Genomic prediction using QTL derived from whole genome sequence data

***R.F Brøndum*, G. Su, L. Janss, G. Sahana, and M.S. Lund.**
Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University

ABSTRACT: This study investigated the gain in accuracy of genomic prediction when a small number of significant variants from single marker analysis based on whole genome sequence data were added to the regular 54k SNP data. Analyses were performed for Nordic Holstein and Danish Jersey animals, using either a genomic BLUP or a Bayesian variable selection model. When using the genomic BLUP model, results showed increases in accuracy of up to two percentage points for production traits in both Holstein and Jersey animals by including the extra variants in the analysis, and an extra 1.5 percentage points for fertility in Jersey animals. When using a Bayesian model accuracies were generally higher, but only small increases in accuracy of up to 0.6 percentage points were observed for the Holstein animals when including the extra markers, while both increases and decreases were observed for Jersey.

Keywords: Custom chip, genomic prediction, QTL.

Introduction

The accuracy of genomic prediction is highly dependent on the linkage disequilibrium (**LD**) between the available markers and actual causative variants (de Los Campos et al. (2013)). In dairy cattle genomic predictions are usually done using the Illumina 54k SNP chip, where the distance between the markers might dictate a low level of LD. Increasing the marker density to 777k has only shown small increases in accuracy of around 1 percent (e.g. (Su et al. (2012)). However, in a study on simulated data it was shown by Meuwissen and Goddard (2010) that even at an already high marker density, a further inclusion of causative variants led to a higher accuracy of genomic prediction. Based on imputed whole genome sequence data, single marker association studies have been done for the 17 traits considered in the breeding goal for Nordic Holstein, Danish Jersey and Nordic Red (for a full list of the traits see Brøndum et al. (2011)). Based on the importance of the trait the 15, 10 or 5 most significant quantitative trait loci (**QTL**) for each trait and breed were selected and 1,486 markers have been added to the IlluminaLD chip. The chip is aimed at large scale genotyping of cows and these markers could then be imputed with high accuracy into the already genotyped bull populations (Boichard et al. (2012)). These 1,486 markers are expected to be either causative variants or in high LD with causative variants, and including them could potentially increase the accuracy of genomic prediction. In this study we investigate the gain in the accuracy of genomic prediction when these markers are included in the analysis.

Materials and Methods

Data. For the analysis 4,999 Nordic Holstein animals and 1,140 Danish Jersey animals with both phenotype and 54k genotype data were available. Phenotypic data was de-regressed estimated breeding values (**DRP**) for milk yield, protein yield, fat yield, fertility and mastitis. For the Holstein population the animals born before January 1st 2005 were chosen as a reference set, leading to 3,953 animals in the reference and 1,046 for validation. For the Jersey population animals were split by January 1st 2003, leaving 917 animals in the reference and 223 for validation. Genotypes of the 1,486 QTL markers were extracted from imputed sequence data, where sequence data for the available bulls were imputed using combined reference data from the 1,000 bull genomes project (Daetwyler et al. (2014)) and private data from Aarhus University. Imputation was done using a two-step procedure, where animals were first imputed from 54k to HD data using a multi-breed reference of 3,385 animals and subsequently imputed to whole genome sequence data using a multi-breed reference of 531 animals. Imputations were done using IMPUTE2 (Howie et al. (2011)) with pre-phased data from BEAGLE4 (Browning and Browning (2013)). The number of selected QTL markers for each of the analyzed traits is shown in Table 1. Some markers showed significant association for more than one trait, meaning that the number of QTL per trait in some cases exceeded 15.

Table 1: Number of QTL markers selected for the custom low density chip for each of the five analyzed traits.

Trait	Holstein	Jersey
Fat	15	14
Fertility	19	15
Mastitis	16	25
Milk	13	16
Protein	22	14

Statistical analyses. Analyses comprised three marker panel scenarios: (a) only 54k data, (b) 54k data pooled with the imputed QTL markers and (c) 54k data and the imputed QTL markers as separate variance components. All 1,486 QTL markers were included for all traits. A number of statistical models were tested. The first two were a one component GBLUP:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Za} + \mathbf{e}$$

and a two component GBLUP:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}_{\mathbf{54k}}\mathbf{a}_{\mathbf{54k}} + \mathbf{Z}_{\mathbf{QTL}}\mathbf{a}_{\mathbf{QTL}} + \mathbf{e}$$

