Proc. Assoc. Advmt. Anim. Breed. Genet. 20:215-219

WILL SEQUENCE SNP DATA IMPROVE THE ACCURACY OF GENOMIC PREDICTION IN THE PRESENCE OF LONG TERM SELECTION?

I.M. MacLeod[1,2], B.J. Hayes[2,3] and M.E. Goddard[1,2]

[1] Melbourne School of Land & Environment, University of Melbourne, Victoria 3010, Australia
[2] AgriBio, Department of Environment & Primary Industries, Victoria 3086, Australia
[3] Biosciences Research Centre, La Trobe University, Victoria 3086, Australia

SUMMARY

To date genomic prediction (GP) of breeding values in cattle generally exploits either ~50K or ~800K SNP chips. Now that whole genome sequence data is also available, it is important to evaluate its potential to improve the accuracy of GP. SNP chips include only more common SNP while sequence data includes rare and common SNP as well as all causal mutations (QTL). It is expected that sequence data will improve accuracy of GP particularly if QTL are rare because they have been under long-term negative selection. This study evaluates accuracy of GP using sequence data compared with the equivalent of ~800K or ~50K SNP densities. Accuracy of GP was tested in simulated populations (mimicking Holstein cattle) with and without long-term negative selection acting on QTL. GP was implemented with both BLUP (GBLUP) and Bayesian (BayesR) methods. There was not a very marked difference between GP accuracy in scenarios with neutral QTL or selected QTL because the recent low effective population size (N_e) of cattle decreased the proportion of rare causal mutations compared to expectations in larger N_e. Only the BayesR method was able to exploit an advantage from sequence data. We conclude that combining data from more than one breed in training (reference) populations and using Bayesian analyses, will take better advantage of sequence data for GP than using single breed and GBLUP analyses.

INTRODUCTION

Genomic prediction (GP) of breeding values is an efficient method of selecting livestock for traits that are difficult to measure, or traits not expressed in males (Meuwissen *et al.* 2001). To date GP in cattle generally exploits either ~50K or ~800K SNP chips, but soon whole genome sequence data (direct or imputed) could also be used to improve accuracy of GP. The advantage of sequence is that it contains the causal mutations. Furthermore, SNP chips include only common SNP and these may not be in high linkage disequilibrium with causal mutations if the latter are rare because they have been subject to long-term negative selection. In this case SNP chips will not be able to accurately estimate the QTL effects. It is therefore expected that sequence data will improve accuracy of GP, particularly if causal mutations have been under long-term negative selection.

Using a bovine-like neutral model to simulate data, Clark *et al.* (2011) demonstrated a 5-15% advantage for accuracy of GP using sequence compared to 50K SNP chip densities, but did not include a comparison with 800K SNP density. Druet *et al.* (2013) indirectly estimated the potential effect of long term negative selection on GP by simulating QTL effects on a subset of loci with low or very low minor allele frequencies (MAF). They demonstrated a 4-28% advantage in accuracy of GP using sequence data compared to 50K SNP densities, but did not test 800K SNP density. Although it can be argued that simulating QTL on rare mutations mimics the expected effect of long term negative selection, the approach may not reflect the true MAF distribution of loci actually subjected to long-term negative selection because demography also shapes the MAF distribution. For example, in populations with recent bottlenecks in effective population size (N_e), mutations with a deleterious effect on fitness are more likely to be lost, but may also sometimes rise to higher frequencies due to drift, compared to populations with large or expanding N_e. Using

simulations of bovine populations, we evaluate the accuracy of GP using sequence data, ~800K or ~50K SNP chip densities, with and without long term negative selection applied to QTL.

