Genomic Estimation Of Heritability Under Different Gene Action Scenarios

L. Janss[*], M. M. Shariati[*]

Introduction

Different versions of genomic models are proposed to predict breeding values for animals using dense markers. However, so far little attention has been payed to estimating the genomic explained variance from the data, or to compare "genomic" heritability estimates with pedigree-based heritability estimates. One group of genomic models is based on mixed models, using markers to build genomic relationship matrices (e.g., Garrick (2007), VanRaden (2008)). These models could be straightforwardly applied to estimate genomic (co)variances and heritabilities by REML, but have limited possibilities for extensions. For instance, extensions to multi-trait estimation, to add interaction variances, or to estimate genomic (co)variances by chromosome would result in an unmanageably large number of relationship matrices and (co)variance components to deal with. Also, these genomic relationship matrices may be singular, requiring bending which effectively reduces the off-diagonals, and then could be expected to lead to biases in genomic variance estimates.

A second group of genomic models is based on Bayesian approaches that estimate individual marker allele effects and marker variances (e.g. Meuwissen et al. (2001), De Los Campos et al. (2008)). These approaches are in principle much more versatile, but the so far proposed models apply fixed settings for parameters that describe distributions of allele effects or variances of markers. Hence, in these proposed Bayesian methods marker effects or variances are completely determined by prior knowledge (Gianola et al. (2009)) and genomic explained variance cannot be determined.

In this study we present Bayesian genomic models which are extended to estimate variance components for allele-effects from the data and that are used to estimate "genomic" variances and heritabilities. Extensions for a long-tailed allele effects model comparable to BayesA (Meuwissen et al. (2001)) and a mixture model based on Bayesian Variable Selection (George and McCulloch (1993)) comparable to BayesB (Meuwissen et al. (2001)) are shown. This through-development of the Bayesian genomic models can be an interesting alternative to mixed-model approaches for estimation of genomic variance because of the larger versatility and applicability of Bayesian models. Using simulations that also include various interaction effects, and by comparison with pedigree-based REML estimation, these Bayesian models are validated for estimation of narrow-sense heritability, based on estimation of the allele substitution effects at markers.

[*]Aarhus University, Dept. of Genetics and Biotechnology, DK-8830 Tjele, Denmark

Material and methods

Data simulation. Linkage disequilibrium between markers and QTLs was generated by simulation of 100 generations of random mating in a population of 100 individuals, using 5000 bi-allelic markers on 5 chromosomes, and 160 bi-allelic QTLs. Due to drift marker and QTL frequencies exhibited a typical U-shaped distribution at the 100th generation. After generation 100, the population was expanded to 1000 individuals per generation, 500 males and 500 females mated in pairs, and 3 of these generations (3000 individuals) were used in the data analyses. Various additive and non-additive effects, sampled form Gamma distributions, were assigned to the QTLs: two scenarios with pure additive gene actions, with heritability 0.15 and 0.30; and four scenarios with additional interaction effects where broad sense heritability was 0.30. In order to assess the narrow sense heritability in the interaction scenarios REML analyses using the pedigree structure were performed. In the interaction scenarios 80 of the 160 QTLs were assigned dominance effects, or were assigned between-QTL interaction effects in pairs. The interaction scenarios considered dominance, additive by additive, additive by dominance and a general interaction effect between genotypes labelled as "epistatic".

Models. The first genomic model used is a model with a long-tailed distribution for allele effects. This long-tailed distribution is achieved by use of marker-specific dispersion terms, similarly as in Bayesian LASSO (Park and Casella (2008)) and BayesA (Meuwissen et al. (2001)). However, in our implementation, the marker-specific dispersion terms are modelled as scaling factors with the following model:

$$y = \mu + X\Phi a + e \tag{1}$$
$$\phi_i \sim N(0, \sigma_a^2) \tag{2}$$
$$a \sim N(0, I) \tag{3}$$
$$e \sim N(0, I\sigma_e^2) \tag{4}$$
$$\mu, \sigma_e^2, \sigma_a^2 \sim \text{flat} \tag{5}$$

where y is the vector of phenotypes, μ is a general mean, X is a matrix with genotype covariates (coded for SNP genotypes as -1/0/1), a is a vector of allele substitution effects on a standard Normal scale, $\Phi = \text{diag}\{\phi_i\}$ is a diagonal matrix with scaling factors and e is a vector of residuals. The main advantage of the use of scaling factors to model dispersion is that it introduces only one unknown parameter σ_a^2 that determines the average size of allele effects, and this parameter can be quite robustly estimated from the data.

The second genomic model used considered a mixture distribution on the dispersion terms ϕ_i as:

$$\phi_i \sim \gamma_i N(0, \sigma_a^2) + (1 - \gamma_i) N(0, \tau^2) \tag{6}$$

where γ_i is an indicator variable that selects whether the effect for marker i should be large or (negligibly) small. Also in this model the variance components σ_e^2 and σ_a^2 are estimated from the data assigning flat priors for these variances. The prior for indicator variables is considered known as $\Pr(\gamma_i = 1) = \pi_1$ and τ^2 is set to a small enough value so that all "small" effects will not explain more than 1% of the total variance.

