Haplotype-assisted genomic evaluations in Nordic Red dairy cattle

T. Knürr*****, **I. Strandén***, **M. Koivula***, **G.P. Aamand†** and **E.A. Mäntysaari***
*MTT Agrifood Research Finland, Biotechnology and Food Research, Jokioinen, Finland
†NAV Nordic Cattle Genetic Evaluation, Aarhus, Denmark

ABSTRACT: In admixed populations originating from different base breeds, such as Nordic Red dairy cattle, haplotypes of chromosomal segments instead of single SNP are expected to improve prediction accuracy in genomic evaluations, because linkage disequilibrium with QTL is likely to be more consistent for haplotypes than for SNP. The suggested evaluation approach consists of (i) pre-selecting a limited number of chromosomal segments based on a genome-wide QTL-scan with BayesB, (ii) estimating relative variances of haplotype markers with BayesA, and (iii) obtaining solutions for haplotype effects from mixed model equations including pedigree-based animal effects. For three production traits (milk, protein, fat) and fertility, the highest validation test reliabilities R^2 were 0.48, 0.41, 0.42 and 0.33, respectively. For milk, protein and fertility, we observed an improvement over G-BLUP of 3, 1 and 3 %-units, respectively, whereas for fat, a decline of 1 %-unit.

Keywords: Nordic Red dairy cattle; genomic evaluation; haplotype

Introduction

In genomic evaluations, DNA information is exploited to improve reliability of predictions for genetic merit in e.g. breeding programmes of livestock. One of the main benefits from using DNA information is that it becomes available for the evaluation of individual animals earlier in life than most traits can be measured. As a consequence, the need to wait for results from cost-intensive and lengthy progeny testing decreases.

In their pioneering study on genomic selection, Meuwissen et al. (2001) originally formulated BayesA and BayesB in terms of haplotype effects to be estimated. Haplotypes are chromosomes, or chromosome segments, which are jointly inherited from parent to offspring. Yet, high-throughput genotyping based on single nucleotide polymorphism (SNP) arrays has afterwards promoted the development and the implementation of genetic evaluations models in terms of bi-allelic markers such as SNP. Whereas SNP-based genomic evaluations have shown outstanding performance in genetically homogenous populations such as Holstein dairy cattle, the application to heterogeneous populations originating from various base breeds such as Nordic Red dairy cattle (RDC) has been less successful.

The main motivation to use haplotype markers in admixed populations is that identity-by-state of haplotypes instead of SNP is expected to be a better surrogate for identity-by-descent of a chromosomal segment. This is because joint inheritance of markers in different lineages of the population is reflected more accurately in haplotypes. Consequently, linkage disequilibrium with quantitative trait loci (QTL) is expected to be more consistent for haplotypes than for SNP. Further, many genomic prediction models try to improve estimates for genetic relationships between individuals by using genome-based relationships rather than relationships using pedigree information. In genetically heterogeneous populations, however, SNP are not able to trace relationships well enough.

In this study, we aimed at improving genomic prediction in Nordic RDC by exploiting haplotype information. First, the genome was scanned to detect the chromosomal segments with the strongest QTL signals. To improve power to estimate genetic effects and to reduce computational demands, only chromosomal segments harboring the strongest QTL signals were used in the following prediction of genome-enhanced breeding values (GEBV). We considered different alternatives for the number of segments and for the length of the segments and compared validation results with two SNP-based prediction methods. Models were compared in evaluations of three production traits and fertility using real Nordic RDC data.

Materials and Methods

Data. The data included phenotype, genotype and pedigree information for Nordic RDC bulls born between 1971 and 2008. Genotypes were obtained from the Illumina Bovine SNP50 Bead Chip (Illumina, San Diego, CA). After application of editing criteria, 38,194 SNP markers on the 29 bovine autosomes were used in further analysis. The software BEAGLE v3.3 (Browning and Browning (2009)) was used to impute missing genotypes and to phase the SNP data. The phenotype data were obtained from Nordic genetic evaluations in February 2013. The data included deregressed proofs (DRP) complemented by effective daughter contributions (EDC) for three production traits (milk, protein and fat yield) and fertility. DRP were based on standardized estimated breeding values for index traits. The index traits and the standardization procedure are described in detail by Nordic Cattle Genetic Evaluation (2013). The training/reference set comprised bulls born between 1971 and 2005 (4250 for production traits, 4422 for fertility) and the validation/candidate set comprised bulls born between 2006 and 2008 (516 for production traits, 551 for fertility).

