SMARTER

SMAll RuminanTs breeding for Efficiency and Resilience

Research and Innovation action: H2020 – 772787
Call: H2020-SFS-2017-2
Type of action: Research and Innovation Action (RIA)
Work programme topic: SFS-15-2016-2017
Duration of the project: 01 November 2018 – 31 October 2022

A report of an improved method to calculate genomic relationship across individuals of different purebred and crossbred populations

Ricardo Pong-Wong*, Andres Legarra

Roslin Institute, University of Edinburgh

INRAE

* Deliverable leader – Contact: ricardo.pong-wong@roslin.ed.ac.uk

DELIVERABLE D5.2

Workpackage N°5

Due date: M24

Actual date: 27/04/2021

Dissemination level: Public

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 772787

About the SMARTER research project

SMARTER will develop and deploy innovative strategies to improve Resilience and Efficiency (R&E) related traits in sheep and goats. SMARTER will find these strategies by: i) generating and validating novel R&E related traits at a phenotypic and genetic level ii) improving and developing new genome-based solutions and tools relevant for the data structure and size of small ruminant populations, iii) establishing new breeding and selection strategies for various breeds and environments that consider R&E traits.

SMARTER with help from stakeholders chose several key R&E traits including feed efficiency, health (resistance to disease, survival) and welfare. Experimental populations will be used to identify and dissect new predictors of these R&E traits and the trade-off between animal ability to overcome external challenges. SMARTER will estimate the underlying genetic and genomic variability governing these R&E related traits. This variability will be related to performance in different environments including genotype-by-environment interactions (conventional, agro-ecological and organic systems) in commercial populations. The outcome will be accurate genomic predictions for R&E traits in different environments across different breeds and populations. SMARTER will also create a new cooperative European and international initiative that will use genomic selection across countries. This initiative will make selection for R&E traits faster and more efficient. SMARTER will also characterize the phenotype and genome of traditional and underutilized breeds. Finally, SMARTER will propose new breeding strategies that utilise R&E traits and trade-offs and balance economic, social and environmental challenges.

The overall impact of the multi-actor SMARTER project will be ready-to-use effective and efficient tools to make small ruminant production resilient through improved profitability and efficiency.

Table of contents

1 Summary 3

2 Introduction 3

3 Adapting the genomic relationship to improve prediction across divergent populations 5

3.1 Simulation protocol 5

3.1.1 Simulation of the gene pool of the reference population in linkage disequilibrium 5

3.1.2 Simulation of the gene pool for other populations distantly related to the reference population 6

3.1.3 Genetic architecture and population structure 6

3.1.4 Genomic evaluation 7

3.1.5 Genetic distance measures between populations 9

3.2 Effect on the accuracy of predicted additive breeding value when accounting for dominance in the genomic evaluation 10

3.3 Accounting for population divergence in the GRM, when predicting across populations 12

4 Metafounders to improve across population prediction with ssGBLUP **17**

4.1 Resources 18

5 Conclusions 19

6 Deviations or delays 19

7 References 19

1 Summary

This document presents a report on potential benefit from improvements in the estimation of genomic relationship matrices to enhance the quality of the genomic prediction estimates across populations. Simulation was used to test the impact of including non-additive genetic effects to remove nuisance and improve additive breeding value estimates. The value of adaptations in the genomic relationships to take into account the divergence between populations was also assessed. The results from the simulations showed limited benefit in the accuracy of the estimates when including dominance or accounting for population divergence. Finally, we described the method of metafounders to enhance the Numerator relationship matrix when there is missing information on pedigree within or across populations and its potential impact when used with ssGBLUP evaluation. Studies being carried out with this method have shown that the use of metafounders has the potential to enhance the accuracy of ssGBLUP for prediction across populations.

2 Introduction

Genomic Prediction methods use high dense genotyping in the genetic evaluation to improve the accuracy of the predictions. The most popular method is the genomic Best Linear Unbiased predictor (GBLUP) (GARRICK 2007; VANRADEN 2008) as it has the practical convenience that its implementation is similar to traditional pedigree-based BLUP, but the Numerator Relationship Matrix (NRM) is replaced by a Genomic Relationship Matrix (GRM) calculated with dense genotype information. Since genetic relationships are calculated with genomic information the estimation of genomic breeding values (GEBVs) is possible for non-pedigree-related, unrecorded individuals and even from other populations. Moreover, further developments on this method have resulted in the single-step GBLUP (ssGBLUP), which allows to include information of genotyped and non-genotyped individuals on the same analysis, by using a relationship matrix that combines the NRM and the GRM (LEGARRA *et al.* 2009; AGUILAR *et al.* 2010; CHRISTENSEN AND LUND 2010; LEGARRA *et al.* 2014).

The possibility of prediction across populations is of great value in the implementation of genomic prediction in populations, which for some specific reasons -economical or logistic ones- cannot have an adequate training population. The success of prediction across populations, however, has not always been certain -in term of obtaining GEBVs with an acceptable level of accuracy-, and studies showing so have been reported in the literature for several species (e.g.(HAYES *et al.* 2009; RIGGIO *et al.* 2014; ZHOU *et al.* 2014)).

Reasons for the failure to predict across populations are various, but a common one is the level of divergence between the populations in question. Genomic prediction, basically, estimates the effect of mostly neutral SNPs with the intension to capture the effect of linked QTLs which are in linkage disequilibrium (LD) with them. The magnitude and sign of the estimated effects for the neutral SNPs vary according to the strength and pattern of linkage phases between SNPs and linked QTLs. Then, the success of these methods to accurate predict the genetic breeding values of candidates depends on the predicted candidates having similar LD pattern to the training group where the SNP effects were calculated. Using such SNP estimates in another population with very different LD pattern would also be adding some extra noise to the estimated breeding values.

Methods to account for the degree of divergence between populations have been proposed. MAKGAHLELA *et al.* (2013) proposed to adjust the genotype score by using the ancestral allele frequency of the population to calculate a multibreed GRM to be used in the evaluation. Additionally, ZHOU *et al.* (2014) modified the multibreed GRM to weight the relationships between individuals of different populations, based on the persistence of linkage phase between the two populations. However, when testing these methods using real data they showed moderate or no impact on the accuracy of prediction across population. The question which remains to be answered is whether these results are specific to the genetic divergence pattern of the populations in which they were tested or they are an overall trend expected from the method to account for divergence between populations.

Another possible reason for failure to predict across populations could be related to the presence of non-additive genetic effects. The genetic influence on the variation of quantitative traits is of heterogeneous type, with the additive and dominance effects being the main genetic components and to lesser extend their epistatic interaction. Other genetic factors controlling variation in performance may include imprinting and epigenetics. From the selection point of view the additive component is mostly the component of interest as selection act on this component only. The restricted value of dominance and epistasis, partly explains why genetic evaluations are generally implemented with a model assuming additive effect only. But, another reason for ignoring dominance effects in the evaluation has been the difficulty in estimating them. However, genomic prediction using genomic data has made easier to extract information on dominance, partially solving the problem of poor estimation. Methodologies for the construction of GRM to model dominance in the evaluation are already available (e.g.(VITEZICA *et al.* 2013)).

