Genet. Res., Camb. (2011), **93**, *pp.* 125–138. © Cambridge University Press 2011
doi:10.1017/S0016672311000012

Use of genomic models to study genetic control of environmental variance

YE YANG, OLE F. CHRISTENSEN AND DANIEL SORENSEN*
Department of Genetics and Biotechnology, Faculty of Science and Technology, Aarhus University, DK-8830 Tjele, Denmark

(*Received 17 June 2010 and in revised form 23 December 2010; first published online 11 March 2011*)

Summary

Vast amount of genetic marker information is being used to obtain insight into the genetic architecture of complex traits, for locating genomic regions (quantitative trait loci (QTL)) affecting disease and for enhancing the accuracy of prediction of genetic values in selection programmes. The genomic model commonly found in the literature, with marker effects affecting mean only, is extended to investigate putative effects at the level of the environmental variance. Two classes of models are proposed and their behaviour, studied using simulated data, indicates that they are capable of detecting genetic variation at the level of mean and variance. Implementation is via Markov chain Monte Carlo (McMC) algorithms. The models are compared in terms of a measure of global fit, in their ability to detect QTL effects and in terms of their predictive power. The models are subsequently fitted to back fat thickness data in pigs. The analysis of back fat thickness shows that the data support genomic models with effects on the mean but not on the variance. The relative sizes of experiment necessary to detect effects on mean and variance is discussed and an extension of the McMC algorithm is proposed.

1. Introduction

The availability of massive genetic marker information provides new opportunities for understanding quantitative variation, locating genes, molecular classification of disease status, designing genotype-specific regimes (e.g. diets or therapies) or for enhancing the accuracy of prediction of genetic merit in plant and animal breeding. A nice example of novel insight is provided by Visscher *et al.* (2006) and Visscher *et al.* (2007), who studied the genetics of height in humans. The authors infer genetic variation from information arising within families only, exploiting the variation in identity-by-descent shared between relatives, uncovered by marker information. They find that their data are consistent with a uniform spread of trait loci throughout the genome whose effects act additively on height. In animal and plant breeding several selection programmes are now genotyping elite individuals and genetic evaluations based on SNPs are becoming routine (González-Recio *et al.*, 2008; Hayes *et al.*, 2009; VanRaden *et al.*, 2009; Hayes *et al.*, 2009). Considerable research efforts are currently devoted to the development of methods that incorporate massive marker information, and a large variety of models and approaches are becoming available.

The use of massive marker information in a linear regression model to predict genetic values for quantitative traits was first proposed by Meuwissen *et al.* (2001). This model and others discussed in the literature postulate that quantitative trait loci (QTL) affect the mean of the quantitative trait, and assume homogeneity of residual variation. Sorensen (2009) suggests an extension to investigate whether genomic regions also have an effect on the environmental variance. Support for genetic regulation of the environmental variance has been reported for a number of traits in tomato (Weller *et al.*, 1988), litter size in sheep (San Cristobal-Gaudy *et al.*, 2001), litter size in pigs (Sorensen & Waagepetersen, 2003), adult weight in snails (Ros *et al.*, 2004), body weight in poultry (Rowe *et al.*, 2006; Wolc *et al.*, 2009; Mulder *et al.*, 2009), slaughter weight in pigs (Ibáñez *et al.*, 2007), litter size and weight traits in mice (Gutierrez *et al.*, 2006; Ibáñez *et al.*, 2008*a*), litter size in rabbits (Ibáñez *et al.*, 2008*b*), bristle counts in *Drosophila*

* Corresponding author: e-mail: daniel.sorensen@agrsci.dk

(Whitlock & Fowler, 1999; Mackay & Lyman, 2005), a number of traits in maize (Ordas *et al*., 2008) and levels of gene expression in yeast (Ansel *et al*., 2008). With the exception of the first and last two references, inferences were based on models for the residual variance where marker information was not included. Ordas *et al*. (2008) analysed a number of maize recombinant inbred lines and incorporated information on 85 genetic markers. The design made it possible to use a simple analysis, including a fixed effect of genotype in the least-squares linear model for residual variances. Ansel *et al*. (2008) provide convincing evidence for heterogeneity of gene expression in isogenic yeast cells of different genotypes and identify three QTLs involved in the control of heterogeneity.

The purpose of this work is to incorporate marker information to detect genomic regions that have effects on the residual variance. The standard genomic model operating on the mean of a quantitative trait is extended to accommodate marker covariates on the log-environmental variance, and two models are proposed. The first one assumes that marker effects at the level of the mean and variance are *a priori* bivariate normally distributed, with common mean and covariance matrix. The second model is based on stochastic search variable selection (George & McCulloch, 1993). It assumes that marker effects at the level of the mean and variance are independent *a priori* and that their distributions are two-component normal mixtures. The models are implemented using Markov chain Monte Carlo (McMC) and their ability to detect QTL is studied using simulated data. Subsequently the models are fitted to back fat thickness data in pigs.

This paper is organized as follows: Section 2, introduces the genomic models and describes the McMC algorithm. Section 3 describes the types of data simulated and the inferences that are possible from the various models. Section 4 contains results of the application to back fat data. Section 5 discusses issues related to the relative sizes of experiment to detect marker effects at the level of mean and variance and proposes an extension of the McMC algorithm.

2. Methods

Four models are studied that differ in the structure of the residual variance of the likelihood and in the prior distributions of marker effects at the level of the mean and variance. Two classes of prior distributions of marker effects are considered, and the residual variance is assumed to be either homogeneous or genetically heterogeneous. The first two models assume that marker effects are independent and identically normally distributed. Model GHOM includes marker effects at the level of the mean only with identical distribution and homogeneous residual variance. The heterogeneous variance model GHET assumes marker effects on the mean and on the log-variance of the trait, and for marker j, the pair of marker effects (affecting mean and variance) is bivariate normally distributed. The third and fourth models assume that marker effects originate from identical two-component mixture distributions. In the model labelled GHOMMIX, marker effects operate at the level of the mean only and the variance is homogeneous. The final heterogeneous variance model is GHETMIX, where the pair of marker effects on mean and variance for marker j are independently distributed, and each originates from a two-component mixture.

(i) *Likelihood*

The sampling distribution of the data is Gaussian of the form

$$\mathbf{y}|\mu, \mathbf{a}, (\sigma^2_{i,M})_{i=\mathbf{1},\ldots,n} \sim N(\mathbf{1}\mu + \mathbf{X}\mathbf{a}, \operatorname{diag}(\sigma^2_{i,M}, i = 1, \ldots, n)), \quad (1)$$

where $\mathbf{y}$ is the data vector of length n, $\mathbf{1}$ is a vector of ones, the scalar μ is the mean, $\mathbf{a}$ is a vector of marker effects on the mean and $\sigma^2_{i,M}$ is the environmental variance for the ith observation under model M. The matrix $\mathbf{X}$ is an $n \times N$ matrix where X_{ij} is an observable indicator for the jth marker locus of the ith individual, coded as -1 for genotype 11, 0 for genotype 12 and 1 for genotype 22, and N is the number of marker loci. The conditional likelihood is proportional to (1). Here and elsewhere we use $N(\cdot,\cdot)$ to denote both a normal distribution and its density function.

In the genomic model with homogeneous variance (GHOM and GHOMMIX), $\sigma^2_{i,1} = \exp(\mu^*)$ (Meuwissen *et al*., 2001). Genetic heterogeneity of environmental variance (in models GHET and GHETMIX) is incorporated by assuming that

$$\sigma^2_{i,2} = \sigma^2_{i,1} \exp(\mathbf{x}_i' \mathbf{a}^*), \quad (2)$$

where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$, and $\mathbf{a}^*$ is the column vector with N marker effects on the variance (Sorensen, 2009).

(ii) *Prior specifications*

The mean μ and the environmental variance when $a^* = 0$, $\exp(\mu^*)$, are assigned improper uniform distributions. Depending on the model, two possible distributions are assigned to vector $\mathbf{a}$ or vectors ($\mathbf{a}$, $\mathbf{a}^*$). The first one is a common Gaussian distribution for all marker effects. A mechanistic justification for this distribution is to assume that markers capture the effects of loci with which they are in disequilibrium. These effects are of similar magnitude across loci and therefore can be approximated by a common Gaussian model. The second distribution is a mixture,

https://doi.org/10.1017/S0016672311000012 Published online by Cambridge University Press

such that a small proportion of loci originate from a normal distribution with relatively large variance, allowing a broad range of marker effects, that are captured by the markers. The rest of the loci have normally distributed effects with zero mean and very small variance. These loci can be interpreted as having no detectable effects on the trait.

(a) *No mixture models*

In models GHOM and GHET, marker effects are assumed to be independent realizations from the same normal process. In the GHOM model, the vector of marker effects **a** is assumed to have the normal prior distribution

$$\mathbf{a}|\sigma_a^2 \sim N(0, \mathbf{I}\sigma_a^2), \tag{3}$$

where **I** is the identity matrix and σ_a^2 is the variance of marker effects representing their prior uncertainty. This variance is assigned a scaled inverse chi-square distribution with degree of freedom ν and scale parameter $S_{\sigma_a^2}$. This prior specification was used by Legarra *et al.* (2008). The method known as BayesA (Meuwissen *et al.*, 2001) instead assigns marker-specific variances, which are assumed to be realizations from a common scaled inverted chi-square distribution with known hyperparameters. This generates a problem of identifiability as noted by Gianola *et al.* (2009).

