Genomic Predictions Using Whole Genome Sequence Data and Multi-breed Reference Populations

O.O.M. Iheshiulor[1], J.A. Woolliams[1,2], X. Yu[1], R. Wellmann[3] and T.H.E. Meuwissen[1].
[1]Norwegian University of Life Sciences, Ås, [2]The Roslin Institute and R(D)SVS, University of Edinburgh, Midlothian, United Kingdom, [3]Institute of Animal Husbandry and Breeding, University of Hohenheim, Germany.

ABSTRACT: The availability of whole-genome sequence data (WGS data) on large number of livestock's provides new opportunity for genomic selection. We investigated how much accuracy is gained by using WGS data in diverged cattle populations, using simulation. Relative performance of genomic BLUP and a Bayesian (BayesP) method with a mixture of normal distributions were compared. WGS data increased accuracy (3-7%) of within population predictions for moderate – lowly heritable traits. The advantage of WGS data (18-24%) was more pronounced with reference populations (RP) combined across breeds and when using BayesP. Extending the RP to multiple-breeds resulted in a 10-22% increase in accuracy with WGS data. BayesP outperformed GBLUP at 45 QTL/M, although in real data both methods have been shown to perform quite similar. Genomic predictions in numerically minor cattle populations would benefit from a combination of WGS data, multi-breed RP, and Bayesian estimation methods.
Keywords: genomic prediction; whole-genome sequence; multi-breed

Introduction

Genomic selection (GS) is today implemented in most livestock populations especially that of dairy cattle. Also different commercial panel SNP densities ranging from low to high are available with 777k being the largest in the dairy sector. Dairy cattle populations such as Holstein have benefited a lot from GS due to its large reference population (RP) and low effective population size (Ne). On the other hand, the impact of GS on numerically small breeds with large Ne (e.g. Norwegian Red) with special emphasis on functional traits with low heritability's is yet to be maximized even when 54K and or 777K (HD) SNP panels are used (Solberg et al., 2011; Su et al., 2012). The latter studies compared the use of HD to 50K SNP chips and reported small or no gain in reliability. Still, the use of whole-genome sequence data (WGS data) for genomic predictions in these populations might be a way to improve accuracy.

WGS data differs fundamentally from current dense SNP-chip data in that it includes all causal polymorphisms. With all causal polymorphism captured, WGS data could provide better signals for causative mutations within and across families and predictions would no longer have to rely on LD between SNP and QTL. The size of the RP plays a key role in the accuracy of genomic predictions and presently, aggregation of RP is used as a means to increase the RP sizes. Across breed predictions using SNP chips is heavily dependent on SNP-QTL association, however with WGS data there would be no or at least less need to rely on such association which may not persist across breeds.

Using simulated data, Meuwissen and Goddard (2010) demonstrated an advantage for WGS data over the densest SNP chip they simulated, but didn't look into a situation of diverged small populations and across breed predictions. The objective of this study was to ascertain how much accuracy is gained by using whole-genome sequence data as compared to different SNP densities with emphasis on diverged breeds of small populations with large effective population sizes (>100). Also compared was the relative performance of linear (GBLUP) and non-linear (BayesP; Yu and Meuwissen, 2011) methods.

Materials and Methods

Simulation of whole-genome sequence data: A forward simulation approach was used in the simulation of whole-genome sequence of 1 chromosome with a length of 100 cM. The Fisher-Wright idealized population model was assumed (Falconer and Mackay, 1996) with a mutation rate of 10^{-8}/bp/meiosis, assuming 1 000 000 base pairs per cM. Historical effective population size was Ne = 200 and the forward simulation were conducted for 1,990 generations in-order to create a steady-state population. The mating system was based on random union of gametes. Therefore mutation and drift were the only two evolutionary forces considered. Recombinations were sampled according to Haldane mapping function. After 1,990 generations, the population was split into two to represent diverged breeding populations (i.e. Population A and Population B). Each of the populations was further simulated for 10 generations using the same effective population size mentioned above but the number of individuals in the last generation (i.e. generation 2000) which is the generation of interest was increased to 500. Of the 500 individuals from generation 2000, 200 were randomly sampled and designated as reference population (RP) while the remaining 300 individuals were designated as validation.