Accounting for dominance in the genetic evaluation is not only an academic argument, but it may have some practical benefits. Whilst the dominance effects would be ignored during the selection decision process, they can be considered as a nuisance and their inclusion in the model of analysis may improve the precision of the additive effects themselves. However, the results from some studies using real data show little benefit. Additionally, the proportion of the total genetic variance explained by dominance tends to be much smaller than that explained by the additive variance, so ignoring it in the evaluation has a little impact on the evaluation. The objective here was to test whether including dominance in the evaluation may yield benefit in the accuracy of the additive effect.

The original implementation of ssGBLUP may also have some drawbacks when using it for across population genomic prediction. The main feature of ssGBLUP is that it incorporates information from genotype and ungenotyped individuals in a single analysis by using a relationship matrix which combines the NRM with the GRM. To make these two matrices compatible, the GRM is rescaled so it has the same scale as the submatrix of the NRM for genotyped individuals (VITEZICA *et al.* 2011; LEGARRA *et al.* 2014). However, when considering multiple populations, there is an extra complexity that

populations will not be linked via pedigree, so the NRM will have zero relationship between individuals of different populations. Nevertheless, the genomic information will allow to estimate a degree of relatedness between individuals from different populations, which will be used during the across population prediction. Such inconsistence between the NRM and GRM would have an impact on the prediction. Modifying the GRM to account for this inconsistence would remove useful information affecting the estimates, so instead the NRM should be modified. LEGARRA *et al.* (2015) proposed a method to modify the NRM in order to take into account the missing pedigree information by assigning metafounders with some degree of relationships between themselves (including themselves). Such approach would modify the NRM, allowing for relationship between individuals of different populations and making it consistent with the GRM.

In this report we tested the proposed methods to account for population divergence in the GRM under a wide range of scenario to determine the full potential of these methods to improve prediction across populations. We also tested the impact of accounting for dominance in the estimation of additive breeding values, by using simulation. Finally, we described the use of metafounders to adjust the multi-population NRM so that it can be used in ssGBLUP for prediction across populations.

3 Adapting the genomic relationship to improve prediction across divergent populations

3.1 Simulation protocol

3.1.1 Simulation of the gene pool of the reference population in linkage disequilibrium

The gene pool of the population to be used as reference in the genomic prediction was simulated by creating a founder population in linkage disequilibrium (LD) and, thereafter, expanded it to create a larger population still representative of the smaller one, but with less closely related individuals. This final expanded gene pool population was then used to sample the population for each replicate.

In the first step, the founder population in LD was simulated using a mutation-drift-equilibrium algorithm as suggested by MEUWISSEN *et al.* (2001). Briefly, an initial population of *N* individuals is allowed to reproduce, with each individual producing two offspring (one male and one female). Their genome is composed of several chromosomes with biallelic loci mutating at a given rate. As the population develops across the generations, new mutations appear which are lost or increase in their frequency due to drift. After a large number of generation, the resulting population reaches an equilibrium with a genome containing segregating linked loci in LD. The simulation can be tuned to yield a specific LD pattern by adjusting the population size and mutation rate parameters. This population in equilibrium will be referred here as the founder population. In the second step, the founder population is allowed to reproduce by further extra generations with a low expansion rate and no mutation rate, and individuals of the last generation are taken as the gene pool population. By sampling the individuals to be used in each replicate from a much enlarged gene pool population, it allows independence between replicates while ensuring that they share a similar LD pattern.

In order to simulate the genome with similar LD pattern as a typical commercial sheep population, the initial population to create the LD (step 1) was composed of 100 individuals (50 males and 50 females). The genome consisted of 26 autosomal chromosomes of 1 Morgan, each with 1,000,000 loci (all fixed to one allele) with their mutation rate set at 10^{-7}. After 10,000 generations, over 9,000 loci were segregating at different frequency in each chromosome (around 250,000 segregating SNPs were

simulated across the whole genome). Individuals at generation 10,000 were considered to be the founder population. For the expansion step, the founder population was further reproduced by four extra generations at a 4X expansion rate (i.e. a male/female was randomly mated with several mates to produce 8 offspring each). Finally, 5,000 individuals from generation 10,004 were selected to form the gene pool. In order to further reduce close relationships among individuals from the gene pool, both the LD creation and expansion steps for each chromosome were done independently. This approach means that a given pair of individuals could have a half sib relationship at a given chromosome only but not for the rest.

3.1.2 Simulation of the gene pool for other populations distantly related to the reference population

To create a distantly related population, the reference population was allowed to continue to evolve in two further extra periods. First, a 'divergent period' was carried out where the founder population (from generation 10,000) was reproduced for a shorter number of generations to allow drift and mutation to change the gene frequencies and LD pattern of all segregating loci. Second, an 'expansion period' without mutation followed to create the enlarged gene pool for the new population.

The degree of the divergence of the population was controlled by varying the number of extra generations and mutation rate assumed. The number of extra generations considered ranged from 10 to 75 and the mutation rate between 10^{-6} and 10^{-3}. The upper range of the mutation rate may appear unrealistically high, but this was done with the purpose of improving the computational efficiency for generating populations with different degrees of divergence from the reference population. In reality, a population may need to evolve over a larger number of generations to achieve similar level of divergence we simulated with higher mutation rate. The expansion period to create the gene pool had the same characteristics as the one used to create the gene pool for the reference population.

3.1.3 Genetic architecture and population structure

Genome simulation: For a given replicate, the *n* individuals used in the reference population were obtained by sampling 2n haplotypes from their associated gene pool, with each chromosome done independently. A total of 1,100 loci that were still segregating in the sampled population were randomly selected for each chromosome, to become QTLs (100) or SNPs (1,000) which belong to the chip array used to calculate the genomic relationship matrices (GRMs) needed in the evaluation. The total of number of QTLs and SNPs used across the whole genome were 2,600 and 26,000, respectively. This protocol of selecting individuals from the expanded gene pools and the QTLs/SNPs from the 9,000 available loci/chromosome ensured that the sampled population to be used in a replicate avoided closely related individuals and replicates were truly independent among themselves. When the scenario includes other(s) population(s), the genome of the individuals were sampled from their respective gene pool, but the sets of QTLs and SNPs were the same as the ones selected in the reference population.

Genetic effects and phenotype simulation: Once the genome of the reference population was sampled, the additive gene effect (a) for each QTL was sampled from a normal distribution with mean zero and variance 1. When the scenario under study assumed a model with dominance, this effect (d) was sampled in a similar fashion assuming independence between the additive and dominance effect. Once a and d were sampled, the total genetic effect, the additive breeding value and the dominance deviation were calculated for each QTL given the individual genotype for the QTL in question. For a QTL k, these values for individuals with genotypes AA, AB and BB were: -a_k, d_k and a_k for total genetic

effect; $-2p_k\alpha_k$, $(q_k - p_k)\alpha_k$ and $2q_k\alpha_k$ for the additive breeding values; and $-2{p_k}^2 d_k$, $(q_k - p_k)d_k$ and $-2{q_k}^2 d_k$ for the dominance deviation, where p_k and q_k are the frequency for alleles A and B and α_k is the allele substitution effect, equal to $a_k + (q_k - p_k)d_k$. Thereafter, these values for all QTLs are summed over within each individual to obtain their overall genetic effects.