In the GHET model, marker effects (**a**, **a***) also have a common prior distribution of the form

$$(\mathbf{a}, \mathbf{a}^*)'|\mathbf{G} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \mathbf{G}\otimes\mathbf{I}\right),$$

where **G** is the 2×2 variance–covariance matrix

$$\mathbf{G} = \begin{pmatrix}\sigma_a^2 & \rho\sigma_a\sigma_{a^*}\\ \rho\sigma_a\sigma_{a^*} & \sigma_{a^*}^2\end{pmatrix}. \tag{4}$$

Thus, the pair of scalars (a_j, a_j^*) for the jth marker, is bivariate normally distributed. In eqn (4), ρ is the correlation between marker effects at the level of mean and variance, and σ_a^2 and $\sigma_{a^*}^2$ are variances for marker effects **a** and **a***, respectively. The parameter ρ is assigned a uniform prior bounded in $(-1,1)$, and σ_a^2 and $\sigma_{a^*}^2$ are assigned scaled inverted chi-square distributions with degrees of freedom ν and scale parameters $S_{\sigma_a^2}$ and $S_{\sigma_{a^*}^2}$, respectively.

(b) *Mixture models*

In model GHOMMIX, the two-component normal mixture prior for marker effect a_j is

$$P(a_j|p, c^2, \tau^2) = pN(0, c^2\tau^2) + (1-p)N(0, \tau^2),\ j = 1, \ldots, N,\ c > 1, \tag{5}$$

where p is the probability that the effect is a realization from the normal component with variance $c^2\tau^2$, and the complement, $(1-p)$ is the probability that it originates from the normal component with variance τ^2. The term τ^2 is chosen to be small, which results in $a_j's$ very close to zero. The distribution with larger variance $c^2\tau^2$, allows for the effects to depart markedly from the mean of zero and this is interpreted as a signal of the existence of a QTL in the proximity of the marker. The larger variance is obtained by setting c large.

The variance of the marker effect is

$$\mathrm{Var}(a_j|p, c^2, \tau^2) = pc^2\tau^2 + (1-p)\tau^2. \tag{6}$$

As in George & McCulloah (1993), the mixture prior is implemented augmenting with a set of independent binary indicator variables $\boldsymbol{\delta}' = (\delta_1, \ldots, \delta_N)$, such that when $\delta_j = 1$, a_j has density $N(0, c^2\tau^2)$, and when $\delta_j = 0$, a_j has density $N(0, \tau^2)$. Then

$$P(a_j|\delta_j, c^2, \tau^2) = \delta_j N(0, c^2\tau^2) + (1-\delta_j)N(0, \tau^2),\ c > 1, \tag{7}$$

and

$$P(\delta_j = 1|p) = 1 - P(\delta_j = 0|p) = p.$$

Since the $\delta_j's$ are independent with the same distribution,

$$P(\delta_1, \ldots, \delta_N|p) = \prod_{j=1}^{N} P(\delta_j|p) = \prod_{j=1}^{N} p^{\delta_j}(1-p)^{1-\delta_j}.$$

The joint prior distribution of all marker effects **a** and binary indicator variables $\boldsymbol{\delta}$ is then

$$P(\mathbf{a}, \boldsymbol{\delta}|p, c^2, \tau^2) = \prod_{j=1}^{N} P(a_j|\delta_j, c^2, \tau^2)P(\delta_j|p).$$

In the implementations that follow, the parameters c^2, τ^2 and the probability p are treated as constants and must be tuned by the user.

The GHOMMIX model is in the same spirit as the so-called BayesB method (Meuwissen *et al.*, 2001). However, in BayesB each marker is assigned a specific variance which as discussed in Gianola *et al.* (2009) leads to the same identifiability problem as BayesA. Further, in BayesB the mixture structure is specified at the level of the marker variances rather than at the level of the marker effects as in GHOMMIX.

The GHETMIX model involves also marker effects at the level of the environmental variance. It is assumed that marker effects at the level of mean a_j and variance a_j^* are independent *a priori*. At the level of the mean, marker effects are as in eqn (7), and at the level of the residual variance, the normal mixture distribution of marker effects a_j^* is implemented augmenting with a set of independent

https://doi.org/10.1017/S0016672311000012 Published online by Cambridge University Press

binary indicator variables $\boldsymbol{\delta}^{*\prime}=(\delta_1^*, \ldots, \delta_N^*)$, such that when $\delta_j^*=1$, a_j^* has density $N(0, c^{*2}\tau^{*2})$, and when $\delta_j^*=0$, a_j^* has density $N(0, \tau^{*2})$. Then

$$P(a_j^*|\delta_j^*, c^{*2}, \tau^{*2}) = \delta_j^* N(0, c^{*2}\tau^{*2}) + (1-\delta_j^*)N(0, \tau^{*2}),\ j=1, \ldots, N,\ c^*>1,$$

with

$$P(\delta_1^*, \ldots, \delta_N^*|p^*) = \prod_{j=1}^{N} P(\delta_j^*|p^*) = \prod_{j=1}^{N} p^{*\delta_j^*}(1-p^*)^{1-\delta_j^*},$$

where

$$P(\delta_j^*=1|p^*) = 1-P(\delta_j^*=0|p^*) = p^*,$$

and p^* is the probability that the marker effects at the level of variance is a realization from the normal component with variance $c^{*2}\tau^{*2}$. On the basis of the above prior specifications, the joint prior distribution for (a_j, a_j^*) conditional on (δ_j, δ_j^*) in model GHETMIX can be written as

$$(a_j, a_j^*|\delta_j, \delta_j^*, c^2, \tau^2, c^{*2}, \tau^{*2}) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\alpha_j^2\tau^2 & 0\\ 0 & \alpha_j^{*2}\tau^{*2}\end{pmatrix}\right),$$

where

$$\alpha_j = \begin{cases}1 & (\delta_j=0)\\ c & (\delta_j=1)\end{cases}, \quad \alpha_j{}^* = \begin{cases}1 & (\delta_j^*=0)\\ c^* & (\delta_j^*=1)\end{cases}.$$

In this GHETMIX model, the parameters c^2, c^{*2}, τ^2, τ^{*2} and the probabilities p and p^* are treated as constants and must be tuned by the user. A way of choosing c and τ^2 is described in the next subsection.

The parameterization in terms of the indicator functions δ_j and δ_j^* provides a simple way of studying the ability of the GHOMMIX and GHETMIX models to detect QTL signals. For each marker, the posterior probabilities $P(\delta_j=1|\mathbf{y})$ and $P(\delta_j^*=1|\mathbf{y})$ provide evidence for the presence of a QTL in the proximity of marker j. In section 3, we provide evidence for the presence of QTL via the Bayes factor, computed as

$$\frac{P(\delta_j=1|p, \mathbf{y})/P(\delta_j=0|p, \mathbf{y})}{P(\delta_j=1|p)/P(\delta_j=0|p)}.$$

Monte Carlo estimates of these posterior probabilities are obtained using the draws from the Markov chain.

(iii) *Choice of user-specified tuning parameters and an overview of effect on inferences*

In the GHOMMIX model, the parameters c^2, τ^2 and the probability p are treated as constants and must be tuned by the user. The following rule has been followed to choose values for these parameters. First, p is set to 0·10, so that 10 % of the markers are assumed to have a detectable effect on the mean of the trait, *a priori*. A rough candidate value for τ^2 is derived as follows. First, an analysis based on the classical infinitesimal model with homogeneous residual variance (using pedigree information only) yields an estimate of the additive genetic variance for the trait, σ_u^2. Let ν_j be the frequency of marker allele j, and assume that markers account for the component of genetic variation equal to $2\sum_j \nu_j(1-\nu_j)a_j^2$. Then as shown by Habier *et al*. (2007) and Gianola *et al*. (2009) a value for $\mathrm{Var}(a_j)$ can be obtained from

$$\mathrm{Var}(a_j) = \frac{\sigma_u^2}{2\sum_j \nu_j(1-\nu_j)}.$$

The value of τ^2 is then set to two orders of magnitude smaller than $\mathrm{Var}(a_j)$. The parameter c^2 is finally obtained from eqn (6). The idea behind this way of choosing τ^2 and $c^2\tau^2$ is to obtain components of the mixture that have the ability to discriminate between markers whose effects on the trait are barely detectable, from those with clearer effects.

The specification of c^{*2}, τ^{*2} and p^* on the GHETMIX model involves setting $p^*=0{\cdot}10$ and fitting first the infinitesimal genetically heterogeneous variance model described in Sorensen & Waagepetersen (2003) to obtain an estimate of $\sigma_{u^*}^2$, the additive genetic variance at the level of the log-variance. The remaining parameters τ^{*2}, c^{*2} are then obtained as in the GHOMMIX model. In the simulation study below we used values of c and c^* approximately equal to 45.

