Genet. Sel. Evol. 31 (1999) 193–211

Original article

Behaviour of the additive finite locus model

Ricardo Pong-Wong*, Chris S. Haley, John A. Woolliams

Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, Scotland, UK

(Received 7 September 1998; accepted 2 April 1999)

Abstract – A finite locus model to estimate additive variance and the breeding values was implemented using Gibbs sampling. Four different distributions for the size of the gene effects across the loci were considered: i) uniform with loci of different effects, ii) uniform with all loci having equal effects, iii) exponential, and iv) normal. Stochastic simulation was used to study the influence of the number of loci and the distribution of their effect assumed in the model analysis. The assumption of loci with different and uniformly distributed effects resulted in an increase in the estimate of the additive variance according to the number of loci assumed in the model of analysis, causing biases in the estimated breeding values. When the gene effects were assumed to be exponentially distributed, the estimate of the additive variance was still dependent on the number of loci assumed in the model of analysis, but this influence was much less. When assuming that all the loci have the same gene effects or when they were normally distributed, the additive variance estimate was the same regardless of the number of loci assumed in the model of analysis. The estimates were not significantly different from either the true simulated values or from those obtained when using the standard mixed model approach where an infinitesimal model is assumed. The results indicate that if the number of loci has to be assumed a priori, the most useful finite locus models are those assuming loci with equal effects or normally distributed effects. © Inra/Elsevier, Paris

finite locus model / gene effect distribution / Gibbs sampling / infinitesimal model

Résumé – Comportement des modèles additifs à nombre fini de loci. On a utilisé, via la méthode de l'échantillonnage de Gibbs, des modèles à nombre fini de loci pour estimer les variances génétiques additives et les valeurs génétiques. On a considéré quatre distributions différentes des effets de gènes sur l'ensemble des loci : i) distribution uniforme avec loci à effets variables, ii) distribution uniforme avec loci à effets égaux, iii) distribution exponentielle, et iv) distribution normale. La simulation stochastique a été utilisée pour étudier l'influence du nombre de loci et de

* Correspondence and reprints
E-mail: ricardo.pong-wong@bbsrc.ac.uk

la distribution supposée de leurs effets. L'hypothèse d'effets différents et uniformément distribués a entraîné le fait que la variance génétique augmentait quand le nombre supposé de loci augmentait, ce qui a causé des biais dans l'estimation des valeurs génétiques. Quand les effets de gènes ont été distribués exponentiellement, l'estimée de la variance génétique additive a été encore dépendante du nombre de loci supposé, quoiqu'à un moindre degré. Quand on a supposé que tous les loci avaient les mêmes effets de gènes ou quand ils ont été normalement distribués, l'estimée de la variance génétique additive a été la même, quel que soit le nombre de loci supposé dans l'analyse. Les résultats indiquent que si le nombre de loci est supposé d'après des considérations a priori, les modèles à nombre fini de loci les plus utiles sont ceux qui supposent des loci à effets égaux ou à distribution normale.

modèle fini / distribution d'effets / échantillonnage de Gibbs / modèle infinitésimal

1. INTRODUCTION

Genetic evaluation in livestock has traditionally been carried out using an infinitesimal genetic model, where the trait is assumed to be influenced by an infinite number of genes, each with an infinitesimally small effect. Although such a model is biologically incorrect, its use has been justified because it allows the handling of the total additive genetic effect as a normally distributed variable so that standard statistical mixed model techniques can be applied. Indeed, solutions from the normal approximation appear to be robust enough for practical selection purposes, provided the trait is not controlled by a small number of loci, few generations are considered (so that there are no substantial changes in the alleles frequencies due to selection or drift) and the additive genetic effect alone is considered [17].

The arguments justifying the use of the infinitesimal model are, however, being weakened by the increasing knowledge about the genetic architecture of quantitative traits. Single genes that have a relatively large effect on quantitative traits (e.g. Booroola gene, double muscle gene, Callipyge gene) are expected to have a rapid change in allele frequency due to selection. Under these circumstances, the infinitesimal model would wrongly predict the evolution of the genetic variance even when the selected trait is also affected by a large number of loci with small effects [8]. Moreover, the assumptions required to describe dominance with the infinitesimal model are unclear [25]. Thus, alternative approaches to incorporating the extra knowledge about the genetic make-up of quantitative traits should be considered.

In this paper, an additive finite locus model is defined and implemented using Gibbs sampling. The effects of the assumptions about the number of loci and the distribution of the size of their effects are studied, extending the results previously reported by Pong-Wong et al. [24]. The results obtained with the finite locus model are compared with those obtained using the mixed model where an infinitesimal genetic model is assumed.

2. MATERIALS AND METHODS

2.1. Finite-locus genetic model

A quantitative trait is assumed to be genetically controlled by L unlinked biallelic loci. Following the same notation as Falconer [4], each locus l, has an additive (a_l) effect with a frequency of the favourable allele in the base population of p_l. The additive variance explained by locus l is then $2p_l(1-p_l)a_l^2$. Since the loci are assumed to be unlinked and in linkage equilibrium the total additive variance (σ_a^2) is the sum over all the loci. The trait is also assumed to be affected by an environmental deviation which is normally distributed with mean zero and variance σ_e^2. Other environmental fixed and random effects may also be included in the model but, for simplicity, they are not considered here.

In matrix algebra the linear model is expressed as:

$$\mathbf{y} = 1\mu + \mathbf{W}_a\mathbf{a} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is the $(n \times 1)$ vector of phenotypic records, μ the overall mean, $\mathbf{a}$ the $(L \times 1)$ vector of additive (a) effects for each locus, $\mathbf{e}$ the $(n \times 1)$ vector of environmental deviation, and $\mathbf{W}_a$ is the $(n \times L)$ matrix of additive effects associated to the individual's genotype. Assuming that the genotypes are denoted as AA, AB and BB (BB the least favourable genotype), the value in column l of $\mathbf{W}_a$ would be 1, 0 or -1, for a phenotypic observation from an individual with genotype (at the l locus) AA, AB or BB, respectively. The vector $\mathbf{a}_{-l}$ is defined the same as $\mathbf{a}$ but excluding the effect at the locus l.