To obtain targeted genetic additive and dominance variances (i.e. σ^2_a and σ^2_d), the values sampled for a_k and d_k (when non-zero) needed to be rescaled, and the approach used varied depending on whether the genetic model was fully additive or included dominance. When the genetic model was fully additive, the rescaling was done by: (i) calculating the additive breeding value for all individuals sampled in the reference population, (ii) estimating their variance $\hat{\sigma}^2_a$ and then (iii) rescaling a_k by the constant $\sqrt{\sigma^2_a / \hat{\sigma}^2_a}$. After rescaling of a_k, the variance of the breeding values for all animals in the reference population matches the targeted σ^2_a.

For a model with dominance, the rescaling was done first on the dominance effect, and thereafter, the additive effect was rescaled using a recursive approach. Similarly as with a fully additive model, the dominance deviations for each QTL were calculated given the sampled d_k, they were summed over to calculate the individuals' overall dominance effect and its variance $\hat{\sigma}^2_d$ calculated to obtain the rescaling constant as $\sqrt{\sigma^2_d / \hat{\sigma}^2_d}$. Thereafter, following a similar protocol as before: (i) α_k were estimated (with the current value for a_k and the already rescaled d_k) and (ii) the additive breeding values calculated and summed over all QTLs to obtain the individuals' total additive breeding values; (iii) the variance of breeding values in the population $\hat{\sigma}^2_a$ is calculated and (iv) used to obtain a new the scaling constant equal to $\sqrt{\sigma^2_a / \hat{\sigma}^2_a}$; (v) α_k and the additive breeding values are recalculated as well as their $\hat{\sigma}^2_a$. If $\hat{\sigma}^2_a$ is higher (or lower) than the targeted σ^2, the scaling constant is slightly decreased (or increased); a_k are rescaled again and the additive breeding values and $\hat{\sigma}^2_a$ reestimated. Then an iterative process of increasing or decreasing the scaling constant when $\hat{\sigma}^2_a$ does not match the targeted σ^2_a is carried out and the recursion stops once $\hat{\sigma}^2_a$ is equal to σ^2_a.

Once a_k and d_k were properly rescaled to yield the right magnitude for σ^2_a and σ^2_d, the phenotypic record for an individual was calculated by adding a random term to their total genetic effect which is sampled from a Normal distribution with mean zero and variance σ^2_e. When the scenario under study included other populations, their genetic effects were calculated using the same value for a_k and d_k sampled for the reference population. Because the QTL allele frequencies for these populations deviated from the one observed in the reference population, slight changes in the magnitude for σ^2_a and σ^2_d were observed in these extra populations.

3.1.4 Genomic evaluation

The genomic estimated breeding values (GEBVs) were calculated based on the genomic best linear unbiased prediction (GBLUP) method, where the covariance matrix to model the genetic effects of all individuals included in the analysis are calculated using genomic information. Since this simulation study addressed two main factors influencing the efficiency of the genomic evaluation across

populations, the GBLUP implementation was based on two main models, with some variants within them:

GBLUP with only additive effects

The following mixed model was used to estimate the GEBVs:

$$\mathbf{y} = \mu + \mathbf{Zu} + \mathbf{e} \quad [1]$$

where **y** is the vector of observations, μ is the overall mean effect, **Z** is the incidence matrix linking the individuals to their phenotype, **u** is the vector of genomic additive breeding value effects distributed as N(**0**, **G**σ^2_a) and **e** is the vector of environmental deviation distributed N (**0**, **I**σ^2_e), with **G** and **I** being the genomic relationship matrix and an identity matrix, and σ^2_a and σ^2_e the genetic additive and the residual variance, respectively.

The **G** matrix was obtained using genotype information from the 26,000 SNPs included in the SNP array. The **G** matrix was calculated using a method commonly known as the VanRaden's method 2 (VANRADEN 2008) and it is based on the cross product of (rebased and rescaled) genotype scores. Hence, for a pair of individuals (*i,j*) their relationship is:

$$G_{i,j} = \frac{1}{m}\sum_{k=1}^{m}\frac{(x_{ik}-2p_k)(x_{jk}-2p_k)}{2p_k\,(1-p_k)} \quad [2]$$

where x_{ik} and x_{jk} are the number of copies of B alleles in the genotype of *i* and *j* at SNP *k*, equal to 0,1,2 for individuals with genotype AA, AB and BB, respectively; p_k is the observed frequency of allele B in the genotyped individuals and *m* is the number of SNPs in the SNP array (i.e. *m*=26,000).

In order to take into account that, for some scenarios, the analysis was done including individuals from different populations, and GBLUP was performed using three different variations of **G** and denoted as:

- GBLUP$_O$: where the relationship matrix is calculated assuming that p_k is the overall allele frequency across all individuals included in **G**, without distinction on which population they belong to.
- GBLUP$_P$: where p_k refers to the allele frequency of the population to which the individuals belong to (MAKGAHLELA *et al.* 2013). Hence, for a pair of individuals from the same population, their relationship is the same as [2] but using their population specific allele frequencies. But for pairs from two different populations $G_{i,j} = \frac{1}{m}\sum_{k=1}^{m}\frac{(x_{ik}-2p_{ik})(x_{jk}-2p_{jk})}{2\sqrt{p_{ik}\,(1-p_{ik})p_{jk}\,(1-p_{jk})}}$, where p_{ik} and p_{jk} are the allele frequencies from population of individual *i* and *j*.
- GBLUP$_W$: where the relationship matrix is also calculated using population specific allele frequency, but it considers that the LD patterns may differ across populations, and it takes it into account by adding a weighing factor to the relationships between two individuals of different populations (ZHOU *et al.* 2014). Hence, their relationship is $G_{i,j} = \frac{1}{m}\sum_{k=1}^{m}\frac{(x_{ik}-2p_{ik})*w_{i,j,k}}{2\sqrt{p_{ik}\,(1-p_{ik})p_{jk}\,(1-p_{jk})}}$, where $w_{i,j,k}$ is the weighing factor at SNP *k* for these two specific populations. The weighing factor used was the persistence of linkage phase for SNP *k*, between these two populations (see below for description how this weight is calculated).