A little intuition for how inferences may be affected by the choice of c can be obtained as follows. The model specifies that when $\delta_i=1$, the SNP effect is a realization from $N(0, \tau^2)$ and when $\delta_i=1$, from $N(0, c^2\tau^2)$. It is readily seen that the ratio of the heights of $N(0, \tau^2)$ and $N(0, c^2\tau^2)$ at $a_j=0$ is equal to c (George & McCulloch, 1993). That is, c can be interpreted as the prior odds of allocating the SNP to the distribution $N(0, \tau^2)$ when a_j is very close to zero. Insight about posterior inferences can be gained by assuming a simple scenario, whereby the model for the data when $\delta_i=0$ is of the form $y_{ij}|a_i, \sigma^2 \sim \mathrm{N}(a_i, \sigma^2)$ and $a_i|\tau^2 \sim N(0, \tau^2)$, $i=1, \ldots, n$, $j=1, \ldots, n_i$, where n_i is the number of records with genotype i whose effect is a_i. On the other hand, when $\delta_i=1$, $a_i|c^2\tau^2 \sim N(0, c^2\tau^2)$. Standard calculations show that the posterior distribution $[a_i|\boldsymbol{y}, \delta_i=0]$ is $N\left(\widehat{a}_i, \frac{\sigma^2}{n_i+\sigma^2/\tau^2}\right)$, and the posterior distribution $[a_i|\boldsymbol{y}, \delta_i=1]$ is $N\left(\widehat{a}_i, \frac{\sigma^2}{n_i+\sigma^2/c^2\tau^2}\right)$, where $\widehat{a}_i$ is BLUP of a_i. The Bayes factor for testing the null hypothesis H_0 of $a_i=0$ against the alternative H_1 that $a_i \neq 0$, computed using the Savage–Dickey ratio (Verdinelli & Wasserman,

https://doi.org/10.1017/S0016672311000012 Published online by Cambridge University Press

1995), given $\delta_i = 1$, is

$$B_{01} = \frac{p(a_i = 0|c^2\tau^2, H_1)}{p(a_i = 0|c^2\tau^2, H_0)} = \frac{N\left(\widehat{a}_i, \frac{\sigma^2}{n_i + \sigma^2/c^2\tau^2}\right)}{N(0, c^2\tau^2)},$$

where both $N(0, c^2\tau^2)$ and $N\left(\widehat{a}_i, \frac{\sigma^2}{n_i + \sigma^2/c^2\tau^2}\right)$ are evaluated at $a_i = 0$. For example, when $\hat{a}_i = 0{\cdot}30$, $\tau^2 = 0{\cdot}00005$, $\sigma^2 = 4$, for values of $c = (40;400)$, B_{01} takes values $(0{\cdot}051; 0{\cdot}072)$, $(1{\cdot}08 \times 10^{-3}; 1{\cdot}79 \times 10^{-3})$, $(5{\cdot}4 \times 10^{-6}; 9{\cdot}1 \times 10^{-6})$, when $n_i = 100$, 500, 1000, respectively. The evidence against the null hypothesis increases with the amount of data n_i, and the tuning parameter c has a modest effect on this inference.

(iv) *McMC algorithm*

The models are implemented using McMC algorithms, where the components in each model are updated sequentially. In general, the McMC algorithm for each model is based on a combination of Gibbs updates, updates based on random walk proposals and updates based on Langevin–Hastings proposals. In addition, a reparameterization described in Sorensen & Waagepetersen (2003) and Waagepetersen *et al.* (2008) is made to improve the mixing of the algorithms. The vector $(\mathbf{a}, \mathbf{a}^*)$ is transformed in its prior distribution to an independently distributed vector (γ, γ^*), with the intention of reducing the posterior correlation. In the GHET model, using the factorization $\mathbf{G} = \mathbf{L_G}\mathbf{L}'_\mathbf{G}$, where $\mathbf{L_G} = \begin{pmatrix} \sigma_a & 0 \\ \rho\sigma_{a^*} & \sqrt{\sigma^2_{a^*}(1-\rho^2)} \end{pmatrix}$ is the lower triangular Cholesky factor of the variance–covariance matrix $\mathbf{G}$, $(a_j, a_j^*)'$ is reparameterized into $\mathbf{L_G}\,(\gamma_j, \gamma_j^*)'$, $j = 1, \ldots, N$, leading to

$$P(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*|\mathbf{G}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma}'\boldsymbol{\gamma} + \boldsymbol{\gamma}^{*\prime}\boldsymbol{\gamma}^*)\right\}, \tag{8}$$

the density of a multivariate normal distribution with mean vector 0 and variance matrix equal to the identity matrix.

In model GHETMIX, $(a_j, a_j^*)'$ is reparameterized into $\begin{pmatrix} \alpha_j\tau & 0 \\ 0 & \alpha_j^*\tau^* \end{pmatrix}(\gamma_j, \gamma_j^*)'$, where (γ_j, γ_j^*) has the *a priori* density as in eqn (8).

In summary, the McMC algorithms for the four models are as follows:

1. For model GHET, μ, exp (μ^*), (γ, γ^*) and $(\sigma^2_a, \sigma^2_{a^*}, \rho)$ are updated in turn using Gibbs updates for μ and exp(μ^*), Metropolis–Hastings updates with Langevin–Hastings proposals for (γ, γ^*), and random walk proposals for σ^2_a, $\sigma^2_{a^*}$ (on the log scale) and ρ.
2. The same algorithm is used for model GHOM, with γ^*, $\sigma^2_{a^*}$ and ρ omitted from the algorithm.
3. For model GHETMIX, μ, exp (μ^*), δ_j and δ_j^* are updated using Gibbs steps, with fully conditional posterior distributions parameterized in terms of $(\mathbf{a}, \mathbf{a}^*)$. These are transformed into (γ, γ^*), which is updated using a Langevin–Hastings proposal.
4. Model GHOMMIX is based on the same algorithm as for model GHETMIX with γ^* and δ^* omitted from the algorithm.

The McMC algorithms run for 1 800 000 cycles after discarding the first 400 000 cycles as burn-in period. The chains were thinned (saved one iteration every 140 cycles), so that the total number of samples kept was 10 000 for all models. Convergence was checked informally looking at traceplots of chosen parameters (data not shown). The algorithms showed good mixing properties. The smallest effective chain size (Sorensen *et al.*, 1995) corresponding to $\sigma^2_{a^*}$ for the GHET model was equal to 1080. This resulted in a 95% Monte Carlo interval equal to 1·4% relative to the posterior mean.

(v) *Model comparison*

The models were evaluated with three criteria, using both simulated data (where the true state of nature is known) and real data. Firstly, interest focused on the ability of GHETMIX and GHET models to detect QTL signals at the level of the mean and variance. Secondly, a measure of the quality of the global fit of the models is reported. It is relevant to study whether this can be used to discriminate among the models' ability to capture the true state of nature. The third criterion is the predictive ability of the models. Additive genetic values were predicted and compared with the true ones using simulated data and cross-validation was used with the real data.

The quality of the global fit of the models was compared using the pseudo marginal probability of the data (Gelfand, 1996) that is defined and computed as follows. Consider data vector $\mathbf{y}' = (y_i, \mathbf{y}'_{-i})$, where y_i is the ith datum, and $\mathbf{y}_{-i}$ is the vector of data with the ith datum deleted. For a particular model M, the conditional predictive distribution is

$$P(y_i|\mathbf{y}_{-i}, M) = \int P(y_i|\boldsymbol{\theta}, \mathbf{y}_{-i}, M)P(\boldsymbol{\theta}|\mathbf{y}_{-i}, M)\mathrm{d}\boldsymbol{\theta}, \tag{9}$$

and can be interpreted as the likelihood of each datum given the remainder of the data. The actual value of $P(y_i|\mathbf{y}_{-i}, M)$ is known as the conditional predictive ordinate (CPO) for the ith observation, where $\boldsymbol{\theta}$ is the vector of model parameters. A Monte Carlo approximation to the CPO in eqn (9) for observation i is given by (Gelfand, 1996)

$$\widehat{P}(y_i|\mathbf{y}_{-i}, M) = \left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{P(y_i|\boldsymbol{\theta}^{(k)}, M)}\right]^{-1}, \tag{10}$$

where K is the number of McMC draws and $\boldsymbol{\theta}^{(k)}$ is the kth draw from the posterior distribution of $\boldsymbol{\theta}$ under model M.

https://doi.org/10.1017/S0016672311000012 Published online by Cambridge University Press

For a given model, the log-pseudo-marginal probability of the data is $\sum_{i=1}^{n}\log(\mathrm{CPO}_i)$, which is a surrogate for the Bayes factor (Geisser & Eddy, 1979; Gelfand *et al*., 1992). A larger value of $\sum_{i=1}^{n}\log(\mathrm{CPO}_i)$ indicates a better relative fit. The log-pseudo-marginal probability of the data is used to compare different genomic models.

3. Simulation study

The aim of the simulation study is to examine properties of the proposed methods in their ability to detect QTL signals and predicting breeding values. A third objective is to study whether the models are capable of capturing the true state of nature. This is done by fitting the four models to simulated data and computing the log-pseudo-marginal probability of the data under each model.

