Haplotype-assisted genomic evaluations in Nordic Red Dairy Cattle

Timo Knürr[1)], Ismo Strandén[1)], Minna Koivula[1)], Gert Pedersen Aamand[2)], Esa A. Mäntysaari[1)]

[1)] *MTT Agrifood Research Finland, Biotechnology and Food Research, 31600 Jokioinen, Timo.Knurr(at)mtt.fi, Ismo.Stranden(at)mtt.fi, Minna.Koivula(at)mtt.fi, Esa.Mantysaari(at)mtt.fi*
[1)] *NAV Nordic Cattle Genetic Evaluation, Agro Food Park 15, 8200 Aarhus N, Denmark, GAP(at)vfl.dk*

Abstract

In admixed populations originating from different base breeds, such as the Nordic Red Dairy Cattle, identity by state of haplotypes instead of single nucleotide polymorphisms (SNP) is a better surrogate for identity by descent. Therefore, haplotypes are expected to be more useful in recovering genetic relationships among animals and linkage disequilibrium between markers and quantitative trait loci (QTL). The objective of this study was to improve the prediction accuracy in genomic evaluations by the use of haplotypes of short chromosomal segments.

In the first step, around 38,000 SNPs from a 50K chip were simultaneously scanned for QTL signals with BayesB. Based on these results, the SNPs with the strongest QTL signals were then pre-selected and haplotype blocks were constructed around these. In the second step, we estimated the relative variances for the pre-selected haplotypes with BayesA. In the final step, mixed model equations of the evaluation model were solved to estimate haplotype effects. Here, a pre-defined proportion of the genetic variance was assigned to a pedigree-based animal effect and the rest to haplotypes. The accuracies of several combinations for the number and the length of haplotype blocks were assessed by calculation of genome-enhanced breeding values and validation test reliabilities (R^2).

For three production traits (milk, protein, fat) and fertility, the highest R^2 observed were 0.48, 0.41, 0.42 and 0.33, respectively. For milk and protein, we observed an improvement over G-BLUP of 3 and 1 percentage points, respectively. For fat and fertility, the highest R^2 were the same as for GBLUP. Further analysis using only single SNPs instead of haplotype blocks yielded similar results. This may indicate a need to choose an alternative approach to pre-select haplotype blocks.

Keywords

Nordic Red Dairy Cattle, genomic evaluation, prediction accuracy, haplotype, SNP

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 their prediction models 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 SNPs. 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 has been less successful.

The main motivation to use haplotype markers in admixed populations is that identity-by-state of haplotypes instead of SNPs 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 (LD) with quantitative trait loci (QTL) is expected to be more consistent for haplotypes than for SNPs. 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, SNPs are not able to trace relationships well enough.

In this study, we aimed at improving genomic prediction in Nordic Red dairy cattle 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. 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. We evaluated prediction of three production traits and fertility using real Nordic Red dairy cattle data.

Materials and Methods

The data included phenotype, genotype and pedigree information for Nordic Red dairy cattle (RDC) bulls born between 1971 and 2008. The bulls were split into a training and a validation set by birth year: bulls born between 1971 and 2005 were defined as training or reference bulls and bulls born between 2006 and 2008 as validation or candidate bulls.

Marker data and haplotype phasing

Genotypes were obtained from the Illumina Bovine SNP50 Bead Chip (Illumina, San Diego, CA). After application of exclusion criteria, 38,194 SNP markers on the 29 bovine autosomes were available for further analysis. The software BEAGLE v3.3 (Browning and Browning 2009) was used to impute missing genotypes and to phase the SNP data.

Phenotype data

The phenotype data were obtained from Nordic genetic evaluations for RDC. The data included deregressed proofs (DRP) complemented by effective daughter contributions (EDC). The deregressed proofs 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). Three production traits (milk, protein and fat yield) and fertility were selected for this study and a summary for these traits is given in Table 1.

Table 1. Summary of phenotype data (deregressed proofs DRP and effective daughter contributions EDC) for the four index traits analyzed. DRP are based on national breeding value indices which are standardized such that cows born 2008-10 have a mean of 100 and bulls born 1997-98 have a standard deviation of 10.

			DRP			EDC		
Trait	Group	N	Mean	SD	Range	Mean	SD	Range
Milk	Reference	4250	89.9	12.1	44.9-128.1	281	207	28-916
	Candidate	516	102.7	9.5	72.5-126.4	126	60	19-295
Protein	Reference	4250	86.7	13.7	32.1-129.4		(a)	
	Candidate	516	103.9	9.1	77.1-133.6			
Fat	Reference	4250	90.3	11.7	37.6-130.0			
	Candidate	516	103.5	8.5	76.8-129.5			
Fertility	Reference	4422	100.3	13.5	19.7-151.7	860	1807	26-9771
	Candidate	551	98.3	20.4	-6.9-195.2	170	107	25- 519

(a) EDCs were identical for the three production traits (milk, protein and fat).