GBLUP with additive and dominance effects

Following derivation from VITEZICA *et al.* (2013), the model used in the GBLUP to account for dominance effect is equal to:

$$\mathbf{y} = \mu + \mathbf{Zu} + \mathbf{Zd} + \mathbf{e} \quad [3]$$

where $\mathbf{d}$ is the vector of dominance deviation distributed N(0, $\mathbf{D}\sigma^2_d$), with **D** being the dominance genomic relationship matrix and σ^2_d the dominance variance. Following similar approach as VanRaden's method 2, the elements in the matrix are equal to $D_{i,j} = \frac{1}{m}\sum_{k=1}^{m} \frac{(w_{ik})(w_{jk})}{[2p_k(1-p_k)]^2}$, where w_{ik} and w_{jk} are the dominance deviation score of *i* and *j* at SNP *k* equal to $-2p_k^2$, $2p_kq_k$ and $-2q_k^2$ for individuals with genotypes AA, AB and BB, respectively. The model including dominance can also be implemented using a genotypic parameterisation where the effects a and d are included in the model instead. Then the dominance relationship matrix is constructed by redefining w_{ik} and w_{ik} as $\frac{-2p_kq_k}{[2p_kq_k(1-2p_kq_k)]}$, $\frac{(1-2p_kq_k)}{[2p_kq_k(1-2p_kq_k)]}$ and $\frac{-2p_kq_k}{[2p_kq_k(1-2p_kq_k)]}$ for individuals with genotype AA, AB and BB, respectively. Additionally the dominance variance estimated with the genotypic parameterisation changes to $\sum_{k=1}^{m}[2p_kq_k(1-2p_kq_k)]\sigma^2_d$. For further details of the parameterisation for fitting dominance in the GBLUP see (VITEZICA *et al.* 2013).

3.1.5 Genetic distance measures between populations

The degree of divergence between the reference population and the other ones was quantified using two criteria: the Euclidean distance and their persistence of linkage phase.

Euclidean distance between populations: To calculate the genetic distance within population and between two populations, first the additive GRM including all individuals from both populations is calculated and an Eigen decomposition is used to decompose it into its Eigenvalues and Eigenvectors. Since the Eigenvectors are orthogonal vectors, the Euclidean distance (δ) between a pair of individuals (*x,y*) is calculated as $\delta_{xy} = \sqrt{\sum_{i=1}^{n} \lambda_i (v_{xi} - v_{yi})^2}$, where δ_{xy} is the Euclidaen distance between x and y, v_{xi}, v_{yi} are their loading in the eigenvector *i* respectively, λ_i the eigenvalue associated to *i* and *n* is the total number of the eigenvectors with non-zero eigenvalue. Hence, the average distance within a population is the mean distance between all pairwise comparisons for all individuals within the population. Similarly, the Euclidean distance between two populations is the mean distance of all pairs involving one individual from each population.

Persistence of linkage phase: The linkage phase between two linked loci represents the degree of excess/deficit of the haplotypes relative to their expected frequency given the frequencies of the alleles in the haplotype in question.

Let two linked loci A and B, the linkage phase between alleles A_1 and B_1, is $r_{A1,B1} = \frac{f_{A1,B1} - f_{A1} * f_{B1}}{\sqrt{f_{A1} * f_{A2} * f_{B1} * f_{B2}}}$,

where $f_{x,y}$ is the frequency of allele/haplotype y at locus x, observed in the population. Hence, a positive (or negative) value for $r_{A1,B1}$ means that allele A_1 is more (or less) associated to B_1 than what is expected due to random association given the allele frequencies. Note that the square of $r_{A?,B?}$ ($r^2_{A,B}$) is more commonly used as a measure of LD between two linked loci, but it does not indicate the direction of the phase.

Persistence of linkage phase between two populations is the degree of concordance of the linkage phases between the two populations. It is measured as the Pearson's correlation between the linkage phases observed in haplotypes of consecutive loci in both populations, where a high and positive correlation means that the allele association between linked loci is similar in both populations.

Here the persistence of phase between two populations is estimated with only the loci included in the SNP chip array used for the genetic evaluation. While persistence relates to the whole genome or to a genomic region, a persistence value was given to each SNP by calculating the correlation in rolling intervals of 21 SNPs. For a given SNP, the closest 10 SNPs at each side were taken to calculate the correlation of the measures across the two populations (in average, the 21 SNPs used for each interval cover a genomic interval of 2 cM, see above). The rolling correlations (or their squared values) were used as in the calculation of the GRM to weight the relationship between individuals of different populations (see above). When the correlation for a SNP was negative, the weighing factor was set to zero, representing that the SNP does not affect the relationship across population.

3.2 Effect on the accuracy of predicted additive breeding value when accounting for dominance in the genomic evaluation

The objective here was to assess the impact of including dominance in the evaluation as a nuisance parameter to improve the accuracy of the additive effect.

The simulation in this study assumed a population of 1,500 individuals, all genotyped for the 26,000 SNPs in the chip array and 1,000 also having performance record for a given trait. The genetic architecture for the trait was assumed to have an additive and a dominance component.

The protocol for simulating the genomic data and the performance records is as described above. At each replicate, the genotypes for the individuals were sampled from the gene pool from the reference population and thereafter a set of 1,100 segregating loci per chromosome selected randomly to be QTL (100) or part of the SNP array (1000) used in the evaluation. The QTL effects (additive and dominance) were sampled from a normal distribution and rescaled to yield the targeted additive, dominance and environmental variance (i.e. σ^2_a, σ^2_d and σ^2_e).

Three different GBLUP analyses were carried out, defined by the genetic model considered in the evaluation: (i) only additive effects were fitted, (ii) additive and dominance effects were fitted assuming a breeding value/dominance deviation parameterisation and (iii) additive and dominance effects were fitted assuming a genotypic parameterisation.

Three different scenarios were simulated here. All assumed σ^2_a = 20 and σ^2_e = 80, but σ^2_d changed across the scenarios taking values of 10, 20 and 30, representing a substantial proportion of the total genetic variance (i.e. 33.3%, 50% and 60% for when σ^2_d was 10, 20 and 30, respectively).

GBLUP evaluation was carried out and the accuracy of the additive genetic effect was calculated as the Pearson's correlation between the true additive breeding values and their estimates obtained from the GBLUP analysis. The GBLUP was done assuming the true genetic variance being known or they were estimated from the reference animals prior the GBLUP evaluation.

The results shown are the average of 100 replicates for each scenario.

Table 1 shows the estimated variance components obtained from the REML analysis fitting the three different models of analysis (i.e. one additive model and two models with dominance). As it can be seen, the estimates for σ^2_a were very close to the true value (i.e. 20) with all three models of analysis and when dominance is not fitted, this variance is absorbed by the environmental variance. These results suggest that the additive and the dominance effects may not be confounded between each

other, which may have some consequences on the accuracy of the estimates if not fitting dominance when it exists.