Haplotype-assisted genomic prediction. The approach for haplotype-assisted genomic prediction is briefly summarized as follows: (i) all SNP were simultaneously screened for QTL signals; (ii) a certain number of chromosomal segments ("blocks") of pre-defined length containing the SNP with the strongest QTL signals were pre-selected for further analysis; (iii) the pre-selected blocks were jointly evaluated in a multi-locus model to obtain block-specific variances of haplotype effects; (iv) in the genomic evaluation model, the effects of haplotypes were re-estimated, using the variance estimates obtained in the previous step and including a pedigree-based polygenic term; (v) GEBV were then calculated for the candidate bulls and validated using DRP of candidate bulls.

Screening for QTL signals. The DRP were modeled by generalized BayesB (Strandén et al. (2011)) with model equation

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}^{\mathrm{SNP}}\mathbf{g}^{\mathrm{SNP}} + \mathbf{e}.$$

Here, $\mathbf{y}$ is the vector of N DRP observations, μ the common intercept, $\mathbf{Z}^{\mathrm{SNP}}$ the $N \times M$ genotype matrix holding codes 0, 1, and 2 for the three possible genotypes at each of M SNP markers, $\mathbf{g}^{\mathrm{SNP}}$ the vector of M additive marker effects, and $\mathbf{e}$ the vector of N residuals. Observations were weighted by EDC. The model parameters were estimated using Markov chain Monte Carlo (MCMC) approximation with 200,000 samples, of which the first 20,000 were discarded as burn-in.

Pre-selection of haplotype blocks. The absolute values of the posterior means of marker effects ($|\hat{g}_m|$) were used to rank QTL signals and to pre-select M^B(=1500 or 750) haplotype blocks for further analysis. The first block was chosen including the SNP with largest $|\hat{g}_m|$ (denote its index m_1). The SNP with indices $m_1 - s, \dots, m_1 + s$ formed the first block in case all these SNP were on the same chromosome. Otherwise, i.e. if not enough flanking markers were available at the start or the end of a chromosome, the indices were shifted forward or backward such that all $2s + 1$ SNP were chosen from the same chromosome. The following $M^B - 1$ blocks were chosen likewise, but with the restriction that any two blocks were allowed to share at most one SNP. The values for s were 1 and 2, thus forming haplotype blocks of length 3 and 5 SNP, respectively.

Estimation of haplotype block variances. Once the haplotype blocks had been pre-selected, the variance of the effects in each block (σ_{gj}^2) was estimated using BayesA (Meuwissen et al. (2001)). Here, the regression equation for the DRP was

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}^{\mathrm{HAP}}\mathbf{g}^{\mathrm{HAP}} + \mathbf{e}$$

with observations weighted by EDC. Denoting the number of distinct haplotypes in block $j = 1, \dots, M^B$ with M_j and the indices of these haplotypes with $j_1, \dots, j_{M_j}$, the haplotype effects in block j were $g_{j_1}^{\mathrm{HAP}}, \dots, g_{j_{M_j}}^{\mathrm{HAP}}$. Matrix $\mathbf{Z}^{\mathrm{HAP}}$ had $M^H = \sum_{j=1}^{M^B} M_j$ columns and N rows. Its elements were the numbers of copies (0, 1 or 2) of a given haplotype for an individual. In the case that blocks comprised five adjacent SNP, the upper limit for M_j was $2^5 = 32$, and in the case of three SNP $2^3 = 8$. The lengths of the MCMC chains were 200,000 iterations, of which the first 20,000 were discarded as burn-in.

Evaluation model. The final evaluation model for the DRP was

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{a} + \mathbf{Z}^{\mathrm{HAP}}\mathbf{g}^{\mathrm{HAP}} + \mathbf{e},$$

