Norwegian University of Life Sciences
Faculty of Veterinary Medicine and Biosciences
Department of Animal and Aquacultural Sciences

Master Thesis 2015
30 credits

The accuracy of across population genomic prediction using a Bayesian variable selection model

S. van den Berg

The accuracy of across population genomic prediction using a Bayesian variable selection model

S. van den Berg

Supervisor WUR: Y.C.J Wientjes & M.P.L Calus

Supervisor NMBU: T.H.E Meuwissen

Major thesis European Master of Animal Breeding and Genetics

June 2015

Acknowledgement

This thesis would not have been made without the help and support of many. I would like to express my appreciation to all those who made it possible.

I would like to thank my supervisor Yvonne Wientjes for sharing her knowledge with me. She provided me with valuable feedback from the early phases of proposal writing onto the finalization of the report.

I would also like to express my gratitude to Mario Calus for his supervision and useful feedback .Further I would like to thank Theo Meuwissen for his supervision.

I would like to thank Thomas for his unconditional support and patience. He gave me feedback on ideas and written texts and inspired me by sharing his thoughts.

Finally, I would like to thank my friends and family for being there for me when I needed it.

June, 2015

Sanne van den Berg

Table of Contents

Abstract

Background

The use of information across populations is an attractive approach to increase the accuracy of genomic prediction for numerically small populations. However, accuracies of across population genomic prediction, in which reference and selection individuals are from different populations, are currently disappointing. The objective of this study is to estimate the accuracy of across population genomic prediction using a Bayesian variable selection model and compare the obtained accuracies with the accuracy obtained by a GBLUP model. In this study, high density genotypes of 1033 HF, 147 MRY and 105 GWH are used. The phenotypes are simulated for two changing variables: (1) the number of QTL underlying the trait (3000, 300, 30, 3) and (2) the genetic correlation across populations (0.4, 0.8, 1.0).

Results

The accuracy of across population genomic prediction obtained by a Bayesian variable selection model was the highest for a small number of QTL underlying the simulated trait. When the number of QTL increased, the accuracy of genomic prediction declined and eventually reached a plateau. This trend is stronger for across population genomic prediction than for within genomic prediction. Comparing the accuracy of across population genomic prediction obtained by Bayesian variable selection and GBLUP, it is shown that Bayesian variable selection has an advantage over GBLUP when the number of QTL underlying the simulated trait is small. However this advantage diminishes when the number of QTL underlying the simulated trait increases. The point where the accuracy of the Bayesian variable selection model and GBLUP becomes equal can be approximated by the independent number of chromosome fragments (M_e).

Conclusion

Bayesian variable selection performs better than GBLUP when the number of QTL underlying the trait is smaller than M_e. Across populations M_e is approximately ten times larger than within populations. So if the actual number of QTL is smaller than M_e, the use of Bayesian variable selection models can help to improve the accuracy of across population genomic prediction.

Keywords

Genomic prediction, within population, across population, Bayesian variable selection, GBLUP, accuracy

Background

In genomic prediction, a reference population consisting of animals with known phenotypes and marker genotypes is used to build a prediction equation with estimated single nucleotide polymorphism (SNP) effects to predict the quantitative trait loci (QTL) underlying a trait. With this prediction equation, genomic estimated breeding values (GEBVs) can be estimated for selection candidates with an unknown phenotype and a known genotype [1,2]. The accuracy of estimating GEBVs is dependent on several factors, such as: the size of the reference population [3-5], the heritability of the trait [6,7], the level of linkage disequilibrium (LD) between SNPs and QTL [1,8] and the additive genetic relationship between individuals [9,10].

In numerically small populations, e.g. lines or breeds with a low number of individuals, the size of the reference population is limited which restricts the potential accuracy of genomic prediction [5]. An attractive approach to increase the size of the reference population for a numerically small population is to add individuals from other populations to the reference population. Genomic prediction based on a reference population including individuals from multiple populations is known as multi population genomic prediction. A special case of multi population genomic prediction is across population genomic prediction where the populations in the reference population differ from the populations of the selection candidates. Simulation studies have shown that the accuracies of genomic prediction can be increased by adding individuals from other populations to the reference population [6]. However, several empirical studies show that applying across population genomic prediction to a numerically small population did not result in a significant increase in accuracy compared to within population genomic prediction [11-15].

The challenge for multi population genomic prediction is to deal with biodiversity and genetic variation across populations. It is a simple fact that individuals differ. Individuals of the same breed have more in common than individuals of different breeds and individuals from the same family have even more in common. Across populations, differences in LD, allele frequencies [16,17] and allele substitution effects [18,19] can be observed. These differences and the absence of close family relationships across populations [20] restrict the accuracy of multi population genomic prediction.

Another factor that is influencing the accuracy of genomic prediction, is the breeding value estimation model. Nowadays, two breeding value estimation models are commonly used, i.e. a linear GBLUP model and a nonlinear Bayesian variable selection model [21,22] . These models differ in their assumption about the distribution of the SNP variance. The GBLUP model assumes a homogeneous variance among SNPs; each SNP contributes equally to the total SNP variance. A Bayesian variable selection model assumes heterogeneous variances among SNPs, i.e. some SNPs have a large contribution to the variance and some SNPs have a small or zero contribution.

The difference in accuracy between GBLUP and a Bayesian variable selection model is dependent on the genetic architecture underlying the investigated trait and characteristics of the population. A study that has compared the accuracy of within population genomic

prediction for both a Bayesian variable selection model and GBLUP model, has shown that Bayesian approaches have an advantage over GBLUP when the number of QTL is smaller than the number of independent chromosome fragments (M_e) [22]. However when the number of QTL is equal or larger than M_e, the accuracy of both statistical methods become equal or, in some cases, GBLUP outperforms the Bayesian variable selection model [22].To our knowledge, to date the relationship between the number of QTL and M_e and the difference in accuracy between a Bayesian variable selection model and a GBLUP model has not been evaluated for across population genomic prediction. Wientjes *et al* [10] reports that M_e is substantially larger across populations than within a populations. Therefore it is more likely that the actual number of QTL underlying a trait is smaller than M_e across populations than within populations. It is therefore hypothesized that Bayesian variable selection models will be more accurate than GBLUP in case of across population genomic prediction, if the actual number of QTL is indeed smaller than M_e across populations.

The objective of this study is to estimate the accuracy of across population genomic prediction using a Bayesian variable selection model and compare the obtained accuracies with accuracy obtained by a GBLUP model, which are presented by Wientjes *et al*. [23]. This study uses the same dataset as in Wientjes *et al.* [23] which consists of high density genotypes of three dairy cattle breeds, i.e. Holstein Friesian, Meuse-Rhine-Yssel and Groninger White Headed. The phenotypes are simulated for two changing variables: (1) the number of QTL underlying the simulated trait (3, 30, 300, and 3000) and (2) the correlation between allele substitution effects across populations (0.4, 0.8, and 1.0). Simulated phenotypes are used to get a theoretical understanding of the factors that that are acting on the accuracy of across population genomic prediction.

Methods

Data

The dataset used in this study was retrieved from previous research of Wientjes *et al.* [23], containing the genotypes of 1285 Dutch dairy cows. The cows originated from three different breeds; 1033 Holstein Friesian (HF), 105 Groninger White Headed (GWH) and 147 Meuse Rhine Yssel (MRY) cattle. At least 87.5% of an individual's genotype originated from one of the three breeds and therefore all individuals were considered to be pure-bred.

The HF individuals were genotyped with the Illumina BovineSNP50 Beadchip (50k, Illumina, San Diego, CA). The genotypes were imputed to high-density (777k) using a reference population of 3150 HF individuals by Pryce *et al* [24]. The GWH and MRY individuals were genotyped with the Illumina BovineHD Beadchip (777k, Illumina, San Diego, CA). To increase the power of the analyses, only the SNPs on *Bos Taurus* chromosome (BTA) 13, 23, and 28 were considered. These three chromosomes are a good representation of the *Bos Taurus* genome because the LD pattern of chromosome 13, 23 and 28 is comparable to the LD pattern of the entire genome [25,26]. Non-segregating SNPs from the whole dataset were deleted, i.e. SNPs with a minor allele frequency equal or lower than 0.5%. After passing the quality control and editing, a total of 31,503 SNPs located on the three chromosomes remained. More details on the genotypes, quality control and editing of the SNP data are described in Wientjes *et al*.[23].

Phenotypes were simulated for different scenarios using two changing variables [23]: (1) the number of QTL underlying the trait, and (2) the correlation between allele substitution effects of the QTL across populations, which represents the genetic correlation between populations [27]. From all 31,503 SNPs in the dataset, 5000 SNPs were randomly selected as candidate QTL. Of these 5000 candidate QTL 3000, 300, 30 or 3 QTL were randomly selected, regardless of the chromosome and allele frequency, to have an effect on the simulated trait. The allele substitution effects of the QTL were sampled from a multi-normal distribution, assuming a genetic correlation of 1.0, 0.8 or 0.4 across all combinations of the three breeds. The remaining 26,503 (31,503-5000) SNPs were used as the group of markers for all analyses.

The simulated phenotypes were calculated as the sum of the true breeding values (**TBV**) and the environmental effect. The TBV was calculated by multiplying the QTL genotype of the 3000, 300, 30, or 3 QTL with the corresponding allele substitution effect, that is sampled from a multi normal distribution assuming a genetic correlation of 1.0, 0.8 or 0.4 [23]. The environmental effect was sampled from a normal distribution with mean zero and the variance equal to $(\frac{1}{h^2} - 1)$ * (variance of TBV corrected for mean TBV within population). The simulations of the phenotypes were replicated 100 times for each scenario and for each number of QTL underlying the trait, assuming a heritability of 0.95. More details about the simulations of the phenotypes are described in Wientjes *et al*.[23]

Scenarios

The accuracy of genomic prediction was evaluated for five different scenarios. An overview

of the scenarios is given in Table 1. The first scenario represents a within population scenario, where HF animals were used as reference population to predict GEBVs for HF selection candidates. Since the selection candidates and the reference population were selected from the same population, a 20-fold cross-validation was used to estimate GEBVs. The cross-validations were performed by randomly dividing the HF population in 20 groups where each group consisted of 51 or 52 individuals. In each cross-validation, one group was used as selection candidates and the other 19 groups were used as reference population. In the other four scenarios, the GEBVs were estimated for selection candidates of one population using a reference population of one or two other populations, i.e. applying across population genomic prediction. In all across population scenarios the HF population was included in the reference population.

