Genomic selection in mink yield higher accuracies with a Bayesian approach allowing for heterogeneous variance than a GBLUP model.

T. Villumsen, G. Su, Z. Cai, B. Guldbrandtsen, T. Asp G. Sahana, M.S. Lund.

Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark
tmv@mbg.au.dk (Corresponding Author)

Summary

The accuracy of genomic prediction for mink was compared for single-trait and multiple-trait GBLUP models and Bayesian models that allowed for heterogeneous (co)variance structure over the genome. The mink population consisted of 2,103 brown minks genotyped with the method of genotyping by sequencing. Four live grading traits and four traits on dried pelts for size and quality were analysed. GWAS analysis detected significant SNPs for all the traits. The single-trait Bayesian model resulted in higher accuracies for the genomic predictions than the single-trait GBLUP model, especially for the traits measured on dried pelts. We expected the multiple-trait models to be superior to the single trait models since the multiple-trait model can make use of information when traits are correlated. However, we did not find a general improvement in accuracies with the multiple-trait models compared to the single-trait models.

Keywords: GWAS, GBLUP, BayesAS, heterogeneous (co)variances

Introduction

The primary objective of mink breeding is to produce large and high quality skins. Parental animals for future generations are usually selected based on size of the animal and secondly quality at the live grading whereas the breeding goal is size and quality in the dried skins. However, size and quality traits on dried skins are difficult and expensive to measure. Body weight in the live grading is highly correlated with skin size, which is under intense selection in the mink-breeding program. We expect genomic selection in mink can increase accuracy of breeding value, and the accuracy of genomic prediction can increase when using a multiple-trait model where information from genetically correlated traits can be utilized (Calus and Veerkamp, 2011).

The mink genome was sequenced in 2016 (Cai *et al.*, 2016), which has opened opportunities for genomic selection in mink. However, no SNP-chip is developed for mink so far. Therefore, genomic selection in mink requires mink to be genotyped with the method of genotyping by sequencing (GBS).

The standard model for genomic selection is a GBLUP model, which implicitly assumes equal variance and covariance for all SNP-markers across the genome. For traits where a large part of the variance is explained by relatively few SNP-markers with major effects, this may not be a desired approach. Janss (2014) has developed a Bayesian model (BayesAS) that can model a heterogeneous (co)variance structure over the genome. This model accounts for specific local genetic (co)variance between the traits. Therefore, we expect the BayesAS model to be superior to the GBLUP model in the estimation of genomic breeding values. In this study, we compared the predictive ability of single-trait and multiple-trait GBLUP models (ST-GBLUP and MT-GBLUP) with the predictive ability of single-trait

and multiple-trait Bayesian random regression BLUP models (ST-BayesAS and MT-BayesAS).

Material and methods

We analysed size and quality traits for 2,103 genotyped brown mink (*Neovison vison*) from the research farm at Aarhus University in Denmark. The genotyped mink were primarily breeding animals born in 2010-2014. The mink were genotyped with the method of genotyping by sequencing (GBS). The genomic data consisted of 28,336 SNP-markers of good quality. In genomic prediction using BayesAS, we treated a scaffold as a genome region. Scaffolds with five or less markers was joined to one "pseudo scaffold" with 767 markers, leaving us with 391 scaffolds with 6-831 markers in the analyses (mean=72.5, SD=123.6).

We used corrected phenotypic observations (Yc) of size and quality traits as response variables for the GBLUP and BayesAS models. There were two categories of traits. One included four traits from the live grading: body weight (weight), overall general impression of pelt (quality), under wool density (density) and silky appearance of the pelt (silky). The other included four traits measured on dried skins: pelt length (p_length), pelt quality (p_quality), pelt density (p_density) and pelt silkiness (p_silky). We had 2,095 observations of weight, 1,325 observations on the other traits from the live grading and about 1,830 observations of each pelt trait.

The Yc were obtained by correcting the original phenotypic value for birth year, sex, house after weaning, and for the pelt traits correcting also for age at pelting (i.e., first year or older) . We estimated the correction factors for each trait using a BLUP model containing all records for brown mink born 2013 -2016 at the research farm. In total, there were 17,598 to 20,638 records for each trait. The pedigree file for the BLUP estimation included 85,132 mink born from 2001 to 2016. Within each trait, the Yc were scaled to a mean of zero, and a variance of 1.

