Proc. Assoc. Advmt. Anim. Breed. Genet. 22:285-288

STRATEGIES TO USE WHOLE GENOME SEQUENCE DATA FOR GENOMIC PREDICTION IN DAIRY CATTLE

I. van den Berg[1], I.M. MacLeod[2], P.J. Bowman[2,3], T. Wang[2] and M.E. Goddard[1,2]

[1] Faculty of Veterinary & Agricultural Science, University of Melbourne, Victoria, Australia
[2] AgriBio, Department of Economic Development, Jobs, Transport & Resources, Victoria, Australia
[3] Biosciences Research Centre, La Trobe University, Victoria, Australia

SUMMARY

Sequence data may potentially increase prediction accuracy compared to medium or high density (HD) SNP markers, by containing causative mutations directly rather than relying on linkage disequilibrium between markers and causative mutations. Besides causative mutations, sequence data contains a much larger number of variants that have no effect on the analysed trait. A Bayesian variable selection model could be used to assign large effects only to the causative mutations. In practice, however, analysing millions of sequence variants is computationally challenging. Therefore, we tested an approach to split up the analysis per chromosome, correcting for all other chromosomes using HD estimates, in a simulation study, using a faster, hybrid version of the Bayes R variable selection model. While directly computing breeding values based on effects estimated per chromosome resulted in a reduced accuracy, reanalysing all variants that were selected per chromosome resulted in a similar accuracy to analysing all variants simultaneously, especially when HD variants were included.

INTRODUCTION

Sequence data can potentially increase prediction accuracy compared to medium or high density SNP markers, by containing causative mutations directly rather than relying on linkage disequilibrium (LD) between markers and causative mutations. However, the majority of sequence variants have no effect and can introduce noise into the prediction. In theory, Bayesian variable selection models could assign larger effects to the causative mutations, and zero effects to the rest. In practice, correctly estimating the effects of millions of variants in high LD with each other is computationally challenging, and the results of genomic prediction using sequence data so far have been variable. Using a Bayesian variable selection model to analyse all variants simultaneously, van Binsbergen *et al*. (2015) reported a slightly lower accuracy with full sequence data than with high density (HD) genotypes. Other approaches, using various methods to preselect variants, show sometimes an increase in accuracy (Brøndum *et al*. 2015; Macleod *et al*. 2016; van den Berg *et al*. 2016), while others found no increase in accuracy but increased bias (Calus *et al*., 2016; Veerkamp *et al*., 2016).

Our objective was to find an approach to approximate genomic prediction of whole genome sequence data with a Bayesian variable selection model. To parallelise the analysis, we tested analysing chromosomes separately after correcting the phenotypes for all other chromosomes using HD estimates. Results obtained per chromosome were either directly used to compute breeding values, or used to preselect variants for subsequent analysis with all chromosomes together. The analysis with all selected variants was performed either with or without the HD variants. A dataset with a limited number of realised imputed sequence variants and simulated phenotypes was used, to enable comparison with prediction using all sequence variants at once.

MATERIALS AND METHODS

The dataset used was the AUS-Sim simulated dataset described in more detail by Macleod *et*

al. (2016). The dataset contained realised genotypes for 3,047 Holstein bulls, 4,942 Holstein cows, 770 Jersey bulls, 1,553 Jersey cows, 869 Red Holstein bulls, 741 Australian Red cows and 114 Australian Red bulls. The data was split up in a reference population containing all Holstein and Jersey individuals, and a validation population containing all Australian Red and Red Holstein individuals. Pedigree information for all individuals was obtained from the Australian Dairy Herd Improvement Scheme (ADHIS) and Interbull.

Two sets of genotypes were used, the HD set containing genotypes for 600,641 SNP on the Illumina BovineHD beadChip, and the SEQ set, containing 994,019 imputed sequence variants selected based on their function annotations. The HD genotypes were either obtained by direct genotyping, or imputation from the Illumina BovineSNP50 chip. The SEQ set contained 45,026 non-synonymous coding variants, 578,734 variants within 5 Kb upstream and downstream of genes, or in three/five prime untranslated genic regions, and 370,259 variants on the HD chip.