2.1.1. Distribution of the size of gene effects

Since the size of the effects across the different loci are assumed to be different, an assumption about how the gene effects are distributed is required. Here, three possible distributions to model the gene effects are examined: i) uniform, ii) exponential, and iii) (folded-over) normal.

The probability density functions for the distribution of the size of the additive effects ($\psi(a)$) when assuming the uniform, exponential and the (folded-over) normal distributions, respectively, are:

$$\psi(a) = \text{constant} \qquad \text{for} \quad a \geq 0 \tag{2}$$

$$\psi(a) = \frac{1}{\lambda_a} \exp\left\{\frac{-a}{\lambda_a}\right\} \quad \text{for} \quad a \geq 0 \tag{3}$$

and

$$\psi(a) = \frac{2}{\sqrt{2\pi\lambda_a}} \exp\left\{-0.5\frac{a^2}{\lambda_a}\right\} \quad \text{for} \quad a \geq 0 \tag{4}$$

where λ_a is the scale parameter for the exponential and the normal distribution. The density function $\psi(a)$ is defined only for the range of the positive numbers (including zero) since a is, by definition, the effect of the favourable homozygote genotype. The assumption that the gene effects are either normally

or exponentially distributed is consistent with the general belief that most of the loci affecting a given quantitative trait would have a small effect, while only a few genes have a major effect on the trait in question.

2.2. Implementation of the finite locus model using Markov chain Monte Carlo

Genetic analyses assuming the proposed finite locus model involve the estimation of the gene effect at each locus, the parameter defining the distribution of the gene effects, the genotype probability for each individual at all the loci and their allele frequencies. In the model of analysis, the number of loci affecting the trait in question as well as the distribution of their effects are assumed known. The total additive variance is estimated as a linear function of the effect and allele frequency across all the loci (i.e. $\sigma_a^2 = \sum_l 2p_l(1-p_l)a_l^2$). A graphical representation of the finite locus model is presented in *figure 1.*

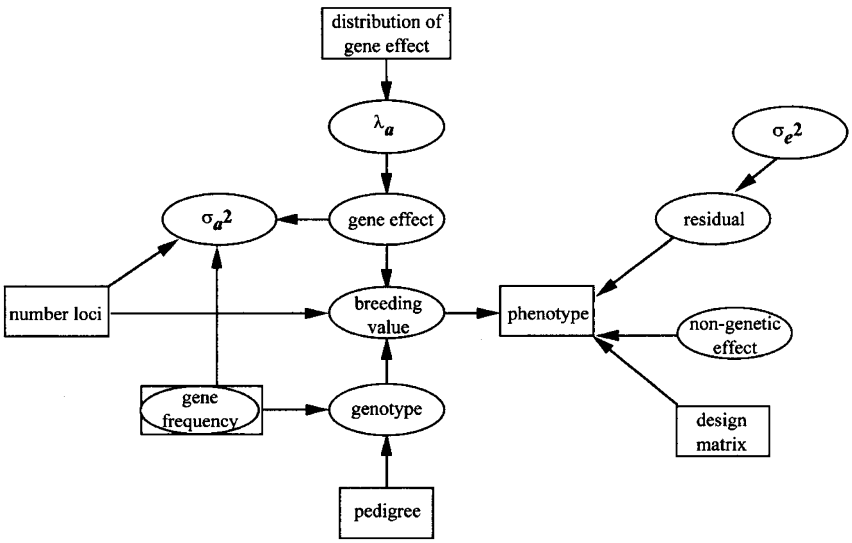

Figure 1. Graphical representation showing the relationship between the different variables in a finite locus model. Variables enclosed in squares are known prior to the analysis (data input or assumptions), variables in circles are estimated (sampled) during the analysis (output). The gene frequency at each locus may be estimable but it was assumed to be 0.5 in our analyses.

The main problem in implementing a finite locus genetic model using a standard likelihood approach is the calculation of the genotype probability for all the loci. In practice this task is computationally very difficult because of the large number of possible genotype combinations that need to be considered, a number which rapidly increases with the number of individuals. This problem becomes further exacerbated with complex pedigree structures involving loops and, especially, when assuming multiple loci are present in the model.

In order to avoid this problem, the finite locus model proposed is implemented using a Markov chain Monte Carlo (MCMC) approach based upon Gibbs sampling algorithms previously suggested for segregation studies of untyped single genes in complex pedigree structures (e.g. [16, 18]). These algorithms are simply extended to include L loci accounting for the entire genetic effects. Because all loci are assumed to be unlinked the sampling of the genotype at each locus is performed independently.

A sampling protocol for updating the relevant parameters (conditional on the others) of a finite locus model in the Markov chain would then be as follow:

1) sample overall mean;
2) sample the genotype configurations locus by locus;
3) sample the gene effects locus by locus;
4) sample the scale parameter of the assumed distribution of gene effects (not needed when assuming a uniform distribution);
5) sample all other environmental fixed and random effects (not included here);
6) sample non-permanent environmental variance and variance for other random effects.

The sampling of the allele frequencies for each locus may also be added in the sampling scheme. In this study, however, they were not estimated but they were fixed to be 0.5.

The full conditional distributions for the gene effects and the scale parameter for the distribution of gene effects, needed during the sampling process, are presented below. The conditional distributions of other parameters (e.g. genotype configuration, environmental variance, other random and fixed effects) are not shown here since they have been described in previous studies reported in the literature. For the description of the algorithms used to sample genotypes see Guo and Thompson [16] and Janss et al. [18] (the latter algorithm was used here, since it allows a better mixing in pedigrees with large family sizes). For the use of Gibbs sampling in more general genetic evaluations and the conditional distributions of other environmental effects, see Firat [7] and Wang et al. [29, 30].