Table 1. Estimated variance components for the three scenarios, obtained with the three models of analysis: (i) a fully additive, (ii) with dominance using the breeding value parameterisation and (iii) with dominance using the genotypic parameterisation. Log-likelihood (LogL) and Log-likelihood ratio test (LRT) are also reported. Results shown are average of 100 replicates.

model	Estimate						
	σ_a^2	σ_d^2	σ_e^2	σ_p^2	LogL	LRT	σ_a^2 = 0*
			True σ_d^2=10				
Additive	19.1		80.4	99.5	-2794.8		
Dominance (breeding value parameterisation)	18.2	12.1	72.5	102.8	-2794.0	1.5	20%
Dominance (genotypic effect parameterisation)	16.0	11.4	73.3	100.8	-2794.1	1.3	19%
			True σ_d^2=20				
Additive	18.9		91.1	109.9	-2845.5		
Dominance (breeding value parameterisation)	17.3	24.0	73.5	114.8	-2843.4	4.3	3%
Dominance (genotypic effect parameterisation)	12.9	22.4	74.8	110.0	-2843.3	4.5	1%
			True σ_d^2=30				
Additive	18.1		101.8	119.8	-2889.4		
Dominance (breeding value parameterisation)	15.8	34.4	76.6	126.7	-2886.1	6.5	1%
Dominance (genotypic effect parameterisation)	9.3	33.7	77.5	120.5	-2885.7	7.4	0%

*: percentage of replicates where σ_d^2 was estimated to be zero.

Table 2 shows the accuracy of the genomic additive breeding values for the three models of analysis considered here and under the different scenarios defined by the magnitude of the dominance variance in the true model which was used to simulate the data. The accuracies are shown for individuals with phenotypic records (i.e. the training individuals) and individuals without phenotypic records, and the GBLUP was performed using the estimated variances obtained from the REML analysis or assuming the true variances. The results suggest that ignoring to fit the dominance effects in the model of analysis does not have a negative impact on the accuracy of the additive breeding value. Similar results have been reported previously using real data (e.g (VITEZICA *et al.* 2013)). But it is rather surprising that this was also the case for our simulation study, as our data were simulated assuming very large magnitude for dominance variance. Our expectation was that as dominance starts explaining large proportion of the total variance, it should become an important nuisance factor and not fitting it into the model of analysis would affect the prediction of the other effects. As described before, for the

scenarios considered here, the dominance effects explained 33%, 50% and 66% of the total genetic variance.

The no benefit from fitting dominance seems to be related to a lack of confounding in the information (for additive and dominance) contained in the data. Results from the REML analyses when dominance is not fitted, showed that the proportion of the variance explained by this effect is recognised as environmental variance (which explains why the quality of predicted additive effects was not hindered if dominance is not fitted). The data used as the training population is adequate enough to detect dominance (i.e. see the estimated for dominance variance in Table 1), so the explanation seems to be more a true lack of confounding between the effects rather than a lack of information in the data. Whilst the better separation of additive and dominance effect may be an added benefit from the genomic prediction, the experimental design may also have some influence on it. The simulation protocol used in this study aimed at ensuring that the replicates were independent and the population with replicate was composed of truly unrelated animals. Further inspections may still be required to determine if the results (i.e. that not fitting dominance does not hinder the accuracy of additive effect) extrapolate to other situations with close relationships between individuals of the training population.

The size of the dominance variances tested here are relatively large compared to real values commonly observed in commercial traits in livestock species. This suggests that perhaps, fitting dominance effect to improve the genomic prediction would not have a substantial impact in practical breeding programmes.

Table 2. Accuracy of the genomic estimated additive breeding value for the scenarios where the true value for σ_d^2 is 10, 20 and 30. Estimates were obtained with GBLUP with three model of analyses, using the variance calculated from the REML analysis or the true values used to simulate the data.

	Individuals with records			Individuals without records		
	Model of analysis			Model of analysis		
True σ_d^2	Additive only	Dominance breeding value parameterisation	Dominance genotypic parameterisation	Additive only	Dominance breeding value parameterisation	Dominance genotypic parameterisation
GBLUP using variance component estimated with REML analysis						
10	0.509	0.509	0.509	0.302	0.302	0.304
20	0.485	0.485	0.483	0.282	0.282	0.282
30	0.467	0.467	0.466	0.273	0.272	0.269
GBLUP using the true variance						
10	0.510	0.510	0.509	0.303	0.302	0.302
20	0.486	0.486	0.484	0.282	0.283	0.282
30	0.486	0.486	0.484	0.282	0.283	0.282

3.3 Accounting for population divergence in the GRM, when predicting across populations

The objective here was to assess the impact on the accuracy of across population genomic prediction when modifying the GRM to account for divergence between populations.

The simulation mimics scenarios with two populations with an 'observable' level of genetic divergence (labelled as the "reference" and the "divergent" populations). Genomic prediction analyses are carried out with a training set composed of individuals from the reference population and the candidates are from both populations. The scenarios considered here varied on the degree of genetic divergence between the two populations.

A total of 1,500 individuals belonging to the reference population (i.e. 1,000 to be training individuals and 500 to be candidates) were sampled from their gene pool and 500 individuals from the divergent population. The genetic effects for all individuals and the phenotypes for the training group were simulated as described above assuming to be fully additive genetic model with σ^2_a = 20 and σ^2_e = 80.

Four different GBLUP evaluations were performed varying on the GRM to account for the population divergence:

(i) $GBLUP_O$: divergence is not accounted for and GRM was calculated assuming a single population using the overall allele frequency in both populations.
(ii) $GBLUP_P$: divergence is accounted by calculating the genotype score of individuals using the specific allele frequency of their own population.
(iii) $GBLUP_{W1}$: same as $GBLUP_P$ but also using the rolling correlation of the persistence of linkage phase as weighing factor (see above).
(iv) $GBLUP_{W2}$: as $GBLUP_{W1}$ but using the square of the correlation as the weighing factor.

Additionally, the genomic evaluation analyses were carried out using the true heritability or the estimate from a REML analysis on the reference animals.

The results shown are the average of 200 replicates for each scenario.

A total of twelve divergent populations were considered here. Eleven populations were created with different degree of genetic differentiation from the reference population (sorted according to their degree of differentiation and labelled from A to K). The extra twelfth population (labelled as AxJ) was simulated by combining the even chromosomes from population A (closely related to the reference one) and the uneven chromosomes from population J (very distant from the reference one). This latter population is unrealistic, but it helps to understand the impact of the different approaches suggested to account for the divergence in the genomic evaluation across populations. The degree of differentiation of these populations is observed in Figure 1, measured as their average persistence of linkage phase with the reference population.

