59th Annual Meeting of the European Association for Animal Production, Vilnius, Lithuania, August 24th-27th 2008, Session 16 "Free communications on Animal Genetics", Abstract number 2560

Impact of missing genotypes on the estimation of genetic parameters and breeding values in MA-BLUP models

S. Neuner[1,*], C. Edel[1], R. Emmerling[1], G. Thaller[2] and K.-U. Götz[1]

[1]Bavarian State Research Center for Agriculture, Institute of Animal Breeding, D-85580 Grub

[2]Christian-Albrechts-University, Institute of Animal Breeding and Husbandry, D-24098 Kiel

Introduction

Advances in molecular genetics have led to the identification of several genes and of genetic markers associated with genes that affect quantitative traits in livestock (quantitative trait locus, QTL). Once QTL are detected, the aim of animal breeders is to integrate linked markers for QTL into the breeding program, in so called marker assisted selection schemes (MAS). The statistical model for using marker information in BLUP (best linear unbiased prediction) genetic evaluations (MA-BLUP) was developed by Fernando and Grossman (1989). MA-BLUP methodology allows to estimate QTL- and polygenic effects simultaneously.

Practical implementations of MA-BLUP in dairy cattle are most often based on approaches where only the genotyped animals and their close relatives are used in MAS schemes. Animals genotyped are usually proven bulls, bull dams and selection candidates for progeny testing. A best case scenario for MA-BLUP evaluations would be that in addition to selection candidates several generations of ancestors are genotyped. But in reality in many cases one will encounter only sparse genotyping of ancestors of selection candidates, because genotyping increases costs and DNA-samples are not available. These observations were made for real data during the implementation of MA-BLUP for Simmental cattle in Germany and Austria. At the same time the questions came up, how strongly MA-BLUP is affected by missing genotypes, and whether a large pedigree with incomplete data should be preferred over a short pedigree with complete genotyping. The aim of our study is to examine these questions with respect to bias and standard error of estimated variance components and accuracies of MA-BLUP breeding values by means of simulation.

Material and Methods

A stochastic simulation model was used to generate the data. Each simulation cycle consists of two phases: data generation and analysis of simulated data sets.

Data Generation

In the simulation, data was generated for a conventional dairy cattle breeding scheme. The general procedure is described by Neuner et al. (2008). In contrast to Neuner et al. (2008) the time horizon for data generation was 34 years in the current study. A single trait model for 305-day milk yield with a heritability of 0.36 and an additive genetic variance of 260,100 kg² was chosen. Genetic parameters were in agreement with the actual first lactation parameters of German Simmental cattle (Interbull, 2007). The overall breeding value of each animal was the sum of a 'residual polygenic breeding value' and a 'QTL breeding value'. A single biallelic QTL with an allele frequency of 0.5 was assumed and the QTL was bracketed by two marker loci located 3 cM and 2 cM apart, each with 10 alleles but different allelic distributions. Allele frequencies for the marker 3 cM apart from the QTL were 40, 19, 15, 12, 7, 2, 2, 1, 1 and 1% (polymorphic information content, PIC = 0.732), and 60, 20, 8, 4, 2, 2, 1, 1, 1 and 1% (PIC =

*Corresponding author: Stefan.Neuner@LfL.bayern.de

0.555) for the marker 2 cM apart, respectively. All calculations assumed a QTL accounting for 20% of the overall additive genetic variance of the trait investigated.

Analysis of simulated data sets

In routine genetic evaluations of dairy cattle all pedigreed animals are included. However, when applying MAS, only a small fraction of animals might be genotyped at genetic markers. As only genotyped animals provide information for the estimation of QTL variance components and breeding values in MA-BLUP models, the 'two-step approach' as described by Liu et al. (2004), Druet et al. (2006) and Neuner et al. (2008) was used in this study. Phenotypic observations in MA-BLUP were daughter yield deviations (DYD, VanRaden and Wiggans, 1991) of bulls together with yield deviations (YD, VanRaden and Wiggans, 1991) of cows derived in routine genetic evaluations for the entire population. The different amount of information available for the calculation of DYD was accounted for by applying the weighting factors EDC (effective daughter contributions, Fikse and Banos (2001)) and γ (Neuner et al., 2008) to twice the DYD. YD were not weighted, because each cow had only one record in the current study. Genetic parameters for MA-BLUP models and MA-BLUP EBV were estimated with the ASREML package (Gilmour et al., 1995) using the MA-BLUP model of Fernando and Grossman (1989). The QTL effect is accounted for in the genetic model as an extra random effect with covariance structure proportional to the IBD (identity by descent) matrix at the QTL position given the linked markers.

Two different pedigrees were derived for MA-BLUP. The difference between them was the depth of the pedigree. The short pedigree was spanning over 3 generations and contained 1,821 animals, whereas the deep pedigree was over 4 generations with 2,671 animals. In each of the pedigrees all animals in the youngest generation had no phenotypic information.