TABLE. OVERVIEW OF THE DIFFERENT SCENARIOS

	REFERENCE POPULATION		SELECTION CANDIDATE	
SCENARIO	BREED(S)	NUMBER OF INDIVIDUALS	BREED	NUMBER OF CANDIDATES
BASE	HF	981 - 982[1]	HF	51 - 52[1]
1	HF	1033	GWH	105
2	HF & MRY	1180	GWH	105
3	HF	1033	MRY	147
4	HF & GWH	1138	MRY	147

HF = HOLSTEIN FRIESIAN; GWH = GRONINGER WHITE HEADED; MRY = MEUSE RHINE YSSEL; 1GENOMIC PREDICTION IS BASED ON A 20-FOLD CROSS VALIDATION USING 20 GROUPS OF 51 OR 52 SELECTION CANDIDATES

Genomic prediction

In this study a Bayesian stochastic search variable selection model (Bayes SSVS) [8,28] was used to perform genomic prediction. The following general model was applied for *n* individuals and *m* markers:

$$\mathbf{y} = \mathbf{1}_n\mu + \sum_{j=1}^{m}\mathbf{X}_j\beta_j + \mathbf{e}\ ;$$

where $\mathbf{y}$ is the vector of phenotypic records for all *n* individuals; μ is the mean; $\mathbf{1}_n$ is a vector of ones of length *n*; $\mathbf{X}_j$ is a vector of indicator variables referring to the genotypes for SNP *j* (*j*=1..*m*) for all individuals, β_j is the allele substitution effects associated with SNP *j* and $\mathbf{e}$ is a vector of residuals. The residuals were assumed to be normally distributed, $\mathbf{e} \sim N(0,\mathbf{I}\sigma_e^2)$ [1,2].

A uniform prior distribution is assigned to μ. The allele substitution effects β_j are assumed to be from a mixture of a normal distributions. From which distribution the allele substitution effects were sampled was determined by an indicator variable γ. The indicator variable reflects whether the SNP can be included in the model as a large effect, γ=1, or not, γ=0. For γ=1, β_j is sampled from $N(0, \sigma_\beta^2)$. For γ=0, β_j is sampled from $N(0,\frac{\sigma_\beta^2}{100})$ so that it has a very

small effect and therefore these SNP are not included in the model. As such, the prior distribution for each SNP effect is $\beta_i|\gamma_i, \sigma_\beta^2 \sim (1-\gamma_i)\mathbf{N}\left(0, \frac{\sigma_\beta^2}{100}\right) + \gamma_i \mathbf{N}(0, \sigma_\beta^2)$, with σ_β^2 sampled from an inverse chi-square distribution.

The prior distribution of the indicator variable γ was a Bernoulli distribution for prior probability p_i: $\gamma_i \sim bernoulli(p_i)$. Variable p_i reflects on the proportion of SNPs that have a large effect compared to the total number of SNPs. In this study p_i was set to 0.01. The posterior probability of the indicator variable can be sampled directly from its posterior distribution [28]: $p(\gamma = 1|\beta_j, \sigma_i^2, \gamma_{-i}, u, y) \sim Bernoulli\left(\frac{p(\beta_j|\gamma_{-i}, \gamma_i=1)p_i}{p(\beta_j|\gamma_{-i}, \gamma_i=1)p_i + p(\beta_j|\gamma_{-i}, \gamma_i=0)(1-p_i)}\right)$; where γ_i is the indicator variable and γ_{-i} refers to all indicator variables except γ_i. The posterior mean of the indicator variable refers to the frequency that the SNP is included in the model and is commonly referred to as the posterior probability. The higher the posterior probability, the more frequent the SNP was included in the model and therefore the higher the likelihood that the SNP is linked to a QTL [28].

A Monte Carlo Markov Chain (MCMC) algorithm was used to perform the analysis. For each analyses a Gibbs sampling chain with 5000 iterations was run. The first 1000 iterations were discarded as burn-in. A Monte Carlo Markov Chain (MCMC) algorithm was used to perform the analysis. For each analyses a Gibbs sampling chain with 5000 iterations was run. The first 1000 iterations were discarded as burn-in. For the first replicate of each scenario initially a Gibbs sampling of 100,000 iteration with 20,000 burn-in was run. The GEBVs obtained with 100,000 iterations had approximately a correlation of 1 with the GEBVs obtained with 5000 iterations. Therefore, in all remaining analyses a Gibbs sampling chain of 5000 iteration was considered.

Accuracy of genomic prediction

The accuracy of the genomic prediction was calculated as the Pearson correlation coefficient between the GEBV and TBV across all selection candidates per replicate, since the TBV was known for the selection candidates. An average accuracy and corresponding standard error were calculated for each scenario by averaging the accuracies for all 100 replicates. The average accuracy was used for further analysis and comparisons.

Model comparison

For each of the scenarios, the achieved average accuracy of genomic prediction from the Bayes SSVS model was compared with the average accuracy obtained by the GBLUP model, using the same data, estimated by Wientjes *et al.* [23]. It was investigated if also in across population genomic prediction the accuracies of both model becomes equivalent when the number of QTL is equal to M_e, as has been shown to be the case for within population genomic prediction [22]. To do this, M_e estimates were calculated by Wientjes *et al*. [23]: $M_e = \frac{1}{Var(\mathbf{G}_{RP_i,SK_j} - \mathbf{A}_{RP_i,SK_j})}$; where $\mathbf{G}_{RPi,SKj}$ refers to the genomic relationship between reference individual *i* and selection candidate *j*, $\mathbf{A}_{RPi,SKj}$ refers to the pedigree relationship between reference individual *i* and selection candidate *j* and the variance is take over all pair-

wise relationships between the individuals in the reference population and the selection candidates. An overview of the estimates of M_e is given in Table 2. More details on the calculation of M_e are described in Wientjes *et al*. [23].

TABLE 2. THE NUMBER OF INDEPENDENT CHROMOSOME FRAGMENTS (M_E) FOR EACH SCENARIO

Scenario[1]	M_e
Base	185
1	1809
2	1891
3	2435
4	2462

1. BASE SCENARIO: REFERENCE = HF, SELECTION CANDIDATES = HF; SCENARIO 1: REFERENCE = HF, SELECTION CANDIDATES = GWH; SCENARIO 2: REFERENCE = HF & MRY, SELECTION CANDIDATES = GWH; SCENARIO 3: REFERENCE = HF, SELECTION CANDIDATES = MRY; SCENARIO 4: REFERENCE = HF & GWH, SELECTION CANDIDATES = MRY.

Results

Equal allele substitution effects across populations

The accuracies of genomic prediction obtained with Bayes SSVS are shown in Figure 1 for all scenarios assuming equal allele substitution effects across the three breeds. The accuracy of the base scenario, which refers to within population genomic prediction, is high and increases slightly as the number of QTL reduces. The standard errors is very small for the base scenario. Accuracies of the other four scenarios, in which across population genomic prediction was applied, are lower than the accuracies for the base scenario. Standard errors for the across population scenarios are low as well and range from 0.009 to 0.02. The accuracy decreases significantly as the number of QTL is increasing. The effect of changing the number of QTL is much stronger for the across populations scenarios than for the within population scenario and the difference between 30 and 3 QTL is much smaller than the difference between 3000 and 300 QTL. The largest difference in accuracy can be observed between 300 and 30 QTL. Altogether, our results show that there is an effect of the number of QTL on the accuracy of across population genomic prediction using a Bayesian variable selection model.

The accuracy was slightly higher for selection candidates originating from GWH population than for those originating from the MRY population. For both breeds the accuracies somewhat increased when the other breed was added to the HF reference population.

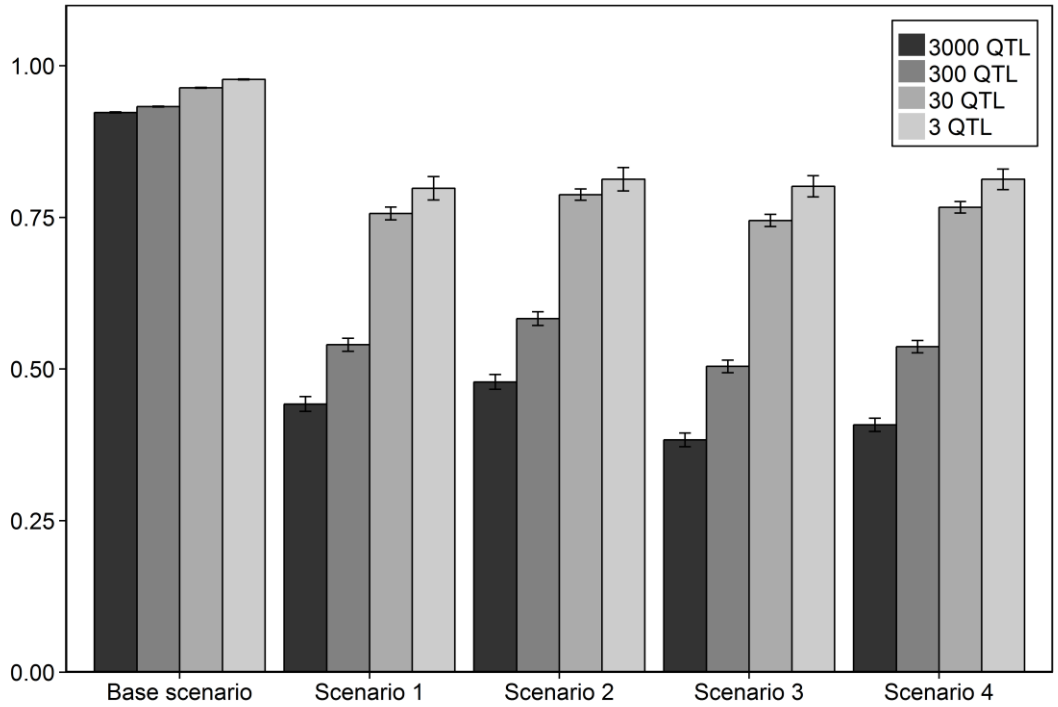


FIGURE 1. ACCURACIES OF GENOMIC PREDICTION ASSUMING EQUAL ALLELE SUBSTITUTION EFFECTS ACROSS POPULATIONS. MEAN ACCURACIES OF GENOMIC PREDICTION (± STANDARD ERROR) OBTAINED BY BAYES SSVS ASSUMING EQUAL ALLELE SUBSTITUTION EFFECTS ACROSS THE THREE BREEDS FOR THE FIVE DIFFERENT SCENARIOS; BASE SCENARIO: REFERENCE = HF, SELECTION CANDIDATES = HF; SCENARIO 1: REFERENCE = HF, SELECTION CANDIDATES = GWH; SCENARIO 2: REFERENCE = HF & MRY, SELECTION