The data were simulated mimicking two scenarios. In scenario 1, a total of eight QTLs were placed in five chromosomes with length 100 cM each. Of these eight QTLs, four affected mean only, and four affected variance only. In both cases, three QTLs were assumed to have relatively substantial effects and one had a relatively little effect. In scenario 2, 80 QTLs were placed along the five chromosomes. Of these 80 QTLs, 30 had the effects only on the mean, 30 only on the variance and 20 were pleiotropic with the effects on both. In the two scenarios 4779 biallelic SNP markers evenly distributed in five chromosomes at a distance of approximately 0·1 cM between adjacent markers were available to detect the QTL.

The data were simulated as follows. Initially, a population size of 100 individuals was created, of which half were females and the other half males. Each of these individuals was allocated a genotype with 5000 biallelic SNP markers evenly placed in the five chromosomes. In addition, at this stage, 100 QTLs were randomly placed in the five chromosomes. Recombination between adjacent loci was generated using Haldane's mapping function (with no interference). A total of 50 generations of random mating were simulated. At generation 51, the population size was incremented to 100 males and 1000 females, and kept constant at a size of 1100 individuals for two extra generations. Each of the 100 males mated to 10 females. Thus, at generation 52, the pedigree consisted of full-sibs (one male and one female) and of half-sib families. A similar strategy was used to produce individuals of generation 53. The data for analysis belong to generations 51–53. At this final stage, the QTLs were allocated an effect and markers with a gene frequency smaller than 0·05 were discarded. A total of 4779 marker loci satisfied this criterion and were included in the final data set. Among these 3300 individuals of generations 51, 52 and 53, the average squared correlation of gene frequencies between adjacent pairs of marker loci (approximately 0·1 cM apart) was 0·20. The results of Calus *et al*. (2008) and Meuwissen *et al*. (2001) suggest that this level of linkage disequilibrium is sufficient to achieve high accuracies of prediction of genomic breeding values.

Genetic variation was generated as follows. QTL j was allocated an effect b_j and its additive genetic variance at the level of the mean was computed as $\sigma^2_{b_j}=2q_j(1-q_j)b_j^2$, where q_j is the observed frequency of the favourable allele at QTL j. In scenario 1, the values of b_j (in absolute values) range from 0·2 to 1·5 units, and in scenario 2 from 0·2 to 0·45 units. At the level of the variance, these figures are 0·3–1·5 for scenario 1, and 0·20–0·45 for scenario 2. The total additive genetic variance at the level of the mean, σ^2_b, was defined as the sum of the contributions $\sigma^2_{b_j}$ from each QTL, ignoring the correlation structure due to linkage disequilibrium. The total additive genetic variance at the level of the environmental variance, σ^{*2}_b, was generated in a similar manner, by summing contributions $\sigma^2_{b_j^*}=2q_j(1-q_j)b_j^{*2}$ from each QTL.

In the two simulated scenarios, the overall mean was $\mu=50$ and the total additive genetic variances (ignoring the covariance due to linkage disequilibrium) at the level of the mean and variance (σ^2_b and σ^{*2}_b) are equal to 1·98 and 1·85. The term exp (μ^*) was set equal to 2·00, which resulted in a heritability equal to 0·28. The value of $\sigma^2_{b^*}$ used in these scenarios is rather large compared to the estimates reported in the literature (summarized in Mulder *et al*., 2007). The implications for the probability of detecting effects on mean and variance are elaborated in the discussion.

The 2200 individuals from generations 51 and 52 are allocated a single phenotypic record, whereas the 1100 from generation 53 have only genotypic values, determined by the sum of the effects of the individual QTL.

(i) *The detection of QTL*

For simulated scenario 1, the ability of the GHETMIX model to detect signals is displayed at the top of Fig. 1. For the three QTLs of relatively large effect on the mean, in chromosomes 1, 3 and 4, the Monte Carlo estimates of the posterior probabilities $P(\delta_j|\mathbf{y},M)$ of the markers closest to these QTL are in the vicinity of 1. However, the model fails to detect the signal in chromosome 2 due to the QTL of small effect. The picture concerning detection at the level of the variance (top, right of Fig. 1) is very similar. The bottom of Fig. 1 shows posterior means of marker effects from the GHET model plotted against their position in the genome. There is overall agreement between the results from both models.

https://doi.org/10.1017/S0016672311000012 Published online by Cambridge University Press

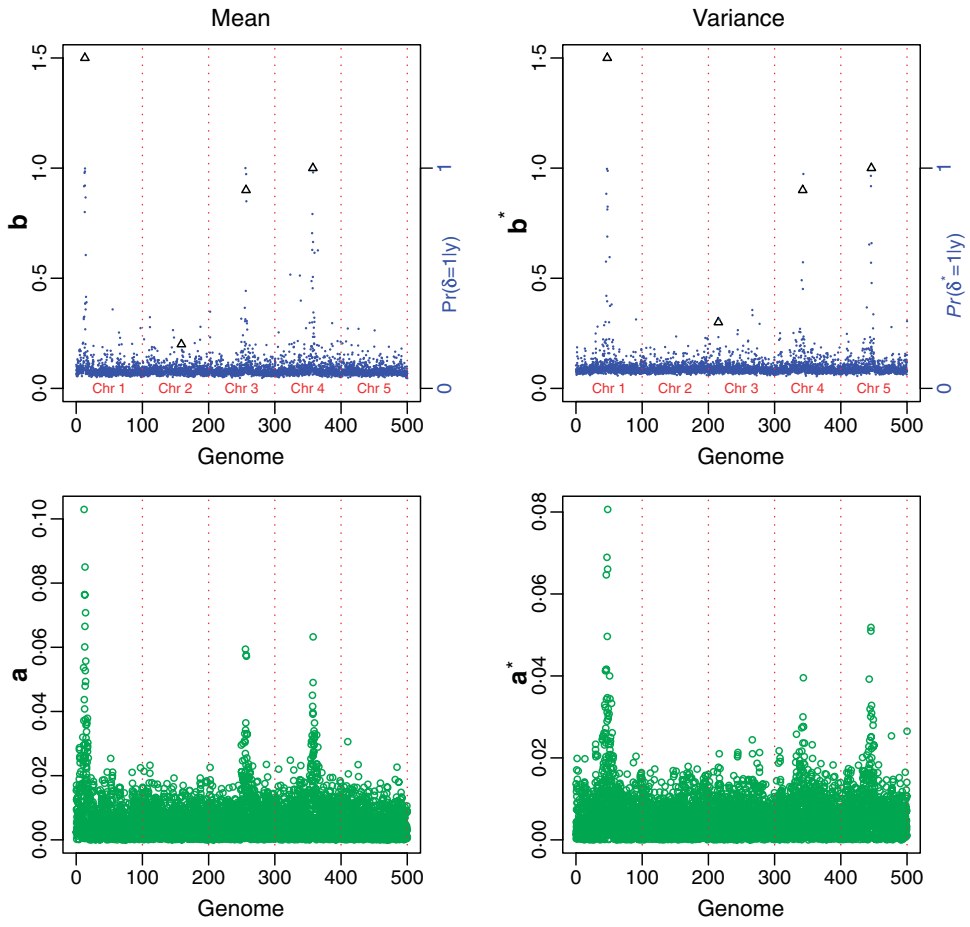

Fig. 1. Simulated scenario 1. Top: results from the GHETMIX model, with $p=p^*=0{\cdot}1$ and $(\tau^2, c^2\tau^2, \tau^{*2}, c^{*2}\tau^{*2})=(0{\cdot}5\times10^{-5}, 0{\cdot}1\times10^{-1}, 0{\cdot}5\times10^{-5}, 0{\cdot}1\times10^{-1})$. The true QTL effects (**b** and **b***, black triangle pointing upwards) and posterior probabilities of marker indicators (blue solid circles) are plotted against marker locations along the genome with effects on mean (left) and on environmental variance (right). Bottom: similar results from the GHET model, with posterior means of marker effects **a** and **a*** in the *Y*-axis.

The mixture models have the attractive property that they readily provide Monte Carlo estimates of Bayes factors as evidence for the presence/absence for a QTL at each SNP via the ratios of posterior to prior odds of detection. For example, for the SNP on chromosome 1 affecting mean, the Bayes factor is $(0{\cdot}97/0{\cdot}03)/(0{\cdot}10/0{\cdot}90)=291$, which is decisive evidence for the presence of a QTL associated with the SNP (see Kass & Raftery (1995), for guidelines for interpreting actual values). A posterior probability of 0·5 results in a Bayes factor equal to $(0{\cdot}5/0{\cdot}5)/(0{\cdot}10/0{\cdot}90)=9$, which is interpreted as substantial evidence for the presence of a QTL associated with the SNP.

The performance of GHETMIX model under simulated scenario 2 is shown in Fig. 2. The model can detect QTL successfully at the level of the mean. Indeed, for many of the markers close to the QTL, the posterior probabilities of the indicator variable is in the proximity of 0·5, and these are scattered along the genome, in agreement with the genetic model. Similar results hold at the level of the variance with several markers associated with posterior probabilities of the indicator variable around 0·5 and a few at higher values. Inferences from GHET model are shown at the bottom of Fig. 2. The signals are also scattered along the genome in agreement with the true genetic model, with a few markers showing larger effects than the rest. The pattern is similar as with the GHETMIX model. Arguably, the size of the estimates of SNP effects from the GHET model, as a source of evidence for the presence of regions affecting the trait, is not as clear to interpret as the posterior probabilities generated by the mixture models.