2.2.1. Joint posterior density (conditional on the genotype structure)

The full conditional density for the effect at each locus as well as the scale parameter of the distribution of gene effects are obtained from their joint posterior density by extracting the terms containing the variable in question. The joint posterior density of σ_e^2, $\mathbf{a}$ and λ_a conditional on the genotype structure (considered as known to simplify the expression) is of the form:

$$P(\mathbf{a}, \lambda_a, \sigma_e^2 | \mathbf{G}, p, \mathbf{y}) \propto (\sigma_e^2)^{-n/2} \exp\left[-\frac{1}{2\sigma_e^2}(\mathbf{y} - \mathbf{W}_a\mathbf{a})'(\mathbf{y} - \mathbf{W}_a\mathbf{a})\right] \times \prod_{l=1}^{L} \psi(a_l) \times P(\lambda_a) \times P(\sigma_e^2) \quad (5)$$

where $\mathbf{W}_a$ depends on the current genotype structures, $\psi(a)$ is the probability density function of the gene effect given the assumed distribution, and $P(\lambda_a)$

and $P(\sigma_e^2)$ are the prior distributions of λ_a and σ_e^2, respectively. The respective conjugate prior distribution for λ_a when assuming the gene effects being exponentially and normally distributed is proportional to $(\lambda_a)^{-v-1}\exp(-vs/\lambda_a)$ and $(\lambda_a)^{-v/2-1}\exp(-0.5\,vs/\lambda_a)$, where v is the degree of belief and s the prior value of λ_a. Assuming that v is equal to zero (i.e. there is no belief in any particular value of s) gives the 'naive' prior, which is proportional to $1/\lambda_a$. This prior denotes a lack of prior knowledge about the parameter and it has been used as a prior for variance components including some animal breeding implementations [9, 29]. In this study 'naive' priors were used for both λ_a and σ_e^2.

2.2.2. Conditional distributions for the (size of the) gene effects

The conditional distribution of the gene effects depends on the assumption of how they are distributed.

2.2.2.1. Uniform and independent

When the additive effects are assumed to be uniformly distributed, the conditional density depends only on the first term of equation (5) (i.e. the second term is a constant). Thus, the conditional distribution for the effect of the locus l is proportional to:

$$P(a_l|\mathbf{a}_{-l}, \mathbf{G}, \lambda_a, \sigma_e^2, \mathbf{y}) \propto \exp\left\{-0.5\,\frac{(a_l - \widehat{a}_l)^2}{\sigma^2}\right\} \tag{6}$$

which is equivalent to a truncated normal distribution with mean $\widehat{a}_l$ and variance σ^2 evaluated in the range of positive values. The value for $\widehat{a}_l$ is the solution from the linear model equal to $\left(\sum y_{AA} - \sum y_{BB}\right)/(n_{AA} + n_{BB})$, and σ^2 its error variance equal to $\sigma_e^2/(n_{AA} + n_{BB})$, where y_g is the adjusted phenotype of individuals with updated genotype g, and n_g is the number of records from individuals with such a genotype. The solution of the linear model $\widehat{a}_l$, is equivalent to the coefficient from the regression (passing through the origin) of the phenotype (adjusted for the effect of other loci and any other environmental effects) on the genotype value (i.e. 1, 0 or -1 for the record from an individual sampled to have genotype AA, AB or BB, respectively). The conditional distribution resulting from assuming a uniform distribution has been generally used to sample the major gene effect in mixed inheritance models (e.g. [18]).

2.2.2.2. Uniform and constant

During the estimation of the gene effects, an extra assumption may also be taken to consider that all loci have the same effect (as assumed in a previous study by Fernando et al. [6]). For this case, the full conditional distribution is similar to equation (6), but $\widehat{a}$ and σ^2 are the regression coefficient and its error variance, estimated from the regression (passing through the origin) of the adjusted phenotype on the combined genotype value across all loci (i.e. the

regression is on the number of loci sampled as AA minus the number of loci sampled as BB for the individual contributing to the record).

2.2.2.3. Exponential

The full conditional distribution of the effect of locus l is proportional to:

$$P(a_l|\mathbf{a}_{-l}, \mathbf{G}, \lambda_a, \mathbf{y}) \propto \exp\left\{-0.5\,\frac{(a_l - \widehat{a}_l)^2}{\sigma^2}\right\} \times \exp\left\{-\frac{a_l}{\lambda_a}\right\}$$

where $\widehat{a}_l$ and σ^2 are defined as in equation (6). Rearranging the previous equation results in the following:

$$P(a_l|\mathbf{a}_{-l}, \mathbf{G}, \lambda_a, \mathbf{y}) \propto \exp\left\{-0.5\,\frac{\left[a_l - (\widehat{a}_l - \sigma^2\lambda_a^{-1})\right]^2}{\sigma^2}\right\} \times \exp\left\{\frac{1}{\lambda_a}\left[\widehat{a} - \frac{2\,\sigma^2}{\lambda_a}\right]\right\} \tag{7}$$

where the first term is proportional to a normal distribution with mean $\widehat{a}_l - \sigma^2\lambda_a^{-1}$ and variance σ^2, and the second term is a constant. Substituting the values $\widehat{a}_l$ and σ^2 as defined in equation (6), the full conditional distribution is a truncated normal defined for the positive values with mean $\left(\sum y_{AA} - \sum y_{BB} - \sigma_e^2\lambda^{-1})/(n_{AA} + n_{BB}\right)$ and variance $\sigma_e^2/(n_{AA} + n_{BB})$.

2.2.2.4. Folded-over normal

Extracting the terms containing a_l in equation (5), its conditional distribution is proportional to:

$$P(a_l|\mathbf{a}_{-l}, \mathbf{G}, \lambda_a, \mathbf{y}) \propto \exp\left\{-0.5\,\frac{(a_l - \widehat{a}_l)^2}{\sigma^2}\right\} \times \exp\left\{-0.5\,\frac{a_l^2}{\lambda_a}\right\}$$

and when substituting the values of $\widehat{a}_l$ and σ^2, the previous expression can be rearranged as

$$P(a_l|\mathbf{a}_{-l}, \mathbf{G}, \lambda_a, \mathbf{y}) \propto \exp\left\{-0.5\,\frac{\left[a_l - \left(\sum y_{AA} - \sum y_{BB}\right)(n_{AA} + n_{BB} + \sigma_e^2\lambda_a^{-1})^{-1}\right]^2}{\left[(n_{AA} + n_{BB} + \sigma_e^2\lambda_a^{-1})^{-1}\,\sigma_e^2\right]}\right\} \tag{8}$$

which is proportional to a truncated normal with mean $\left(\sum y_{AA} - \sum y_{BB}\right)$ $(n_{AA} + n_{BB} + \sigma_e^2\lambda_a^{-1})^{-1}$ and variance $(n_{AA} + n_{BB} + \sigma_e^2\lambda_a^{-1})^{-1}\,\sigma_e^2$.

