Heritability Estimation Based On Small Sample Size Using SNP Markers

K. Krag[*], L.L. Janss[*], M.M. Shariati[*], and A.J. Buitenhuis[*]

Introduction

Heritability is a central parameter for breeding, dating back to the foundation of quantitative genetics (Fisher (1918)). The current state-of-art to estimate heritabilities, as well as genetic correlations, is the use of REML (Thompson (2008)) with an "animal model". Even though these estimation methods are now highly advanced, their applicability is limited to populations with recorded pedigrees, revealing an informative family structure, and preferably large size. In all other situations, i.e. populations with weak family structure, un-recorded pedigrees or small sample size, estimation of heritabilities and genetic correlations is impossible or very inaccurate.

Use of genetic markers to estimate heritabilities and genetic correlations could provide new applications of quantitative genetics and breeding. Marker-based heritabilities are expected to be more precise because they also capture Mendelian sampling variation within families, which cannot be captured using pedigree information. Estimation methods based on markers have been proposed through the construction of marker-based relationships (Visscher, P.M., Medland, S.E., Ferreira, M.A.R. *et al*. (2006)). Such methods, however, are based on Identity by Descent (IBD), which still requires some minimal population structure to reconstruct haplotypes and to trace allele inheritance.

With recent year's development and improvement of large-scale genotyping techniques for bi-allelic SNP markers, the use of genomic information for breeding is actively investigated. The methods employed for these genomic predictions (Meuwissen, T.H.E. and Goddard, M.E. (2001)) are based on association and, implicitly, on allele-sharing or Alike in State (AIS). However, extension of such association/AIS methods to estimate variance components from the data and to estimate "genomic heritabilities" is not yet considered, and the properties of such association/AIS methods estimation of "genomic heritability" are not yet known.

Today the 54 K SNP array has become a common tool for genomic prediction in cattle populations. In the near future a 600 K bovine SNP array will become available. This array offers a more dense distribution of markers along the cattle genome. A question that comes along in connection with the introduction of this array is if this array is able to obtain more precise estimates of variance and heritability in small size datasets.

The aim of this paper is to use association-based genomic models for estimation of variance and heritability in small sample size datasets by use of simulation studies. Two types of genomic models are used, one with and one without a mixture distribution, and both models were extended to estimate their "tuning variances" from the data. Furthermore, we test for

[*] Department of Genetics and Biotechnology, Faculty of Agricultural Sciences, Aarhus University, DK-8830 Tjele, Denmark

differences in precision of genetic parameters based on the density of the 54 K bovine SNP array and the 600 K bovine SNP array.

Material and methods

Scenarios. To test the capabilities of different genomic models for estimation of variance and heritability four scenarios was created. The four scenarios were composed of a combination of two different levels of effective population size (Ne) – 50 and 100 – and two different levels of marker density corresponding to a 54 K bovine SNP array and a 600 K bovine SNP array. Scenario I and II consisted of a Ne of 50 and with the two different marker densities. Scenario III and IV consisted of a Ne of 100 and again with the two marker densitites.

Simulation. A population was simulated to investigate the performance of genomic selection for small sample sizes. The length of the bovine genome is estimated to be 3,160 centimorgan (cM) (Ihara, N., Takasuga, A., Mizoshita, K. *et al*. (2004)). Simulations were based on a genome corresponding to between one-eighth and one-seventh of the bovine genome to reduce computation time. The genome was assumed to consist of 4 chromosomes each approximately 400 cM in length, making a total length of approximately 1,600 cM. For the scenarios I and III, bi-allelic markers (SNPs) of about 3,664 were equally spaced over the genome with a space of 0.44 cM, which corresponds to the 54 K SNP array. For the scenarios II and IV, SNP markers of about 40,704 were equally spaced over the genome with a space of 0.04 cM, which corresponded to the 600 K SNP array. To create linkage disequilibrium (LD) a historic population with a Ne of 50 or 100 consisting of equal numbers of sires and dams was simulated for 100 generations. For each scenario 20 replicates were simulated. All populations were equal in size consisting of a base population from which 100 sires were crossed to 400 dams – one sire to four dams and with one daughter from each cross resulting in a population of 400 cows. The cows were assumed to have direct phenotypes. The phenotype had a heritability of 0.3 and having 200 QTL of different effect size equally distributed over the genome.

Analysis models. All analyses of genomic prediction models were carried out with the software iBay v. 1.47. The technique included in the software follows a version of "Bayesian Variable Selection Method" in which "off" markers are set to have a small effect/variance instead of no effect/variance (George, E. I. and Mcculloch, R.E. (1993); Janss, L. (2009)). In all, three models were tested (model I, II, and III) – a truncated normal distributed model and two mixture models in which 30 % and 75 % of the markers were "switched off". For all populations three levels of minor allele frequencies (MAF) were included – 0.05, 0.02, and 0.01. The prediction models make use of both phenotypes and genotypes. A Markov Chain Monte Carlo (MCMC) sampler is integrated to obtain the variance estimates. For each analyze MCMC chains were run with 50,000 cycles, discarding the first 3,000 as burn-in.

Statistics. Simulation results were analyzed with the statistical software R v.2.10.1. Within each scenario and for each model and MAF level, heritability estimates was tested for being normal distributed. Prediction capabilities of each model, array and MAF level was examined for deviation from the correct level of heritability at 0.3 by means of a t-test.

Deviation between MAF levels within each model was examined with t-test. All tests were corrected for multiple testing by false discovery rate.