MATERIALS AND METHODS

We simulated sequence data with FREGENE (Chadeau-Hyam *et al.* 2008) using a demographic model demonstrated to mimic Holstein sequence data (Macleod *et al*. 2013), in which the effective population size (N_e) reduces from ancestrally very large to very small in recent times. For computational efficiency we simulated a genome size of 50Mb under the scaling argument demonstrated by Meuwissen & Goddard (2010): i.e. GP accuracy is proportional to the number of training individuals/Morgan (M) length of the genome. Therefore to achieve similar accuracies with a ~30M bovine genome, the training population size would need to increase by a factor of ~60. Simulations were either a neutral model (NEUT) or with long-term negative selection imposed on QTL (SEL). In the SEL model, 0.1% of new mutations were subject to an additive selection coefficient: $s = -2x10^{-4}$, and those still segregating at the end of the simulation were used as QTL. In both NEUT and SEL scenarios we simulated 20 replicates, each with 5000 individuals.

We created a "Medium Density" (MD) and "High Density" (HD) SNP panel for each replicate, by selecting a subset of 1000 and 10,000 SNP loci respectively: representing a density of 60K and 600K SNP across the whole bovine genome (the latter is equivalent to an 800K SNP panel because often after quality control in real data there are ~600 usable SNP). To mimic the ascertainment bias of commercial panels, SNP were only selected if MAF > 0.1 and SNP positions were then selected uniformly at random. We generated HD and MD SNP genotypes for all individuals in addition to the sequence data (SEQ). For each replicate, additive QTL effects were simulated from a normal distribution with two different QTL densities: number of QTL=50 or 15. In the NEUT populations, QTL were randomly selected from SNP loci, while in SEL scenarios the QTL were chosen from polymorphic loci subjected to selection. In 5 of the 20 replicate SEL populations, there were only 49, 47, 46, 46 and 41 selected loci still segregating, therefore for the scenario with QTL=50 the remaining QTL were drawn from neutral loci with MAF < 0.1. QTL effects were summed to give True Breeding Values (TBV_j) for each individual. Phenotypes were generated by adding a residual term to the TBV_j of each individual, drawn from a normal distribution to produce a trait heritability of 0.1. We randomly selected 3750 "training" individuals to calculate the genomic prediction equations (using genotypes and phenotypes). We used the remaining 1250 individuals from the same population (genotypes only) to validate the prediction equations (Gen=0, "validation" individuals). After both 10 and 15 further generations of random breeding, genotypes were again sampled for 2000 validation individuals (Gen=10 and Gen=15 validations).

We implemented both GBLUP and BayesR analyses to generate Genomic Estimated Breeding Values (GEBV). GBLUP was implemented in ASReml (Gilmour *et al.* 2005): $\mathbf{y} = \mu\mathbf{1} + \boldsymbol{Zg} + \boldsymbol{e}$, where μ is the population mean, $\mathbf{1}$ is a vector of 1s, $\boldsymbol{Z}$ is the incidence matrix for random individual effects. The $\boldsymbol{g}$ and $\boldsymbol{e}$ are vectors of GEBV and residuals, assumed normally distributed as $N(0, \boldsymbol{G}\sigma^2_g)$ and $N(0, \boldsymbol{I}\sigma^2_e)$, where $\boldsymbol{G}$ is the genomic relationship matrix (GRM) estimated either from MD, HD or SEQ genotypes (eg. Erbe *et al.* 2010). Our BayesR implementation (Erbe *et al.* 2012) omitted a polygenic effect because individuals were randomly bred with no close pedigree structure: $\mathbf{y} = \mu\mathbf{1} + \boldsymbol{Wu} + \boldsymbol{e}$, where μ is the mean, $\boldsymbol{e}$ is the vector of random residuals and $\boldsymbol{W}$ is the design matrix allocating records to the vector of marker effects, $\boldsymbol{u}$. The accuracy of GP was determined as the correlation between the $GEBV_j$ and the TBV_j in i=1…N validation individuals, averaged the across the 20 replicate simulations for each scenario.

RESULTS AND DISCUSSION

The marked reduction in recent effective population size (N_e) used in our simulation to mimic

the Holstein breed demography, resulted in a relatively flat derived allele frequency (DAF) distribution for neutral alleles compared to the expectation in a larger constant or expanding N_e. The recent reduction in N_e results in random drift very quickly purging low frequency loci as well as increasing linkage disequilibrium (LD) compared to larger N_e. Among neutral loci in our simulations, 19% had DAF < 0.1 while this figure increased to 31% for loci subjected to long term negative selection. This indicates that selection had a significant impact on allele frequency distribution while not being so strong as to immediately purge new mutations. The impact of the selection coefficient (s) is generally significant if: $|sN_e| >> 1$ and in our large ancestral bovine population $|sN_e| = 12$ which is similar to some estimates in humans (Keightley & Halligan 2009).