2.2.3. Conditional distribution of the scale parameter of the gene effect distribution

The conditional density of the scale parameter depends only on the second term of equation (5) and varies according to which distribution of the gene

effects is being assumed. The estimation of this parameter is not required when assuming that the gene effects are uniformly distributed.

The conditional density of λ_a under the assumption that the gene effects are exponentially distributed and with 'naive' prior is:

$$P(\lambda_a|\mathbf{a}) \propto (\lambda_a)^{-(L-2)} \times \exp\left(-\frac{\sum_{l=1}^{L}|a_l|}{\lambda_a}\right) \times (\lambda_a)^{-1}$$

which is equivalent to:

$$P(\lambda_a|\mathbf{a}) \propto \sum_{l=1}^{L}|a_l|/\gamma_{(1,L)} \tag{9}$$

where $\gamma_{(1,L)}$ is a gamma distribution with scale and shape parameters equal to 1 and L, respectively.

Similarly, when the gene effects are normally distributed, the conditional distribution of λ_a assuming a 'naive' prior is:

$$P(\lambda_a|\mathbf{a}) \propto (\lambda_a)^{-(L/2-2)} \times \exp\left(-\frac{\sum_{l=1}^{L}|a_l^2|}{2\,\lambda_a}\right) \times (\lambda_a)^{-1}$$

which is a scaled inverted chi-squared of the form:

$$P(\lambda_a|\mathbf{a}) \propto \sum_{l=1}^{L}|a_l^2|/\chi_L^2 \tag{10}$$

2.3. Simulated population

2.3.1. Population structure

The structure of the simulated population consisted of a base population of 80 unrelated individuals (40 males and 40 females) plus five other discrete generations. At each generation five males and 20 females were chosen and randomly mated to produce four offspring (two males and two females) per female. Selection of parents was at random unless otherwise noted in the results. All individuals had one phenotypic record.

2.3.2. Genetic model

The total genetic effects were accounted for by 20 independent and diallelic loci. All loci were assumed to be completely additive and their initial allele

frequency was 0.5. The genotype at each locus of the base individuals was sampled from the expected genotype frequency of a locus in Hardy-Weinberg equilibrium. The genotype of individuals from further generations were sampled assuming Mendelian inheritance. The total genetic effects of an individual are the sum of all the genotype effects over all loci.

2.3.3. Parameters used

For all the cases the environmental variance was assumed to be 80, the additive genetic variance 20. In order to account for the total genetic variance, the effect of each locus was simulated in two ways: i) assuming that all the 20 loci have the same effect (i.e. $a = \sqrt{2}$); or ii) that each effect was sampled from an exponential distribution with scale parameter equal to 1 (which is expected to yield the correct total genetic variance).

2.4. Situations compared

Data sets simulated using the population structure explained above were used to study the behaviour of the finite locus model (FIN) in genetic evaluations. Each data set (replicate) was analysed with several FIN approaches varying in the assumptions about the distribution of gene effects and the number of loci taken in the model of analysis.

These variations in assumptions were the following.

i) The distribution of the gene effects: effects of loci uniformly and independently (FIN-UNI), uniformly but constant (i.e. equal effects; FIN-CON), exponentially (FIN-EXP) or normally (FIN-NOR) distributed.

ii) The number of loci: 5, 10, 20 or 30.

As previously stated, the allele frequencies in the base population for each locus were not estimated in the analysis. Instead they were fixed at 0.5.

The case when all loci have the same effects (FIN-CON) is similar to the finite locus model proposed by Fernando et al. [6].

The same data sets were also analysed using the standard mixed model approach (MM) where an infinitesimal genetic model is assumed. In order to make the results comparable with those obtained with the FIN analyses, the MM was also performed using a Gibbs sampling approach to obtain the marginal posterior density of each variance component. From a Bayesian perspective, the variance estimates from MM using a restricted maximum likelihood (REML) approach are the mode of their joint posterior distribution, which are not expected to coincide with the mode of their marginal distributions [11]. The implementation of the mixed model using Gibbs sampling and its differences from REML approaches have been much studied (e.g. Wang et al. [30]).

2.4.1. Criteria of comparison

The criteria of comparison were the estimates of the variance components (σ_a^2, σ_e^2) and the correlation between the estimated breeding values (EBV).

3. RESULTS

3.1. Gibbs sampling implementation

The results presented below are the summaries of 50 replicates. The variance estimates of each evaluation within a replicate is the mean of a Markov chain of 1 000 realisations sampled every 50 cycles after a burning period of 5 000 cycles (i.e. total length of the chain = 55 000 cycles). This sampling protocol ensured that the autocorrelation between consecutive realisations was less than 0.1 for all the parameters studied here.

3.2. True model: the same gene effects across all loci (random selection)

3.2.1. FIN-UNI

The estimates of the variance components assuming that all loci have different effects and are uniformly distributed are shown in *table I.* These results were highly dependent on the number of loci assumed in the model of analysis. The estimate of the additive variance increased when more loci were assumed in the model of analysis. This trend was consistently observed across all the replicates. The additive variance estimate closest to the true simulated value was produced when only five loci were assumed in the model of analysis, which is substantially less than the true number used to simulate the data.

Table I. Variance component estimates (SE in brackets) obtained from the mixed model (MM) and the finite locus model approaches assuming different numbers of loci of different effects and uniform distribution (FIN-UNI).