The mutation-drift process resulted on average in 4,565 polymorphism evenly distributed across a chromosome of 100 cM with a minor allele frequency (MAF) >0.02 and the standard deviation of this number was 129. We refer to all these polymorphism as SNPs. Out of the SNPs, 45 were randomly sampled and designated as QTL (Causative SNPs) leaving 4,520 markers. All polymorphism including causative SNPs represents WGS data. By randomly sampling (without replacement) the full marker data, dataset 3000 and dataset 2000 were created and each contained evenly distributed 3,000 and 2,000 SNPs respectively. The marker data represents different marker density panels and would be equivalent to 90K and 60K SNP chip for a 30 Morgan cattle genome. The above simulation was repeated 30 times.

Genetic and phenotypic values: Two traits with heritability of 0.30 and 0.07 were simulated. Assuming the additive genetic model, allelic effect (a_j') were assigned to reference allele (allele "1") of every QTL by sampling effects from the normal distribution. After sampling, their effects were standardized (i.e. $a_j = a_j'/\sqrt{\sum_i 2p_i(1-p_i)(a_i')^2}$, where subscripts *i* and *j* denote the *i* and *j* QTL; summation is over all QTL; p_i is the frequency of allele "1" of the *i*th QTL). Then total genetic value was calculated as $g_i = \sum_{j=1}^{N_{QTL}} x_{ij}a_j$

Where x_{ij} is the number of alleles "1" that individual *i* carries at locus *j*. Phenotypes were generated by adding an environmental effects drawn from the normal distribution with mean equal to zero. The variance of the environmental effects was chosen such that the heritability for the traits was 0.30 and 0.07 respectively.

Estimation methods and data analysis: Firstly, genomic predictions were within population A (i.e. RP comprised of 200 individuals from population A while 300 individuals from the same population were used for validation). Secondly, predictions were based on a multi-breed RP and validation remained in breed A. In this case, equal number of individuals (200) from both populations were used to set up the RP (totaling 400 individuals) while 300 individuals from population A were used for validation. The second situation would occur in practice when the breeders of population A were able to combine their RP with those of the breeders of population B.

Genomic best linear unbiased prediction (GBLUP) and a Bayesian approach (BayesP; Yu and Meuwissen, 2011) were used to estimate SNP effects in the RP. GBLUP estimates SNP effects by best linear unbiased prediction assuming that every SNP explains an equal proportion of the total genetic variance (Meuwissen et al., 2001). BayesP assumes that SNP effects come from a mixture of normal distributions. SNPs with small effects are assumed to have effects sampled from a normal distribution with a small variance (σ_1^2) while the SNPs with big effect are assumed sampled from a normal distribution with big variance (σ_2^2). The distribution of the total genetic variance (Vg) over the 'big' SNPs and the 'small' SNPs is according to the Pareto principle (hence the P in BayesP) i.e. x% of the SNPs with the largest effects cause $(100-x)$% of the genetic variance. Given that prior ($\pi = x/100$) and using the Pareto principle, the variances of the large and small SNP effects are respectively: $\left.\begin{matrix}\sigma_1^2 = \frac{(1-\pi)Vg}{\pi M} \\ \sigma_2^2 = \frac{\pi Vg}{(1-\pi)M}\end{matrix}\right\}$

Where M is the number of SNP, such that $M(\pi\sigma_1^2 + (1-\pi)\sigma_2^2) = V_g$.

The model used by GBLUP and BayesP to estimate SNP effects was: $y = \mu + \sum_{j=1}^{N_m} X_j b_j + e$

Where **y** is a Nx1 vector of phenotypes; μ is overall mean; N_m is total number of genotyped SNPs; X_j is a Nx1 vector of the N standardized SNP genotypes, i.e. $X_j = \frac{-2p_j}{\sqrt{2p_j(1-p_j)}}, \frac{1-2p_j}{\sqrt{2p_j(1-p_j)}}, or\ \frac{2(1-p_j)}{\sqrt{2p_j(1-p_j)}}$ for SNP genotype "0 0", "1 0", or "1 1", respectively and p_j is the allele frequency of SNP *j*; b_j is the effect of the jth SNP genotype; and **e** is a Nx1 vector of environmental effects.