CANDIDATES = GWH; SCENARIO 3: REFERENCE = HF, SELECTION CANDIDATES = MRY; SCENARIO 4: REFERENCE = HF & GWH, SELECTION CANDIDATES = MRY

Different genetic correlation across populations

The accuracies of genomic prediction are shown in Figure 2 assuming a genetic correlation across the populations of 0.8 (A.) or 0.4 (B.). The standard errors range from 0.01 to 0.05 for all scenarios. When there were 3 QTL underlying the simulated trait the standard errors are larger than when there were 30, 300 or 3000 QTL underlying the simulated trait. Compared to the scenarios with equal allele substitution effects across populations, the accuracy of the scenarios with different allele substitution effects across populations is decreasing proportional to the correlation in allele substitution effects, i.e. the genetic correlation. So, when the genetic correlation is 0.8, the accuracy is approximately 80 percent of accuracy obtained with a genetic correlation across populations of 1, and when the genetic correlation is 0.4, the accuracy is approximately 40% of the accuracy obtained with genetic correlation across populations of 1.

The effect of the number of QTL on the accuracy is the same for the scenarios that use a different genetic correlation across populations as for the scenarios with an equal genetic correlation; the accuracy is increasing when the number of QTL underlying the trait is decreasing. Remarkably, the accuracies for scenario 1 and 2 with 3 simulated QTL is smaller than the accuracies for the scenario that used 30 QTL when the genetic correlation was 0.8. This can be explained by the higher standard error of the accuracies for the scenarios using 3 QTL.

Again, the accuracies for the selection candidates originating from the breed GWH were slightly higher than the accuracies for the selection candidates from the breed MRY. Adding another breed to the HF reference population increased the accuracy of genomic prediction for the selection candidates.

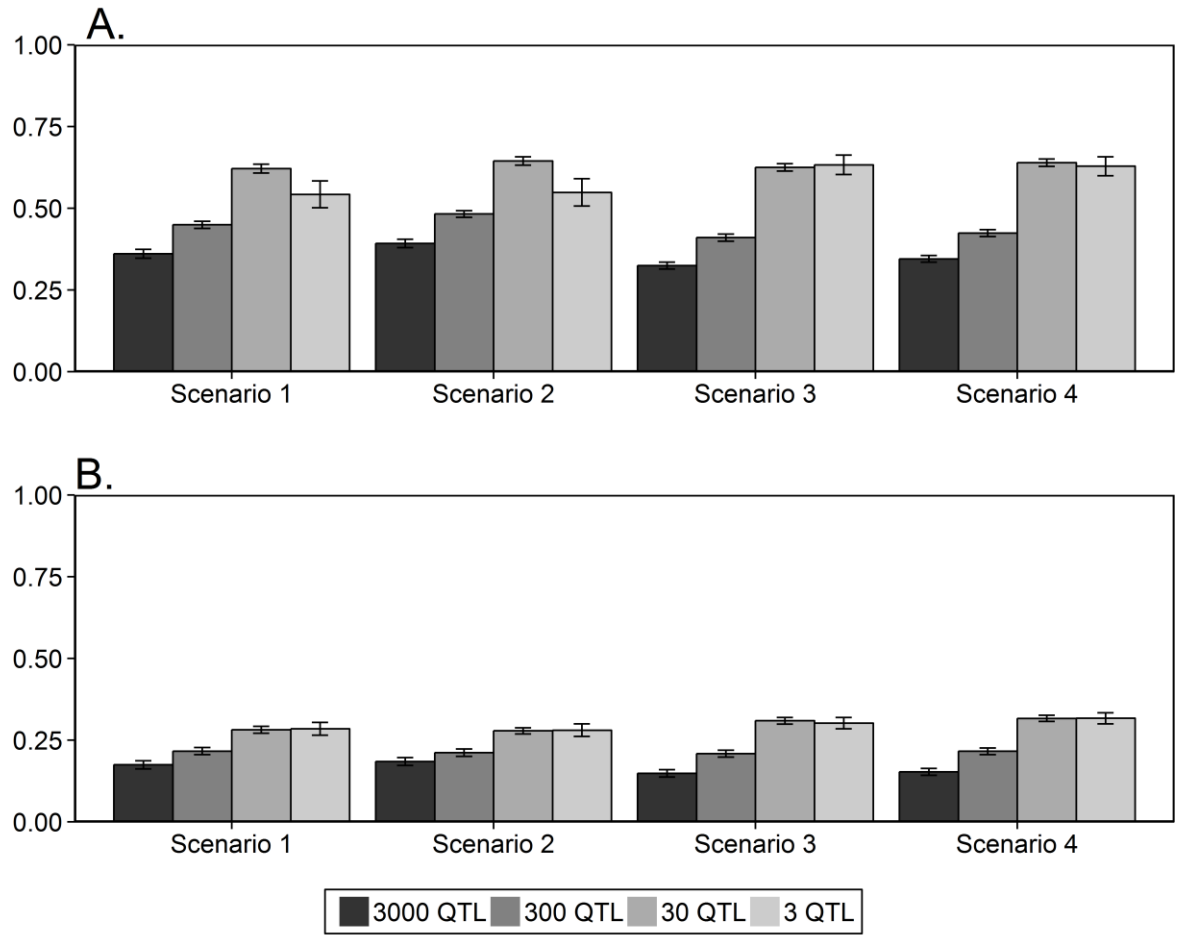


FIGURE 2. ACCURACIES OF GENOMIC PREDICTION ASSUMING DIFFERENT ALLELE SUBSTITUTION EFFECTS ACROSS POPULATIONS. MEAN ACCURACIES OF GENOMIC PREDICTION (± STANDARD ERROR) OBTAINED BY BAYES SSVS ASSUMING GENETIC CORRELATIONS OF (**A**) 0.8 OR (**B**) 0.4 ACROSS THE THREE BREEDS FOR FOUR DIFFERENT SCENARIOS; SCENARIO 1: REFERENCE = HF, SELECTION CANDIDATES = GWH; SCENARIO 2: REFERENCE = HF & MRY, SELECTION CANDIDATES = GWH; SCENARIO 3: REFERENCE = HF, SELECTION CANDIDATES = MRY; SCENARIO 4: REFERENCE = HF & GWH, SELECTION CANDIDATES = MRY

QTL detection

Whether or not a QTL is detected can be evaluated by a Manhattan plot that shows the posterior probabilities for all SNP across the chromosomes. The posterior probability refers to the likelihood that a SNP has a large effect. So a high posterior probability indicates that it is highly likely that the SNP has a large effect on the simulated trait and therefore it is likely that the SNP is in LD with a QTL. The level of LD is expected to be stronger when the SNPs and QTL are close to each other. So the SNPs that have a high posterior probability and therefore have a large effect, are close to the QTL and can indicate the position of the QTL.

The Manhattan plots for the three different reference populations are shown in Figure 3 for a genetic correlation of 1.0 and Figure 4 for a genetic correlation of 0.8 assuming 3 QTL underlying the simulated trait. The Manhattan plots for a genetic correlation of 0.4 are shown in Figure S1 in the supplementary files. The three reference population are: (A) reference population 1 which consist of only HF, (B) reference population 2 which consists of HF and

GWH and (C) reference population 3 which consists of HF and MRY. For illustration purposes, only the posterior probabilities for the first replicate were evaluated.

When a genetic correlation of 1.0 is assumed, all QTL are surrounded by SNPs that have a high posterior probability. In the analyses with a genetic correlation across populations of 0.8, not all QTL were surrounded by a couple of SNPs with a high posterior probability. Near the second QTL, the SNPs do not show a high posterior probability. This is probably due to the small effect of this QTL, i.e. the sampled QTL effect was approximately 0.1. Therefore the QTL is hard to detect. SNPs near the first and third QTL have a high posterior probability. This indicates that there is a strong association between these SNPs and QTL and that the QTL effect was picked up by the SNPs.