A GWAS analysis was carried out on the Yc records, to identify significant SNP-effects of the traits, the model for GWAS was a standard single SNP regression model, and will not be described here.

We evaluated the predictive ability of the genomic single-trait and multiple-trait GBLUP and BayesAS models using 5-fold cross validation. In each fold of validation, 1/5 of animals in 2014 were taken as test set and their Yc were discarded, the remaining 4/5 records formed the training set. Prediction accuracies for each model was measured as: $cor(Y_c, GEBV)/\sqrt{h^2}$, where GEBV were estimated for the test animal in each fold validation.

For the multiple-trait analyses, weight was always as secondary trait in a bivariate model, because this live grading measure of size is under intense selection in the mink breeding and have a correlation with many traits of interest.

Models

The MT-GBLUP model was:

(1)

where was vector of corrected phenotypes of trait 1 and 2; 1_1 and 1_2 were the vector with element 1, was the overall means of trait 1 and 2; was the vector of additive genomic breeding values of trait 1 and 2; Z_1 and Z_2 were the design matrices linking to and to was a vector of random residual effects of the two traits; respectively. It was assumed that ~where G is genomic relationship matrix, H= was the (co)variance matrix of the additive genomic breeding values for the two traits, and ~, where R= was the residual (co)variance matrix of the two traits (Guo *et al.*, 2014). The ST-GBLUP model was a simplified model of equation 1.

The MT-BayesAS model was:

(2)

Where , and were the same as in model 1 for trait *i*; *nscaffold* was the total number of scaffolds; W_j was a matrix with SNP genotypes of scaffold *j*; a_{ij} was a vector of SNP effects for SNPs within scaffold *j* in trait *i*. To handle the correlation between SNP effects of two traits for scaffold *j*, a hierarchical model was formulated:

(3)

In the hierarchical model, s_o and s_1 were vectors and were the same for all traits. Thus s_o modeled the overall covariance between traits and s_1 modeled the scaffold specific covariance between traits in scaffold *j*. r_i and r_{ij} were overall and scaffold specific regression coefficients; were residual SNP effects for SNPs within scaffold *j* in trait *i*.

The prior distributions were assumed as

,

,

~Scale-inv-X^2 (v, 1/v)

For the single trait model the hierarchical model to handle correlation between SNPs reduces to

(4)

Li et al. (2017) described how genomic (co)variance and correlations for marker, region (here scaffold) and total genome were calculated.

The GBLUP analyses were carried out using the DMU software (dmu.agrsci.dk/DMU/Doc/Current/dmuv6_guide.5.2.pdf) The BayesAS analyses were carried out using the BAYZ software (www.bayz.biz). For the BayesAS analyses the Markov chains were run for 50,000 cycles of Gibbs sampling, the first 10.000 samples were discarded as burning in. For the remaining 40,000 cycles, every 20th cycle was saved for the posterior analyses. The mean was considered as the estimate of each unknown parameter

Results and discussion

The GWAS resulted in the detection of significant SNPs for all traits ($p < 10^{-5}$). The highest

number of SNPs significantly associated was observed for weight and p_length. These traits also had the highest h^2 (Table 1). In general, SNPs explain a larger proportion of the genetic variance for pelt traits than for traits from the live grading (Table 1). More SNPs were associated with pelt quality traits than with quality traits measured during live grading. SNPs associated with size traits explain more than half of the genetic variance. The variance explained by SNPs may be overestimated as variance from one gene may be associated with several scaffolds. We expect traits where a larger proportion of the variance is explained by significant SNPs to benefit more from the BayesAS model that accounts for SNP variance heterogeneity.

For the single-trait models, BayesAS model with heterogeneous variance improved accuracy relative to the standard GBLUP model for most traits (Table 3). The increase was most profound for p_length. The average increase in accuracy from the ST-GBLUP to the ST-BayesAS model was 6 % for live grading traits, but 27 % for the pelt traits. The increase in accuracy supported that the BayesAS model better captures the heterogeneous variance between SNPs with large effects and other SNPs. However, we also had many significant SNPs for weight, but the increase in accuracy from the ST-GBLUP to the ST-BayesAS model was much lower.

Genomic correlations between weight and other traits were estimated using multiple-trait models. Correlations varied from almost zero between weight and quality to more than 0.8 between weight and p_length (Table 2). Genetic correlations between weight and the live graded quality traits were close to zero or positive, whereas the correlations between weight and the pelt quality traits were close to zero or clearly negative. Correlations estimated with MT-BayesAS tended to be further from zero than those estimated by the MT-GBLUP model.