Haplotype-assisted genomic prediction

The approach for haplotype-assisted genomic prediction can be summarized as follows:

1. All SNPs were simultaneously screened for QTL signals.
2. A certain number of chromosomal segments ("blocks") of pre-defined length containing the SNPs with the strongest QTL signals were pre-selected for further analysis.
3. The pre-selected blocks were jointly evaluated in a multi-locus model to obtain block-specific variances of haplotype effects.
4. 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.
5. Genome-enhanced breeding values (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.

The prior distribution for the common intercept μ was uniform on the entire real line, i.e. $p(\mu) \propto 1$. The marker effects $g_1^{\mathrm{SNP}}, \ldots, g_M^{\mathrm{SNP}}$ were assigned mutually independent prior distributions, which were specified by

$$g_m^{\mathrm{SNP}} \sim \begin{cases} 0 & \text{with probability } \pi, \\ N(0, s_m^2 \sigma_t^2) & \text{with probability } (1-\pi). \end{cases}$$

Here, $s_m^2 \sim \nu/\chi_\nu^2$ and $p(\sigma_t^2) \propto 1$. In our analysis, we set degrees of freedom $\nu = 4.01$ and proportion of markers with zero effect $\pi = 0.9$, i.e., 10% of the SNP markers were assumed to have non-zero effect. The prior distribution for the vector of residuals was multivariate normal with $\mathbf{e} \sim N(0, \sigma_e^2 \mathbf{R})$, where $p(\sigma_e^2) \propto 1$. The weight matrix $\mathbf{R}$ was a diagonal matrix with the inverses of EDC as diagonal elements.

The prior distribution for a marker effect differed from the one in original BayesB (Meuwissen et al. 2001): under original BayesB, the unconditional prior of a marker effect, when different from 0, can be represented as a non-standardized centered Student's t-distribution with fixed dispersion parameter, whereas in generalized BayesB, the dispersion parameter σ_t^2 of the Student's t-distribution was random.

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 and building 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 N_B haplotype blocks for further analysis. The first haplotype block was chosen including the SNP with largest $|\hat{g}_m|$ (denote its index m_1). The SNPs with indices $m_1 - s, \dots, m_1 + s$ formed the first haplotype block in the case that all these SNPs 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 by one or two positions such that the five SNPs were chosen from the same chromosome. The following $N_B - 1$ haplotype blocks were chosen likewise, but with the restriction that haplotype blocks were allowed to share one SNP at most. The values for s were 1 and 2, thus forming haplotype blocks of length 3 and 5 SNPs, respectively.

Estimation of haplotype block variances

Once the haplotype blocks had been pre-selected for further analysis, the variance of effects in each haplotype block was estimated using BayesA (Meuwissen et al. 2001) with block-specific variances. Here, the regression equation for the DRP was

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

where the same prior distributions as above were assumed for the common intercept μ and the vector of residuals $\mathbf{e}$. Denoting the number of distinct haplotypes in haplotype block j with N_j and the indices of these haplotypes with $j_1, \dots, j_{N_j}$, $\mathbf{Z}^{\text{HAP}}$ had $N_H = \sum_{j=1}^{N_B} N_j$ 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 haplotype blocks comprised five adjacent SNPs, the upper limit for N_j was $2^5 = 32$, and in the case of three SNPs $2^3 = 8$.

A normal distribution with mean 0 and variance σ_{gj}^2 was assigned as to each of the N_j haplotype effects $g_{j_1}^{\text{HAP}}, \dots, g_{j_{N_j}}^{\text{HAP}}$in block j, i.e. they shared a common variance. The prior distribution for σ_{gj}^2 was a scaled inverse-chi-square distribution with 4.01 degrees of freedom. Two separate MCMC estimation runs were conducted for each trait by setting the expected value of σ_{gj}^2 to 0.01 and alternatively to 0.1. 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

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

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. The random residuals were assumed to have variance-covariance $\widehat{\sigma_e^2}\mathbf{R}$, where the weight matrix $\mathbf{R}$ was defined as above. Both 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 $g_{j_1}^{\text{HAP}}, \dots, g_{j_{N_j}}^{\text{HAP}}$ shared variance $(1-\omega)\widehat{\sigma_a^2}\widehat{\sigma_{gj}^2}/S$. Here, $\widehat{\sigma_{gj}^2}$ is the posterior mean of σ_{gj}^2, estimated as described 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 $S = \mathrm{tr}(\mathbf{Z}'\mathrm{var}(\mathbf{g}^{\mathbf{HAP}})\mathbf{Z})/N_H$, with $\mathbf{Z}$ being $\mathbf{Z}^{\mathrm{HAP}}$ centered to have column means 0.

