.05</td>
<td>-</td>
<td>5.03*10^2</td>
<td>-</td>
<td>-</td>
<td>3.08*10^1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>SelDiv</td>
<td></td>
<td>3 168.00</td>
<td>-</td>
<td>1.06</td>
<td>-</td>
<td>5.03*10^2</td>
<td>-</td>
<td>-</td>
<td>3.07*10^1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td></td>
<td>3 165.02</td>
<td>[3 163.84 ; 3 166.19]</td>
<td>1.06</td>
<td>[1.06 ; 1.06]</td>
<td>5.04*10^2</td>
<td>[5.04<em>10^2 ; 5.05</em>10^2]</td>
<td>3.07*10^1</td>
<td>[3.07<em>10^1 ; 3.07</em>10^1]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2000</td>
<td>Sel</td>
<td></td>
<td>3 178.40</td>
<td>-</td>
<td>1.05</td>
<td>-</td>
<td>5.03*10^2</td>
<td>-</td>
<td>-</td>
<td>3.07*10^1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>SelDiv</td>
<td></td>
<td>3 163.26</td>
<td>-</td>
<td>1.06</td>
<td>-</td>
<td>5.03*10^2</td>
<td>-</td>
<td>-</td>
<td>3.07*10^1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td></td>
<td>3 166.19</td>
<td>[3 165.13 ; 3 167.24]</td>
<td>1.06</td>
<td>[1.06 ; 1.06]</td>
<td>5.04*10^2</td>
<td>[5.04<em>10^2 ; 5.04</em>10^2]</td>
<td>3.07*10^1</td>
<td>[3.07<em>10^1 ; 3.07</em>10^1]</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 2: Trends of changes throughout the ten generations of simulation for each of the three updating strategies and four variables.

<table>
<thead>
<tr>
<th></th>
<th>Generation trend</th>
<th>Standard error</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Breeding value</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sel</strong></td>
<td>173.77</td>
<td>7.07*10^-1</td>
<td>[172.38 ; 175.15]</td>
</tr>
<tr>
<td><strong>SelDiv</strong></td>
<td>167.69</td>
<td>7.05*10^-1</td>
<td>[166.30 ; 169.07]</td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td>167.93</td>
<td>7.07*10^-1</td>
<td>[166.55 ; 169.32]</td>
</tr>
<tr>
<td><strong>Prediction bias</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sel</strong></td>
<td>4.03*10^-2</td>
<td>1.30*10^-3</td>
<td>[3.77<em>10^-2 ; 4.28</em>10^-2]</td>
</tr>
<tr>
<td><strong>SelDiv</strong></td>
<td>3.40*10^-2</td>
<td>1.29*10^-3</td>
<td>[3.14<em>10^-2 ; 3.65</em>10^-2]</td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td>6.57*10^-2</td>
<td>1.30*10^-3</td>
<td>[6.31<em>10^-2 ; 6.82</em>10^-2]</td>
</tr>
<tr>
<td><strong>Inbreeding</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sel</strong></td>
<td>1.19*10^-3</td>
<td>2.96*10^-5</td>
<td>[1.13<em>10^-3 ; 1.24</em>10^-3]</td>
</tr>
<tr>
<td><strong>SelDiv</strong></td>
<td>1.14*10^-3</td>
<td>2.96*10^-5</td>
<td>[1.08<em>10^-3 ; 1.20</em>10^-3]</td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td>1.16*10^-3</td>
<td>2.96*10^-5</td>
<td>[1.10<em>10^-3 ; 1.22</em>10^-3]</td>
</tr>
<tr>
<td><strong>Observed Heterozygosity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sel</strong></td>
<td>-2.10*10^-3</td>
<td>2.17*10^-5</td>
<td>[-2.14<em>10^-3 ; -2.06</em>10^-3]</td>
</tr>
<tr>
<td><strong>SelDiv</strong></td>
<td>-2.02*10^-3</td>
<td>2.16*10^-5</td>
<td>[-2.06<em>10^-3 ; -1.97</em>10^-3]</td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td>-2.33*10^-3</td>
<td>2.17*10^-5</td>
<td>[-2.38<em>10^-3 ; -2.29</em>10^-3]</td>
</tr>
</tbody>
</table>