In total four IBD matrices were calculated for each simulated data set. For each pedigree size one IBD was calculated for the situation of complete genotyping. To analyze the effect of missing genotypes another two matrices were generated in order to reflect two different genotyping structures in the deep pedigree: moderate and extensive gaps. We generated the missing genotypes in order to mimic practical conditions, i.e. old animals at the top of the pedigree are more often not genotyped than animals at the bottom of the pedigree, and missing genotypes occur more often for females than for males. In total, the proportion of missing genotypes is about 41% for moderate and 61% for extensive gaps. All IBD matrices applied for MA-BLUP evaluations were calculated using the package LOKI (Heath, 1997).

Parameters studied

Parameters considered for the estimation of variance components were the bias of estimated variance components and their asymptotic standard errors. The deviation of estimates from the simulated parameters was used to check for bias due to the pedigree depth and/or missing genotypes. Standard errors of the estimates were used to assess the precision of estimates between different models. In order to assess the fit of the genetic model the likelihood ratio test was calculated as described by George et al. (2000). To examine the impact of the investigated factors on the estimation of MA-BLUP breeding values, the correlation between true and estimated breeding value was calculated for young bull candidates.

Results

Variance Component Estimation

Findings for the estimation of variance components are summarized in table 1. The values of the estimated parameters are nearly the same whether the short or deep pedigree is applied, and

whether genotypes are missing or not. Bias of variance components could not be observed for any of the applied models.

Table 1: Simulated and estimated parameters for the estimation of variance components. Parameters shown are the additive genetic variance ($\hat{\sigma}_a^2$), the residual variance ($\hat{\sigma}_e^2$), the genetic variance explained by one QTL ($\hat{\sigma}_{qtl}^2$), the log likelihood ratio (log *LR*), the ratio of $\hat{\sigma}_a^2$ and $\hat{\sigma}_{qtl}^2$, and the estimated standard errors for the estimated variance components (*s.e.*). The values are averages over 100 replicates.

Pedigree depth	Missing genotypes	$\hat{\sigma}_a^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_{qtl}^2$	log *LR*	$\hat{\sigma}_{qtl}^2 / \hat{\sigma}_a^2$	$s.e.(\hat{\sigma}_a^2)$	$s.e.(\hat{\sigma}_e^2)$	$s.e.(\hat{\sigma}_{qtl}^2)$
short	none	259,493	459,265	52,924	2.539	0.204	20,033	31,899	31,892
deep	none	260,677	458,426	50,293	7.284	0.193	16,100	25,221	19,392
deep	moderate	260,738	458,418	50,462	6.400	0.194	16,109	25,246	20,996
deep	extensive	260,899	458,271	51,020	5.373	0.196	16,133	25,284	23,333
Simulated parameters		*260,100*	*462,400*	*52,020*		*0.200*			

The standard errors of estimated variance components were used to asses the precision of the estimates. From lines 1 and 2 in table 1 follows, that increasing the pedigree depth leads to more accurate estimates. In contrast, standard errors increase for missing genotypes, especially for the genetic variance explained by the QTL. The observed LRT values indicate the same as the standard errors: The best fit was observed for the model with the deep pedigree and no missing genotypes.

Accuracy of MA-BLUP EBV

To evaluate the consequences of the different conditions on the estimation of MA-BLUP breeding values, accuracies were calculated for young bull candidates. Correlations between simulated and estimated breeding values were assessed for the overall MA-BLUP breeding value, the residual polygenic breeding value, and the breeding value at the QTL.

Table 2: Accuracies of estimated breeding values in MA-BLUP evaluation models. Accuracies are shown for the overall breeding values of MA-BLUP evaluations, the residual polygenic breeding value and breeding value for the QTL position (QTL-EBV). In breeding value estimation, the estimated variance components were used. The results are averages over 100 replicates per scenario.

Pedigree depth	Missing genotypes	Young bull candidates		
		MA-BLUP	Residual polygenic	QTL-EBV
short	none	0.556	0.497	0.348
deep	none	0.566	0.501	0.437
deep	moderate	0.563	0.501	0.418
deep	extensive	0.560	0.500	0.390

Average MA-BLUP accuracies were hardly affected by the model, pedigree depth or missing genotypes. The increase in pedigree depth caused a slight improvement of accuracy for young bulls, because the gametic effects could be estimated more accurately. However, this slight increase got gradually lost as the amount of missing genotypes increased. From table 2 follows, that mainly QTL-EBV are affected by changes in the data structure.

Discussion

Pedigree depth

The advantage of a more extensive pedigree of genotyped animals for MA-BLUP is obvious. Similar to the effect of having more offspring for progeny tested bulls, a deeper pedigree implies more phenotypic and genotypic data and more informative matings for the estimation of QTL effects. As a consequence deeper pedigrees improve the estimation of genetic and residual variances compared to shorter pedigrees.

The effect of a more parsimonious pedigree was also shown by George et al. (2000). They altered the number of offspring per mating from 1.8 to 14.3 offspring per mating. As a consequence, there were more progeny per parent providing information to estimate genetic parameters and MA-BLUP EBV. Even if the approach of George et. al [8] was different from the one in our study, the effect of a larger pedigree was the same as in our approach: more accurate estimates and increased power.