(ii) *Model comparison*

The results of the model comparison are shown in Table 1. The four models were fitted to the data simulated under scenarios 1 and 2, and the log-marginal probability of the data under each model is computed (third column of Table 1). In both scenarios, the $\sum_i \log(\mathrm{CPO}_i)$ were relatively larger under the models postulating QTL effects at the level of mean and variance (models GHET and GHETMIX), and

https://doi.org/10.1017/S0016672311000012 Published online by Cambridge University Press

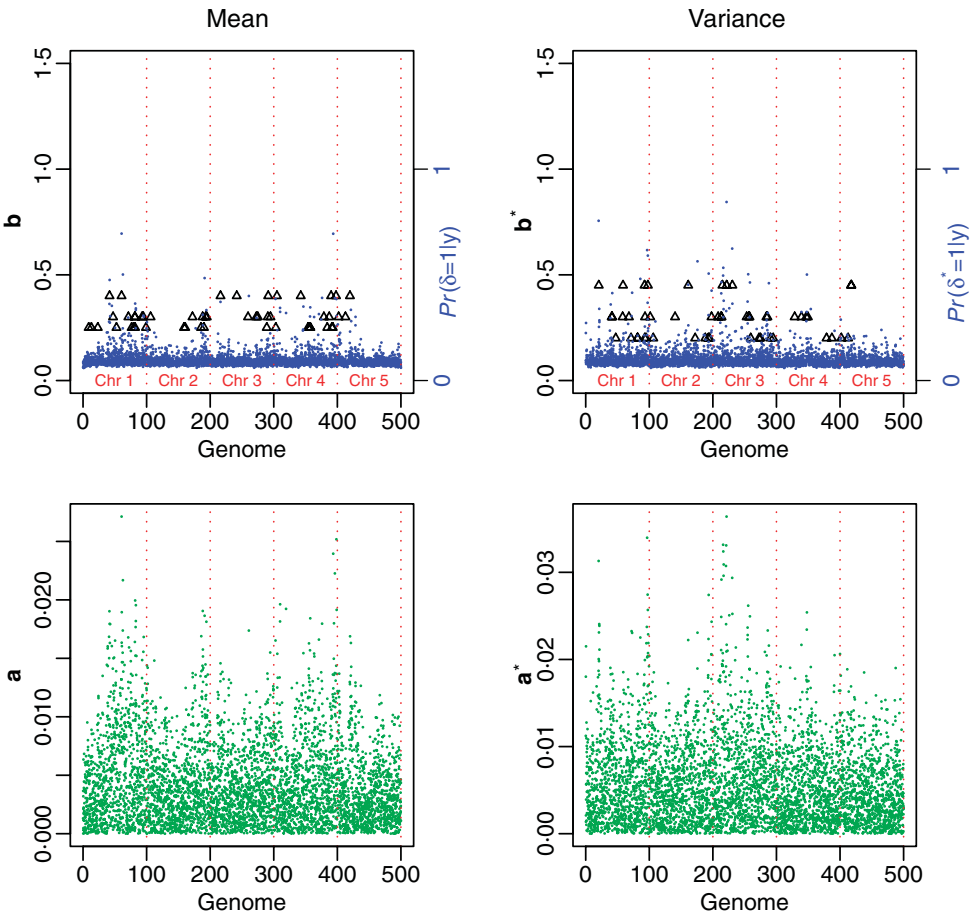

Fig. 2. Simulated scenario 2. Top: results from the GHETMIX model, with $p = p^* = 0{\cdot}1$ and $(\tau^2, c^2\tau^2, \tau^{*2}, c^{*2}\tau^{*2}) = (0{\cdot}5 \times 10^{-5}, 0{\cdot}11 \times 10^{-1}, 0{\cdot}5 \times 10^{-5}, 0{\cdot}01)$. The true QTL effects (**b** and **b***, black triangle pointing upwards) and posterior probabilities of marker indicators (blue solid circles) are plotted against marker locations along the genome with effects on mean (left) and on environmental variance (right). Bottom: similar results from the GHET model, with posterior means of marker effects **a** and **a*** in the *Y*–axis.

the best overall fit is obtained with the GHETMIX model.

(iii) *Genomic prediction*

The predictive ability of the four models is studied computing the correlation between the true breeding values and the predicted genomic breeding values. The predicted genomic breeding value at the level of the mean for the *i*th individual is

$$\widehat{g}_i = \mathbf{w}_i' \widehat{\mathbf{a}},$$

and a similar predicted genomic breeding value at the level of the variance (with obvious notation) is

$$\widehat{g}_i^* = \mathbf{w}_i' \widehat{\mathbf{a}}^*,$$

where $\mathbf{w}_i$ is the column vector of marker genotypes for the *i*th individual in generation 53, coded as -1 for genotype 11, 0 for genotype 12 and 1 for genotype 22. We distinguish the genotypic indicator matrix **X** of individuals belonging to generation 51 and 52 from **W**, associated with those of 53. The $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{a}}^*$ are the vectors of posterior means of marker effects operating at the level of mean and variance among individuals belonging to generation 53, estimated from a given model using phenotypes belonging to generations 51 and 52.

Results are shown in the last two columns of Table 1. At the level of the mean, model GHETMIX produces the largest correlations for scenarios 1 and 2, and the difference in favour of model GHETMIX is relatively more visible in the first scenario. The GHOM model results in the smallest correlations in both scenarios.

At the level of the variance, the difference in favour of model GHETMIX relative to model GHET is also more pronounced in scenario 1.

The results from the genomic models can be placed in perspective by comparing with the correlations achievable using pedigree information only, ignoring marker information. An additive model with homogeneous environmental variance resulted in a correlation between true and predicted breeding values, at

https://doi.org/10.1017/S0016672311000012 Published online by Cambridge University Press

Table 1. *The log-pseudo-marginal probability of the data and correlation* (Corr(TBV, PGBV)) *between true* (TBV) *and predicted genomic breeding value* (PGBV)*, at the level of the mean and variance, obtained from the four models fitted to data simulated under scenarios 1 and 2. In both scenarios,* $p = 0{\cdot}1$ *and* $p^* = 0{\cdot}1$*. In scenario 1,* $(\tau^2, c^2\tau^2, \tau^{*2}, c^{*2}\tau^{*2}) = (0{\cdot}5 \times 10^{-5}, 0{\cdot}1 \times 10^{-1}, 0{\cdot}5 \times 10^{-5}, 0{\cdot}1 \times 10^{-1})$*, and in scenario 2,* $(\tau^2, c^2\tau^2, \tau^{*2}, c^{*2}\tau^{*2}) = (0{\cdot}5 \times 10^{-5}, 0{\cdot}11 \times 10^{-1}, 0{\cdot}5 \times 10^{-5}, 0{\cdot}01)$

			Corr(TBV, PGBV)	
Data	Model	$\sum_{i=1}^{n} \log(\text{CPO}_i)$	Mean	Variance
Scenario 1	GHOM	−3245·9	0·77	
	GHOMMIX	−3203·2	0·86	
	GHET	−2520·8	0·85	0·73
	GHETMIX	−2335·6	0·93	0·85
Scenario 2	GHOM	−4136·4	0·55	
	GHOMMIX	−4135·5	0·57	
	GHET	−3496·7	0·69	0·76
	GHETMIX	−3480·7	0·71	0·79

the level of the mean, equal to 0·516. The additive model with genetically structured environmental variance produced a value of 0·522, and at the level of the environmental variance, the correlation between true and predicted breeding values was 0·400. For this simulated data set, there is a clear difference in favour of the genomic models. The predictive performance of a model with both polygenic and marker effects was not included in this study. Calus & Veerkamp (2007) show that very little is gained using such a model unless the average squared correlation coefficient between adjacent markers is lower than 0·10 for low heritability traits and lower than 0·14 for high heritability traits. On the other hand when the focus is detection of genomic regions rather than prediction, omission of a polygenic effect may affect inferences in at least two ways. Firstly, not accounting for the correlated error structure induced by the polygenes can lead to underestimation of uncertainty. Secondly, effects of the omitted polygenes may be captured by the SNPs leading to overestimation of their effects. These consequences are likely to be more pronounced at low SNP marker densities. This has not been investigated in the present work.

4. Real data analysis – back fat thickness in pigs

Results of a pilot study are reported using back fat thickness measurements taken on 960 Landrace boars from the Danish nucleus breeding herds. The objective is to illustrate the application of the methods developed rather than studying details of the genetic architecture of the trait.

(i) *Data*

SNP marker genotypes of 960 boars were obtained using a 6 K Illumina bead chip, from which 2011 SNPs had good quality and minor allele frequency larger than 5 %. Each of the 960 boars has also a back fat record, which was corrected for weight prior to the analysis. The square root of back fat was used since in this scale the posterior distribution of the coefficient of skewness is symmetrical. The heritability based on a classical infinitesimal model was estimated to be 0·24.

A glance at the pedigree file constructed using two-generation data revealed that the genotyped offspring consisted of 225 full-sibs, 405 half-sibs, and the remaining individuals were unrelated. This pedigree information is not incorporated into the genomic models used in the present study. Similar data were used by Janss *et al.* (2009).