	Model of analysis*				
	MM	FIN05	FIN10	FIN20	FIN30
σ_a^2	20.58	29.64	38.73	55.95	72.47
	(1.47)	(1.23)	(1.21)	(1.25)	(1.34)
σ_e^2	81.07	75.61	71.05	63.66	57.56
	(1.49)	(1.23)	(1.15)	(1.06)	(1.01)

* MM, mixed model; FINnn, finite locus model assuming nn loci.

The increase in the estimated additive variance when assuming more loci in the model of analysis was also accompanied by a decrease in the estimated environmental variance. However, this reduction did not completely compensate for the extra estimated additive variance, thus resulting in an overestimate in the total phenotypic variance. The estimated total variance increased from 105 when assuming five loci to 129 when the analysis was carried out assuming 30 loci (the simulated value was 100).

The excess of additive variance which appeared when increasing the number of loci had repercussions on the estimated breeding values. As expected, the increased additive variance resulted in a higher dispersion of the EBV, so

individuals with extreme EBV became even more extreme when more loci were assumed in the model of analysis. Additionally, the prediction error variance associated with the EBV also tended to increase with the number of loci. The mean prediction error variance of the EBV when using five loci was 16 compared with 25 when the EBV were obtained assuming 30 loci (for the MM the mean prediction error variance was 12.5). Nevertheless, it is important to note that although the EBV were very sensitive to the number of loci, the correlation between the different estimates was always greater that 0.9 (*table II*). Thus, the ranking of individuals was little affected.

Table II. Correlation between the estimated breeding values obtained with the mixed model (MM) and the finite locus model (FIN) approach assuming different numbers of loci with different uniformly distributed effects (FIN-UNI).

	Model of analysis*			
	MM	FIN05	FIN10	FIN20
FIN05	0.973			
	(0.132)			
FIN10	0.960	0.994		
	(0.131)	(0.135)		
FIN20	0.932	0.976	0.991	
	(0.127)	(0.133)	(0.134)	
FIN30	0.905	0.955	0.978	0.995
	(0.123)	(0.130)	(0.133)	(0.135)

* MM, mixed model; FINnn, finite locus model assuming nn loci.

3.2.2. FIN-CON

The variance estimates when assuming all loci had the same effects is summarised in *table III*. Under this assumption the estimates of the additive variance were the same regardless of the number of loci assumed in the model of analysis. The results from FIN-CON were not significantly different from the simulated values or from those obtained with the MM. The EBV and their prediction error variance were also insensitive to the number of loci assumed in the model of analysis (results not shown).

Table III. Variance component estimates (SE in brackets) obtained from the mixed model (MM) and the finite locus model approach assuming different numbers of loci of equal effects (FIN-CON).

	Model of analysis*				
	MM	FIN05	FIN10	FIN20	FIN30
σ_a^2	20.58	22.94	22.96	22.96	23.06
	(1.47)	(1.37)	(1.36)	(1.36)	(1.36)
σ_e^2	81.07	79.48	79.44	79.50	79.39
	(1.49)	(1.37)	(1.37)	(1.36)	(1.37)

* MM, mixed model; FINnn, finite locus model assuming nn loci.

3.2.3. *FIN-EXP*

Table IV shows the summary of the variance components estimated assuming the gene effects being exponentially distributed. Increasing the number of loci used in the model of analysis yielded a slight increase in the estimated additive variance. However, this trend was very small compared with the results from FIN-UNI. In contrast to the case of FIN-UNI, the estimate of the total phenotypic variance remained constant. The correlation among the EBV obtained with the FIN-EXP analyses with different numbers of loci was always higher that 0.95 (results not shown).

Table IV. Variance component estimates (SE in brackets) obtained from the mixed model (MM) and the finite locus model (FIN) approach assuming different numbers of loci with exponentially distributed effects (FIN-EXP).

	Model of analysis*				
	MM	FIN05	FIN10	FIN20	FIN30
σ_a^2	20.58	20.59	21.57	23.39	25.17
	(1.47)	(1.42)	(1.32)	(1.23)	(1.19)
σ_e^2	81.07	81.00	80.10	78.66	77.31
	(1.49)	(1.45)	(1.35)	(1.26)	(1.21)

* MM, mixed model; FINnn, finite locus model assuming nn loci.

3.2.4. *FIN-NOR*

The results when the gene effects were assumed to be normally distributed appeared not to be affected by the number of loci used in the model of analysis (*table V*). The EBV were also the same regardless of the number of loci used in the model of analysis. The results obtained with FIN-NOR were similar to those observed with standard mixed model.

Table V. Variance component estimates (SE in brackets) obtained from the mixed model (MM) and the finite locus model (FIN) approach assuming different numbers of loci with normally distributed effects (FIN-NOR).

	Model of analysis*				
	MM	FIN05	FIN10	FIN20	FIN30
σ_a^2	20.58	20.03	20.21	20.28	20.25
	(1.47)	(1.51)	(1.52)	(1.50)	(1.48)
σ_e^2	81.07	81.42	81.23	81.15	81.17
	(1.49)	(1.53)	(1.53)	(1.50)	(1.50)

* MM, mixed model; FINnn, finite locus model assuming nn loci.

3.3. True model: gene effects simulated as exponentially distributed

The main purpose of using simulated data assuming the gene effects to be exponentially distributed was to test whether the observed behaviour of FIN-EXP is the same even when it corresponds to the true model.

3.3.1. Population under random selection

Table VI summarises the results when the population was under random selection. The estimated additive variance showed the same trend to increase when more loci were assumed in the model of analysis. The best estimates were obtained when using 20 loci, which corresponds to the true genetic model used to simulate the data. For this case, the variance component estimates were the same as the values used to simulate the data. The correlations between EBV obtained when using different numbers of loci have a correlation greater than 0.99 (results not shown).

Table VI. Variance component estimates (SE in brackets) obtained from the mixed model (MM) and the finite locus model (FIN) approach assuming different numbers of loci with exponentially distributed effects (FIN-EXP) from data simulated assuming 20 loci with exponentially distributed effects and random selection.