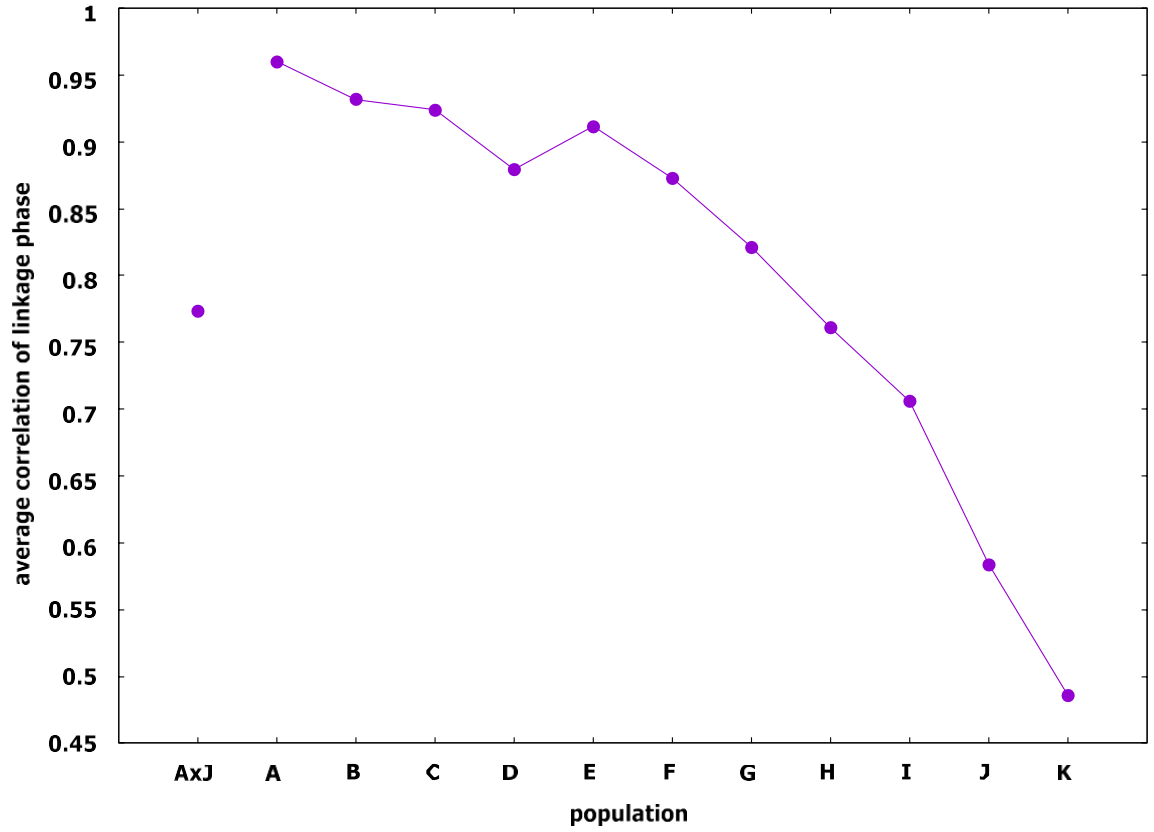

Figure 1. Persistence of linkage phase between the reference population and the 12 divergent populations (labelled from A to K, plus AxJ). The linkage phase parameter is the mean of the rolling correlation for all 26,000 SNPs in the SNP chip array, the value shown in the X axis is the average of 200 replicates.

The accuracies of the GEBVs obtained when evaluations were carried out using the REML estimates on the training set are shown in Figure 2. The results when using the true variance are very similar to those found when using the REML estimates, so they are not shown here. The accuracy of the phenotyped individuals included in the training group was 0.497 (average over 2,400 replicates) and 0.272 for unphenotyped candidates from the same reference population as the training one, representing an approximately 45% drop in accuracy when the individuals have no available phenotype at the time of the evaluation. As expected the accuracy for candidates from the divergent population decreased according to their degree of divergence. When their GEBVs were calculated without accounting for them to be from different populations (i.e. $GBLUP_O$), their accuracy ranged from 0.06 to 0.19, equivalent to 24% to 72% of the accuracy observed in unphenotyped candidates from the same population as the training set.

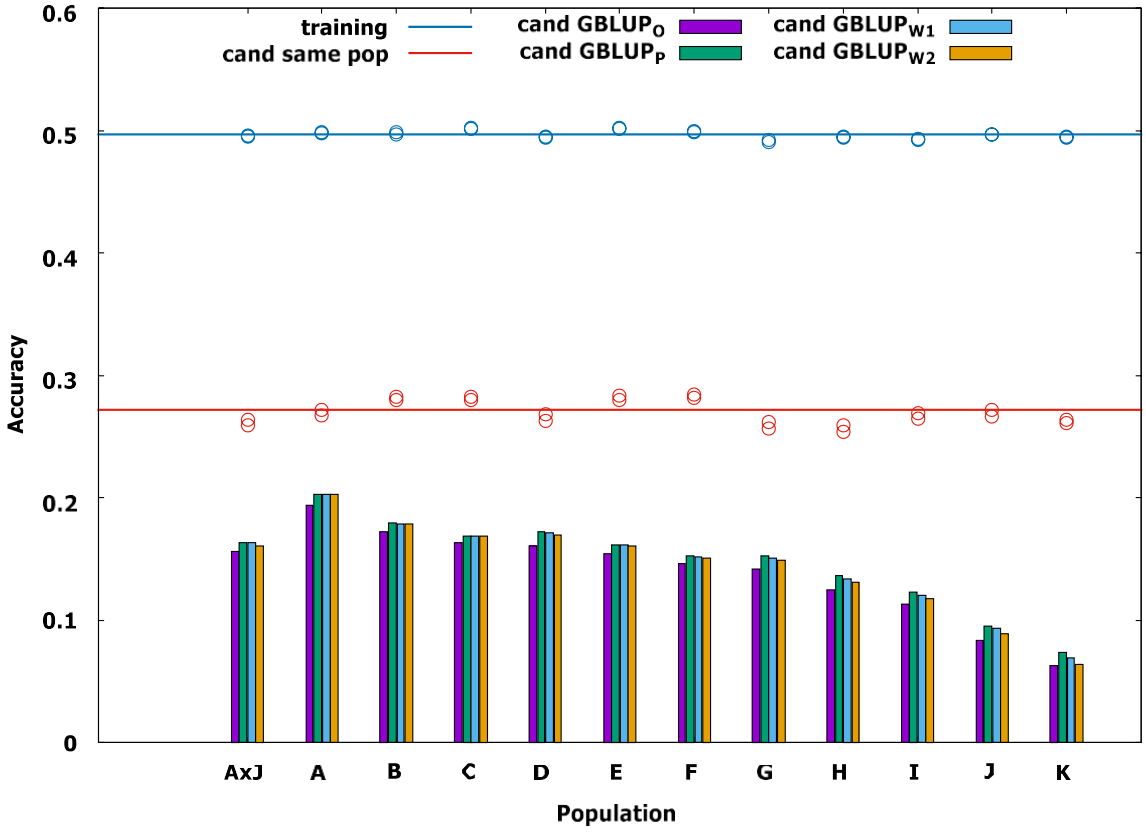

Figure 2. Accuracy of the GEBVs for training (blue line) and candidates (red line) in the reference population and candidates from the twelve divergent populations (bars) estimated using the four GBLUP approaches varying on their GRM to account for population divergence.

