Genomic GxE approaches modelling heterogeneous SNP variances: applied to simulated data

B. Gredler-Grandl[1] and M.P.L. Calus[1]

[1] Wageningen University & Research Animal Breeding and Genomics, PO Box 338, 6700 AH Wageningen, The Netherlands

Abstract

Genotype by environment interaction (GxE) can be modelled using a multi-trait approach where the same trait measured in different environments is considered a different, but correlated trait. An alternative is to model GxE with reaction norm models where the breeding values are modeled as a function of the environment defined as a continuous variable. Genomic implementations of both models can be parameterized such that homogeneous (co)variances are assumed for all SNP across the genome. Since specific regions in the genome may harbor QTL and others may not or loci may have a large effect in one environment and a zero effect in another, the assumption of equal (co)variances across the genome is violated. We have developed an analysis protocol based on readily available BLUP software packages to allow for heterogeneous SNP (co)variances in genomic GxE models. The analysis protocol consists of a two-step approach, where the data set of interest is split in two subsets. One subset is used to estimate SNP effects and derive weights for each SNP, which are subsequently used to upweight SNP in the analysis of the second subset. We have carried out a simulation study that showed a small increase in accuracy of genomic breeding values when allowing for heterogeneous SNP (co)variances compared to homogeneous SNP (co)variances.

Key words: Genotype by environment interaction, genomic GxE model, heterogeneous SNP variances

Introduction

Genotype by environment interaction (GxE) is typically modelled by a multi-trait approach, where the same trait measured in different environments is considered being a genetically different, but correlated trait (e.g. Falconer, 1952). As an alternative, GxE can be modelled with reaction norm models, where the breeding values are modelled as a function of the environment defined as a continuous variable (Kolmodin et al., 2002; Calus and Veerkamp, 2003). Both models can be implemented in genomic prediction models by replacing the pedigree based relationship matrix by the genomic relationship matrix. Both, genomic multi-trait models or reaction norm models, implicitly assume the same (co)variance matrix for every SNP. Since certain regions in the genome may contain QTL, the assumption of equal (co)variances across the genome may be violated. To overcome this limitation, we have developed an analysis protocol allowing for heterogeneous SNP (co)variances across the genome in genomic GxE models. The analysis protocol can be implemented using standard BLUP software packages. The objective of the study was to evaluate the accuracy of genomic GxE models allowing for heterogeneous SNP (co)variances across the genome in simulated data.

Materials and Methods

Simulation

The analysis protocol was tested on simulated data. For this purpose, we have simulated two populations and crossed these to produce F1 crossbred individuals. The simulations were performed using the QMSim

software (Sargolzaei and Schenkel, 2009). A historical population spanning 1 000 generations, consisting of 10 000 individuals in the base population (generation -1 000), was simulated. The population size decreased linearly to 400 individuals across 980 generations until generation -20. This bottleneck was used to achieve LD. The population continuously increased to a size of 4 100 individuals in generation 0. The last generation of the historical population consisted of 100 males and 4 000 females, and was randomly divided in two groups of 4 100 individuals (100 males and 4000 females) forming two separated populations (A and B). In both populations A and B, random mating was applied for 210 generations to produce 1 000 male and 1 000 female offspring in each generation. A crossbreeding program started in generation 206 where 200 male and 500 female individuals were randomly selected to produce 1 000 crossbred offspring (500 males and 500 females) until generation 210. The simulated genome consisted of 30 chromosomes with length of 100 cM each. In total 51 000 markers were simulated equally distributed across the whole genome, which is similar to the 50k Bovine BeadChip. Table 1 shows the parameters used for the selection design and the simulated genome.

Table 1. Parameters used in the simulation

Item	
Litter size	1
Proportion of male progeny	0.5
Mating design	random
Selection design	random
Sire replacement	0.5
Dam replacement	0.25
Culling criteria	Age
Genome	
No. Chromosomes	30
Chromosome length (cM)	100
No. markers per chromosome	1 700
No. QTL per chromosome	150
QTL effects	Sampled from normal distribution
Marker mutation rate	2.5×10^{-5} (recurrent)
QTL mutation rate	2.5×10^{-5}
Position of markers and QTL	random

Phenotypes were simulated to follow a linear reaction norm model with a custom Fortran program and calculated per individual for its assigned environment as the sum of environmental value and true breeding values and a residual error. Environmental values were drawn from a normal distribution with N(0,1) and ranged between -2.063 and 2.063. The genetic variances for intercept and slope were assumed to be 0.3 and 0.025, respectively. The genetic covariance between intercept and slope was 0.05 leading to a genetic correlation of 0.577. The residual variance was set to 0.5. The resulting simulated heritability across environments is shown in Figure 1. The individuals were randomly assigned to the environments.