Missing genotypes

If marker information was complete and could be used to infer the transmission of QTL alleles, the IBD matrix would only contain 1s and 0s. At the other extreme, if there was no marker information, the IBD matrix would become identical to the numerator relationship matrix. Several approaches exist to deal with the problem that non-genotyped animals do not contribute information for QTL models (George et al., 2000). A well-known approach is the multiple-site segregation sampler LOKI (Heath, 1997) that was used in this study. We found that missing genotypes did not lead to biased variance components. Moreover, for a combined analysis of pedigree depth and missing genotypes our results show that using a deep pedigree with many gaps is preferable over a short but complete pedigree. In contrast to our results, George et. al (2000) reported that the QTL variance was overestimated, residual polygenic variance was underestimated and bias increased the more genotypes were missing. The main reason for these disagreeing results could be that the structure of pedigree and missing data in our study allowed a much better reconstruction of missing genotypes by LOKI. In the pedigree of George et al. (2000) the number of progeny and grand progeny was less than for the cattle breeding program in this research. Thus, fewer descendants are available to contribute information for the reconstruction of their ancestors' genotypes. Furthermore, the amount of phenotypic information per sire is very much different in both studies. Compared to George et al. (2000) both the better ability to reconstruct missing genotypes and the higher amount of phenotypic information for MA-BLUP result in unbiased estimates in our study.

Effects on accuracy of MA-BLUP

Our results show that the overall accuracy of MA-BLUP breeding values is hardly affected by neither the pedigree depth nor missing genotypes. With respect to accuracy of EBV in MA-BLUP models, the benefit of increased pedigree and marker information was investigated by Villanueva et. al (2002). They simulated four additional generations of random selection for extending their data set. The increased amount of marker genotype information significantly increased the accuracy of the estimation of the QTL effects from 0.54 to 0.65. Parameters for their study were 0.25 for the heritability and 0.24 for the ratio of genetic variance explained by the QTL. Spelman (1998) also concluded that more animals genotyped in each generation and more generations of genotypic information for MAS will result in an increase in accuracy of the estimation of QTL effects and therefore in MAS superiority.

Conclusions

To estimate variance components and breeding values in MA-BLUP models deep pedigrees have to be preferred over short pedigrees because they provide more phenotypic and genotypic information. As a consequence the estimation of all variance components is improved.

Missing genotypes are always a loss of information and can reduce possible benefits. To minimize this problem programs that allow for genotype reconstruction should be applied. But it has to be noted that genotype reconstruction requires a good data structure.

The main conclusion of this study is that deep pedigrees with incomplete genotyping perform better than short pedigrees with complete genotyping.

Acknowledgements

The authors gratefully acknowledge financial support from of the German Federal Ministry of Education and Research (project FUGATO MAS.-Net, grant no. 0313390F) and of the Förderverein Biotechnologieforschung, Bonn.

References

Boichard, D., S. Fritz, M. N. Rossignol, M. Y. Boscher, A. Malafosse, und J. J. Colleau. 2002. Implementation of Marker Assisted Selection in French Dairy Cattle Breeding. Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France, Communication No. 22–03.

Druet, T., S. Fritz, D. Boichard, und J. J. Colleau. 2006. Estimation of genetic parameters for quantitative trait loci for dairy traits in the French Holstein population. J. Dairy Sci. 89:4070–4076.

Fernando R. L., und M. Grossman. 1989. Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. 21:467–477.

Fikse, W. F., und G. Banos. 2001. Weighting factors of sire daughter information in international genetic evaluations. J. Dairy Sci. 84:1759–1767.

George, A. W., P. M. Visscher, und C. S. Haley. 2000. Mapping Quantitative Trait Loci in Complex Pedigrees: A Two-Step Variance Component Approach. Genetics 156:2081–2092.

Gilmour, A. R., R. Thompson, und B. R. Cullis. 1995. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51:1440–1450.

Heath, S. 1997. Markov Chain Monte Carlo Segregation and Linkage Analysis for Oligogenic Models. Am. J. Hum. Genet. 61:748–760.

Interbull. 2007. Description of National Genetic Evaluation System and Trend Validation for Production Traits. http://www-interbull.slu.se/national_ges_info2/framesida-ges.htm, Accessed Nov. 22, 2007.

Liu, Z., F. Reinhardt, J. Szyda, H. Thomsen, und R. Reents. 2004. A marker assisted genetic evaluation system for dairy cattle using a random QTL model. Interbull Bulletin 32:170–174.

Neuner, S., R. Emmerling, G. Thaller, und K.-U. Götz. 2008a. Strategies for Estimating Genetic Parameters in Marker-Assisted BLUP Models in Dairy Cattle. J. Dairy Sci. (Accepted June 30, 2008).

Spelman R.J. 1998. Major factors in marker-assisted selection genetic response in dairy cattle populations, Proc. 6th World Congress on Genetics Applied to Livestock Production,11–16 January 1998, Vol. 26, University of New England, Armidale, pp. 365–368.

VanRaden, P. M., und G. R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 74:2737–2746.

Villanueva B., R. Pong-Wong, und J. A. Woolliams. 2002. Marker assisted selection with optimised contributions of the candidates to selection. Genet. Sel. Evol. 34:679–703.