The approaches modifying the GRM to take into account the population divergence (i.e. GBLUP$_P$, GBLUP$_{W1}$, GBLUP$_{W2}$) have low to moderate benefit in improving the accuracy. Calculating the GRM using population specific allele frequencies (GBLUP$_P$) increases the accuracy by 7.4% relative to GBLUP$_O$, ranging between 3.4% and 16.2% across all the divergent populations. However, the scenarios with the highest advantage percentage-wise may overemphasize the benefit of GBLUP$_P$ as they are associated to distantly related populations where the performance of GBLUP$_O$ is low (hence, the relative advantage appears to be higher). Overall, the extra accuracy of GBLUP$_P$ over GBLUP$_O$ in absolute magnitude ranges between 0.006 and 0.012.

The accuracies observed with GBLUP$_{W1}$ and GBLUP$_{W2}$ were slightly lower than that achieved with GBLUP$_P$, though they were always higher than with GBLUP$_O$. This is rather surprising as GBLUP$_{W1}$ and GBLUP$_{W2}$ also account for the difference in frequencies as with GBLUP$_P$, so the net effect from accounting for the persistence of linkage phase between populations was negative or at best none at all (the reduction in accuracy relative to GBLUP$_P$ was very small, ranging between 0.0013 and 0.0041, so they should be probably considered as zero).

Genomic prediction methods estimate the effect of neutral SNPs with the intention to capture the effect of linked QTLs which are in LD with them. The magnitude and sign of the estimated effects for

the neutral SNPs vary according to the strength and pattern of linkage phases between SNPs and linked QTLs. Then, the success of these methods to accurate predict the genetic breeding values of candidates depends on the predicted candidate having similar LD pattern to the training group where the SNP effects were calculated. Using such SNP estimates in another population with very different LD pattern would also be adding some extra noise to the estimated breeding values.

Intuitively, using the correlation of persistence of linkage phase to weight the SNP contribution on relationships between individuals across populations, would control this noise. It would not add further information but it should reduce the noise so the prediction should improve (relative to whennot using weights). The fact that the accuracy does not improve, when accounting for the linkage phase, means that the problem of poor predictions on distantly related individuals is probably relatedto the amount of information available, rather than to the amount which is retrievable. Hence, the solution to improve poor predictions of some distant candidates may only be achieved by extending the training set with individuals more related to them. A training set with a mixture of individual from both populations has been frequently shown to be a better solution (e.g. RIGGIO *et al.* 2014).

The results from this study suggest that the correlation of persistence of linkage phase may have limited value (or none) to improve prediction across populations. However, it can provide a measure indicating the degree of genetic divergence which can be allowed (between training and candidates) before the prediction deteriorates below a given threshold. For instance, the results on the accuracies observed in the populations and SNP density studied here showed that the persistence of linkage phase between training and the distant candidate sets should have a correlation of *circa* 0.75 in orderto achieve an accuracy which is, at least, half the accuracy when predicting in the same population (Figure 3). This parameter will be of great value to determine when the training set (or the SNP chip) needs to be updated to ensure that a minimum accuracy is to be achieved.

In conclusions, the approaches modifying the GRM to account for the degree of divergence between populations showed little impact of the prediction accuracy for distantly related populations. Adjustment to use population specific allele frequencies has a small benefit on the accuracy, but the approach to account for the degree of persistence of the linkage phase yielded no benefit at all.

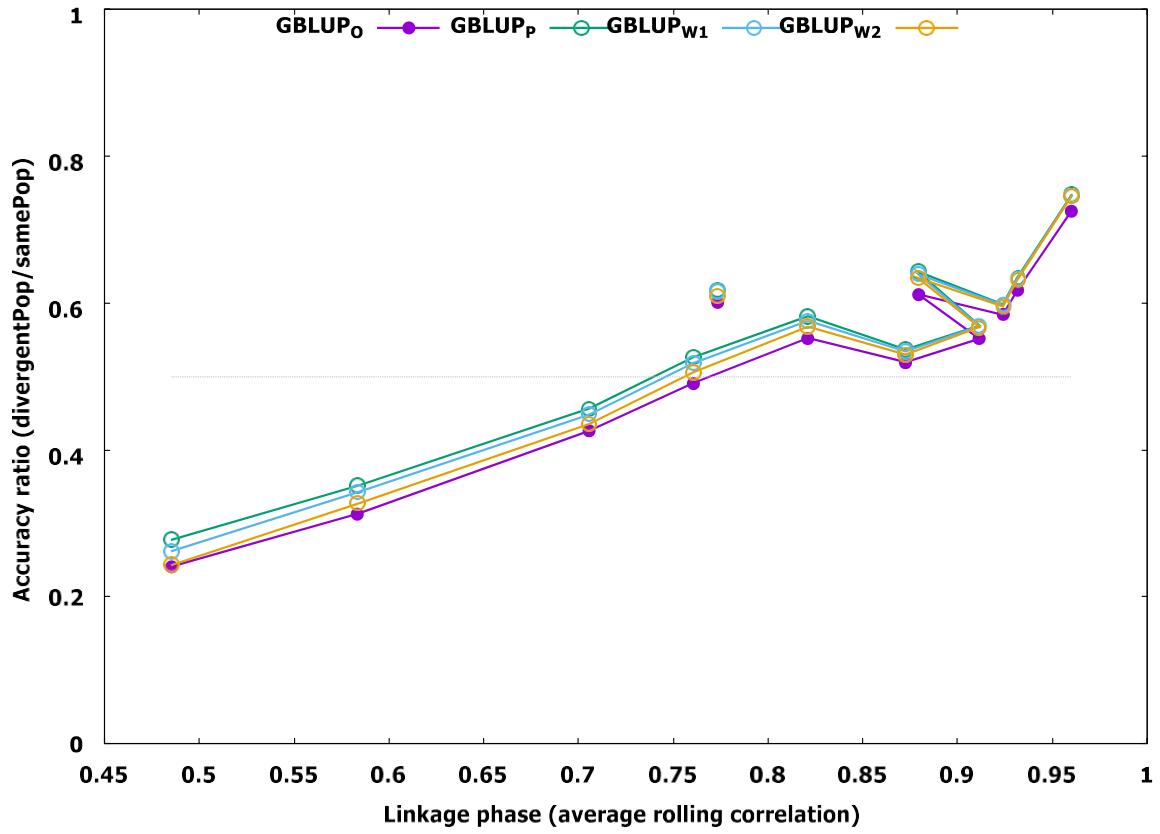

Figure 3. Effect of the persistence of linkage phase between the training and the candidate sets on the accuracy of the GEBV of candidates from other populations. The X axis shows the ratio of the GEBV accuracy in distant candidates over the accuracy in candidates from the same population as the training set. The points outside the line are the values for the population AxJ.

4 Metafounders to improve across population prediction with ssGBLUP

The following section provides a description of the use of metafounders to improve compatibility of the NRM and GRM due to absence of pedigree information across populations. For a more detailed explanation of this metafounders approach as well as it theoretical support, see LEGARRA *et al.* (2015) and GARCIA-BACCINO *et al.* (2017).

