\Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Combining SNPs in latent variables to improve genomic prediction.

H.C.M. Heuven[1,2], G.J.M. Rosa[2] and L.L.G. Janss[3].
[1]Faculty of Veterinary Medicine, Utrecht University NL, [2]University of Wisconsin Madison USA, [3]Aarhus University DK

ABSTRACT: The objective of this study was to develop and test hierarchical genomic models with latent variables that represent parts of the genomic values. An interaction model and a chromosome model were compared with a model based on variable selection in a simulated and real dataset. The program Bayz was used to calculate the parameters which were subsequently used to predict breeding value or the pre-corrected phenotypes in a cross validation.

The predictive value did not vary much for the simulated dataset among models and was in line with earlier results. Correlations between predicted and true breeding were around 0.9. For the mice dataset cross validation correlations were around 0.5. Using latent vectors to combineSNPs in genomic prediction models allows for estimation of non-linear effects such as interaction among SNPs and the use of prior biological information regarding the SNPs.
Keywords: Hierarchical genetic model; Predictive value; Gibbs sampling; Variable selection

Introduction

Breeding programs consist of five parts: definition of a breeding goal; collection of phenotypic, pedigree, and genotypic records of pure bred and hybrid individuals; calculating estimated breeding values (EBVs) and ranking the individuals on a combination of phenotypes; selection of the superior individuals as parents; and combination of these animals to produce the next superior generation. In the coming years more records will become available especially genotype data sets will spectacularly increase in size due to the decreasing cost of genotyping/sequencing. The challenge will be to extract information from these data in order to improve current breeding programs. It also applies to prediction of phenotypic performance, e.g. disease risk, in plants, animals and humans, using DNA-data. It will require optimization of genetic and statistical models. Since genotypes usually can be obtained earlier the goal is to use genotypes to predict phenotypes (De los Campos et al. (2010)).

The current methods vary from single trait – single marker analysis to neural networks including 1000's of SNPs; from simple chi-square analysis (e.g. Plink; Purcell et al. (2007)) to non-linear models that capture (non)-linear relationships between predictors and responses (Gianola et al. (2011); Li et al. (2013)).

The objective of this study was to develop hierarchical genetic models, using gibbs sampling to estimate parameters, and test their predictive ability.

Materials and Methods

Data. A simulated and a real dataset were used to compare the predictive value of three different models. The simulated dataset was obtained from the QTLmas workshop in Rennes (Le Roy et al. (2011)). The population comprised 3000 individuals issued from 20 sires and 200 dams. Within each family, 10 progenies belonged to the experimental population and were assigned phenotypes and marker genotypes and 5 belonged to the selection population, only known on their marker genotypes. A total of 10,000 SNPs carried by 5 chromosomes of 1 Morgan each were simulated. Eight QTL were created (1 quadri-allelic, 2 linked in phase, 2 linked in repulsion, 1 imprinted and 2 epistatic). Random noise was added giving an heritability of 0.30. The marker density, LD and MAF were similar to real life parameters. For the real data a publicly available dataset on mice (http://gscan.well.ox.ac.uk/) was used. This outbred mice population that descended from eight inbred progenitor strains was created for fine-mapping QTL and high-resolution GWAS analysis of body weight among other traits. After editing the data set contains the genotypic information from 1,843 fully pedigreed mice and 12226 SNPs. A full description of this population and other details can be found in Valdar et al. (2006). Since the pedigree and the effects of cage were partly confounded body weight was corrected for effects of sex and cage using a standard animal model.

Statistical analyses. Three models: an interaction model (INT) and a chromosome model (CHR) were compared with a model based on variable selection (BVS) in both datasets. The program Bayz 2.4.1 (Janss (2014)) was used to calculate the parameters which were subsequently used to predict breeding value of the animals without phenotypes (simulation dataset) or the corrected phenotype in the cross validation (mice dataset).

BVS model:

$$y = \text{mean} + \sum_j b_j * SNP_j + e$$

INT model:

$$y = \text{mean} + \beta1*L1 + \beta2*L2 + \beta3*L1*L2 + e$$
$$L1 = \sum_j b_j * SNP_j + \varepsilon1$$
$$L2 = \sum_j b_j * SNP_j + \varepsilon2$$

CHR model:

$$y = mean + \sum_k \beta_k * L_k + e$$
$$L_k = \sum_j b_{k:j} * SNP_{k:j} + \varepsilon_k$$

where y is the simulated trait or the pre-corrected body weight. L1, L2 and L_k are latent vectors ('animal effects') which, at the next level are modelled with the SNP-effects (j= number of SNPs, i.e. 9990 for the simulated dataset and 12226 for the mice dataset). In the INT model the interaction among SNPs is modeled by the interaction among latent vectors. For the CHR model a latent vector for each chromosome was used (k=1,5 or k=1,19 and k:j= number of SNPs per chromosome which was 1998 for the simulated dataset and variable for the mice dataset).
The following priors were used:

mean ~ uniform
$e \sim N(0, I\,\sigma^2_e)$
$\sigma^2_e \sim$ uniform
$\beta_k \sim$uniform [0,2]
$\beta 3 \sim$uniform [-3,3] (for the INT model)
L_k = latent variable
$\varepsilon_k \sim N(0, I*.1)$
$b_k \sim N(0, I\,\sigma^2 b_{k0})$ z=0
$\sim N(0, I\,\sigma^2 b_{k1})$ z=1
$\sigma^2 b_k \sim InvX(scale_k, 10)$
$scale_k \sim$ uniform
$z_k \sim Bernoulli(1-\pi, \pi)$
$\pi_k \sim Beta(10,1)$

Posterior means were calculated from 50000 iterations (after 1000 iterations of burn in) based on 200 samples. TheSNPs were fitted using a Bayesian variable selection method were the probability (π) that a SNP is sampled from the distribution with large effects is also determined by the data using a slightly informative Beta-prior. The scale of the inverted-Chi-square distribution used as prior for the SNP variance components was also estimated using an uniform prior. The prior residual variance of latent vectors was set to a fixed value in order to make the model identifiable.