(ii) *The detection of QTL*

Fig. 3 shows the results from fitting the GHETMIX (top) and GHET (bottom) models to back fat data. The top figure on the left is suggestive of the presence of genomic regions with effects on the mean. Five of these are associated with posterior probabilities of the indicator function larger than 0·45, and for one of these five, the probability is larger than 0·65. Results from fitting the GHET model lead to similar conclusions. Both models fail to produce signals suggesting detectable effects on the variance (right panels).

Another way of viewing the results is displayed in Fig. 4, that shows the distribution of Monte Carlo estimates of the posterior probabilities of the indicator function across the markers at the level of the mean (left), $P(\delta_j = 1|\mathbf{y})$, and at the level of the variance, $P(\delta_j^* = 1|\mathbf{y})$, (right) obtained from model GHETMIX. The Monte Carlo estimate of the posterior probability associated with each marker, is obtained by summing the Monte Carlo draws of the marker's indicator function over the Monte Carlo samples, and dividing by the number of samples. The left figure indicates that most of the markers are associated with very small probabilities, essentially reproducing the prior probability (0·10), indicating absence of QTL in their proximity. However, the figure uncovers the five markers with relatively larger effects on the mean that are vaguely discernible at the larger end of the probability scale. The histogram on the right reflects what is to be expected if the data are uninformative about the marker indicator $\boldsymbol{\delta}^*$, so that $P(\delta_j^* = 1|\mathbf{y}) = P(\delta_j^* = 1|p) = p$. In this case, the value of 1 for the indicator function is randomly assigned to the markers with probability p. The histogram represents the sampling distribution of the Monte Carlo estimator of $P(\delta_j^* = 1|\mathbf{y})$, $\mu_j = 1/K \sum_{t=1}^{K} I^{(t)}(\delta_j = 1|p)$, where $I^{(t)}(\delta_j = 1|p)$ is the value of the indicator function

https://doi.org/10.1017/S0016672311000012 Published online by Cambridge University Press

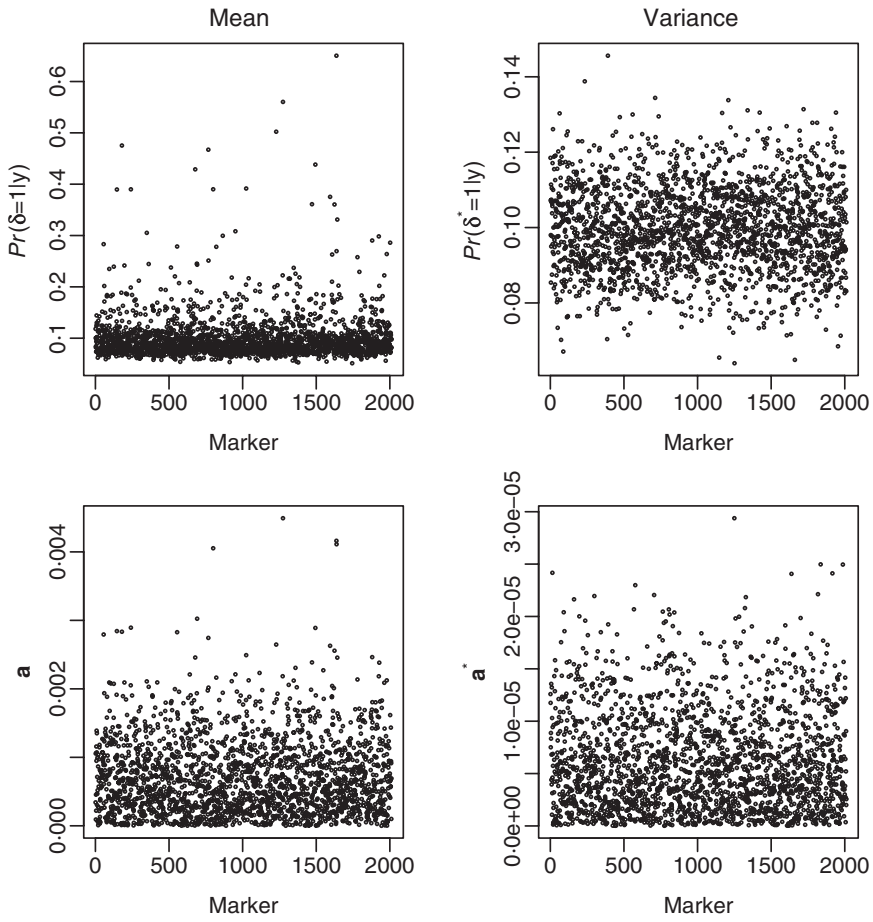

Fig. 3. Back fat data. Top: results from the GHETMIX model, with $p=p^*=0{\cdot}1$ and $(\tau^2, c^2\tau^2, \tau^{*2}, c^{*2}\tau^{*2})=(0{\cdot}5\times10^{-7}, 0{\cdot}1\times10^{-3}, 0{\cdot}5\times10^{-9}, 0{\cdot}2\times10^{-5})$. Posterior probabilities of the indicator function plotted against marker number for QTL effects on mean (left) and on environmental variance (right). Bottom: results from the GHET model, with posterior means of marker effects **a** affecting mean (left) and variance **a*** (right) in the *Y*–axis.

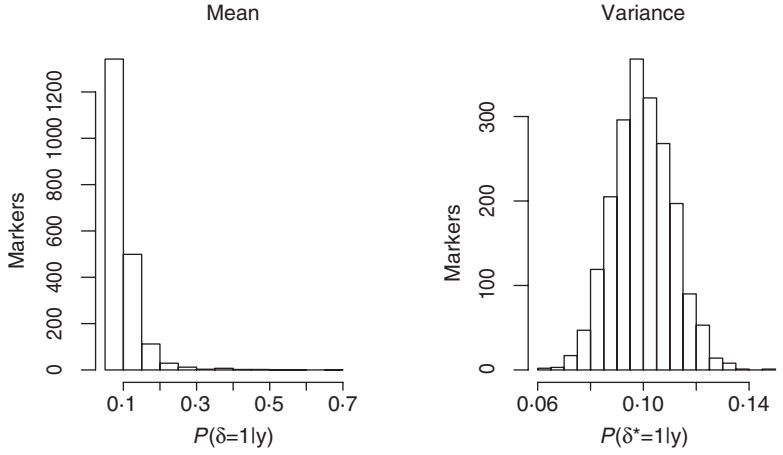

Fig. 4. Histograms of posterior probabilities of marker indicators from the GHETMIX model, across number of markers, at the level of the mean (left) and variance (right) for back fat data.

for marker j in round t, and $K=10\,000$ is the length of the McMC chain. Estimator μ_j is asymptotically normally distributed with mean $p=0{\cdot}1$ and standard deviation $\sqrt{V_{\text{asymp}}/K}$, where $V_{\text{asymp}}=\lim_{N\to\infty}\text{Var}\left(\sqrt{K}\mu_j\right)$, the limiting variance of $\sqrt{K}\mu_j$ (Geyer, 1992). With independent draws, the SD of μ_j is $\sqrt{p(1-p)/K}=0{\cdot}003$. In our case, using the estimator of the asymptotic variance proposed by (Geyer, 1992) , the SD of μ_j is

https://doi.org/10.1017/S0016672311000012 Published online by Cambridge University Press

Table 2. *Monte Carlo estimates of* $\sum_i \log(\text{CPO}_i)$*, of the correlation between observed and predicted data* $\text{Corr}(\mathbf{y}, \hat{\mathbf{y}})$ *obtained from the cross-validation study, and of the measure of predictive ability at the level of the variance given by the average of expression (11), D, for the four genomic models fitted to back fat data, and different values of p.* $(\tau^2, c^2\tau^2, \tau^{*2}, c^{*2}\tau^{*2})$ *is* $(0{\cdot}5\times10^{-7}, 0{\cdot}1\times10^{-3}, 0{\cdot}5\times10^{-9}, 0{\cdot}2\times10^{-5})$ *for* $p=p^*=0{\cdot}1$, $(\tau^2, c^2\tau^2)$ *is* $(0{\cdot}1\times10^{-7}, 0{\cdot}55\times10^{-4})$ *for* $p=0{\cdot}2$ *and* $(0{\cdot}5\times10^{-7}, 0{\cdot}22\times10^{-4})$ *for* $p=0{\cdot}5$

$P(\boldsymbol{\delta}=1\|p)$ or $P(\boldsymbol{\delta}^*=1\|p)$	Model	$\sum_i \log(\text{CPO}_i)$	$\text{Corr}(\mathbf{y}, \hat{\mathbf{y}})$	D
$p=0{\cdot}1$	GHOM	1205·0	0·30	6·13
	GHOMMIX	1204·9	0·31	6·11
	GHET	1204·6	0·30	6·15
$p=p^*=0{\cdot}1$	GHETMIX	1204·4	0·31	6·13
$p=0{\cdot}2$	GHOMMIX	1206·4	0·31	6·11
$p=0{\cdot}5$	GHOMMIX	1207·3	0·31	6·11

approximately 0·0097 (similar figure for all j). The histogram reproduces very well this null distribution, suggesting that markers have no detectable effects at the level of the variance of back fat.

The posterior means of marker effects obtained from model GHET were investigated and found to be in good agreement with those from model GHETMIX. For example, the two largest posterior means from model GHETMIX are also the two largest from model GHET. The third, fourth and fifth largest from model GHETMIX correspond to the 13th, 12th and 9th largest from model GHET.