Quantitative trait loci (QTL) were simulated by randomly sampling 4,000 variants from all SEQ variants. QTL effects were sampled from three normal distributions with a mean of zero and variances of $0.0001\,\sigma_g^2$, $0.001\,\sigma_g^2$ and $0.01\,\sigma_g^2$ for 3,485 small, 500 medium and 15 large QTL, respectively, where σ_g^2 is the additive genetic variance. Subsequently, the true breeding value (TBV) of individual *j* was computed as $TBV_j = \sum_{i=1}^{4000} x_{ij} a_i$, where x_{ij} is the standardised genotype of individual *j* for QTL *i*, and a_i the additive effect of QTL *i*. An environmental effect was sampled from a normal distribution and added to the TBV to obtain a phenotype with a heritability of 0.6. A Holstein breed effect was sampled from $N(10,1)$ and added to the TBV for all Holstein individuals.

Genomic prediction was done using the hybrid version of the Bayes R mixture model described by Wang *et al*. (2016). This assumes that variant effects were drawn from four distributions with $N(0,0\sigma_g^2)$, $N(0,0.0001\sigma_g^2)$, $N(0,0.001\sigma_g^2)$ and $N(0,0.01\sigma_g^2)$. The hybrid model first uses an Expectation-Maximization (EM) model to estimate variant effects, the proportion of variants assigned to each of the four distributions, fixed effects (breed and sex), polygenic effects and residual variance. Subsequently, the converged estimates from the EM module were used as starting values for a Monte Carlo Markov Chain (MCMC) module that was run for 10,000 iterations. The analysis either included all variants for the full MCMC chain, or dropped a proportion of variants directly after the EM part, after 200 MCMC iterations, or after 10,000 MCMC iterations based on their probability to be included in any of the non-zero distributions. After some variants were dropped from the model another 10,000 iterations of the MCMC chain were performed. The mixing proportions at the moment of dropping were added to the prior of the mixing proportions for the remaining analysis. The analysis was done with either all variants together (FULL), split up per chromosome (CHR), with variants selected per chromosome but rerun with all chromosomes together (KEPT), and KEPT but including the HD variants (KEPT+HD). For FULL and CHR, the prior for the mixing proportions was [1,1,1,1], while for the KEPT and KEPT+HD, the posterior estimate of the mixing proportions obtained by FULL was used. Accuracies were calculated as the correlation between TBVs and GEBVs, and bias was calculated as the regression of TBVs on GEBVs.

RESULTS AND DISCUSSION

The accuracy of all scenarios using sequence data was higher than that using the HD genotypes, as shown in Table 1. For all scenarios, the accuracy for Red Holstein was larger than

that for Australian Red. This was expected, as Red Holstein individuals are closer related to the Holstein population in the reference population than the Australian Red individuals. The accuracy obtained with the hybrid, analysing all sequence variants simultaneously, corresponded with the accuracy Macleod *et al*. (2016) obtained with the same dataset, using Bayes R. This confirms that, in agreement with Wang *et al*. (2016), the hybrid is a good, faster alternative for Bayes R. Table 1 shows the results of dropping some of the variants after 10,000 MCMC iterations and then running 10,000 additional iterations. When all variants were analysed together, dropping up to 90% of the variants had minimal effect on the accuracy. However, as shown in Figure 1, when variants were dropped immediately after the EM or after only 200 MCMC iterations the reduction in accuracy increased when more variants were dropped, especially for Australian Red. Dropping variants after 10,000 MCMC iterations does, however, increase the computing time because 10,000 additional iterations were run after dropping some variants. Therefore, these results show that while this strategy can be used to select variants associated with a trait, it does not help to reduce the computing time.

Table 1. Accuracy and bias of genomic prediction

			Accuracy		Bias	
Data	**Analysis**	**DropProp**	**AusRed**	**RedHol**	**AusRed**	**RedHol**
HD	FULL	0.0	0.45	0.64	0.83	0.99
SEQ	FULL	0.0	0.60	0.66	1.07	0.97
		0.7	0.59	0.66	1.03	0.96
		0.9	0.59	0.66	1.01	0.96
SEQ	CHR	0.0	0.56	0.65	0.97	0.93
		0.7	0.56	0.65	0.97	0.93
		0.9	0.56	0.65	0.96	0.93
SEQ	KEPT	0.7	0.59	0.65	1.03	0.94
		0.9	0.59	0.65	1.01	0.93
SEQ+HD	KEPT+HD	0.7	0.60	0.66	1.04	0.94
		0.9	0.60	0.67	1.02	0.95