Highly correlated traits were expected to gain more from a multiple-trait model as it utilizes information from the correlated trait. For traits with no genetic correlation with weight, we did not expect the multiple-trait models to perform better in accuracy than the single-trait models. We did expect an increase in accuracy in the multiple-trait models when the two traits in the model were highly correlated such as weight and p_length. For both live graded and dried skin traits neither going from ST-GBLUP to MT-GBLUP nor going from the ST-BayesAS to MT-BayesAS model improved accuracies (Table 3).

Conclusion

Prediction accuracies indicates that BayesAS models allowing heterogeneous variance over the genome were superior to GBLUP models, which implicitly assumed equal variance and covariance for all SNP-markers across the genome. The GWAS analyses uncovered many significant SNP-effects for all traits. This support the superiority of the BayesAS approach. In general, the accuracies of BayesAS GEBV increased more for pelt traits than for live grading traits.

We expected increased accuracies for the multiple-trait models compared to accuracies from single-trait models when the two traits in the model were correlated, since the multiple-trait models can gain from correlation between traits. In this study, we did not observe a clear tendency of this pattern, neither for the GBLUP nor for the BayesAS models. The reason behind this is not clear.

Table 1. Number of significant SNP-markers and number of scaffolds with significant SNPs from GWAS for size and quality traits, the explained variance is sum of most significant SNP-marker in each scaffold.

	Trait	Significant SNPs	Scaffolds	Explained variance	h^2
Liv	Weight	25	13	0.293	0.532
e	Quality	2	2	0.019	0.303
gra	Density	4	4	0.012	0.163
d.	Silky	1	1	0.005	0.297
Dri	P_length	23	9	0.330	0.457
ed	P_quality	2	2	0.136	0.327
ski	P_density	10	8	0.023	0.222
ns	P_silky	5	4	0.014	0.181

Table 2. Genetic correlations between weight and the other trait estimated from MT-GBLUP and MT-BayesAS analyses, all genotyped mink were included.

	Trait	Cor GBLUP	Cor BayesAS
Liv	Quality	0.0003	0.028
e	Density	0.498	0.374
gra	Silky	-0.019	-0.081
d.			
Dri	P_length	0.710	0.837
ed	P_quality	-0.318	-0.452
ski	P_density	-0.286	-0.395
ns	P_silky	-0.019	-0.016

Table 3. Accuracy of prediction in the 5-fold cross validation.

	Trait	No of Animals	ST-GBLUP	MT-GBLUP	ST-BayesAS	MT-BayesAS
Liv	Weight[1]	752	0.490	0.500[1]	0.516	0.523[1]
e	Quality	752	0.685	0.694	0.695	0.706
gra	Density	752	0.393	0.462	0.450	0.459
d.	Silky	752	0.820	0.818	0.844	0.843
Dri	P_length	639	0.480	0.489	0.618	0.624
ed	P_quality	665	0.227	0.234	0.270	0.258
ski	P_density	665	0.299	0.302	0.357	0.362
ns	P_silky	665	0.138	0.135	0.197	0.194

[1]Weight: analysed with p_length in the multiple-trait model

Acknowledgements

This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark (grant 0603-00519B).

List of References

Cai, Z., F. Panitz, B. Petersen, G. Sahana , B. Thomsen, C. Bendixen, M.S. Lund, 2016. The draft genomic sequence of the American mink (Neovison vison) opens new

opportunities of genomic research in mink. Proc XIth Intl Sci Congr in Fur Anim Prod, Helsinki, Finland. Scientifur, 40 (3/4).

Calus, M.P.L., R.F. Veerkamp, 2011. Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol. 43: 26.

Guo, G., F. Zhao, Y. Wang, Y. Zhang, L. Du, G. Su, 2014. Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet. 15: 1–7.

Janss, L., 2014. Disentangling Pleiotropy along the Genome using Sparse Latent Variable Models. Proc 10th World Congr Genet Appl to Livest Prod: 0–3.

Li, X., M.S. Lund, L. Janss, C. Wang, X. Ding, Q. Zhang, G. Su, 2017. The patterns of genomic variances and covariances across genome for milk production traits between Chinese and Nordic Holstein populations. BMC Genet. 18:26.