Fig 1A shows the results for the realised accuracy of GP when the number of QTL=50 (equivalent to ~3000 QTL affecting a trait genome wide) while results in Fig 1B are for QTL=15 (equivalent to 900 QTL genome wide). QTL densities were chosen to reflect realistic models based on recent mammalian estimates (eg. Kemper *et al*. 2012). In all scenarios there was an advantage for sequence (SEQ) over MD SNP (up to 11.8%) as for previous studies (Clark *et al.* 2011, Meuwissen & Goddard 2010, Druet *et al*. 2013), particularly with BayesR and an increasing number of generations separating training and validation populations (Gen=10 and 15). With GBLUP analyses there was generally no advantage for SEQ compared to HD SNP, except in the SEL scenario with QTL=15. With BayesR there was a modest advantage for SEQ over HD SNP (up to 3.6%), particularly in the Gen=15 validation and was consistently higher in SEL compared to NEUT scenarios. Furthermore, there was less decay in the BayesR accuracy compared to GBLUP when the number of generations separating training and validation individuals increased.

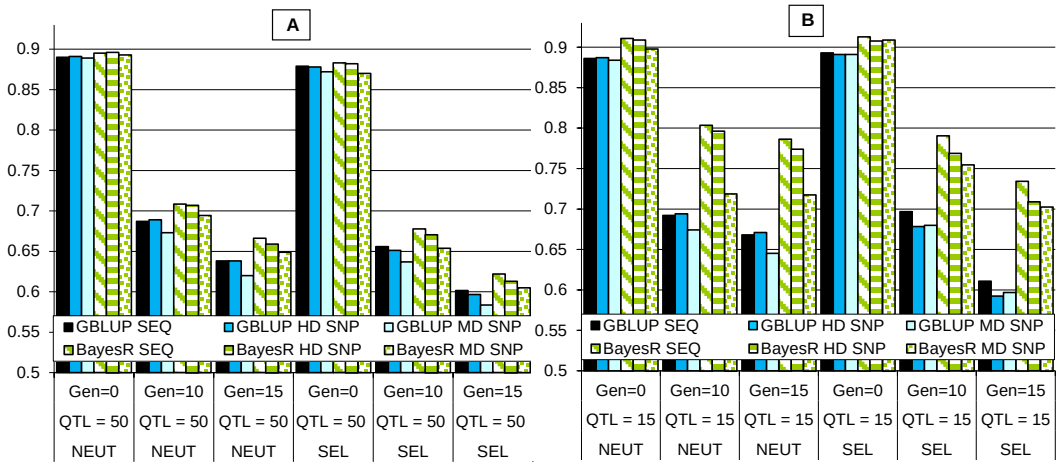

Figure 1A and B. **Genomic prediction accuracy in populations with QTL under a neutral (NEUT) or negative selection model (SEL), using GBLUP or BayesR analysis, with two contrasting QTL densities: number of QTL=50 (A) or QTL=15 (B). Zero, 10 or 15 generations separated training and validation individuals (Gen=0, 10 or 15).**

GBLUP assumes a quasi infinitesimal model with each sequence SNP assumed to contribute an additive effect sampled from a single normal distribution. BayesR method could be expected to perform better with sequence data because it sets a prior expectation that many SNP will have no effect, while the remaining effects will be sampled from a mixture of distributions, with many small effects and up to some rare large effects. However the recent reduction in N_e within *Bos taurus* cattle breeds has resulted in high (but variable) LD across relatively long chromosome

segments and therefore GBLUP will tend to "spread" the estimate of each QTL effect across a number of loci on chromosome segments in which SNP are in high LD. We estimated the number of "effectively independent chromosome segments" (M_e , see Goddard 2009) is ≈ 85 on our 50 Mb genome. Therefore, when the number of QTL=50, GBLUP works as well as BayesR in Gen=0, because nearly all segments contain a QTL and so the prior assumption that chromosome segment effects are normally distributed is approximately correct. Also, when animals are relatively closely related (Gen=0) there was no advantage for SEQ because HD and MD SNP are dense enough to predict the QTL effects given the low M_e.