The main feature of of ssGBLUP is its ability to incorporate genotyped and ungenotyped individuals in a single analysis by using a relationship matrix (**H**), which combines the NRM with the GRM. To avoid the introduction of bias it requires that both the NRM and the GRM are at the same scale (for the subset of genotyped individuals), so the GRM is rescaled prior combining them to create **H** (VITEZICA *et al.* 2011; LEGARRA *et al.* 2014). However, when considering multiple populations, there is an extra complexity that populations are generally not linked via pedigree information, so the NRM assigns zero relationship between individuals of different populations. However, the genomic information will allow to estimate a non-zero degree of relatedness between individuals from different populations,

which will be used during the across population prediction. Such inconsistence between the NRM and GRM would have a negative impact on the prediction across populations. Modifying the GRM to account for this inconsistence would remove useful information affecting the estimates, so instead the NRM needs to be modified.

LEGARRA *et al.* (2015) proposed the method of metafounders for enhancing the NRM to '*fill the gap*' on relationship when pedigree information is missing (within or across populations). Let consider a group of individuals with missing parent information, then a pseudo-individual labelled as metafounder is assumed such as that all individuals in the group are offspring of this metafounder via self-mating (i.e. the metafounder is the sire and dam for all individuals). By adjusting the self-relationship of the metafounder (γ) it allows to assign a given degree of relationship among all individuals of the group. This self-relationship in the metafounders generally implies a negative inbreeding (i.e. $\gamma = 1 + f$) and when f = -1 (i.e γ=0) it is equivalent to the model assuming genetic groups (or unknown parent groups) which generally is used to correct for difference in mean of the group of individuals with missing parent information. The metafounder approach can be extended to consider several groups, each associated to different metafounders, with their relationship among themselves defining the degree of relatedness between individuals within groups and across them. Hence, for the specific case of multi-population ssGBLUP, the metafounders would allow to have non-zero degree of relatedness across populations in the NRM, making it compatible with the GRM.

Under the metafounder representation, the enhanced NRM can be calculated using the tabular method or its fast inversion using Henderson's method (HENDERSON 1976; QUAAS 1976). However, before this can be done, it does requires to calculate the matrix $\boldsymbol{\Gamma}$, containing the relationship among metafounders. This matrix $\boldsymbol{\Gamma}$ can be obtained using marker information with each element being equal to 8 times the covariance between the marker frequencies in their base populations in question (LEGARRA *et al.* 2015).

Hence, the main task when implementing metafounders in ssGBLUP is the estimation of the SNP allele frequencies for the different groups associated to metafounders. GARCIA-BACCINO *et al.* (2017) compared four different methods to calculate these frequencies based on the genotype information of the sampled animals to be included in the analysis. They concluded that the most reliable methods are those accounting for (pedigree) relatedness between the genotyped animals.

Studies testing the benefit of including metafounders in the ssGBLUP evaluation have shown a significant improvement of the accuracy of the estimates. The beneficial effect of adding metafounders has been observed even when assuming one single metafounder in a single population with complete pedigree information (GARCIA-BACCINO *et al.* 2017). Furthermore, situations with larger degree of missing pedigree information has shown that ssGBLUP with metafounders is more accurate and has less bias than when using unknown parent group (MACEDO *et al.* 2020). Hence, ssGBLUP for predicting across population should benefit from the inclusion of metafounders, as it will allow to account for the unrecorded relationship between the populations due to the lack of pedigree information linking them.

4.1 Resources

Free software (binary files) can be found in http://nce.ads.uga.edu/wiki/doku.php?id=distribution

A detailed tutorial with example files can be found in http://genoweb.toulouse.inra.fr/~ale-garra/ThreeWayDist/

Comprehensive notes can be found in http://genoweb.toulouse.inra.fr/~alegarra and http://nce.ads.uga.edu/wiki/doku.php?id=course_information_-_uga_2018

5 Conclusions

The results from the simulation studies showed little or no benefit in the additive GEBV when accounting for dominance effects into the model of analysis. The results from the REML analyses suggested that the additive and dominance effects have little confounding, which explain why accounting for dominance as a nuisance factor did not affect the predictions. Similarly, modification of the GRM to account for divergence between populations has a modest effect for improving prediction across populations. Within the ssGBLUP framework, the method of metafounders toenhance the NRM has a positive impact on the accuracy of the GEBV. The beneficial effect can be observed even in situations with a single population. Because relationships across populations are generally not available/recorded, the use of Metafounders to calculate these unknown values has a potential to improve the prediction across populations.

6 Deviations or delays

There was a delay of six months. This delay was due to COVID, however it is considered a minor delay which should not impact any further work still to be done in WP5.

7 References

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta *et al.*, 2010 Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. Journal of Dairy Science 93**:** 743-752.

Christensen, O. F., and M. S. Lund, 2010 Genomic prediction when some animals are not genotyped. Genetics Selection Evolution 42.

Garcia-Baccino, C. A., A. Legarra, O. F. Christensen, I. Misztal, I. Pocrnic *et al.*, 2017 Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations. Genetics Selection Evolution 49**:** 34.

Garrick, D. J., 2007 Equivalent mixed model equations for genomic selection. Journal of Animal Science 85**:** 376-376.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla and M. E. Goddard, 2009 Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41.

Henderson, C. R., 1976 A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. Biometrics 32**:** 69-83.

Legarra, A., I. Aguilar and I. Misztal, 2009 A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92**:** 4656-4663.

Legarra, A., O. F. Christensen, I. Aguilar and I. Misztal, 2014 Single Step, a general approach for genomic selection. Livestock Science 166**:** 54-65.

Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar and I. Misztal, 2015 Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. Genetics 200**:** 455.

Macedo, F. L., O. F. Christensen, J.-M. Astruc, I. Aguilar, Y. Masuda *et al.*, 2020 Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. Genetics Selection Evolution 52**:** 47.

Makgahlela, M. L., I. Strandén, U. S. Nielsen, M. J. Sillanpää and E. A. Mäntysaari, 2013 The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. J Dairy Sci 96**:** 5364-5375.

Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157**:** 1819-1829.

Quaas, R. L., 1976 Computing the Diagonal Elements and Inverse of a Large Numerator Relationship Matrix. Biometrics 32**:** 949-953.

Riggio, V., M. Abdel-Aziz, O. Matika, C. R. Moreno, A. Carta *et al.*, 2014 Accuracy of genomic prediction within and across populations for nematode resistance and body weight traits in sheep. Animal 8**:** 520-528.

VanRaden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91**:** 4414-4423.

Vitezica, Z. G., I. Aguilar, I. Misztal and A. Legarra, 2011 Bias in genomic predictions for populations under selection. Genetics Research 93**:** 357-366.

Vitezica, Z. G., L. Varona and A. Legarra, 2013 On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. Genetics 195**:** 1223-1230.

Zhou, L., M. S. Lund, Y. Wang and G. Su, 2014 Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. Journal of Animal Breeding and Genetics 131**:** 249-257.