Where, μ is an overall mean, **a**, $\mathbf{a}_{\mathbf{54k}}$ and $\mathbf{a}_{\mathbf{QTL}}$ are the vectors of additive genetic values and **e** is the vector of residual errors. **Z** is an incidence matrix allocating additive genetic values for the 54k data or the pooled 54k and QTL data to phenotypes and $\mathbf{Z}_{\mathbf{54k}}$ and $\mathbf{Z}_{\mathbf{QTL}}$ are incidence matrices for respectively the 54k data and the QTL data. It was assumed that $\mathbf{a}_{\mathbf{i}} \sim N(\mathbf{0}, \mathbf{G}_{\mathrm{i}}\sigma_{gi}^2)$, where the genomic relationship matrix $\mathbf{G}_{\mathrm{i}}$ was constructed according to method 1 in VanRaden (2008) based on the 54k data, the QTL marker data or the pooled marker data. Heterogeneous residual variances of DRP were accounted for by exerting a weight according to the reliability of DRP. Two Bayesian two-distribution mixture models were also tested. The first was given as:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\mathbf{g} + \mathbf{e}$$

and the second as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}_{\mathbf{54k}}\mathbf{g}_{\mathbf{54k}} + \mathbf{X}_{\mathbf{QTL}}\mathbf{g}_{\mathbf{QTL}} + \mathbf{e}$$

Where μ and **e** were as defined above, **g**, $\mathbf{g}_{\mathbf{54k}}$, $\mathbf{g}_{\mathbf{QTL}}$ were vectors of effects for the corresponding set of SNP, **X** is a genotype matrix for the 54k data or pooled 54k and QTL data, and $\mathbf{X}_{\mathbf{54k}}$ and $\mathbf{X}_{\mathbf{QTL}}$ are genotype matrices for respectively 54k and QTL data. Prior distributions for SNP effects are given as $\mathbf{g} \sim \pi N(\mathbf{0}, \sigma_1^2) + (1-\pi)N(\mathbf{0}, \sigma_2^2)$. In this study, σ_1^2 was set to 1e^{-4}, σ_2^2 and π were estimated from data with the assumption that $\sigma_2^2 \propto$ constant and $\pi \sim \beta(10,1)$. Estimates were obtained as posterior means from a MCMC sampler run for 50,000 iterations where the first 10,000 were discarded as burn-in. For all models the accuracy of genomic breeding values (**GEBV**) was calculated as the correlation of GEBV and DRP in the test population.

Results and Discussion

Holstein. Accuracies for Holstein are shown in Table 2. Here gains in accuracy ranged from 0.4% to 2% for production traits when pooling data and using the regular GBLUP model, whereas only small advantages are observed for fertility and mastitis. No further advantages are seen when modeling the 54k data and QTL data as two separate components. Looking at the Bayesian methods accuracies are generally higher but smaller or no gains in accuracy are seen when adding in QTL markers. The only notable exception is the large gain for fat yield which might come from a better estimation of the DGAT1 mutation as well as a few relatively large QTLs, for example, on chromosomes 5 and 20. When modeling the 54k and the QTL data as two separate components, accuracies are lower than for a simple pooling of the data.

Jersey. Results for the Jerseys are shown in Table 3. Gains in accuracy, when pooling 54k and QTL data and using a GBLUP model, are also found with improvements of 1.5-2% for milk yield, fat yield and fertility. Separating 54k and QTL markers gives larger improvements for milk yield and especially fertility, but decreases the accuracy for mastitis and protein yield. When using the Bayesian mixture model, increases in accuracy are only found for fat yield and fertility, whereas accuracies decrease for the other three traits.

Table 2: Accuracy of genomic prediction for Holstein.

	GBLUP			Bayesian		
Scenario	a	b	c	a	b	c
Milk	0.670	0.679	0.677	0.683	0.685	0.683
Fat	0.666	0.686	0.685	0.684	0.690	0.690
Protein	0.651	0.655	0.655	0.661	0.662	0.659
Fertility	0.518	0.519	0.513	0.521	0.523	0.518
Mastitis	0.548	0.551	0.551	0.557	0.557	0.552

a: only 54k
b: 54k pooled with QTL
c: 54k and QTL with separate variance components.

Table 3: Accuracy of genomic prediction for Jersey.

	GBLUP			Bayesian		
Scenario	a	b	c	a	b	C
Milk	0.608	0.623	0.631	0.636	0.619	0.630
Fat	0.446	0.469	0.470	0.460	0.470	0.482
Protein	0.585	0.586	0.576	0.592	0.585	0.589
Fertility	0.302	0.317	0.378	0.297	0.320	0.402
Mastitis	0.528	0.528	0.524	0.528	0.529	0.526

a: only 54k
b: 54k pooled with QTL
c: 54k and QTL with separate variance components.

Model. Overall accuracies are higher when using the Bayesian methods in both populations but adding in the QTL markers brings accuracies from the GBLUP closer to the accuracies from the Bayesian mixture model. The mixture model allows for different variances at different loci and facilitates an effect of some loci close to zero and thus allows for more emphasis on markers in close LD with causal variants. This could explain its superiority when QTL markers are not included. However, when adding the major QTL from sequence data it seems the genomic relationship used in the GBLUP model gets closer to the functional relationship, putting the accuracy of GEBV closer to that from the Bayesian model.