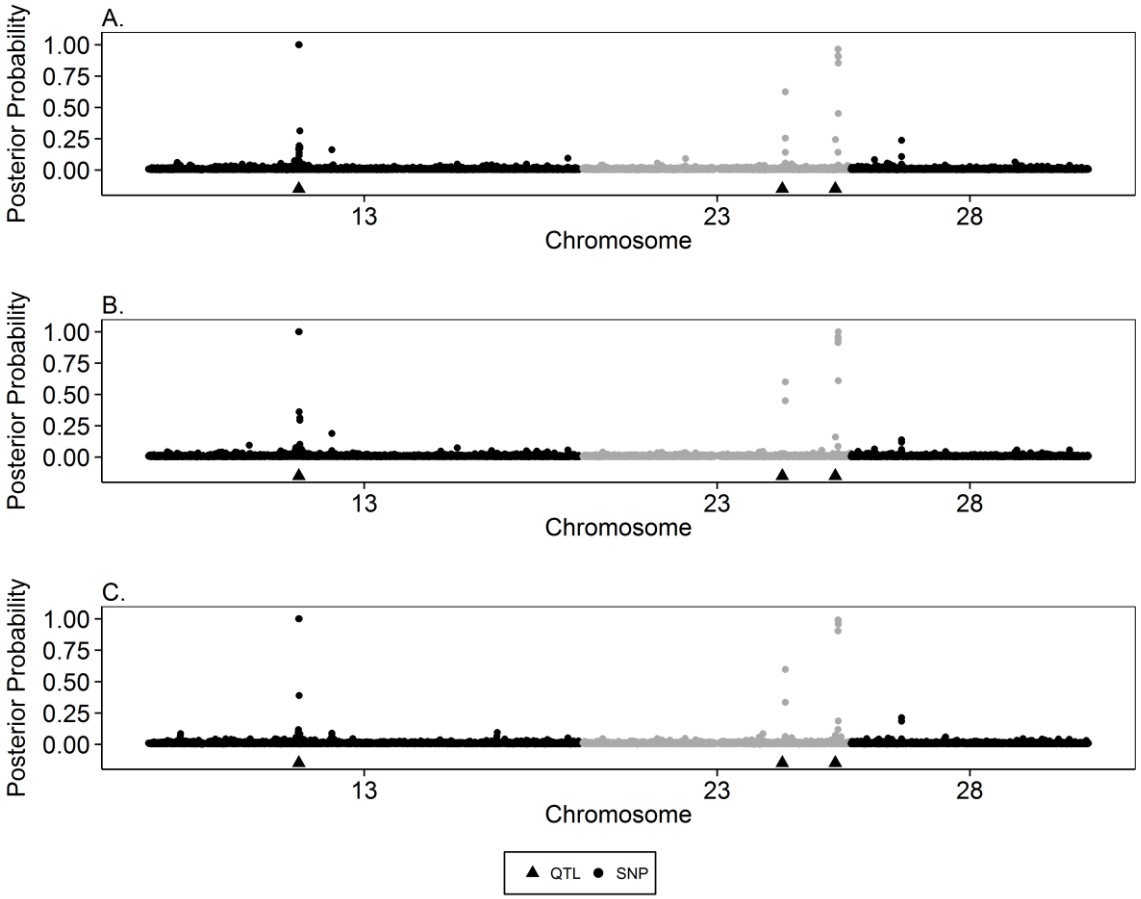


FIGURE 3 MANHATTAN PLOT FOR REPLICATE 1 WITH GENETIC CORRELATION 1.0 AND 3 QTL. DISTRIBUTION OF THE POSTERIOR PROBABILITY OF 26,503 SNPS (**POINTS**) ACROSS CHROMOSOME 13, 23 AND 28 ASSUMING EQUAL ALLELE SUBSTITUTION EFFECTS ACROSS POPULATIONS USING THREE DIFFERENT REFERENCE POPULATIONS; (**A**) REFERENCE 1: HF; (**B**) REFERENCE 2: HF & MRY; (**C**) REFERENCE 3: HF & GWH. THE **TRIANGLES** INDICATES THE POSITIONS OF THE THREE QTL.

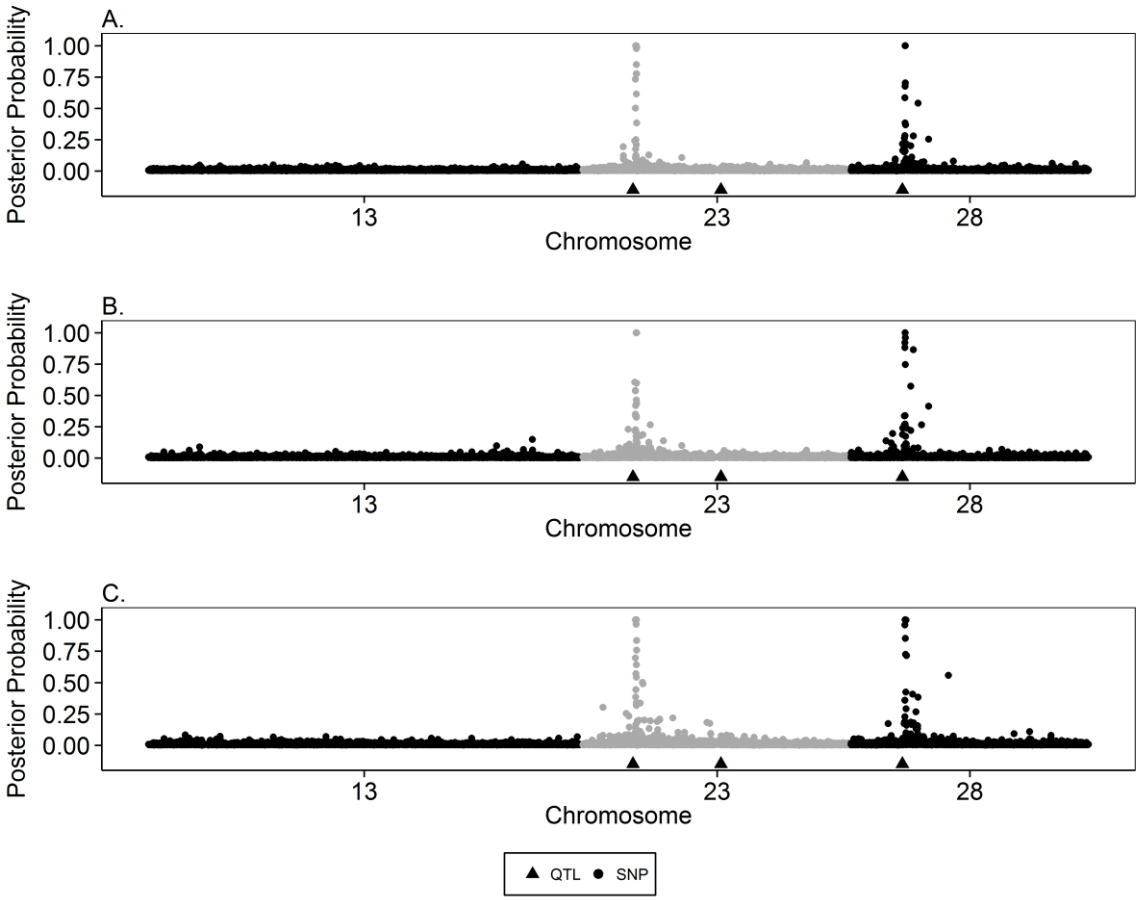


FIGURE 4 MANHATTAN PLOT FOR REPLICATE 1 WITH GENETIC CORRELATION OF 0.8 AND 3 QTL. DISTRIBUTION OF THE POSTERIOR PROBABILITY OF 26,503 SNPS (**POINTS**) ACROSS CHROMOSOME 13, 23 AND 28 ASSUMING A GENETIC CORRELATION OF 0.8 ACROSS POPULATIONS USING THREE DIFFERENT REFERENCE POPULATIONS; REFERENCE 1 (**A**): HF; REFERENCE 2 (**B**): HF & MRY; REFERENCE 3 (**C**): HF & GWH. THE **TRIANGLES** INDICATES THE POSITIONS OF THE THREE QTL.

Assuming a genetic correlation of 1.0, the addition of an extra breed to the reference population, as in reference population 2 and 3, results in less SNPs with an elevated posterior probability close to the QTL. When the genetic correlation across populations was different from 1.0 (Figure 4), adding an extra breed to the reference population did not always result in a decrease of the number of SNPs with an elevated posterior probability close to the QTL. For example, for a genetic correlation of 0.8 adding MRY to the reference population (C.) seems to lead to an increase in the number of SNPs with a high posterior probability close to the QTL instead of a decrease. To further investigate this trend, the region of hundred SNPs before and after the first QTL is evaluated in Figure 5 for an equal genetic correlation across populations and in figure 6 for a genetic correlation of 0.8 across populations. The zoomed plot for a genetic correlation of 0.4 can be found Figure S2 in the supplementary files. Only the first QTL is evaluated for illustration purposes. The below described trends are also found for the second and third QTL. The upper triangle refers always to the QTL effect of the HF. The lower triangle refers to the QTL effect of the additional breed, which is either GWH in reference population 2 (B.) or MRY in reference population 3 (C.), relative to the QTL effect of HF.