Results and discussion

Estimates of heritability for the four investigated scenarios are summarized in table 1. For the inbred population (Ne 50) we did not for any of the three tested models find any significant differences compared to the real level of 0.3. This shows that there is no gain in implementing the 600 K array compared to the 54 K array in a population with an inbreeding level as this. For both arrays we found a general tendency that for each of the three models heritability estimates were obtained very close to the real level. For the less inbred population (Ne 100), we obtained a significant difference, when testing for differences between real and estimated heritability (model III in scenario IV). For this case differences were found to deviate significantly and the heritability was underestimated within each of the three MAF levels for this model. This show that with less inbreeding and together with data obtained by the 600 K SNP array the last mixture model is not useful for precise heritability estimation.

For the models I-III in scenario III and models I and II in scenario IV, all models were able to estimate an acceptable level of heritability. However, in each case the estimated heritability levels showed indication to be underestimated, although not significantly. A closer look at the three models in the two scenarios – III and IV – showed a tendency towards that model I and II obtain similar heritability estimates followed by model III. Model III obtain estimates a bit below the other two models. In scenario II, model II were the only time, where heritability for each MAF level were overestimated, although not significantly.

Within all of the four scenarios we did not find any significant changes with regard to the different levels of MAF (results not shown). It is generally accepted to use a level of minor allele frequency at 0.05, however for the case, where the heritability estimates would be too low, we were interested to identify a level of minor allele frequency, where a sufficient heritability estimate could be obtained. With the results obtained from this study it does although not seem necessary to ease on the level of minor allele frequency from the acceptable for these models to estimate heritability in a population size and composition as the one used here.

In general, comparing the performance of each model between the two levels of inbreeding reveals a tendency towards that all models are better suited for heritability estimation when the level of inbreeding is higher. However, iBay leaves the possibility, especially for the mixture models, to modify the models in a direction, which possibly can improve the estimates.

To summarize we did obtain the best estimates, when inbreeding within the population were high (Ne 50), whereas results became less precise with a less inbred population (Ne 100).

The procedure for variance estimation in this study is based on both genotypic and phenotypic information in a small sample of the population. Depending on the trait of interest, the phenotypic information can be difficult or expensive to obtain and therefore only a small number of animals can be tested. Here we show that it is possible to obtain precise estimates in a relatively small population and that there is no gain in using a denser SNP array (54 K SNP array *vs* 600 K SNP array) if the purpose is to estimate heritability.

Table 1: Mean values of heritability estimates for scenarios I-IV. Test for significant deviation of estimates from real heritability level was obtained by means of a t-test. Model I: truncated distribution model, model II: mixture model (75 % markers excluded), model III: mixture model (30 % markers excluded). 54 K = simulated marker density of 54 K SNP array, 600 K = simulated marker density of 600 K SNP array, minor allele frequency: 0.05, 0.02, and 0.01. All tests was corrected for multiple testing by False Discovery Rate (*K = 9*)

	Scenario I (Ne 50 / 54 K)			**Scenario II (Ne 50 / 600 K)**		
Model	**0.05±SE**	**0.02±SE**	**0.01±SE**	**0.05± SE**	**0.02±SE**	**0.01±SE**
I	0.29±0.06	0.29±0.06	0.29±0.06	0.29±0.07	0.29±0.06	0.29±0.08
II	0.3±0.05	0.3±0.05	0.3±0.05	0.31±0.08	0.31± 0.08	0.31± 0.07
III	0.3±0.06	0.3±0.06	0.3±0.06	0.28±0.08	0.27±0.07	0.28± 0.07
	Scenario III (Ne 100 / 54 K)			**Scenario IV (Ne 100 / 600 K)**		
Model	**0.05± SE**	**0.02±SE**	**0.01±SE**	**0.05± SE**	**0.02±SE**	**0.01±SE**
I	0.27±0.08	0.27±0.09	0.27±0.08	0.28±0.07	0.27±0.09	0.29±0.07
II	0.26±0.07	0.26±0.08	0.26± 0.07	0.28±0.08	0.28±0.08	0.28±0.07
III	0.27±0.08	0.27± 0.08	0.28±0.08	0.24±0.08*	0.25±0.08*	0.21±0.07***

**P*<0.05; significant after correction with False Discovery Rate
***P*<0.001; significant after correction with False Discovery Rate

Conclusion

The results obtained from this study show that it is possible for both of the two SNP arrays (54K SNP vs. 600K SNP) to obtain precise estimates of heritability from a dataset based on a population with a relatively low sample size. Models obtained from this study also show that estimates of heritability are most precise with these models when inbreeding is high (Ne 50). Further our results show, that there is no gain in the inclusion of a SNP array with a denser marker distribution as long as the sample size of the population is small.

References

Fisher, R. A. (1918). *Trans. R. Soc. Edinburgh.,* 52:399-433.

George, E. I. and Mcculloch, R.E. (1993). *J. Am. Stat. Assoc.,* 88:881-89.

Ihara, N., Takasuga, A., Mizoshita, K. *et al*. (2004). *Genome Res.*, 14:1987-98.

Janss, L. (2009). iBay manual version 1.47, http://www.lucjanss.com

Meuwissen, T.H.E. and Goddard, M.E. (2001). *Gene. Sel. Evol.,* 33:605-634

Thompson, R. (2008). *Proc. Biol. Sci.,* 275:679-86.

Visscher, P.M., Medland, S.E., Ferreira, M.A.R. *et al*. (2006). *PLoS Genet.,* 2:316-325.