Although BayesR analysis estimates an effect for each SNP with many set to zero, the method still has difficulty defining which SNP within a segment of high LD is the true QTL, and several SNP effects are estimated as contributing to part of the QTL effect, particularly with dense SNP. With QTL=15 the BayesR method showed an advantage over GBLUP even in Gen=0, and in all scenarios the advantage of BayesR becomes more pronounced in Gen=10 and 15. This implies that even with many SNP in high LD, BayesR is superior to GBLUP in accurately attributing SNP effects to a more precise chromosome region harbouring the real QTL. Recombination is therefore less likely to occur between the true QTL and the SNP to which BayesR has attributed part of the QTL effect and accuracy of GP is more persistent across generations. The decay in accuracy is more rapid with GBLUP than BayesR because more SNP effects over longer segments are contributing to predicting the individual QTL effects and therefore there is a much higher chance that recombination will disrupt the LD between QTL and SNP alleles.

Druet *et al.* (2013) tried to indirectly estimate the effect of negative selection on accuracy of GP by simulating QTL only on loci with MAF<0.1 compared to their neutral model allocating QTL randomly across all loci. They observed ~10% reduction in SEQ accuracy of GP with BayesR when QTL MAF<0.1. However, our simulation demonstrates that the MAF distribution of QTL subjected to long term negative selection is unlikely to be as extreme as assumed in Druet *et al.* (2013). There was a consistent reduction in the accuracy of GP due to the effect of selection, but only when there were 10 or more generations separating the training and validation populations. If a gamma distribution of QTL effects had been used in this study, the difference between BayesR and GBLUP accuracies might have been more pronounced, particularly when the number of QTL=15 because this is closer to BayesR assumed distribution of QTL effects. However no further differences in the results would be expected.

To gain more advantage from sequence, we conclude that training data should be combined from more than one breed to reduce the LD between more distant SNP (equivalent to an increase in the N_e). This will also require an increase in the size of training populations but should ensure better persistency of GP accuracy across generations with SEQ, provided that a reasonable proportion of QTL are segregating in both breeds. It is also likely to be more beneficial to use a Bayesian analyses and to select a subset of potentially more biologically active SNP from sequence data prior to analysis.

REFERENCES.

Chadeau-Hyam M., Hoggart C., O'Reilly P., Whittaker J., De Iorio M. and Balding D. (2008) *BMC Bioinformatics* **9**: 364.

Clark S., Hickey J. and van der Werf J. (2011) *Genet. Sel. Evol.* **43**: 18.

Druet T., MacLeod I.M. and Hayes B.J. (2013) *Heredity*. http://dx.doi.org/10.1038/hdy.2013.13

Erbe M., Hayes B.J., Matukumalli L.K, Goswami S., Bowman P.J., Reich C.M., Mason B.A. and Goddard M.E. (2012) *J. Dairy Sci.* **95**: 4114.

Goddard M.E. (2009) *Genetica* **136**: 245.

Gilmour A.R., Cullis B.R., Gogel B.J., Welham S.J. and Thompson R. (2005) *ASReml User Guide Release 2.0*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Keightley P. & Halligan D. (2009) *Genetica* **136**: 359.
Kemper K.E., Visscher P.M. and Goddard M.E. *Genome Biology* 2012, 13:244.
Macleod I.M., Larkin D., Lewin H., Hayes B.J. and Goddard M.E. (2013) *Mol. Bio. Evol.* (Published ahead of print: doi:10.1093/molbev/mst125)
Meuwissen T.H.E., Hayes B.J. and Goddard M.E. (2001) *Genetics* **157**: 1819.
Meuwissen T.H.E. and Goddard M.E. (2010) *Genetics* **185:** 623.