	Model of analysis*				
	MM	FIN05	FIN10	FIN20	FIN30
σ_a^2	17.99	17.82	19.22	21.40	23.40
	(1.70)	(1.25)	(1.46)	(1.41)	(1.37)
σ_e^2	81.48	81.43	80.23	78.52	76.95
	(1.35)	(1.54)	(1.19)	(1.13)	(1.10)

* MM, mixed model; FINnn, finite locus model assuming nn loci.

3.3.2. Population under truncation selection

Table VII shows the results of FIN-EXP when the population was undergoing selection. The results showed the same trend for the additive variance, but

Table VII. Variance component estimates (SE in brackets) obtained from the mixed model (MM) and the finite locus model (FIN) approach assuming different numbers of loci with exponentially distributed effects (FIN-EXP) from data simulated assuming 20 loci with exponentially distributed effects and truncation selection.

	Model of analysis*				
	MM	FIN05	FIN10	FIN20	FIN30
σ_a^2	18.22	16.37	16.78	17.63	18.36
	(1.52)	(1.29)	(1.26)	(1.27)	(1.28)
σ_e^2	79.29	79.99	79.46	78.75	78.09
	(0.88)	(0.88)	(0.86)	(0.86)	(0.85)

* MM, mixed model; FINnn, finite locus model (FIN-EXP) assuming nn loci.

surprisingly, the magnitude was smaller than that observed with random selection. The correlation between EBV calculated assuming different numbers of loci was always greater than 0.97 (results not shown).

4. DISCUSSION

In this paper a genetic model assuming a finite number of loci affecting a quantitative trait was implemented using Gibbs sampling. The behaviour of the results when changing the number of loci and the distribution of the gene effects assumed on the model of analysis were studied using stochastic simulation.

The use of genetic models assuming a finite number of loci has so far been hardly studied. Chevalet [3] proposed a genetic model which allows the estimation of the effective number of loci affecting a quantitative trait, but this model is still based upon the same Gaussian assumptions made with the infinitesimal model. A model which does not depend on normal theory was proposed by Fernando et al. [6] and is known as the hypergeometric model [20]. In this model the calculation of the multilocus genotype probability is simplified by not treating the genotype at each locus independently, but by identifying their combined genotypes as the total number of favourable alleles present across all loci. This simplification, however, forces the assumption that all loci must have the same effect, and the model is not strictly consistent with Mendelian transmission [6, 20]. However, the main purpose of the hypergeometric model has been to mimic the results of the infinitesimal model but with a lower complexity when calculating the likelihood, thereby greatly reducing the computational difficulties of segregation and linkage analyses of single major genes [28]. More recently, Goddard [14] and Pong-Wong et al. [24] studied the feasibility of estimating dominance using a finite locus model assuming that the gene effects were uniformly distributed.

The results from this study show a remarkable interaction between the distribution of gene effects and the number of loci assumed in the model of analysis. When the gene effects were assumed to follow a uniform distribution (FIN-UNI), the estimate of the additive variance sharply increased when adding more loci to the model of analysis. A less marked trend was also observed when assuming that the gene effects were exponentially distributed (FIN-EXP). When the model of analysis assumed the allelic effects to be normally distributed (FIN-NOR) or constant over all loci (FIN-CON), the results were the same regardless of the number of loci assumed in the model. However, despite the similarity in the trend of the additive variance, the results from FIN-UNI and FIN-EXP are qualitatively different. The slight increase in the additive variance observed with FIN-EXP was only due to differences in the partition of the total variance, whereas with FIN-UNI there was also an increase in the total phenotypic variance observed in the system. From this point of view, the behaviour of FIN-EXP is more similar to FIN-NOR than to FIN-UNI.

This difference in the behaviour of FIN-UNI compared with FIN-EXP or FIN-NOR is, perhaps, not surprising when examining the statistical meaning of these models. From a strictly statistical point of view, it can be seen that the gene effects in FIN-UNI are treated as fixed effects while with FIN-EXP and FIN-NOR they are considered to be random variables (drawn from an exponential and a normal distribution, respectively). Thus, the estimation of

the gene effects using FIN-UNI is expected to yield different answers to those obtained with FIN-NOR and FIN-EXP, and thereby, the total additive variance which is calculated as a linear combination of the gene effect estimates.

The consequences of adding more loci to the model of analysis when treating their effects as fixed appears to create an overparameterised model. The extra gene effects (fixed effects) that need to be estimated in the model result in some of them being confounded and explaining spurious effects. Thus, the more loci fitted in the model, the more spurious effects are estimated, increasing both the genetic and the total variance. Hence, the reduction in the number of parameters to be estimated by assuming that all the loci have the same effects (only one effect is estimated compared with L effects estimated when assuming all loci having different effects) may avoid this overparameterisation, which would explain why the results of FIN-CON are insensitive to the number of loci. On the other hand, the possibility of spurious effects arising when increasing the number of loci in FIN-EXP or FIN-NOR is better controlled since the estimates of the gene effects (random effects) are regressed towards zero, restricting their dispersion accordingly to their scale parameter (λ_a). The difference between treating a variable as fixed or as random is well known in animal breeding. For example, the variance of the estimated sire effects obtained after treating them as fixed would be greater than the estimated inter-sire variance when assuming them to be random normal variables.

The trend of σ_a^2 when using FIN-EXP was also observed when the true model also assumed the gene effects to be exponentially distributed. Surprisingly, this trend was smaller when the population was undergoing selection. This consistency of results across simulated data sets assuming different genetic models suggests that the overall trend observed with FIN-EXP is more likely to be a true characteristic of this model of analysis rather than being a Monte Carlo error due to a small number of replicates. Another interesting result is the fact that the mean estimate of the genetic variance when assuming ten loci was marginally closer to the true simulated value than when 20 loci were assumed (*table VI*). Intuitively, the latter would be expected to yield better answers as it corresponds exactly to the model used to simulate the data. However, the difference in results is too small to firmly conclude which is the better model of analysis, so their rating should not be based only on the average estimate (across the replicates) relative to the true simulated value. The estimation of the Bayes factor to assess the goodness of fit of these models should also be considered before concluding which one better describes the data.