After estimation of SNP effects, genetic value ($\hat{g}$, GEBV) for the validation individuals (i.e. the individuals having only genotypic records) was predicted as $\hat{g}_i = \sum_{j=1}^{N_m} X_{ij}\hat{b}_j$

Where X_{ij} is the standardized marker genotype of individual *i* for SNP *j*; and $\hat{b}_j$ is the estimated marker effect. The correlation between total genetic value (g_i) and estimated genetic value ($\hat{g}_i$) was used as a measure of the accuracy of the genomic prediction.

Results and Discussion

Genomic predictions within population A: Accuracy of genomic predictions for population A based on WGS data, 2 different marker densities and 2 different heritability (h^2) using GBLUP and BayesP methods are shown in Table 1. WGS data resulted in 3-7% increase in accuracy over the different marker densities across the different traits/h^2. Accuracy increased with higher h^2 and marker density. Table 1 also shows the relative performance of both methods when used on the different datasets. Higher accuracies were observed when using BayesP. Using BayesP for WGS data increased accuracy by 6.0% and 4.4% as compared to GBLUP for the different h^2. Our results follows the same upward trend in accuracy as in Meuwissen and Goddard (2010). The higher accuracy obtained with WGS data could be attributed to not having to rely on LD between markers and the QTL. According to (Meuwissen et al., 2001), GS depends on the number of genetic marker as well as LD between SNPs and QTL in-order to maximize the proportion of genetic variance explained by the SNPs. However, with WGS data, predictions are no longer dependent on SNP-QTL association because all polymorphism are captured and utilized in the analysis (de Roos, 2011). Meuwissen and Goddard (2010) showed that even at higher marker density, an extra gain in accuracy was obtained by inclusion of the causative mutations. Furthermore, WGS data contains all base pairs, therefore no rare alleles would be missing (Daetwyler et al., 2010). Thus no effect is left un-captured be it small or large when using WGS data.

Table 1. Accuracy of genomic predictions (se) in population A obtained with GBLUP and BayesP using WGS data and different marker densities

	GBLUP		BAYESP	
	Trait 1 ($h^2 = 0.30$)			
Dataset	$r_{TBV;GEBV}$	% decrease	$r_{TBV;GEBV}$	% decrease
WGS data	0.596 (0.015)		0.632 (0.018)	
data 3000	0.578 (0.016)	3.0	0.598 (0.018)	5.4
data 2000	0.576 (0.014)	3.4	0.591 (0.015)	6.5
	Trait 2 ($h^2 = 0.07$)			
	$r_{TBV;GEBV}$	% decrease	$r_{TBV;GEBV}$	% decrease
WGS data	0.413 (0.024)		0.431 (0.028)	
data 3000	0.400 (0.023)	3.1	0.407 (0.025)	5.6
data 2000	0.394 (0.021)	4.6	0.404 (0.023)	6.3

Multi-breed genomic predictions: Accuracies of multi-breed genomic predictions using the different datasets and the two methods (GBLUP and BayesP) in a situation where populations have diverged for 10 generations are presented in Table 2. The data type, heritability and method used, all had effect on the accuracy of predictions. Inclusion of 200 individuals from population B to the RP of population A led to higher accuracy of genomic breeding values in population A as compared to within population predictions when using WGS data whilst minimal or no increase in accuracy was observed for the lower marker densities (Table 2). Highest accuracy of 0.710 and 0.525 respectively was observed for $h^2 = 0.30$ and $h^2 = 0.07$, when using BayesP. The higher accuracies obtained when using a multi-breed RP with WGS data is the result of an increased number of reference animals, not having to rely on LD between SNP and QTL, and the presence of the causative mutation for marker effect estimation. With WGS data, all polymorphism are captured and by combining the different breeds/populations to make-up the RP, a high possibility exist for identifying QTLs having an allelic phase that is preserved across the populations (Hayes et al., 2009), hence the increased accuracy. According to de Roos, (2011), the maximum benefit of WGS data is obtained if the number of reference individuals is increased accordingly. Meuwissen, (2009) also reported that large training datasets are needed in order to take full advantage of high density markers. Our study shows that if all genetic variance in a population is captured and given large RP across breeds, high accuracies can be realized also for lowly heritable traits.