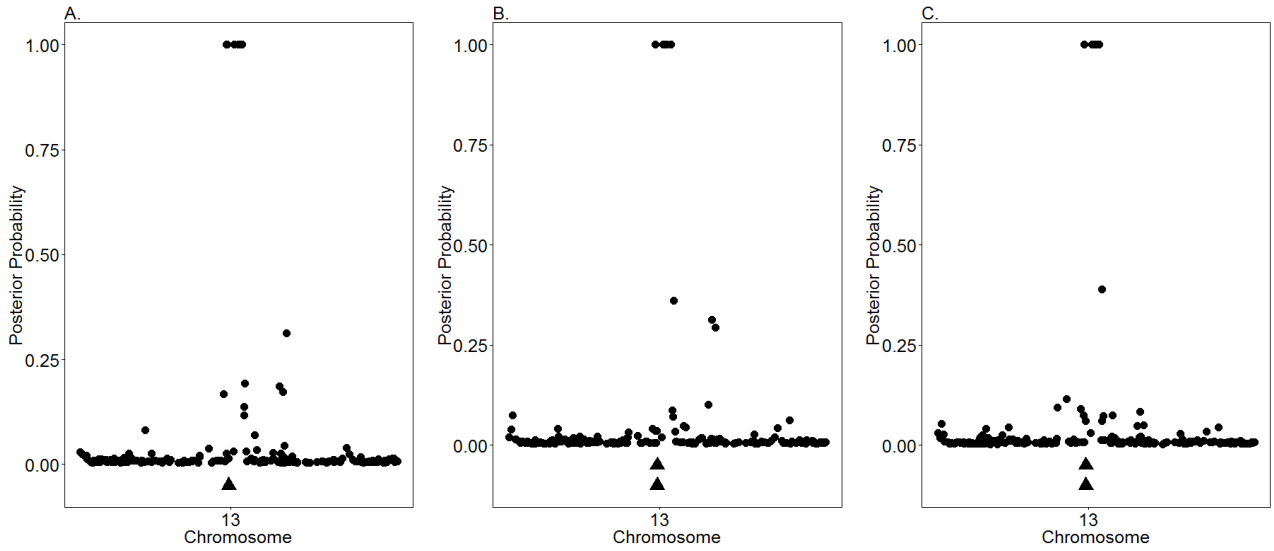


FIGURE 5 MANHATTAN PLOT FOR THE REPLICATE 1 WITH GENETIC CORRELATION 1.0 ZOOMED IN ON A 200 SNP REGION NEIGHBOURING THE FIRST QTL. DISTRIBUTION OF THE POSTERIOR PROBABILITY OF 100 SNPS (**POINTS**) BEFORE AND AFTER THE POSITION OF THE FIRST QTL, WHICH IS INDICATED BY THE **TRIANGLE**, USING THREE DIFFERENT REFERENCE POPULATIONS ASSUMING AN EQUAL GENETIC CORRELATION ACROSS POPULATIONS; REFERENCE 1 (**A**): HF; REFERENCE 2 (**B**): HF & GWH; REFERENCE 3 (**C**): HF & MRY. THE UPPER TRIANGLE REFERS ALWAYS TO THE QTL EFFECT OF THE HF. THE LOWER TRIANGLE REFERS TO THE QTL EFFECT OF THE ADDITIONAL BREED RELATIVE TO THE QTL EFFECT OF HF.

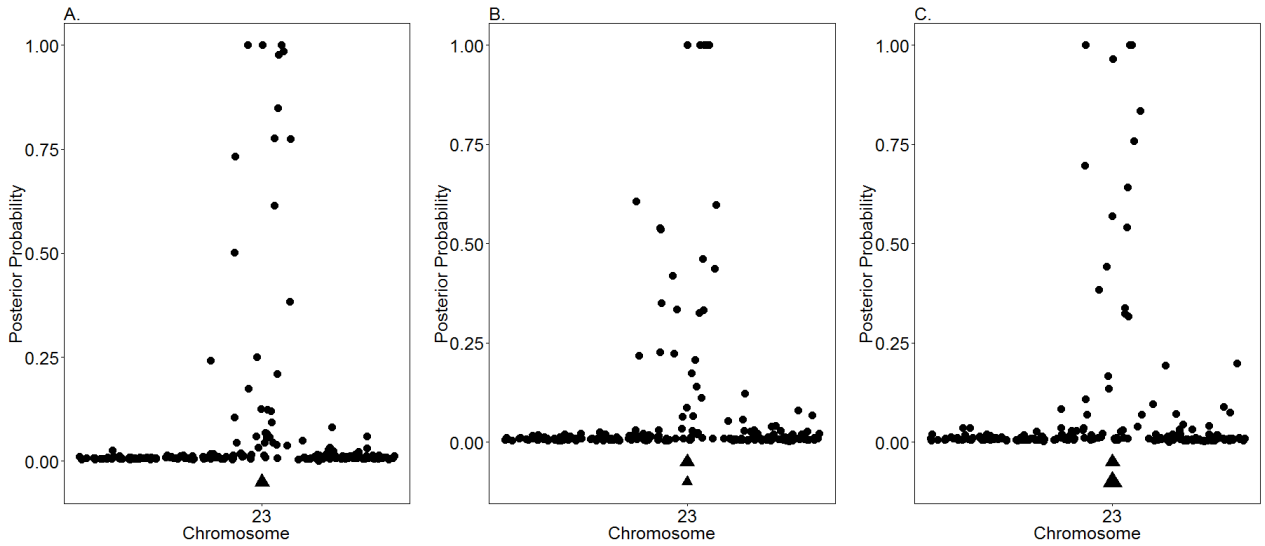


FIGURE 6 MANHATTAN PLOT FOR THE REPLICATE 1 WITH GENETIC CORRELATION 0.8 ZOOMED IN ON A 200 SNP REGION NEIGHBOURING THE FIRST QTL. DISTRIBUTION OF THE POSTERIOR PROBABILITY OF 100 SNPS (**POINTS**) BEFORE AND AFTER THE POSITION OF THE FIRST QTL, WHICH IS INDICATED BY THE **TRIANGLE**, USING THREE DIFFERENT REFERENCE POPULATIONS ASSUMING A GENETIC CORRELATION OF 0.8 ACROSS POPULATIONS; REFERENCE 1 (**A**): HF; REFERENCE 2 (**B**): HF & GWH; REFERENCE 3 (**C**): HF & MRY. THE UPPER TRIANGLE REFERS ALWAYS TO THE QTL EFFECT OF THE HF. THE LOWER TRIANGLE REFERS TO THE QTL EFFECT OF THE ADDITIONAL BREED RELATIVE TO THE QTL EFFECT OF HF.

As expected the number of SNPs that have a high posterior probability becomes smaller when you add an additional breed to the reference population, when the genetic correlation across populations is 1.0 (Figure 5). When a different genetic correlation across populations in assumed, i.e. 0.8, the number of SNP that have a high posterior probability becomes larger

when the additional breed has a larger QTL effect relative to the QTL effect of the HF and the number of SNPs that have a high posterior probability becomes smaller when the additional breed has an equal or smaller QTL effect. For example, for a genetic correlation of 0.8 GWH has a smaller QTL effect compared to HF and adding GWH to reference population (B.) resulted in a decrease of SNP that have a high posterior probability. When MRY, which has a larger QTL effect than the HF, was added to the reference population, the number of SNP with high posterior probability increased.

Altogether, the results indicate that the number of SNPs with a high posterior probability decrease when an extra breed is added to the reference population and the effect of the QTL in that population is equal or smaller than in the first population. When an extra breed is added where the effect of the QTL is higher, the number of SNPs with a high posterior probability increases.

Comparison with GBLUP

Figure 7 shows the comparison between the two methods in relationship to the number of QTL for the within population scenario, i.e. the base scenario, and Figure 8 shows the comparison between Bayes SSVS and GBLUP for the across population scenarios, i.e. scenario 1 (A.), scenario 2 (B), scenario 3 (C.) and scenario 4 (D.). Please note that ln(number of QTL) is plotted against the reliability, since the relationships between number of QTL and reliability can be approximated by an exponential function following the prediction formula of Daetwyler et al (2008).

For both the Bayesian variable selection model and the GBLUP model the reliabilities for the scenarios that are assuming a lower genetic correlation are always smaller than those reliabilities that are estimated for scenarios with higher genetic correlations. Assuming an equal genetic correlation across populations, i.e. the genetic correlation is 1.0, resulted in the highest reliabilities for both Bayes SSVS and GBLUP.

With a low number of QTL underlying the trait, Bayes SSVS performs always better than GBLUP. When the number of QTL becomes higher, the difference between the reliabilities of both approaches becomes smaller and eventually GBLUP results in slightly better reliabilities than Bayes SSVS.

The number of independent chromosome fragments (M_e)

The results show that the Bayes SSVS and GBLUP model have an equal reliability at a much smaller number of QTL for the within population scenario (Figure 7) compared to the across populations scenarios (Figure 8). For the within population scenario, the reliability of the Bayes SSVS and GBLUP become equal at approximately 200 QTL (Figure 7), which is much lower than found across population where both models show the same reliability at approximately 2000 QTL. This is in agreement with the estimated values for M_e, which were much lower within population than across populations (see Table 2.).

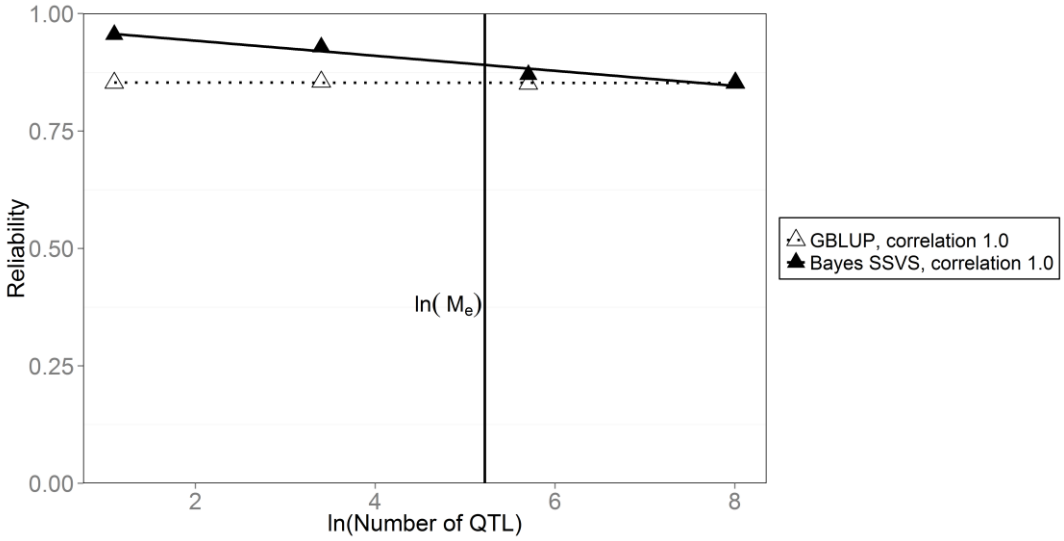


FIGURE 7 COMPARISON OF THE RELIABILITY OF WITHIN POPULATION GENOMIC PREDICTION USING BAYES SSVS OR GBLUP. COMPARISON OF THE MEAN RELIABILITY OF GENOMIC PREDICTION USING BAYES SSVS OR GBLUP FOR THE WITHIN POPULATION SCENARIO. THE VERTICAL LINE INDICATES THE NATURAL LOGARITHM OF THE NUMBER OF INDEPENDENT CHROMOSOMES (M_E). M_E IS ESTIMATED BY WIENTJES *ET AL* [23] ACCORDING TO THE FORMULA OF GODDARD *ET AL* [37].