A protocol consisting of several steps has been developed to allow for heterogeneous (co)variances across the genome in genomic GxE models. Firstly, the data set is split in two subsets: subset 1 is used to estimate SNP effects $\hat{\alpha}$ using a model that assumes equal (co)variances for all SNP. SNP specific variances are then computed as $2p_k(1-p_k)\hat{\alpha}_k^2$. The model applied to subset 2 then considers these SNP specific variances as weights to compute a weighted SNP (co)variance matrix.

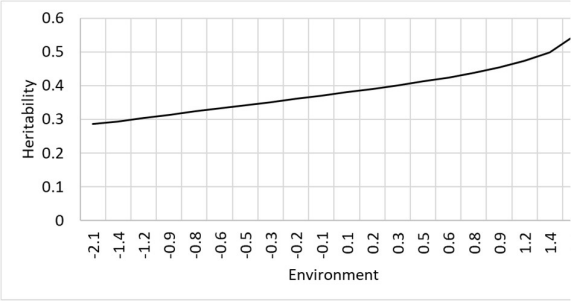

Figure 1. Simulated heritability across environments

Subset 1 consisted of 4 000 individuals of generations 205 and 206 (including crossbred individuals). Subset 2 consisted of individuals of generations 207 to 210 (including crossbred individuals), where all individuals of generations 207 and 208 were used as training population (n=4 000) and all remaining individuals of generations 209 and 210 (n=4 000) were used as validation set.

Analysis subset 1

A univariate linear genomic reaction norm model has been applied to subset 1 using the software package mtg2 (Lee et al., 2016):

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{\beta_0} + \mathbf{Q\beta_1} + \mathbf{e}$$

where y is a vector of simulated phenotypes, μ is an overall mean, $\mathbf{\beta_0}$ and $\mathbf{\beta_1}$ are the vectors of intercept and first order of regression coefficients for the random genetic effects, $\mathbf{1}$ is a vector of ones, $\mathbf{Q}$ is a (diagonal) incidence matrix storing the environmental values for each individual, and $\mathbf{e}$ is the vector of random residuals.

Calculation of SNP specific weights

Allowing for heterogeneous SNP variances (HET), SNP specific weights for each SNP k for each coefficient i of the reaction norm model (i.e. intercept β_0 and the linear regression coefficient β_1) were calculated as:

$$D_{k_i} = \sqrt{2p_k(1-p_k)}\hat{\alpha}_{k_i}$$

where D_{k_i} is diagonal element i of diagonal matrix $\mathbf{D_k}$ that stores the weights for SNP k, p_k is the allele frequency of SNP k, and $\hat{\alpha}_k$ is the estimated effect of SNP k for coefficient i. The SNP effects $\hat{\alpha}_k$ for intercept and linear regression coefficient were obtained by backsolving based on the GEBV for β_0 and β_1 obtained from the genomic reaction norm model. SNP effects were calculated following the approach described in Bouwman et al. (2017) implemented in the companion program compute_SNP_effects of calc_grm (Calus and Vandenplas, 2016).

Analysis subset 2

The following SNP-BLUP model was applied to subset 2 using the MiXBLUP software (ten Napel et al. 2020):

$$\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{Z\gamma_0} + \mathbf{QZ\gamma_1} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of simulated phenotypes of individuals in the training set of each cross validation run, μ is an overall mean, $\mathbf{Z}$ is a matrix including the centered genotypes for each SNP, $\mathbf{Q}$ is a diagonal matrix storing the environmental values for each individual, $\mathbf{\gamma_0}$ and $\mathbf{\gamma_1}$ are vectors of estimated SNP effects for random intercept and linear regression coefficient, respectively, and $\mathbf{e}$ is a random residual term. For HET the following (co)variance matrix is used for SNP k:

$$Var([\mathbf{\gamma_0}, \mathbf{\gamma_1}]') = \mathbf{D_k} * \mathbf{G} * \boldsymbol{D_k'}$$

where $\mathbf{G}$ is the estimated genetic (co)variance matrix between intercept and quadratic regression coefficient obtained from the reaction norm model in the analysis of subset 1. For HOM, homogeneous SNP variances for intercept and linear regression coefficient are provided by $\sigma_g^2/2\sum p_k(1-p)$, where σ_g^2 is the genetic variance for either intercept or linear regression coefficient estimated in subset 1. The GEBV for validation animals were calculated as