The difference in the results between FIN-EXP and FIN-NOR prompts the need for further studies to evaluate the behaviour of finite locus models assuming other distributions of gene effects. The assumption of a normal distribution appears to yield robust/consistent results, but ideally the distribution to be assumed should be one closely reflecting the reality of the trait in question. Although the characterisation of the distribution of gene effects for economically important traits in farm animals is still incomplete, some knowledge in this area may be obtained from studies of mutation effects in Drosophila. For instance, Keightley [19] suggested that the gamma distribution may be suitable for modelling the gene effects since it depends on few parameters (note that the exponential distribution is a special case of gamma distribution) and can

be parameterised to display leptokurtosis. Other alternatives have also been proposed by Caballero and Keightley [2].

An alternative way to avoid the problem of uncertainty on the true distribution of gene effects may be to select the distribution during the analysis. Using the Markov chain framework, Green [15] proposed a technique, called the reversible jump, which allows for model choice during the analysis. For this particular situation, a set of distributions may be predefined and a Markov process built allowing the chain to move among these distributions according to their probabilities. Using the same principle, one may also be able to sample the number of loci [27]. Obviously, the complexity of a Markov chain implementation allowing for model choice in several of the parameters would be higher, and greater care should be taken when assessing the convergence of the chain as well as when interpreting the results. Another alternative means to conclude which set of parameters (e.g. distribution of gene effects, number of loci) fit best the data would be the estimation of Bayes factors [9].

One of the consequences of assuming other distributions of gene effects, such as gamma, is that the resulting full conditional distribution may be of unknown form with no standard sampling routine available. The full conditional density of the gene effects resulting from the three distributions examined in this study are proportional to a truncated normal, for which standard sampling routines are available. The use of techniques such as adaptive rejection sampling [12, 13] and 'slicing-the-density' [23] allow sampling from non-standard distributions, but the computational cost is also expected to increase.

Because of the computational demand of Gibbs sampling implementations, the study of the properties of finite locus models should also be complemented with the proposal of efficient algorithms to improve the mixing and convergence of the Markov chain. Several approaches to improving the efficiency of sampling the genotype structure in complex pedigree are now available (e.g. [10, 21, 22]), and their use may prove beneficial in reducing the computational demand of a finite locus model.

In this study, allele frequencies were not estimated but were assumed to be 0.5. However, the frequencies are only fixed in the base population, so the model is able to account for changes in the genetic level due to drift or directional selection. Although the estimation of the allele frequencies does not add much extra complexity to the model, practical problems in the Gibbs implementation were encountered (unpublished). Allowing variable allele frequencies can result in slow mixing with problems arising from loci becoming temporally fixed. Since inferences from MCMC are valid only when convergence has been obtained, poor mixing requires the length of the chain to be considerably increased, with consequences in computing time. A preliminary analysis of a non-selected population where the allele frequencies were estimated but restricted to between 0.2 and 0.8 (to avoid the Gibbs sampling problem due to fixation) yielded similar results as when analysis was performed assuming the frequencies to be 0.5. Obviously, this restriction on the gene frequency estimation may need to be relaxed when considering populations with deeper pedigree structure and undergoing selection.

A positive characteristic of the finite locus model proposed here is its ability to account for the linkage disequilibrium between loci built up during selection [1]. This disequilibrium creates a correlation between loci in the offspring

generation which results in a reduction of the total observed genetic variance (i.e. the genetic variance in the offspring generation estimated using their total genetic effects across all loci is smaller than the sum of the variances explained by each individual locus). For the case of selection presented in this study, the loss in variance due to disequilibrium for the first three generations from selected parents was 6.3, 7.2 and 9.5 %, respectively.

Conversely, it is also desirable that the genotypes of individuals from the base population being sampled are in linkage equilibrium to avoid potential bias in the estimation of σ_a^2 (as the formula used to estimate it assumes no correlation between loci). Considering the impact on the results if linkage disequilibrium in the base population being built up due to sampling, σ_a^2 was also estimated using the total breeding value of each individual reconstructed given their genotypes and the gene effects across all loci (thus accounting for any correlation appearing due to sampling). With the exception of the cases of FIN-UNI assuming 20 and 30 loci, the estimate of σ_a^2 was the same as when not accounting for any potential correlation. For instance, σ_a^2 estimated from the total breeding values using FIN-EXP assuming 10, 20 and 30 loci were 21.58, 23.36 and 25.12, respectively, which are very similar to the values reported in *table IV*. When the analysis was performed using FIN-UNI with 20 or 30 loci, the estimation of σ_a^2 from the total breeding values yielded smaller results than when assuming the loci being independent (i.e. 53.34 and 67.4, respectively, compared with results from *table I*), but these differences explain less than 10 % of the total overestimation of σ_a^2. Thus, the conclusion that the estimates of both the genetic and the total variance when using FIN-UNI are largely dependent on the assumed number of loci still remains valid (i.e. FIN-UNI is still a questionable model to be used).

Here we considered only the case of a complete-additive genetic model where the results from the mixed model are expected to be robust, and a finite locus model would add little improvement to practical genetic evaluations. However, there are other situations where departing from the infinitesimal model may prove to be beneficial. For instance, a finite locus model may provide a more natural approach to extending marker-assisted selection (MAS) accounting for multiple quantitative trait loci (QTL). Genetic maps in most farm animals are becoming very dense so that the use of MAS to exploit information of only one QTL at a time seems to be a waste of resources and time. Approaches to studying linkage between a QTL and a genetic marker using Gibbs sampling have been suggested in the literature (e.g. [16, 18]), and their implementation in a finite locus model seems to be straight forward. From the mixed model approach, multiple QTL may also be accounted for by extending the MAS method using a BLUP framework [5]. This method, however, presents the problem that it does not account for changes in gene frequency due to selection. Additionally, since each QTL is modelled with two normal variables, the method becomes computationally complex as the rank of the resulting linear model increases by twice the number of individuals per each QTL included in the model.