(iii) *Model comparison*

The third column of Table 2 shows Monte Carlo estimates of $\sum_i \log(\text{CPO}_i)$ for the four models. The similarity of the estimates of $\sum_i \log(\text{CPO}_i)$ for models GHOM and GHOMMIX (1205·0 and 1204·9) does not make it possible to discriminate between them. Increasing the complexity of the models by introducing variance heterogeneity due to the effects of markers does not improve the global fit. This result together with the evidence in Fig. 4 does not lend support to the presence of a detectable genetic component at the level of the residual variance.

The bottom two rows in Table 2 show that global fit is hardly affected by changes in the tuning parameter p. Indeed, setting p equal to 0·2 or to 0·5 has little influence on the estimated values of $\sum_i \log(\text{CPO}_i)$.

In addition to the four genomic models, three other models were fitted to the data: a simple model with only two parameters (mean μ and homogeneous variance σ^2) of the form $\mathbf{y}|\mu,\sigma^2 \sim N(\mu, \sigma^2)$, a pedigree based only infinitesimal homogeneous variance model with normally distributed genetic effects $\mathbf{a}$ of the form $y_i|\mu, u_i, \sigma^2 \sim N(\mu+u_i, \sigma^2)$, $\mathbf{u}|\sigma_u^2 \sim N(0, \mathbf{A}\sigma_u^2)$, and finally a genetically structured heterogeneous variance model with additive genetic effects affecting mean ($\mathbf{u}$) and variance ($\mathbf{u}^*$), of the form $y_i|\mu, u_i, u_i^* \sim N(\mu+u_i, \exp(\mu^*+u_i^*))$, $\mathbf{u}, \mathbf{u}^*|\mathbf{G} \sim N(0, \mathbf{A}\otimes \mathrm{G})$ as in Sorensen & Waagepetersen (2003), also pedigree based only. In this model, $\mathbf{A}$ is the additive genetic relationship matrix and $\mathbf{G}$ is the covariance matrix associated with the joint distribution $[\mathbf{u}, \mathbf{u}^*|\mathbf{G}]$. The $\sum_i \log(\text{CPO}_i)$ for these models were 1156·4, for the simple model, 1171·9 for the second and 1173·2 for the third. The models postulating a genetic component are better supported by the data, but once again there is no additional support for a genetic component at the level of the variance. All these models produce lower measures of global fit than the genomic models.

Estimates of posterior means (95% posterior intervals in brackets) of parameters based on pedigree information only from the genetically structured heterogeneous variance model were as follows. For the additive genetic variance at the level of the mean, 0·0073 (0·0032; 0·012), at the level of the variance, $0{\cdot}10\times10^{-3}$ $(0{\cdot}14\times10^{-4}; 0{\cdot}28\times10^{-3})$ (the estimate of the posterior mode is $0{\cdot}47\times10^{-4}$) and for the genetic correlation, 0·01 (−0·91; 0·98). The modal values of the prior distributions were, for the additive genetic variance at the level of the mean, 0·0034, at the level of the variance, $0{\cdot}43\times10^{-4}$, and the correlation was assumed to be uniformly distributed between −1 and 1. The posterior mode of the additive genetic variance at the level of the variance does not differ from the mode of the prior distribution, and the posterior distribution of the correlation coefficient is centred in the vicinity of zero, with a posterior uncertainty covering almost the whole support of the prior distribution, indicating no Bayesian learning from the data. These results are not in conflict with the absence of genetic variability at the level of the variance.

(iv) *Genomic prediction*

A six-fold cross-validation study was carried out allocating individuals randomly into six folds of equal size. The predicted phenotypes $\hat{\mathbf{y}}_{-f}$ $(f=1, \ldots, 6)$ were obtained using estimates of parameters obtained by fitting the four models to data in which records from the fth fold were excluded. The predictive ability of a model was assessed by the average correlation

https://doi.org/10.1017/S0016672311000012 Published online by Cambridge University Press

(over the six folds) between observed and predicted phenotypes. The predictive ability of a model at the level of the residual variance was obtained by the average over the six folds of the quantity

$$D=\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{S}_i^2-S_i^2\right)^2, \tag{11}$$

where $\widehat{S}_i^2=\widehat{\mu}^*+\sum_{j=1}^{N}X_{ij}\widehat{a}_j^*$ and $S_i^2=\log\left[\left(y_i-\widehat{\mu}-\sum_{j=1}^{N}X_{ij}\widehat{a}_j\right)^2\right]$. In these expressions, $\widehat{\mu}, \widehat{a}_j, \widehat{\mu}^*, \widehat{a}_j^*$ are the posterior means of the model parameters.

The results in the last two columns of Table 2 reveal the same pattern as before, for the measure of global fit (third column of the table). Increasing the complexity of the models by inclusion of markers at the level of the variance does not improve the correlation between observed and predicted phenotypes and does not improve prediction at the level of the variance. These results are not affected by changes in the tuning parameter p.

The overall conclusion from the analysis is that a model postulating genetic heterogeneity of residual variance of back fat is not supported by the data. However, the analysis signals the existence of approximately five genomic regions with detectable effects on the trait at the level of the mean. As mentioned above, a limitation of the present analysis is that a polygenic effect was not included. This could have resulted in an overestimation of marker effects.

5. Discussion

Genomic models designed to detect QTL effects on the mean and variance were developed and McMC algorithms were constructed for their implementation. A study using simulated data with known QTL positions confirmed the ability of the genomic models to detect signals at the level of mean and variance. The strength of the signals clearly depends on the size of the effects and on whether the QTLs operate on mean or variance, with the expectation that detection of effects at the level of the variance may require a larger experiment than detection at the level of the mean. An approximation to the relative sizes of experiment needed can be arrived at as follows. Consider datum on individual i, y_{ki}, carrying marker genotype 1 or 2, with effects a_k on the mean ($i=1, 2, \ldots, n$; $k=1, 2$), n replications per genotype. Individuals with genotype 1 have known residual variance σ^2 and those with genotype 2 have residual variance $\sigma^2\sigma^2_{a^*}=\sigma^2_*$, where $\sigma^2_{a^*}$ is unknown. Thus, individuals carrying genotype 2 have residual variance that is scaled by the factor $\sigma^2_{a^*}$ relative to the variance of individuals carrying genotype 1. Assuming normality of $y_{ki}|a_k$, the variance of the maximum likelihood estimator of a_1 is σ^2/n and of a_2 is σ^2_*/n. The variance of the estimator of the difference $\Delta_m=a_2-a_1$ is $\sigma^2\,(1+\sigma^2_{a^*})/n$, and that of the estimator of $\sigma^2_{a^*}$ is 2 $(\sigma^2_{a^*})^{\,2}/n$ (the asymptotic variance of the variance is used for simplicity). Under the null hypothesis, $\Delta_m=0$ at the level of the mean, and $\sigma^2_{a^*}=1$ at the level of the variance. Using standard calculations (for example, Chow *et al*., 2008), the ratio of sample sizes required to detect QTL effects on mean and variance, assuming the same probabilities of type I error and the same power, is given by

$$\frac{n_v}{n_m}=\left(\frac{\Delta_m}{\sigma}\right)^2\frac{2\left(\sigma^2_{a^*}\right)^2}{\left(1+\sigma^2_{a^*}\right)\left(1-\sigma^2_{a^*}\right)^2},\quad \sigma^2_{a^*}\neq 1, \tag{12}$$

where n_m (n_v) is the sample size required to detect an effect on the mean (variance). The first term on the right-hand side specifies the standardized size of the difference to be detected at the level of the mean, and $\sigma^2_{a^*}$ specifies the size of the effect operating at the level of the variance. For $\Delta_m/\sigma=0{\cdot}20$ and $0{\cdot}83<\sigma^2_{a^*}<1{\cdot}23$, $\sigma^2_{a^*}\neq 1$, the ratio is bigger than 1, indicating that for a large range of scenarios it is harder to detect effects on the variance. For example, setting $\Delta_m/\sigma=0{\cdot}20$ and $\sigma^2_{a^*}=1{\cdot}1$, detecting effects on variance requires an experiment five times larger than on the mean. However, if $\Delta_m/\sigma=0{\cdot}15$ and $\sigma^2_{a^*}=1{\cdot}2$, an approximate representation for simulated scenario 2, the ratio is 0·7, indicating that detection of effects on mean and variance is approximately equally demanding. Alternatively, given the same probabilities of type *I* error and the same power for detection at the level of mean and variance, for a given sample size, setting $n_m=n_v$, sizes of effects to be detected at the level of variance equal to $\sigma^2_{a^*}=1{\cdot}1$, 1·3, 1·5, 2·0, would allow one to detect effects at the level of the mean equal to $\Delta_m/\sigma=0{\cdot}09$, 0·25, 0·37, 0·61, respectively. A detailed analysis on the statistical power to detect loci affecting environmental variance was recently reported by Visscher & Postuma 2010).