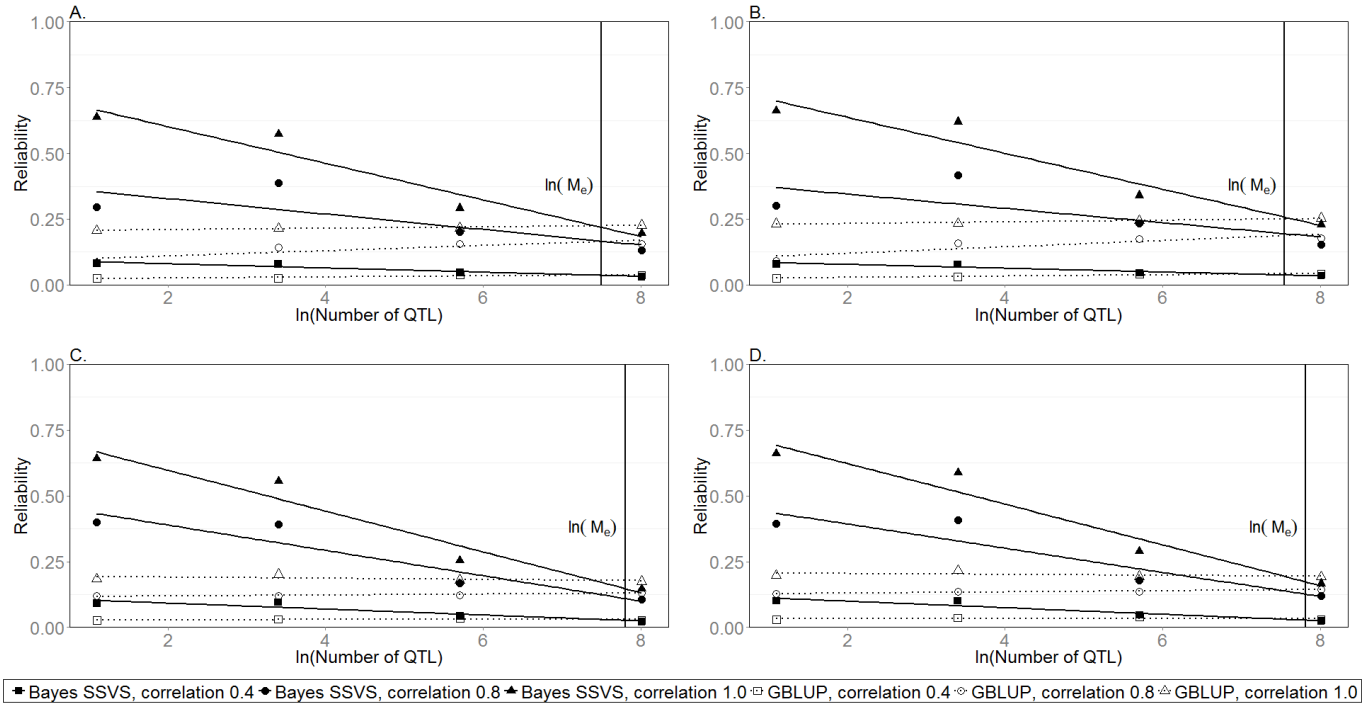


FIGURE 8 COMPARISON OF THE RELIABILITY OF ACROSS POPULATION GENOMIC PREDICTION USING BAYES SSVS OR GBLUP. COMPARISON OF THE MEAN RELIABILITY OF GENOMIC PREDICTION USING BAYES SSVS OR GBLUP FOR THE FOUR ACROSS POPULATION SCENARIOS FOR AN EQUAL GENETIC CORRELATION ACROSS POPULATIONS (TRIANGLE) OR AN DIFFERENT GENETIC CORRELATION OF 0.8 OR 0.4 ACROSS POPULATIONS; SCENARIO 1 (A.): REFERENCE = HF, SELECTION CANDIDATES = GWH; SCENARIO 2 (B.): REFERENCE = HF & MRY, SELECTION CANDIDATES = GWH; SCENARIO 3 (C.): REFERENCE = HF, SELECTION CANDIDATES = MRY; SCENARIO 4 (D.): REFERENCE = HF & GWH, SELECTION CANDIDATES = MRY. THE VERTICAL LINE INDICATES THE NATURAL LOGARITHM OF THE INDEPENDENT NUMBER OF CHROMOSOME FRAGMENTS (M_E). M_E IS ESTIMATED BY WIENTJES *ET AL*. [23] ACCORDING TO THIS FORMULA OF GODDARD *ET AL*.[37].

Discussion

The accuracy of across population genomic prediction

The objective of this study is to estimate the accuracy of across population genomic prediction using a Bayesian variable selection model and compare the obtained accuracies with accuracies obtained by a GBLUP model, which are presented by Wientjes *et al*. [23]**.** In this study real genotypes of 1033 HF, 147 MRY and 105 GWH were used. The phenotypes of the individuals were simulated with two changing variables: (1) the number of QTL underlying the trait, and (2) the genetic correlation across populations.

The accuracies for within population genomic prediction are higher than the accuracies for across population genomic prediction for both the Bayesian variable selection model and the GBLUP model. This is in line with the general observation in literature, e.g. [13,21,29]. The smaller accuracies obtained for across population genomic prediction can be explained by the differences across populations, such as differences in LD patterns, allele frequencies and allele substitution effects. These differences and the absence of family relationships restrict the accuracy of genomic prediction across populations [16-20] .

Wientjes *et al*. [23] pointed out that the value of the genetic correlation is an important factor for the accuracy of across population genomic prediction obtained by GBLUP. A decrease in the genetic correlation resulted in a reduction of the accuracy obtained by GBLUP proportional to the genetic correlation. Our results show that the genetic correlation affects the accuracy of genomic prediction obtained by Bayesian variable selection model in equal way. Chen *et al.* [30] have investigated the effect of the genetic correlation on the accuracy of multi population genomic prediction obtained with a Bayesian variable selection model that was equal to the model used in this study. They have found indeed a reduced accuracy when the genetic correlation decreased.

Interestingly, Chen *et al*. [30] have proposed an approach to deal with the differences across populations by developing multi-task learning of the Bayesian model. They have made the distinction between single-task and multi-task learning of the Bayesian model. Single-task learning refers to the traditional learning mechanism of the Bayesian model where different populations are considered as one population. Multi-task learning refers to simultaneously learning from multiple information sources [31]. The multi-task learning approach distinguishes between the different populations in the reference population by allowing the SNP effects to vary across the populations. Information is shared across populations by a common set of indicator variables. Therefore the multi-task learning model can correct for differences in genetic architecture across populations and is more flexible than a single-task learning model. Thus the accuracy of genomic prediction using a multi-task learning model is improved compared to a single-task learning approach when multiple populations are included in the reference population and the genetic correlation between those population is smaller than one [30]. Using a multi-task learning model that accounts for the differences across populations, will especially be useful for scenarios where two or more populations are included in the reference population. Applied to this study, the obtained accuracies for the scenarios using a reference population of HF with either the MRY or the GWH population

could benefit from multi-task learning. It would be interesting for future research to investigate the accuracy of the multi-task learning approach applied to across population genomic prediction using two or more breeds in the reference population.

In this study it is demonstrated that the accuracy of across population genomic prediction obtained by a Bayesian variable selection model depends on the number of QTL underlying the simulated trait, It was shown that the Bayesian variable selection model obtained the highest accuracies when the number of QTL is small When the number of QTL is increasing, the accuracy obtained by Bayesian variable selection model declines and eventually reaches an asymptote. The dependency of the accuracy of both within and across population genomic prediction obtained by a Bayesian variable selection model on the number of QTL is also found in the literature [17,22,32-34]. For example, Coster *et al*. [32] investigated the effect of the number of QTL on the accuracy of within population genomic prediction. They have found that the accuracy of within population genomic prediction obtained by a Bayesian variable selection model decreased when the number of simulated QTL increased. Similar results are also found by Chen *et al*. [30]. They have used a Bayesian variable selection model equal to the model used in this study to perform multi population genomic prediction and found also that the accuracy of multi population genomic prediction is decreasing when the number of simulated QTL is increasing.

Whereas the accuracy of genomic prediction obtained by a Bayesian variable selection model is affected by the number of QTL, the accuracy obtained by GBLUP is constant and remains unaffected by the number QTL underlying the simulated trait. Although, when the number of QTL underlying the simulated trait becomes very small, the accuracy slightly declines. Several empirical studies have shown that Bayesian variable selection models perform indeed better than GBLUP when across population genomic prediction is performed, e.g. [21,35,36]. Hayes *et al.* [21] have combined Australian HF and Jersey population for genomic prediction. When only Australian HF was used in the reference population and the selection candidates were Jersey, the Bayesian variable selection model resulted in an increase in accuracy compared to the accuracy obtained by GBLUP.

Whether or not the accuracy of genomic prediction is affected by the number of QTL can be explained by the difference in model mechanism. The GBLUP model assumes the infinitesimal model, i.e. each SNP is assumed to explain an equally small amount of the variation. Bayesian variable selection models make a distinction between the SNPs by making a small subset of SNPs that are expected to have a large effect and a subset of the other SNPs are not expected to have an effect. The size of the subset is dependent on p_i, which reflects on the proportion of SNP that has a large effect compared to the total number of SNP. If the number of QTL is substantially smaller than the total number of SNPs, it is clear which subset to choose. If the number of QTL is equal to or larger than the number of SNPs, it is less obvious which subset to choose. Because it seems like each SNP has a small effect on one of the QTL, it is difficult for the model to select the SNPs with a large effect. Therefore the model takes a more random subset of SNPs close to the QTL and assigns an equal amount of variance to each SNP. This approach is equivalent to the assumption of the infinitesimal model that was assumed for GBLUP.

The choice of subset in Bayesian variable selection is also dependent on the distribution underlying the QTL effects. If a normal distribution is assumed, such as in this study, there will be no QTL with extreme effect. If a gamma distribution is assumed, such as in Meuwissen *et al* [1], there will only be a few QTL that have a large effect, because the distribution is positively skewed. The QTL with a large effect have a major contribution to accuracy of genomic prediction. Since only a few QTL have a large effect when a gamma distribution is assumed, the effective number of QTL is smaller than the real number of QTL, resulting in higher accuracies of genomic prediction using a Bayesian model [1].