Another potential use of a finite locus model is the estimation of dominance. Although the mixed model has been used to estimate dominance deviance, the assumptions justifying this approach are not well understood [25]. Despite the substantial increase in complexity to estimate dominance, in some situations

the results of a mixed model analysis may be difficult to interpret (for an example, see [26]). Preliminary results have shown that inclusion of dominance in a finite locus model adds very little extra complexity to the model whilst maintaining a relationship between both the dominance variance and the inbreeding depression [24].

Finally, another benefit of finite locus models is that they offer a more 'biologically appropriate' genetic model which would provide a greater understanding of quantitative traits. For example, it would be possible to examine characteristics of these traits (e.g. the distribution of the gene effects, effective number of loci). Moreover, unlike the infinitesimal model, some of the newly gained knowledge about the architecture of these traits could easily be included in a finite locus model to improve the prediction in genetic evaluations.

ACKNOWLEDGEMENTS

The authors acknowledge financial support from the Ministry of Agriculture, Fisheries and Food, the Biotechnology and Biological Research Council and the Pig Improvement Company. We thank S.C. Bishop and two anonymous readers for useful comments on the manuscript.

REFERENCES

[1] Bulmer M.G., The effect of selection on genetic variability, Am. Nat. 105 (1971) 201–211.

[2] Caballero A., Keightley P.D., A pleiotropic nonadditive model of variation in quantitative traits, Genetics 138 (1994) 883–900.

[3] Chevalet C., An approximate theory of selection assuming a finite number of quantitative trait loci, Genet. Sel. Evol. 26 (1994) 379–400.

[4] Falconer D.S., Introduction to Quantitative Genetics, 3th ed., Longman Scientific and Technical, Essex, 1989.

[5] Fernando R.L., Grossman M., Marker-assisted selection using best linear unbiased prediction, Genet. Sel. Evol. 21 (1989) 467–477.

[6] Fernando R.L. Stricker C., Elston R.C., The finite polygenic mixed model: an alternative formulation for the mixed model of inheritance, Theor. Appl. Genet. 88 (1994) 573–580.

[7] Firat M.Z., Bayesian methods in selection of farm animals for breeding, PhD thesis, University of Edinburgh, 1995.

[8] Fournet-Hanocq F., Elsen M., On the relevance of three genetic models for the description of genetic variance in small population undergoing selection, Genet. Sel. Evol. 30 (1998) 59–70.

[9] Gelman A., Carlin J.B., Stern H.S., Rubin D.B., Bayesian Data Analysis, Chapman and Hall. London, UK, 1995.

[10] Geyer C.J., Thomson E.A., Annealing Markov chain Monte Carlo with application to ancestral inference, J. Am. Stat. Assoc. 90 (1995) 909–920.

[11] Gianola D., Foulley J.L., Variance estimation from integrated likelihood (VEIL), Genet. Sel. Evol. 22 (1990) 403–417.

[12] Gilk W.R., Wild P., Adaptive rejection sampling in Gibbs sampling, Appl. Stat. 41 (1992) 337– 348.

[13] Gilk W.R., Best N.G., Tan K.K.C., Adaptive rejection Metropolis sampling, Appl. Stat. 44 (1995) 455–472.

[14] Goddard M.E., Gene based model for genetic evaluation – An alternative to BLUP ?, Proc. 6th World Cong. Genet. Appl. Livest. Prod. 26 (1998) 33–36.

[15] Green P.J., Reversible jumping Markov chain Monte Carlo computation and Bayesian model determination, Biometrika 82 (1995) 711–732.

[16] Guo S.W., Thompson E.A., A Monte Carlo method for combined segregation and linkage analysis, Am. J. Hum. Genet. 51 (1992) 111–1126.

[17] Hill W.G., Quantitative genetic theory: Chairman's introduction, Proc. 5th World Cong. Genet. Appl. Livest. Prod. 19 (1994) 125–126.

[18] Janss L.L.G., Thompson R., Van Arendonk J.A.M., Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations, Theor. Appl. Genet. 91 (1995) 1137–1147.

[19] Keightley P.D., The distribution of mutation effects on viability in *Drosophila melanogaster*, Genetics 138 (1994) 1315–1322.

[20] Lange K., An approximate model of polygenic inheritance, Genetics 147 (1997) 1423–1430.

[21] Lin S., Thompson E.A., Wijsman E., An algorithm for Monte Carlo estimation of genotype probabilities on complex pedigrees, Ann. Hum. Genet. 58 (1994) 343–357.

[22] Lund M., Jensen C.S., Multivariate updating of genotypes in a Gibbs sampling algorithm in the mixed inheritance model, Proc. 6th World Cong. Genet. Appl. Livest. Prod. 25 (1998) 521– 524.

[23] Neal R.M., Markov chain Monte Carlo methods based on 'slicing' the density function, Technical report 9722, Dept. Stat, Univ., Toronto, 1997.

[24] Pong-Wong R., Shaw F., Woolliams J.A., Estimation of dominance variation using a finite-locus model, Proc. 6th World Cong. Genet. Appl. Livest. Prod. 26 (1998) 41–44.

[25] Robertson A., Hill W.G., Population and quantitative genetics of many linked loci in finite populations, Proc. R. Soc. Lond. B 219 (1983) 253–264.

[26] Shaw F., Woolliams J.A., Variance component analysis of skin and weight data for sheep subjected to rapid inbreeding, Genet. Sel. Evol. 31 (1999) 43–59.

[27] Stephens D.A., Fish R.D., Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo, Biometrics 54 (1998) 1334–1347.

[28] Stricker C., Fernando R.L., Elston R.C., Linkage analysis with an alternative formulation of the mixed model of inheritance: the finite polygenic mixed model, Genetics 141 (1995) 1651– 1656.

[29] Wang C.S., Rutledge J.J., Gianola D., Marginal inferences about variance components in mixed linear model using Gibbs Sampling, Genet. Sel. Evol. 25 (1993) 41–62.

[30] Wang C.S., Rutledge J.J., Gianola D., Bayesian analysis of mixed linear model via Gibbs sampling with application to litter size in Iberian pigs, Genet. Sel. Evol. 26 (1994) 91–115.