The analysis of back fat data does not provide support for a genetic component at the level of the environmental variance. In general, inferences at the level of the variance can be sensitive to the presence of skewness of the conditional distribution of the data. Ros *et al*., (2004) and Mulder *et al*., (2007) show that in a model with genetic components at the level of mean and variance, the skewness of the marginal distribution of the data is directly proportional to the correlation between additive genetic values affecting mean and variance. Therefore, if the distribution of the data is skewed in either direction not necessarily due to the presence of a genetically structured variance heterogeneity, such a model would fit relatively better than a standard model with homogeneous variance, and this would result in spurious inferences. Despite the negative results concerning the detection

https://doi.org/10.1017/S0016672311000012 Published online by Cambridge University Press

of marker effects on the variance, the distribution of residuals was investigated computing Monte Carlo estimates of the posterior distribution of the residual skewness. Results revealed no signs of asymmetry (data not shown).

The genomic models were implemented treating the variances of the mixture distributions and the probability parameter p as known parameters, to be tuned by the user. This was not a severe limitation. Once a rough estimate of the overall additive genetic variance at the level of the mean and variance (fitting pedigree based only infinitesimal models) is available, then the remaining parameters can easily be tuned. Measures of global fit are not sensitive to perturbations in p, as shown in Table 2. However an extension of the McMC algorithm that allows joint inferences of parameters avoiding tuning is in principle straightforward. For example, for the GHETMIX model, let $\sigma^2_a = c^2\tau^2$, $\sigma^2_{a*} = c^{*2}\tau^{*2}$. Then the prior distributions of marker effects at the level of mean a_j and variance a^*_j can be written as $P(a_j|\delta_j, \sigma^2_a) = \delta_j N\ (0,\sigma^2_a) + \ (1-\delta_j)N\ (0, \sigma^2_a/k)$ and $P(a^*_j|\delta^*_j,\sigma^2_{a*}) = \delta^*_j N\ (0,\ \sigma^2_{a*}) + \ (1-\delta^*_j)N\ (0, \sigma^2_{a*}/k)$, where the constant k is chosen to be equal to 1000, say, so that the components of the mixture have good discriminating ability. Assuming that the priors for σ^2_a and σ^2_{a*} are scaled inverse chi-square distributions, and a common beta distribution is assigned as prior for p and p^*, then these two sets of prior distributions are conjugate for the respective fully conditional posterior distributions of σ^2_a, σ^2_{a*}, p and p^* and updates are immediate via Gibbs steps. The remaining parameters are updated using the same strategy used with the GHETMIX model. The algorithm is therefore easy to construct but, in general, the behaviour of the resulting Markov chain will be influenced by the structure of the data and the properties of the trait analysed.

References

Ansel, J., Bottin, H., Rodriguez-Beltran, C., Damon, C., Nagarajan, M., Fehrmann, S., François, J. & Yvert, G. (2008). Cell-to-cell stochastic variation in gene expression is a complex genetic trait. *PLoS Genetics* **4**(4), doi:10.1371/journal.pgen.1000049.

Calus, M. P. L. & Veerkamp, R. F. (2007). Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics* **124**, 362–368.

Calus, M. P. L., Meuwissen, T. H., de Roos, A. P. & Veerkamp, R. F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* **17**, 553–561.

Chow, S.-C., Shao, J. & Wang, H. (2008). *Sample Size Calculations in Clinical Research*. London: Chapman and Hall.

Geisser, S. & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (ed. W. R. Gilks, S. Richardson & D. J. Spiegelhalter), pp 145–161 London: Chapman and Hall.

Gelfand, A. E., Dey, D. K. & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In: *Bayesian Statistics 4* (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith), pp 147–167. Oxford: University Press.

George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **8**, 881–889.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science* **7**, 473–511.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E. & Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* **183**, 347–363.

Gonzalez-Recio, O., Gianola, D., Long, N., Weigel, K. A., Rosa, G. J. M. & Avendaño, S. (2008). Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics* **178**, 2305–2313.

Gutierrez, J. P., Nieto, B., Piqueras, P., Ibáñez, N. & Salgado, C. (2006). Genetic parameters for canalisation analysis of litter size and litter weight at birth in mice. *Genetics, Selection, Evolution* **38**, 445–462.

Habier, D., Fernando, R. & Dekkers, J. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389–2397.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J. & Goddard, M. E. (2009). Genomicselection in dairy cattle: Progress and challenges. *Journal of Dairy Science* **92**, 433–443.

Heffner, E. L., Sorrell, M. E. & Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Science* **49**, 1–12.

Ibáñez, N., Varona, L., Sorensen, D. & Noguera, J. L. (2007). A study of heterogeneityof environmental variance for slaughter weight in pigs. *Animal* **2**, 19–26.

Ibáñez, N., Moreno, A., Nieto, B., Piqueras, P., Salgado, C. & Gutierrez, J. P. (2008*a*). Genetic parameters related to environmental variability of weight of mice; signs of correlated canalised response. *Genetics, Selection, Evolution* **40**, 279–293.

Ibáñez, N., Sorensen, D., Waagepetersen, R. & Blasco, A. (2008*b*). Selection for environmental variation: a statistical analysis and power calculations to detect response. *Genetics* **180**, 2209–2226.

Janss, L. L. G., Nielsen, B., Christensen, O. F., Bendixen, C., Sørensen, K. K. & Lund, M. S. (2009). Genomic prediction for backfat in pigs. In *Book of Abstracts of the 60th Annual Meeting of the European Association for Animal Production*, p. 211. Waginen The Netherlands: Wageningen Academic Publishers.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Legarra, A., Robert-Granie, C., Manfredi, E. & Elsen, J. M. (2008). Performance of genomic selection in mice. *Genetics* **180**, 611–618.

Mackay, T. F. C. & Lyman, R. F. (2005). *Drosophila* bristles and the nature of quantitative genetic variation. *Philosophical Transactions of the Royal Society*, B **360**, 1513–1527.

Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

Mulder, H. A., Bijma, P. & Hill, W. G. (2007). Prediction of breeding values and selection responses with genetic heterogeneity of environmental variance. *Genetics* **175**, 1895–1910.

Mulder, H. A., Hill, W. G., Vereijken, A. & Veerkamp, R. F. (2009). Estimation of genetic variation in residual variance in female and male broilers. *Animal* **3**, 1673–1680.

Ordas, B., Malvar, R. A. & Hill, W. G. (2008). Genetic variation and quantitative trait loci associated developmental stability and the environmental correlation between traits in maize. *Genetical Research* **90**, 385–395.

Ros, M., Sorensen, D., Waagepetersen, R., Dupont-Nivet, M., SanCristobal, M., Bonnet, J.-C. & Mallard, J. (2004). Evidence for genetic control of adult weight plasticity in the snail *Helix aspersa*. *Genetics* **168**, 2089–2097.

Rowe, S., White, I. M. S., Avendano, S. & Hill, W. G. (2006). Genetic heterogeneity of residual variance in broiler chickens. *Genetics, Selection, Evolution* **38**, 617–635.

San Cristobal-Gaudy, M., Bodin, L., Elsen, J. M. & Chevalet, C. (2001). Genetic components of litter size variability in sheep. *Genetics, Selection, Evolution* **33**, 249–271.

Sorensen, D. (2009). Developments in statistical analysis in quantitative genetics. *Genetica* **136**, 319–332, 10.1007/sl0709-008-9303-5.

Sorensen, D. & Waagepetersen, R. (2003). Normal linear models with genetically structured residual variance heterogeneity: a case study. *Genetical Research* **82**, 207–222.

Sorensen, D., Andersen, S., Gianola, D. & Korsgaard, I. R. (1995). Bayesian inference in threshold models using Gibbs sampling. *Genetics, Selection, Evolution* **27**, 229–249.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegaard, T. S., Schnabel, R. R., Taylor, J. F. & Schenkel, F. S. (2009). Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92**, 16–24.

Verdinelli, I. & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association* **90**, 614–618.

Visscher, P. M. & Posthuma, D. (2010). Statistical power to detect genetic loci affecting environmental sensitivity. *Behavior Genetics* **40**, 728–733.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. L, Zhu, G., Cornes, B. K., Montgomery, G. W. & Martin, N. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLOS Genetics* **2**, 316–325.

Visscher, P. M., Macgregor, S., Benyamin, B., Zhu, G., Gordon, S., Medland, S., Hill, W. G., Hottenga, J. J., Willemsen, G., Boomsma, D. L, Liu, Y. Z., Deng, H. W., Montgomery, G. W. & Martin, N. G. (2007). Genome partitioning of genetic variation for height from 11,214 sibling pairs. *American Journal of Human Genetics* **81**, 1104–1110.

Waagepetersen, R., Ibafiez, N. & Sorensen, D. (2008). A comparison of strategies for Markov chain Monte Carlo computation in quantitative genetics. *Genetics, Selection, Evolution* **40**, 161–176.

Weller, J. I., Seller, M. & Brody, T. (1988). Linkage analysis of quantitative traits in an interspecific cross of tomato (*Lycopersicum esculentum* × *Lycopersicum pimpinellifolium*) by means of genetic markers. *Genetics* **118**, 329–339.

Whitlock, M. C. & Fowler, K. (1999). The changes of genetic and environmental variance with inbreeding in *Drosophila melanogaster*. *Genetics* **152**, 345–353.

Wolc, A., White, I. M. S., Avendano, S. & Hill, W. G. (2009). Genetic variability in residual variation of body weight and conformation scores in broiler chickens. *Poultry Science* **88**, 1156–1161.