Unbiasedness. Regression coefficients for the different scenarios are shown in Table 4 and Table 5. The figures are fairly consistent across the scenarios of the 54k data and the pooled data. When modeling the 54k data and the QTL data as two components, the regression coefficients remained the same for the Holsteins animals, while some inflation of the GEBVs was observed for fertility in the Jerseys.

Perspective. The accuracy of imputation for whole genome sequence markers has been shown to be quite low (Daetwyler et al. (2014)). This means that the quality of the imputed QTL markers in this study could be low, which in a previous study has been shown to affect the accuracy of GEBV negatively (Dassonneville et al. (2011)). With the wide use of the custom low density chip (which includes the QTL markers) for genotyping of cows, a sufficiently large reference with genotype data from SNP arrays should however soon become available and this could increase the accuracy of imputation and thus genomic prediction. This also means that the QTL SNPs could be imputed for bulls without the computational burden of whole genome imputation. In addition, in the future cows will be important reference animals, especially in small breeds. Therefore by using the sequence variants accurately genotyped in phenotyped females along with more accurately imputed markers in the bulls, it is expected that a larger gain in reliability of genomic prediction from the QTL markers could be obtained. On the other hand, the effect of using the QTL markers in this study might be overestimated, since all available phenotyped animals were used for identification of the QTL. The markers selected are thus chosen to best describe phenotypic variation in a population containing the validation animals. However, since only a few QTL per trait per breed were selected, and the QTL effect was estimated using a model that accounts for the relationship structure in the population, the probability that these markers are specifically favorable for prediction of the current validation animals could be negligible. However, further studies on an independent validation population are needed to fully conclude on the effect of adding these QTL markers for genomic prediction.

Table 4: Regression coefficients for Holstein.

	GBLUP			Bayesian		
Scenario	a	b	c	a	b	c
Milk	0.909	0.901	0.887	0.921	0.908	0.904
Fat	0.864	0.863	0.842	0.850	0.849	0.842
Protein	0.864	0.860	0.852	0.872	0.864	0.854
Fertility	0.949	0.945	0.933	0.946	0.941	0.934
Mastitis	0.911	0.910	0.911	0.924	0.918	0.909

a: only 54k
b: 54k pooled with QTL
c: 54k and QTL with separate variance components.

Table 5: Regression coefficients for Jersey.

	GBLUP			Bayesian		
Scenario	a	b	c	a	b	C
Milk	0.904	0.897	0.894	0.936	0.869	0.872
Fat	0.698	0.728	0.702	0.710	0.709	0.699
Protein	0.878	0.864	0.826	0.878	0.834	0.826
Fertility	0.993	1.030	1.265	1.001	1.054	1.265
Mastitis	0.892	0.880	0.869	0.885	0.894	0.885

a: only 54k
b: 54k pooled with QTL
c: 54k and QTL with separate variance components.

Conclusion

When adding major QTL to the marker panel for genomic prediction, increases in accuracy of up to 2% were found. When modeling the 54k data and QTL data as two components and using the GBLUP model inconsistent results were found, but simply pooling marker sets always gave similar or higher accuracies. By using cows genotyped with the custom low density chip to impute the QTL markers for the reference bulls, this thus makes for an easy implementable increase in the accuracy of genomic prediction in routine genomic evaluation.

Acknowledgements

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

Literature cited

Boichard, D., Chung, H., Dassonneville, R. et al. (2012). *PLoS One*. 7:e34130.

Browning, B.L., and Browning, S.R. (2013). *Genetics*. 194:459–71.

Brøndum, R.F., Rius-Vilarrasa, E., Strandén, I. et al. (2011). *J. Dairy Sci.* 94:4700–7.

Daetwyler, H.D., Capitan, A., Pausch, H. et al. (2014). *Nat. Genet.* accepted.

Dassonneville, R., Brøndum, R.F., Druet, T. et al. (2011). *J. Dairy Sci.* 94:3679–86.

Howie, B., Marchini, J., and Stephens, M. (2011). *G3 (Bethesda).* 1:457–70.

De Los Campos, G., Vazquez, A.I., Fernando, R. et al. (2013). *PLoS Genet.* 9:e1003608.

Meuwissen, T., and Goddard, M. (2010). *Genetics*. 185:623–31.

Su, G., Brøndum, R.F., Ma, P. et al. (2012). *J. Dairy Sci.* 95:4657–65.

VanRaden, P.M. 2008. *J. Dairy Sci.* 91:4414–23.