Daetwyler *et al*. [22] investigated the difference in factors acting on the accuracy of within population genomic prediction obtained by Bayesian variable selection or GBLUP. They have reported that the accuracy of GBLUP is independent from the number of QTL, but is dependent on genomic parameters of the population, such as the effective population size and LD. The genomic parameter of the population can be summarized with M_e, the number of independent chromosome segments [22]. M_e is a statistical concept that links genomic properties of the population to the statistical analysis. It can be derived from the consistency of variation in LD across the genome and the variation in relationship between relatives [5]. In a wider sense M_e can be interpreted as the independent number of informative markers needed to capture all the variation in QTL effects. Thus the accuracy of genomic prediction obtained by GBLUP is dependent on M_e. The accuracy obtained by a Bayesian variable selection model is dependent on the interaction between the genomic parameters of the population and the characteristics of the trait, i.e. the interaction between the number of QTL and M_e. When the number of QTL is lower than Me, the accuracy obtained by Bayesian variable selection decreases when the number of QTL is increasing. When the number of QTL is above Me, the accuracy is independent from the number of QTL and similar to the accuracy obtained by GBLUP[22]. Our results show that this principle, established by Daetwyler *et al*. [22] for within population genomic prediction, is also correct for across population genomic prediction. Therefore M_e can be considered to be an important parameter for within and across population genomic prediction.

Wientjes *et al*. [23] has shown that M_e is larger across population than within population. They have found that the estimates for M_e were approximately 10 times larger across population than within population [23]. The higher estimates for M_e across populations can be explained by the fact that M_e is dependent on the level of relatedness between individuals [7,10]. When individuals are closely related, LD is strong and less informative markers are needed to explain the variation in QTL effects. Therefore the estimates of M_e are small. However, it is well known that across populations there is an absence of closely related individuals and individuals differ strongly in LD patterns. So more informative markers are needed to explain the variation in QTL effects and the estimate of M_e is higher. Because M_e is higher across populations, it is more likely to have a number of QTL underlying a trait that is smaller than M_e. However the real number of QTL is not known for many traits. So it can be concluded that a Bayesian variable selection model can help to improve the accuracy of across population genomic, if the actual number of QTL underlying a trait is smaller than M_e.

It should be noted that in this study SNPs are representative of the QTL. Therefore the SNPs and the QTL have the same characteristics. However in practice, SNPs and QTL have different characteristics, such the minor allele frequency patterns, and therefore there is a different LD pattern between the SNP and the QTL. Therefore the accuracies and the potential beneficial effect of the Bayesian variable selection model might be overestimated.

QTL detection

In this study it was demonstrated that, when assuming equal allele substitution effects across populations, adding an extra population to the reference population resulted in a decrease of the number of SNPs with a high posterior probability. One explanation might be as follows. The model does not distinguish between the two populations. So if you add an extra population to the reference population, it assumes that the two populations are the same, neglecting the fact that the two populations differ in for example LD pattern and allele substitution effects. In order to pick up QTL in a reference population consisting of two populations, the SNP has to be in LD with the QTL in both populations. For this to occur, the SNP has to be very close to the QTL in order to have strong conserved LD across populations [6,16]. So SNPs that are only in one population in LD with the QTL are filtered and only the SNPs in high LD across populations are used for genomic prediction, resulting in a lower number of SNPs with a high posterior probability. Support for this hypothesis is given by Hayes *et al*. [21]. They have shown that for detection of a QTL with a large effect, only the SNPs that were close to the QTL had an effect that persisted across the populations.

However the hypothesis described above does not hold when a different genetic correlation across populations is assumed. Using a genetic correlation of 0.8 resulted in the interesting observation that the number of SNPs with a high posterior probability increases when an extra population is added to the reference population and the effect of the QTL in that population is larger than the effect of the QTL in the first population. A possible explanation might be as follows. In Bayesian statistics a subset of SNP that have a large effect is selected. Within this subset, the SNPs can vary in how much of the QTL effect it explains. If you have a QTL with an effect that differs across population and the effect of the extra population is larger than the effect of the QTL in the first population, the average QTL effect is higher. Therefore there is more of the QTL effect to explain and the number of SNPs that have a high posterior probability is increased.

Two critical notes need to be made. First, the genetic correlation in the single replicate for which the results were shown was smaller than 0.8. This might cause a slight overestimation of the extent of the trend, however the principle behind the trend will be the same. Second, the size of population that was added to the reference population was quite small, i.e. the size of the HF population is 1033 individuals, while the GWH population size is105 individuals and the MRY population size is 147. The effect of adding a population that has a QTL with a larger effect than the first population is expected to be more pronounced if the population size of the extra population is larger, since the QTL effect of that has a large influence.

Conclusion

The accuracy of across population genomic prediction obtained by a Bayesian variable selection model is dependent on the number of QTL underlying the trait. The Bayesian variable selection model results in the highest accuracy when a small number of QTL is underlying the trait. When the number of QTL underlying the trait is increasing, the accuracy of genomic prediction obtained by a Bayesian variable selection model declines and eventually reaches a plateau, equal to the accuracy obtained by GBLUP. The point where the accuracy obtained by Bayesian variable selection becomes equivalent to the accuracy obtained by GBLUP can be approximated by the independent number of chromosome fragments (M_e). When the number of QTL is smaller than M_e, the Bayesian variable selection model has an advantage over GBLUP. When the number of QTL is equal or larger than M_e, the advantage of the Bayesian variable selection model disappears and the accuracy becomes equivalent to the accuracy obtained by GBLUP. So Bayesian variable selection has an advantage over GBLUP when the number of QTL is smaller than M_e. Across populations M_e is larger than within populations. So if the actual number of QTL is smaller than M_e, a Bayesian variable selection model can help to improve the accuracy of across population genomic prediction.

References

1. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps**. *Genetics* 2001, **157**(4):1819-1829.

2. Meuwissen THE, Hayes BJ, Goddard ME: **Accelerating improvement of livestock with genomic selection**. *Annu Rev Anim Biosci* 2013, **1**(1):221-237.

3. Goddard ME, Hayes BJ: **Mapping genes for complex traits in domestic animals and their use in breeding programmes**. *Nat Rev Genet* 2009, **10**(6):381-391.

4. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS: **Invited review: Reliability of genomic predictions for North American Holstein bulls**. *J Dairy Sci* 2009, **92**(1):16-24.

5. Goddard ME: **Genomic selection: prediction of accuracy and maximisation of long term response**. *Genetica* 2009, **136**(2):245-257.

6. De Roos APW, Hayes BJ, Goddard ME: **Reliability of genomic predictions across multiple populations**. *Genetics* 2009, **183**(4):1545-1553.

7. Hayes BJ, Visscher PM, Goddard ME: **Increased accuracy of artificial selection by using the realized relationship matrix**. *Genet Res* 2009, **91**(01):47-60.

8. Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF: **Accuracy of genomic selection using different methods to define haplotypes**. *Genetics* 2008, **178**(1):553-561.

9. Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G: **The impact of genetic relationship information on genomic breeding values in German Holstein cattle**. *Genet Sel Evol* 2010, **42**(1):5.

10. Wientjes YJC, Veerkamp RF, Calus MPL: **The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction**. *Genetics* 2013, **193**(2):621-631.

11. Olson KM, VanRaden PM, Tooker ME: **Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss**. *J Dairy Sci* 2012, **95**(9):5378-5383.

12. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME: **Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels**. *J Dairy Sci* 2012, **95**(7):4114-4129.

13. Pryce JE, Gredler B, Bolormaa S, Bowman PJ, Egger-Danner C, Fuerst C, Emmerling R, Sölkner J, Goddard ME, Hayes BJ: **Short communication: Genomic selection using a multi-breed, across-country reference population**. *J Dairy Sci* 2011, **94**(5):2625-2630.

14. Brøndum RF, Rius-Vilarrasa E, Strandén I, Su G, Guldbrandtsen B, Fikse WF, Lund MS: **Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations**. *J Dairy Sci* 2011, **94**(9):4700-4707.

15. Karoui S, Carabaño MJ, Díaz C, Legarra A: **Joint genomic evaluation of French dairy cattle breeds using multiple-trait models**. *Genet Sel Evol* 2012, **44**:39.

16. de Roos APW, Hayes BJ, Spelman RJ, Goddard ME: **Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle**. *Genetics* 2008, **179**(3):1503-1512.

17. Zhong S, Dekkers JCM, Fernando RL, Jannink JL: **Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study**. *Genetics* 2009, **182**(1):355-364.

18. Thaller G, Krämer W, Winter A, Kaupe B, Erhardt G, Fries R: **Effects of variants on milk production traits in German cattle breeds**. *J Anim Sci* 2003, **81**(8):1911-1918.

19. Spelman RJ, Ford CA, McElhinney P, Gregory GC, Snell RG: **Characterization of the DGAT1 gene in the New Zealand dairy population**. *J Dairy Sci* 2002, **85**(12):3514-3517.

20. VanRaden PM, Olson KM, Wiggans GR, Cole JB, Tooker ME: **Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss**. *J Dairy Sci* 2011, **94**(11):5673-5682.

21. Hayes BJ, Bowman PJ, Chamberlain AJ, Verbyla K, Goddard ME: **Accuracy of genomic breeding values in multi-breed dairy cattle populations**. *Genet Sel Evol* 2009, **41**(1):51.

22. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA: **The impact of genetic architecture on genome-wide evaluation methods**. *Genetics* 2010, **185**(3):1021-1031.

23. Wientjes YCJ, Veerkamp RF, Bijma P, Bovenhuis H, Schrooten C, Calus MPL: **Empirical and deterministic accuracies of across-population genomic prediction**. *Genet Sel Evol* 2015, **47**:5.