HD = high density genotypes, SEQ = sequence variants, FULL = all chromosomes in a single analysis, CHR = separate analysis for each chromosome, KEPT = variants selected by CHR reanalysed together, KEPT + HD = same as KEPT but including HD genotypes; dropProp = proportion of variants that are dropped after 10,000 MCMC iterations, ausRed = Australian Red, redHol = red Holstein

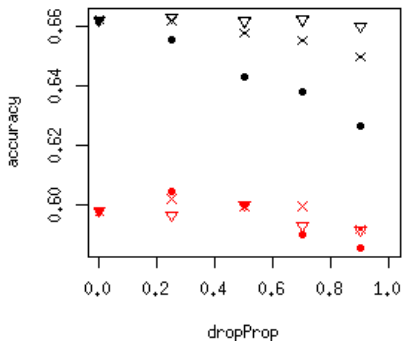

Figure 1. Reduction in prediction accuracy a function of the proportion of dropped variants
Circles = variants dropped after EM, X = variants dropped after 200 MCMC iterations, triangle = variants dropped after 10,000 MCMC iterations, black = Red Holstein, red = Australian Red

Splitting up the analysis per chromosome resulted in a large reduction in elapsed time required to complete the analysis (between 1.9 and 4.5 hours per chromosome, instead of 55 hours when all chromosomes were analysed together), but also reduced the accuracy. Using the HD variants to correct for the rest of the genome assumes independence between effects on chromosomes, while in reality, there could be LD across chromosomes, and the sum of small effects on different chromosomes can contribute to a polygenic effect.

Selecting variants one chromosome at a time and then analysing them all together resulted in an accuracy almost equal to analysing all sequence variants simultaneously (the KEPT row in Table 1). There was, however, still a slight reduction in accuracy compared to the analysis where no variants were dropped. The vast majority of variants that were dropped would have ended up in the distributions with zero or very small effects. Therefore, they may contribute to a polygenic effect rather than be linked to specific QTL. In the analysis including both the variants selected per chromosome, as well as HD variants, the accuracy increased slightly and was equal to that obtained using all sequence variants.

Even though the results in this simulation are rather positive, in reality, the advantage of sequence data is likely to be smaller. For example, in the simulation, it was assumed that all QTL are segregating across breeds and have the same effects across breeds. In reality, only a proportion of variants segregates across breeds (Raven *et al*. 2014), and it is likely that their effects are not exactly the same, for example due to differences in minor allele frequencies (MAF). Another factor that could reduce the advantage of sequence data over high or medium density is imputation accuracy. Most sequences are obtained by imputation rather than direct sequencing, and this introduces errors in the genotypes. Furthermore, sequences used in this simulation were preselected based on functional annotations, strongly reducing the number of variants. Reducing the number of variants made it possible to compare analysing all variants simultaneously with strategies to split up the analyses. When these strategies are applied to datasets containing millions of variants, the large number of variants may induce problems to accurately estimate effects simultaneously, especially for variants that are in high LD with each other.

Our analyses show that preselecting sequence variants with a Bayesian variable selection model per chromosome and subsequently using those variants for genomic prediction, preferably combined with genome wide makers, could be an alternative to analysing full sequence data directly.

ACKNOWLEDGEMENTS

This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research.

REFERENCES

Brøndum R.F., Su G., Janss L., Sahana G., Guldbrandtsen B., Boichard D. and Lund M.S. (2015) *J. Dairy Sci.* **98**:4107.

Calus M.P.L., Bouwman A.C., Schrooten C. and Veerkamp R.F. (2016). *Genet. Sel. Evol.* **48**:49.

Macleod I.M., Bowman P.J., Vander Jagt C.J., Haile-Mariam M., Kemper K.E., Chamberlain A.J., Schrooten C., Hayes B.J. and Goddard M.E. (2016) *Genet. Sel. Evol.* **17**:144.

Raven, L-A., Cocks B.G. and Hayes B.J. (2014) *BMC Genomics* **15**:62.

van Binsbergen R., Calus M.P.L., Bink M.C.A.M., van Eeuwijk F.A., Schrooten C. and Veerkamp R.F. (2015) *Genet. Sel. Evol.* **47**:71.

van den Berg I., Boichard D. and Lund M.S. (2016) *Genet. Sel. Evol.* **48**:83.

Veerkamp R.F., Bouwman A.C., Schrooten C. and Calus M.P.L. (2016) *Genet. Sel. Evol.* **48**:95.

Wang T., Phoebe Chen Y-P., Bowman P.J., Goddard M.E. and Hayes B.J. (2016) *BMC Genomics* **17**:744.