$\mathbf{GEBV} = \mathbf{1}\hat{\mu} + \mathbf{Z}\hat{\mathbf{\gamma}}_{\mathbf{0}} + \mathbf{ZQ}\hat{\mathbf{\gamma}}_{\mathbf{1}}$. The accuracies of GEBV for individuals in the validation set were obtained as the correlation coefficient between the simulated true breeding value and predicted GEBV.

Results & Discussion

The estimated genetic variances for intercept and linear regression coefficient of the genomic reaction norm model in subset 1 were 0.35 and 0.031, respectively. The estimated genetic covariance was 0.04. These results were in good agreement with the underlying simulated genetic covariance structure.

For the HOM scenario, the correlation between GEBV and TBV in the validation set in subset 2 were 0.521 and 0.588 for intercept and the linear regression coefficient, respectively. For HET, where heterogeneous SNP variances were allowed, the correlation between GEBV and TBV for intercept and linear regression coefficient were 0.551 and 0.601, respectively. This results in a small increase in accuracy of HET compared to HOM of 0.03 and 0.013 correlation points for intercept and linear regression coefficient across all environments, respectively. The accuracy was highest in environments where higher genetic variance was observed (environmental value > 0.7) and lowest for environments with smaller genetic variance (environmental value < -0.7).

The increase in accuracy in HET compared to HOM is small. In the current study, SNP effects for intercept and linear regression coefficient were estimated assuming equal (co)variances for each SNP across the genome in subset 1. Bayesian approaches to derive SNP effects in subset 1 and upweight SNP in a following GBLUP or SNP-BLUP analysis could be beneficial. The current implementation of the analysis protocol is based on readily available software allowing for fast and large scale implementations and resulting in an increase in accuracy of GEBV in simulated data.

Conclusions

The aim of this study was to evaluate the accuracy of genomic reaction norm models allowing for heterogeneous SNP (co)variances. We developed an analysis protocol to allow for heterogeneous SNP (co)variances based on readily available BLUP software packages. The results show a small increase in accuracy of genomic GxE models allowing for heterogeneous SNP (co)variances. The analysis protocol allows for fast and large-scale applications in routine genomic evaluations.

Acknowledgments

We acknowledge the GenTORE project, which has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No 727213.

References

Bouwman, A.C., Hayes, B.J., and Calus, M.P.L. 2017. Estimated allele substitution effects underlying genomic evaluation models depend on the scaling of allele counts. Genet. Sel. Evol. 49, 79.

Calus M.P.L., Veerkamp R.F. 2003. Estimation of environmental sensitivity of genetic merit for milk production traits using a random regression model. Journal of Dairy Science86, 3756–64.

Calus M., and Vandenplas, J. 2016. Calc_grm – a program to compute pedigree, genomic, and combined relationship matrices. ABGC, Wageningen UR Livestock Research.

Falconer, D. S. 1952. The problem of environment and selection. Am. Nat. 86:293-298.

Kolmodin, R., E. Strandberg, P. Madsen, J. Jensen, and H. Jorjani. 2002. Genotype by environment interaction in Nordic dairy cattle studied using reaction norms. Acta Agric. Scand. A Anim. Sci. 52:11-24.

Lee, S.H. and van der Werf, J.H. 2016. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics 32, 1420-1422.

Sargolzaei, M. and Schenkel, FS. (2009). QMSim: A large-scale genome simulator for livestock. Bioinformatics, 25(5), 680-681.

ten Napel, J., J. Vandenplas, M. Lidauer, I. Stranden, M. Taskinen, E. Mäntysaari, M. P. L. Calus, and R. F. Veerkamp. 2020. MiXBLUP, the Mixed-model Best Linear Unbiased Prediction software for PCs for large genetic evaluation systems – Manual V2.2 - 2020 - 05, Wageningen, the Netherlands www.mixblup.eu