MCMC and estimation of genomic variance and heritability. The model is fitted using MCMC estimation which follows standard procedures to treat mixed linear models and variance estimation in a Bayesian framework: the first level effects can all be conditionally updated as random regression terms form Gaussian distributions, and variance components are sampled from inverse chi-square distributions. In the mixture model, the indicator variables are sampled as explained in George and McCulloch (1993) using a fast and simple Gibbs update. In order to efficiently construct the vector of observations corrected for all but one effect the technique of residual updating was used (Janss and De Jong (1999)).

From the model the variance in the phenotypes is decomposed as the Var($X\Phi a$) + Var(e), where the first component is interpreted as the "genomic variance" σ^2_g. A posterior statistic for this genomic variance, e.g., posterior mean, is then computed as E[σ^2_g] = E[Var($X\Phi a$)] over the MCMC samples for Φ and a. The same is applied to obtain other posterior statistics, such as posterior variance, and in the same way also the posterior statistics for genomic heritability are obtained. Note that $X\Phi a$ represents fitted genomic values, or genomic breeding values, for individuals, which are therefore an intermediate step in the construction of genomic variance. These individual genomic values include the linkage disequilibrium between markers, which therefore also enters in the estimate of genomic variance.

Results and discussion

Table 1 presents genomic heritability estimates for different gene-action scenarios. These results show that genomic heritability estimates closely match the simulated values for the pure additive cases, and closely match the (narrow sense heritability) REML estimates for the interaction cases.

Figure 1 shows heritabilities with the mixture model when the π_1 parameter (proportion of loci with large effect) is varied with data with additive gene action and heritability 0.30. This figure shows a clear tendency for the mixture models to underestimate heritability when π_1 is reduced, and a small tendency to over-estimate heritability when π_1 is increased to values around and above 0.50. Although theoretically the mixture model with large π_1 approximates the Long-Tail model, such over-estimation was not seen in the Long-Tail model and therefore appears an effect of separating large-effect and negligible small effect loci. The under-estimation of genomic variance in the mixture model for low π_1 can be because tracing of QTL becomes inaccurate when too few markers are allowed to have large effects.

Table 1: Estimates of "genomic heritabilities" under different gene action scenarios[a]

Model	add h^2 15%	add h^2 30%	dom	add x add	add x dom	epistatic
Long-Tail	0.15	0.31	0.19	0.26	0.22	0.23
Mixture	0.15	0.29	0.19	0.25	0.19	0.23
REML	0.15	0.31	0.19	0.27	0.22	0.24

[a]Pure additive gene action at heritabilities of 15 and 30 % (add), and cases with added dominance (dom), additive by additive (add x add), additive by dominance (add x dom) and general gene-gene interactions (epistatic) with a broad sense heritability of 30 %, analysed with additive genomic models with Long-Tail and Mixture distributions of allele effects, and based on pedigree with REML. The mixture model was run with the prior probability for a marker to have a large effect of 0.35.

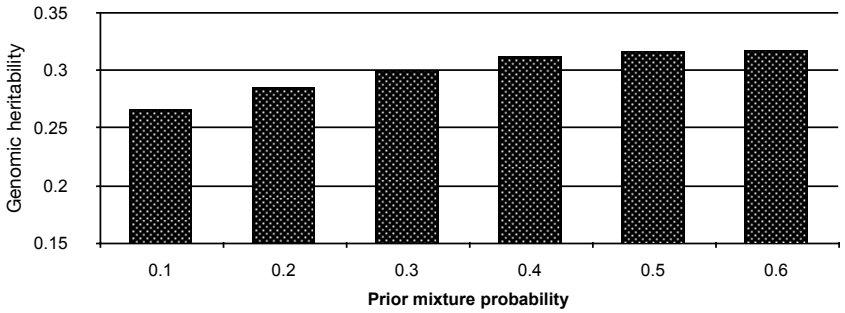

Figure 1: Estimated genomic heritabilities using a mixture model for different prior mixture probability determining the proportion of loci assigned with large effects

Conclusion

This study shows that estimation of heritability from a genomic association model is well feasible, and, under some precautions, is close to the traditional pedigree-based (REML) estimate of heritability. That is, a genomic model that estimates allele substitution effects correctly picks up additive genetic variance (or narrow sense heritability), also when the underlying gene-action is not strictly additive. The main precaution is that in mixture models the number of loci with large effect should not be too small in order to avoid significant under-estimation. Because these genomic association models do not need pedigree, and are expected to be more accurate than pedigree-based heritability estimates, this technique allows to estimate heritabilities in situations where pedigrees are missing, lost (e.g. slaughter house data), or data is small and family structure is weak (e.g., feeding or behavioural trials). Extension to multi-trait estimation is feasible and offers the possibility to estimate genetic correlations between traits that are measured on different and weakly related animals.

References

Garrick, D.J. (2007). *J. Dairy Sci.*, 90(Suppl.1):376.

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

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

De Los Campos, G., H. Naya, D. Gianola, et al. (2008). *Genetics*, 182:375–385.

Gianola, D. G. De Los Campos, W. G. Hill, et al. (2009). *Genetics*, 183:347–363.

George, E. I., and R. E. McCulloch (1993). *J. Am. Stat. Assoc.*, 88(423):881–889.

Park., T. and G. Casella (2008). *J. Am. Stat. Assoc.* , 103(482):681–686.

Janss, L. L. G. and G. De Jong (1999). *Interbull Bull.* , 20:63–68.