for which solutions were obtained from mixed-model equations (MME). The only fixed effect was the common intercept μ. The random term $\mathbf{a}$ was a vector of animal effects with mean 0 and variance-covariance $\omega\widehat{\sigma_a^2}\mathbf{A}$, where ω was fixed to a value in (0,1), $\widehat{\sigma_a^2}$ an estimate for the additive genetic variance and $\mathbf{A}$ the pedigree-based relationship matrix between genotyped animals. The random residuals were assumed to have variance-covariance $\widehat{\sigma_e^2}\mathbf{R}$, where the diagonal matrix $\mathbf{R}$ included the weights based on EDC. The estimates $\widehat{\sigma_a^2}$ and $\widehat{\sigma_e^2}$ had been obtained using a standard animal model without any genomic component. The random haplotype effects in block j shared the same variance $(1 - \omega)\widehat{\sigma_a^2}\widehat{\sigma_{gJ}^2}/S$. Here, $\widehat{\sigma_{gJ}^2}$ was the posterior mean for the haploblock variance estimated in the previous step. Further, S was a constant ensuring that a proportion $1 - \omega$ of the additive genetic variance was assigned to haplotype blocks. It was calculated as $S = \mathrm{tr}(\mathbf{Z}\mathrm{var}(\mathbf{g}^{\mathbf{HAP}})\mathbf{Z}')/M^H$, with $\mathbf{Z}$ being $\mathbf{Z}^{\mathrm{HAP}}$ centered to have column means 0.

Validation of genomic prediction. GEBV were calculated using the equation

$$\mathbf{GEBV} = \hat{\mathbf{a}} + \mathbf{Z}^{\mathrm{HAP}}\widehat{\mathbf{g}^{\mathrm{HAP}}},$$

where $\hat{\mathbf{a}}$ and $\widehat{\mathbf{g}^{\mathrm{HAP}}}$ were the MME solutions obtained from the evaluation model. The model was validated by regressing DRP on GEBV of candidate bulls with observations weighted by EDC. The slope coefficient of this regression (b_1) was used as an estimate for bias of GEBV. Following Mäntysaari et al. (2010), the coefficient of determination $r_{(\mathrm{GEBV,DRP})}^2$ was scaled to obtain an estimate for the validation reliability according to $R^2 = r_{(\mathrm{GEBV,DRP})}^2/\bar{w}$, where $w_i = \mathrm{EDC}_i/(\mathrm{EDC}_i + \lambda)$ with $\lambda = (4 - h^2)/h^2$. Here, the estimates for trait heritability h^2 were the values used in Nordic genetic evaluations: 0.39 for the three production traits and 0.04 for fertility. The resulting scaling factor $\bar{w}$ was 0.92 for the production traits and 0.57 for fertility.

Comparison with SNP-based genomic evaluations. Instead of using haplotype markers as described above, GEBV were also obtained using a limited number of pre-selected SNP markers. Here, we used the results from BayesB to pre-select the SNP with largest effects. The subsequent procedures (estimation of SNP instead of haplotype variances, the evaluation and validation) were altered to accommodate SNP markers. Additionally, GEBV were also calculated with SNP-based GBLUP. Here, a weighted mean of the pedigree-based relationship matrix **A** and the genome-based relationship matrix $\mathbf{G}$ was used instead of a solely genome-based relationship matrix (VanRaden (2008)). Specifically, the variance-covariance matrix for the polygenic effects was calculated as $\mathbf{G}_{0.9} = 0.9\mathbf{G} + 0.1\mathbf{A}$. The genome-based relationship matrix was $\mathbf{G} = \mathbf{Z}^{\mathbf{SNP}}(\mathbf{Z}^{\mathbf{SNP}})' / (2\sum_{m=1}^{M} p_m(1-p_m))$, where $\mathbf{Z}^{\mathbf{SNP}}$ had been centered to have column means 0 and p_m was the frequency of the second allele at SNP m.

Results and Discussion

Table 1 shows validation results for GEBV of candidate bulls for seven models. The number of haplotype or single SNP markers was either 1500 (models HAP 5^{1500}, HAP 3^{1500}, SNP 1^{1500}) or 750 (models HAP 5^{750}, HAP 3^{750}, SNP 1^{750}), whereas all 38,194 SNP markers were used in GBLUP. For the haplotype-based methods, results for haplotype segments of either 5 adjacent SNP (models HAP 5^{1500}, HAP 5^{750}) or 3 adjacent SNP (models HAP 3^{1500}, HAP 3^{750}) are given. The HAP and SNP models were evaluated for proportions $\omega = 0.01, 0.1, 0.2, \ldots, 0.9, 0.99$, but only the results for ω which yielded highest validation reliability R^2 are reported. In the GBLUP model, ω was assumed constant 0.10.

Table 1. Validation results for GEBV of candidate bulls: validation reliability (R^2), bias (b_1) and proportion of genetic variance assigned to pedigree (ω).