24. Pryce JE, Johnston J, Hayes BJ, Sahana G, Weigel KA, McParland S, Spurlock D, Krattenmacher N, Spelman RJ, Wall E: **Imputation of genotypes from low density (50,000 markers) to high density (700,000 markers) of cows from research herds in Europe, North America, and Australasia using 2 reference populations**. *J Dairy Sci* 2014, **97**(3):1799-1811.

25. McKay SD, Schnabel RD, Murdoch BM, Matukumalli LK, Aerts J, Coppieters W, Crews D, Neto ED, Gill CA, Gao C: **Whole genome linkage disequilibrium maps in cattle**. *BMC genet* 2007, **8**(1):74.

26. Khatkar MS, Nicholas FW, Collins AR, Zenger KR, Cavanagh JAL, Barris W, Schnabel RD, Taylor JF, Raadsma HW: **Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel**. *BMC genom* 2008, **9**(1):187.

27. Falconer DS, Mackay TFC: **Introduction to quantitative genetics**. Harlow: Longman; 1996.

28. Verbyla KL, Hayes BJ, Bowman PJ, Goddard ME: **Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle**. *Genet Res* 2009, **91**(05):307-311.

29. Calus MPL, Huang H, Vereijken A, Visscher J, ten Napel J, Windig J: **Genomic prediction based on data from three layer lines: a comparison between linear methods**. *Genet Sel Evol* 2014, **46**(1):57.

30. Chen L, Li C, Miller S, Schenkel F: **Multi-population genomic prediction using a multi-task Bayesian learning model**. *BMC Genet* 2014, **15**(1):53.

31. Caruana R: **Multitask learning**. *Mach Learn* 1997, **28**(1):41-75.

32. Coster A, Bastiaansen JWM, Calus MPL, van Arendonk JAM, Bovenhuis H: **Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance**. *Genet Sel Evol* 2010, **42**(9).

33. Clark SA, Hickey JM, van der Werf JHJ: **Different models of genetic variation and their effect on genomic evaluation**. *Genet Sel Evol* 2011, **43**(18).

34. Bastiaansen JWM, Coster A, Calus MPL, van Arendonk JAM, Bovenhuis H: **Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures**. *Genet Sel Evol* 2012, **44**(3).

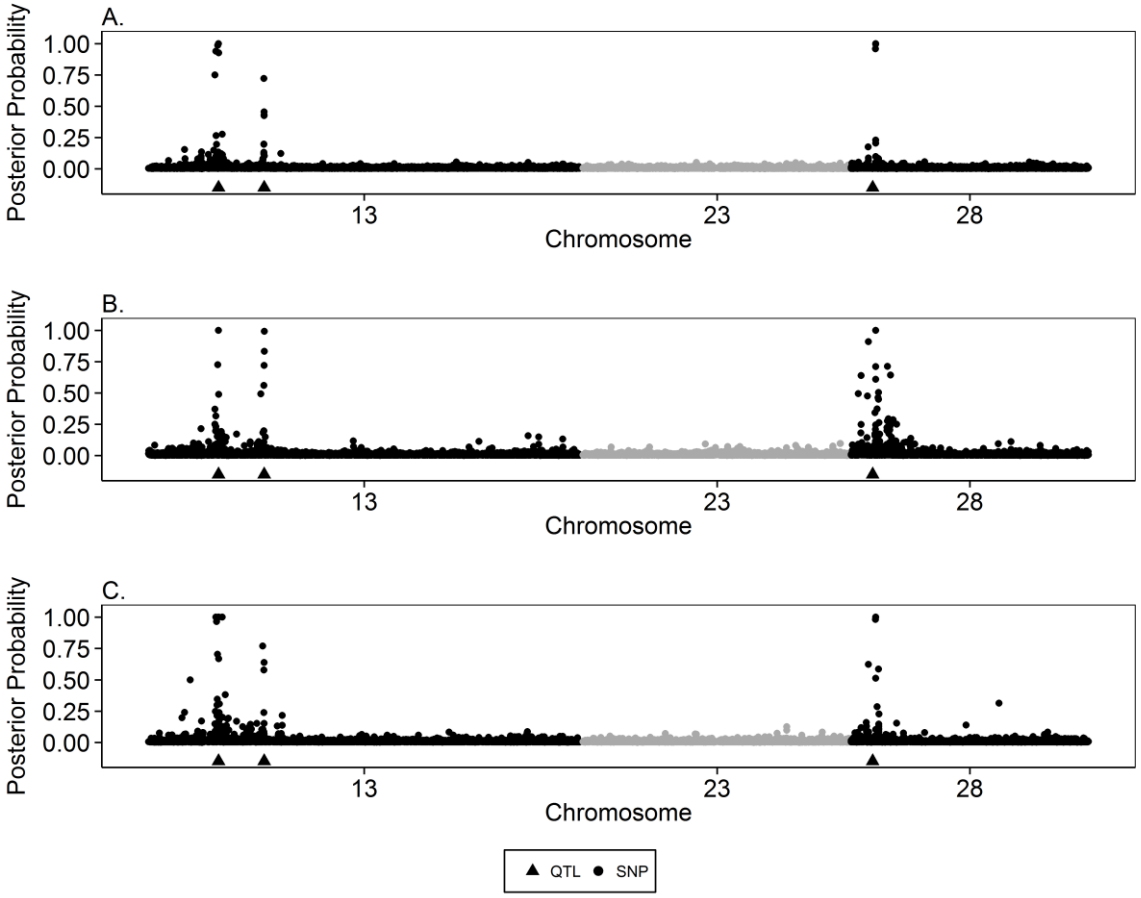
35. Zhou L, Heringstad B, Su G, Guldbrandtsen B, Meuwissen THE, Svendsen M, Grove H, Nielsen US, Lund MS: **Genomic predictions based on a joint reference population for the Nordic Red cattle breeds**. *J Dairy Sci* 2014, **97**(7):4485-4496.

36. Bolormaa S, Pryce JE, Kemper KE, Savin K, Hayes BJ, Barendse W, Zhang Y, Reich CM, Mason BA, Bunch RJ: **Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in, and composite beef cattle**. *J Anim Sci* 2013, **91**(7):3088-3104.

37. Goddard ME, Hayes BJ, Meuwissen THE: **Using the genomic relationship matrix to predict the accuracy of genomic selection**. *J Anim Breed Genet* 2011, **128**(6):409-421.

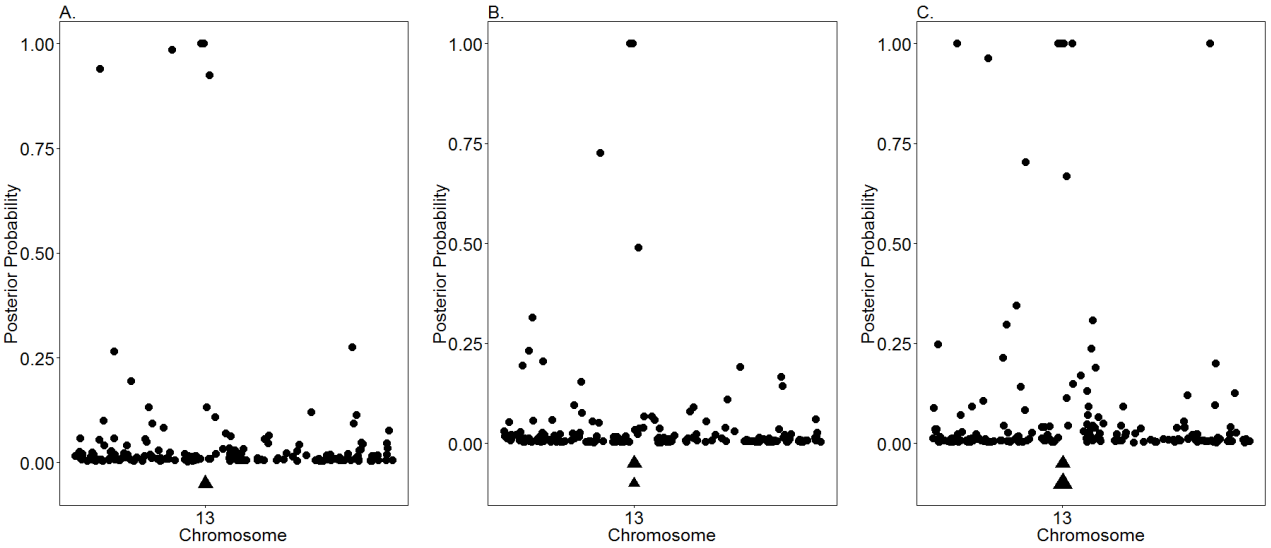
Supplementary files

S1. MANHATTAN PLOT FOR THE REPLICATE 1WITH GENETIC CORRELATION OF 0.4

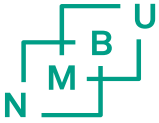


DISTRIBUTION OF THE POSTERIOR PROBABILITY OF 26,503 SNPS ACROSS CHROMOSOME 13, 23 AND 28 ASSUMING A GENETIC CORRELATION OF 0.4 ACROSS POPULATIONS USING THREE DIFFERENT REFERENCE POPULATIONS; REFERENCE 1 **(A)**: HF; REFERENCE 2 **(B)**: HF & GWH; REFERENCE 3 **(C)**: HF & MRY. THE TRIANGLES INDICATES THE POSITIONS OF THE THREE QTL.

S2. MANHATTAN PLOT FOR THE REPLICATE 1 WITH GENETIC CORRELATION 0.4 ZOOMED IN ON A 200 SNP REGION NEIGHBOURING THE FIRST QTL



DISTRIBUTION OF THE POSTERIOR PROBABILITY OF 100 SNPS BEFORE AND AFTER THE POSITION OF EACH QTL, WHICH IS INDICATED BY THE TRIANGLE, USING THREE DIFFERENT REFERENCE POPULATIONS ASSUMING A GENETIC CORRELATION OF 0.4 ACROSS POPULATIONS; REFERENCE 1 **(A)**: HF; REFERENCE 2 **(B)**: HF & GWH; REFERENCE 3 **(C)**: HF & MRY.



Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no