Correlations between predicted genomic values and true breeding values were calculated for the simulation dataset. For the mice dataset cross validation was applied. Five subsets were created by randomly excluding 20% of the phenotypes. The remaining data was used to estimate the parameters and these were used to calculate genomic values for the excluded mice. Correlations between the predicted genomic values and (pre-corrected) bodyweight were calculated.

Results and Discussion

Choice of models. BVS is used as the standard to compare the INT and CHR models. It is expected that with whole genome sequence data some form of variable selection will be required (Meuwissen and Goddard (2010)). The INT model was developed in order to take all (linear) interactions amongSNPs into account. In case the latent vectors would be transformed using functions such as quadratics, square root or tanh, it would also be able to take non-linear interactions into account. In the case tanh would be used it will more or less mimic a neural network as described by Gianola et al. (2011). The CHR model was developed to show thatSNPs can be combined based on some biological information regarding the SNPs. Another combination could be based on SNPs occurring in genes belonging to one or more pathways. It leads to biologically informed genomic models.

Parameters. In table 1 some of the estimated parameters for the simulation dataset are shown for each of the three models. Variances explained by the SNPs in the models were small for latent vector 1 in the INT model and latent vector 4 in the CHR model.

Table 1. Posterior mean (p.mean) and standard error (p.se) for some parameters in simulation dataset for the three models.

model	BVS		INT		CHR	
parameter	p.mean	p.se	p.mean	p.se	p.mean	p.se
σ^2_e	59.04	2.06	54.35	3.00	52.38	3.53
$\beta 1$	--	--	0.81	0.56	1.26	0.47
$\beta 2$	--	--	1.49	0.33	0.93	0.57
$\beta 3$	--	--	0.04	0.38	1.25	0.48
$\beta 4$	--	--	--	--	1.08	0.55
$\beta 5$	--	--	--	--	1.01	0.58
π_1	0.01	0.004	0.08	0.07	0.07	0.09
π_2	--	--	0.01	0.01	0.15	0.17
π_3	--	--	--	--	0.06	0.05
π_4	--	--	--	--	0.09	0.08
π_5	--	--	--	--	0.09	0.08
$\sigma^2 b_1$ z=1	0.37	0.08	0.00003	0.0001	0.30	0.13
$\sigma^2 b_2$ z=1	--	--	0.16	0.06	0.16	0.11
$\sigma^2 b_3$ z=1	--	--	--	--	0.18	0.17
$\sigma^2 b_4$ z=1	--	--	--	--	0.0006	0.0022
$\sigma^2 b_5$ z=1	--	--	--	--	0.01	0.02

σ^2_e: residual variance; β's: regressions on latent variables representing genomic effects, INT has two latent variables, where $\beta 3$ is regression on the interaction between the two latent variables, and CHR has 5 latent variables, one for each chromosome; pi and σ^2bare proportions and variances for large SNP effects in each of the latent variables.

Predictive value. Table 2 shows the correlations of the predicted genomic value and the true breeding value for the 1000 animals without phenotypes in the simulation dataset. The predictive ability was very similar among the models and in line with the results presented at the QTLmas workshop (Le Roy et al. (2012)). Differences were small which was expected given the low number of simulated QTL and the effect size of the QTL. The correlation

between predicted genomic values and body weight of the excluded animals in the mice dataset are also shown in table 2. The predictive ability is in line with results by Felipe et al. (2014), who showed correlations around 0.5 using Bayesian Regularized Artificial Neural Network, Reproducing Kernel Hilbert Spaces and Bayesian Lasso models which included substantial fewer SNPs.

Table 2. Correlations between predicted genomic values and true breeding values (simulation dataset) or with pre-corrected phenotype for the mice dataset.

model	BVS	INT	CHR
dataset			
simulated	0.904	0.903	0.908
mice	0.513	0.518	0.501

Conclusion

Using latent vectors to combine SNPs in genomic prediction models allows for estimation of non-linear effects such as interaction among SNPs and the use of prior biological information regarding the SNPs.

Literature Cited

De los Campos, G., Gianola, D., Allison, D.B. (2010) Nat Rev Genet 11:880-886.

Felipe, V.P., Okut, H.,Gianola, D., et al. (2014) in preparation.

Gianola, D., Okut, H., Weigel, K.A., (2011) BMC Genet. 12:87.

Janss L: Bayz. 2.4.1 (2014). http://www.bayz.biz.

Le Roy, P., Filangi, O, Demeure, O., (2012) BMC Proc., 6(Suppl 2):S3.

Li, F., Zhao, J., Yuan, Z., et al. (2013) BMC Genet. 14:89.

Meuwissen T., Goddard M. (2010) Genet. 185: 623-631

Purcell, S., Neale, B., Todd-Brown, K., et al. (2007) Am J of Hum Genet, 81

Valdar, W., Solberg, L.C.,Gauguier, D., et al. (2006) Nature Genet. 38:879–887.