Model	R^2	b_1	ω
Milk			
HAP 5^{1500}	0.48	0.94	0.4
HAP 3^{1500}	0.47	0.95	0.6
SNP 1^{1500}	0.48	0.93	0.8
HAP 5^{750}	0.45	0.94	0.5
HAP 3^{750}	0.48	0.92	0.6
SNP 1^{750}	0.46	0.88	0.8
GBLUP	0.45	0.79	0.1
Protein			
HAP 5^{1500}	0.41	0.86	0.4
HAP 3^{1500}	0.40	0.88	0.6
SNP 1^{1500}	0.40	0.84	0.8
HAP 5^{750}	0.36	0.83	0.6
HAP 3^{750}	0.39	0.87	0.7
SNP 1^{750}	0.36	0.86	0.9
GBLUP	0.40	0.71	0.1
Fat			
HAP 5^{1500}	0.41	0.81	0.4
HAP 3^{1500}	0.42	0.82	0.5
SNP 1^{1500}	0.43	0.82	0.8
HAP 5^{750}	0.38	0.77	0.4
HAP 3^{750}	0.41	0.83	0.7
SNP 1^{750}	0.41	0.83	0.9
GBLUP	0.43	0.72	0.1
Fertility			
HAP 5^{1500}	0.31	0.82	0.3
HAP 3^{1500}	0.33	0.84	0.4
SNP 1^{1500}	0.29	0.82	0.8
HAP 5^{750}	0.28	0.78	0.5
HAP 3^{750}	0.29	0.84	0.7
SNP 1^{750}	0.29	0.78	0.8
GBLUP	0.30	0.72	0.1

For milk yield evaluated with HAP models, highest R^2 was 0.48 and, thus, higher than R^2 for GBLUP (0.45). However, R^2 was also 0.48 for model SNP 1^{1500}. For protein yield, model HAP 5^{1500} gave highest R^2 (0.41). However, GBLUP and model SNP 1^{1500} performed almost as well ($R^2 = 0.40$). For fat yield, R^2 of all HAP models were below 0.43, the value given by model SNP 1^{1500} and GBLUP. In the case of fertility, highest R^2 with a value of 0.33 was yielded by model HAP 3^{1500}. To summarize, no consistent advantage over SNP-based models or GBLUP was observed for R^2 as yielded by haplotype-based models. In most cases, it was beneficial with respect to R^2 to use 1500 instead of 750 markers in haplotype and single SNP models. The results gave no clear indication if it would be beneficial to use haplotype blocks with 3 or 5 adjacent SNP.

With respect to the bias of GEBV (b_1), the haplotype-based models and the models using a limited number of single SNP markers gave better results, i.e. values closer to 1, than GBLUP. For the proportion of genetic variance assigned to pedigree (ω), a clear trend was observed, as ω generally increased, when the number of markers used was reduced from 1500 to 750. The models with effects of 1500 individual SNP and a weight of only 0.2 relative to the polygenic effect, i.e. $\omega = 0.8$, performed notably well when compared with GBLUP, especially with respect to the variance inflation factor b_1.

Conclusion

According to our results, the haplotype-based method used in this study did not consistently improve genomic prediction when compared to single SNP-based methods or GBLUP. One reason for this could be that the procedure involved a pre-selection step based on a BayesB-type analysis that actually exploited SNP information and not haplotypes. The QTL signals coming up in this part of the analysis may not be representative for QTL-haplotype associations, which the following steps of the method aim to exploit. In other words, effects of important QTL may be missing in the GEBV predicted by haplotype effects, because a “bad” set of chromosomal regions was pre-selected. Therefore, the haplotype-based method may be improved by pre-selection based on screening the genome for QTL-haplotype associations instead of QTL-SNP associations.

Literature Cited

Browning, B.L., and Browning, S.R. (2009). Am. J. Hum. Genet. 84: 210-223.

Mäntysaari, E., Liu, Z. and VanRaden, P. (2010). Interbull Bull. 41: 17-22.

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

Nordic Cattle Genetic Evaluation (2013). http://www.nordicebv.info Accessed Feb 2014.

Strandén, I., Mrode, R. and Berry, D.P. (2011). Book of abstracts of the 62nd annual meeting of the European Federation of Animal Science: 393.

VanRaden, P.M. (2008). J. Dairy Sci. 91:4414-4423.