Validation of genomic prediction

Genome-enhanced breeding values were calculated using the equation

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

where $\hat{\mathbf{a}}$ and $\widehat{\mathbf{g}^{\mathrm{HAP}}}$ are 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 $r^2_{(\mathrm{GEBV,DRP})}$ was scaled to obtain an estimate for the validation reliability according to $R^2 = r^2_{(\mathrm{GEBV,DRP})}/\bar{w}$, 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.

Genomic prediction with pre-selected SNPs

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 the QTL screen described above to pre-select the SNPs with largest effects. The subsequent procedures (estimation of SNP instead of haplotype variances, the evaluation and validation) were altered to accommodate SNP markers.

Genomic prediction with GBLUP

GEBV were also calculated with SNP-based GBLUP. Here, a weighted mean of the pedigree-based relationship matrix $\mathbf{A}$ 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 $= \mathbf{Z}^{\mathrm{SNP}}\left(\mathbf{Z}^{\mathrm{SNP}}\right)' / (2\sum_{m=1}^{M} p_m(1 - p_m))$, where $\mathbf{Z}^{\mathrm{SNP}}$ 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 2 shows validation results for GEBV of candidate bulls for seven models (a-g). The number of haplotype or single SNP markers was either 1500 (a-c) or 750 (d-f), all 38,194 SNP markers were used in GBLUP (g). For the haplotype-based methods, results for haplotype segments of either 5 adjacent SNPs (a, d) or 3 adjacent SNPs (b, e) are reported. The models a-f were evaluated for $\omega = 0.01, 0.1, 0.2, \dots, 0.9, 0.99$, but only the results from ω which yielded highest validation reliability R^2 are reported. In the GBLUP model, the polygenic weight was assumed constant 0.10.

For milk yield evaluated with haplotype models, highest R^2 was 0.48 (models a, e) and, thus, higher than R^2 for GBLUP (0.45). However, R^2 was also 0.48 for the model with 1500 single SNPs (model c). For protein yield, the model with 1500 haplotype blocks of size 5 yielded highest R^2 (0.41). However, GBLUP and the model with 1500 single SNP markers performed almost as well ($R^2 = 0.40$). For fat yield, R^2 of all haplotype models were below 0.43, the value yielded by the model with 1500 single SNPs and GBLUP. In the case of fertility, highest R^2 with a value of 0.33 was yielded by the model with 1500 haplotype blocks of size 3. 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 SNPs.

With respect to the bias of GEBV (b_1), the haplotype-based models and the models using a limited number of single SNP markers yielded 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.

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

	Milk			Protein			Fat			Fertility		
Model	R^2	b_1	ω	R^2	b_1	ω	R^2	b_1	ω	R^2	b_1	ω
a.1500Hap5SNP	0.48	0.94	0.4	0.41	0.86	0.4	0.41	0.81	0.4	0.31	0.82	0.3
b.1500Hap3SNP	0.47	0.95	0.6	0.40	0.88	0.6	0.42	0.82	0.5	0.33	0.84	0.4
c.1500SingleSNP	0.48	0.93	0.8	0.40	0.84	0.8	0.43	0.82	0.8	0.29	0.82	0.8
d.750Hap5SNP	0.45	0.94	0.5	0.36	0.83	0.6	0.38	0.77	0.4	0.28	0.78	0.5
e.750Hap3SNP	0.48	0.92	0.6	0.39	0.87	0.7	0.41	0.83	0.7	0.29	0.84	0.7
f.750SingleSNP	0.46	0.88	0.8	0.36	0.86	0.9	0.41	0.83	0.9	0.29	0.78	0.8
g.GBLUP	0.45	0.79	0.1	0.40	0.71	0.1	0.43	0.72	0.1	0.30	0.72	0.1

Conclusions

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 used in this study 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.

References

Browning, B.L. & Browning, S.R. 2009. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84: 210-223.

Mäntysaari, E., Liu, Z. & VanRaden, P. 2010. Interbull validation test for genomic evaluations. Interbull Bull. 41: 17-22.

Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Nordic Cattle Genetic Evaluation. 2013. NAV routine genetic evaluation of dairy cattle – data and genetic models. http://www.nordicebv.info: 61 p.

Strandén, I., Mrode, R. & Berry, D.P. 2011. Application of generalized BayesA and BayesB in the analysis of genomic data. Book of abstracts of the 62nd annual meeting of the European Federation of Animal Science, Stavanger, Norway 29 August - 2 September 2011: 393.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.