Table 2. Genomic predictions (Se) using multi-breed reference population and single-breed validation when populations are diverged for 10 generations

	GBLUP		BAYESP	
	Trait 1 ($h^2 = 0.30$)			
Dataset	$r_{TBV;GEBV}$	% decrease	$r_{TBV;GEBV}$	% decrease
WGS data	0.654 (0.013)		0.710 (0.015)	
data 3000	0.578 (0.016)	11.6	0.602 (0.018)	15.2
data 2000	0.571 (0.015)	12.7	0.574 (0.016)	19.2
	Trait 2 ($h^2 = 0.07$)			
	$r_{TBV;GEBV}$	% decrease	$r_{TBV;GEBV}$	% decrease
WGS data	0.475 (0.021)		0.525 (0.025)	
data 3000	0.404 (0.023)	14.9	0.414 (0.024)	21.1
data 2000	0.400 (0.022)	15.8	0.414 (0.025)	21.1

GBLUP vs BayesP: The results show that BayesP outperformed GBLUP at 45 QTL/M and also gave higher accuracies with WGS data. However, as QTL density increased (132 QTL/M), accuracy decreased more with BayesP (results not shown). That notwithstanding, BayesP remained superior.

The results presented here assumed 1 chromosome, 45 QTLs, and 4,565 variants, whereas WGS data in cattle would be on 30 chromosomes, containing millions of variants and 1000s of QTLs. We anticipate that with millions of variants LD amongst SNPs will be high and variable selection methods will have difficulties in pinpointing the QTL, but when a SNP with very high LD to the QTL is receiving the QTL effect, predictions will not change much. If that is the case, then no dramatic change is expected in our results given the ideal situation of WGS data. Druet et al. (2014) found smaller difference in accuracy when comparing WGS with SNP panel. The difference in results might be due to the fact that the latter study used a much denser SNP panel and their accuracies were already around 0.9 leaving little room for improvement.

Conclusion

The results of our study suggest that WGS data would lead to increased accuracy of genomic breeding values for moderate – lowly heritable traits in numerically minor dairy populations when RP are combined across breeds. This increase is more pronounced when WGS data is used under multi-breed RP with Bayesian variable selection method such as BayesP. In generally, BayesP yielded higher predictive ability as compared to GBLUP. To take full advantage of WGS data, a large RP and Bayesian variable selections methods are required.

Acknowledgement

This work has been funded by Gene2Farm EU FP7 project (Development of next generation European system for cattle evaluation).

Literature Cited

Daetwyler, H. D., Pong-Wong, R., Villanueva, B, et al. (2010). Genetics 185:1021-1031.

de Roos, A. P. W., Hayes, B. J., and Goddard, M. E. (2009). Genetics 183:1545-1553.

de Roos S., 2011. Animal Breeding and Genomics Centre, Wageningen University, The Netherlands.

Druet, T., Macleod, I. M., and Hayes, B. J. (2014). Heredity 122: 39-47.

Falconer, D. S., and Mackay. T. F. C. (1996). 4th ed. John Wiley and Sons, New York.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J. et al. (2009). J. Dairy Sci. 92: 433-443.

Meuwissen, T. H. E, and Goddard. M. E. (2010). Genetics 185: 623-631.

Meuwissen, T. H. E. (2009). GSE 41: 35-43.

Meuwissen, T. H. E., Hayes, B. J. and Goddard. M. E. (2001). Genetics 157: 1819-1829.

Solberg, T. R., Heringstad, B., Svendsen, M. et al. (2011). INTERBULL BULLETIN 44: 240-243.

Kachman, S. D., Spangler, M. L., Bennett, G. L. et al. (2013). GSE 45:30-38.

Su, G., Brøndum, R. F., Ma, P. et al. (2012). J. Dairy Sci. 95: 4657-4665.

Yu, X., and Meuwissen, T. H. E. (2011). GSE 43: 35-41.