Genetics Selection Evolution

RESEARCH Open Access

Allele coding in genomic evaluation

Ismo Strandén[1*] and Ole F Christensen[2]

Abstract

Background: Genomic data are used in animal breeding to assist genetic evaluation. Several models to estimate genomic breeding values have been studied. In general, two approaches have been used. One approach estimates the marker effects first and then, genomic breeding values are obtained by summing marker effects. In the second approach, genomic breeding values are estimated directly using an equivalent model with a genomic relationship matrix. Allele coding is the method chosen to assign values to the regression coefficients in the statistical model. A common allele coding is zero for the homozygous genotype of the first allele, one for the heterozygote, and two for the homozygous genotype for the other allele. Another common allele coding changes these regression coefficients by subtracting a value from each marker such that the mean of regression coefficients is zero within each marker. We call this centered allele coding. This study considered effects of different allele coding methods on inference. Both marker-based and equivalent models were considered, and restricted maximum likelihood and Bayesian methods were used in inference.

Results: Theoretical derivations showed that parameter estimates and estimated marker effects in marker-based models are the same irrespective of the allele coding, provided that the model has a fixed general mean. For the equivalent models, the same results hold, even though different allele coding methods lead to different genomic relationship matrices. Calculated genomic breeding values are independent of allele coding when the estimate of the general mean is included into the values. Reliabilities of estimated genomic breeding values calculated using elements of the inverse of the coefficient matrix depend on the allele coding because different allele coding methods imply different models. Finally, allele coding affects the mixing of Markov chain Monte Carlo algorithms, with the centered coding being the best.

Conclusions: Different allele coding methods lead to the same inference in the marker-based and equivalent models when a fixed general mean is included in the model. However, reliabilities of genomic breeding values are affected by the allele coding method used. The centered coding has some numerical advantages when Markov chain Monte Carlo methods are used.

Background

There has been growing interest in the use of marker-based models [1] in recent years. In studies using these models, descriptions of the effect of allele coding system on inference and computations are often vague or missing. By allele coding, we refer to the coefficients in the marker matrix of marker-based models. Coefficients, commonly used for allele coding of a marker is 0 when the individual is homozygous for the first allele, 1 when the individual is heterozygous, and 2 when the individual is homozygous for the second allele. Depending on which of the alleles has been chosen as the first allele, the coefficients are different. Thus, this allele coding method does not give unique regression coefficients.

There are other allele coding methods such as the one that use coefficients -1, 0, and 1 instead of 0, 1, and 2, respectively. Different allele coding methods affect coefficients in the statistical models but they do not seem to change the amount of information for statistical inference. Hence, one would expect that the use of different allele coding methods would lead to the same inference. However, allele coding can be of vital importance in computations. First, convergence of iterative methods such as Markov chain Monte Carlo (McMC) methods often used in Bayesian inference can be assumed to be affected by the allele coding method used because different allele coding methods change the correlation structure between marker

* Correspondence: ismo.stranden@mtt.fi
Full list of author information is available at the end of the article

effects. Second, equivalent models have become popular in animal breeding [2,3]. An important concept in these methods is the genomic relationship matrix. Differences in allele coding will yield different genomic relationship matrices [4]. Thus, some elements of the inverse of the coefficient matrix can be different and, consequently, reliabilities may be different.

We investigated effects of different allele coding methods using theoretical derivations and a practical example when restricted maximum likelihood (REML) or Bayesian inference is used. Effects on parameter estimates, reliabilities, and McMC computations were studied. We considered marker models and their equivalent breeding value models.

Methods

Genomic marker model

Let us consider a univariate linear mixed effects model for genomic marker effect estimation, e.g. [1],

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{g} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is $n \times 1$ vector of observations, μ is the general mean, $\mathbf{Z}$ is a $n \times m$ matrix containing a column for each marker locus, $\mathbf{g}$ is a $m \times 1$ vector of random SNP marker effects, and $\mathbf{e}$ is a random residual vector. There can be other fixed or random effects in the model but their inclusion does not change the following derivations.

Allele coding methods

There are several alternatives for coding the coefficients in the $\mathbf{Z}$ matrix. Four allele coding systems are considered. A simple transformation can be made from one allele coding system to another. Our basic allele coding system counts the number of copies of one of the alleles. Depending on which of the alleles is counted, the matrix can be different. In the allele coding system 012, the number of copies of the less frequent allele is counted. Thus, the coefficient is 0 if the individual is homozygous for the more frequent allele, 1 if it is heterozygous, or 2 if it is homozygous for the less frequent allele. In this case, the $\mathbf{Z}$ matrix for the basic allele coding system 012 is denoted by $\mathbf{Z}_0$.

A general form for the allele coding transformation from the basic allele coding system is $\mathbf{Z}_0 - \mathbf{1}_n\mathbf{v}'_m$ where $\mathbf{v}_m$ is a $m \times 1$ vector. This allows many types of allele coding methods. Note that the transformation keeps distances between allele codes within a marker the same. So, the coding 0,1,2 can be changed to - 1,0,1 or 0.5, 1.5, 2.5 by this transformation, but not to -10,0,10.

We define the centered allele coding system as $\mathbf{Z}_c$ = $\mathbf{Z}_0$ - $\mathbf{P}_c$, where each column of the matrix $\mathbf{P}_c$ contains the average allele count for the corresponding marker column. Thus, summing values in each column will give a vector of zeros, i.e., $\mathbf{1}'_n(\mathbf{Z}_0 - \mathbf{P}_c)$ is a vector of zeros. For the centered allele coding system, we have $\mathbf{v}_m = \frac{1}{n}\mathbf{Z}'_0\mathbf{1}_n$, i.e., $\mathbf{Z}_c = \mathbf{Z}_0 - \frac{1}{n}\mathbf{1}_n\mathbf{1}'_n\mathbf{Z}_0$. Note that $\mathbf{v}_m/2$ gives the allele frequencies of the markers in the data.

The allele coding transformation allows shifts in the allele codes. The 101 allele coding system is such that -1 is assigned to the genotype homozygous for the more frequent allele, 0 to the heterozygous individual, and 1 to the individual homozygous for the less frequent allele. For the 101 allele coding system, we have $\mathbf{v}_m = \mathbf{1}_m$. The 101 allele coding system is equal to the centered allele coding system when all allele frequencies are equal to 0.5.

In the following, the derivations will use the general allele coding transformation. The matrix $\mathbf{Z}_0$ is unique. However, in general the decision on which of the alleles to count is arbitrary. Let the 210 allele coding system be such that the more frequent allele is counted. Then the $\mathbf{Z}$ matrix for the 210 allele coding system can be calculated from the 012 coding matrix by $\mathbf{2}\mathbf{1}_n\mathbf{1}'_m - \mathbf{Z}_0$ where $\mathbf{1}_n$ is $n \times 1$ vector of ones. The 210 allele coding system is the opposite to the 012 allele coding system but results in this paper apply to the 210 allele coding system as well, with some modifications mentioned separately.

Different allele coding methods imply different models (1) and different models may lead to different parameter estimates. However, the inference considered in this paper is not affected by allele coding, as we demonstrate below.

Inference in marker-based model

Variance component estimation by restricted maximum likelihood (REML), prediction of random effects, and Bayesian inference are all based on the likelihood after marginalization of the fixed effects, i.e., the fixed effects have been integrated out. Bayesian inference requires even more marginalization. In order to show that inference is not affected by allele coding, it is sufficient to show that the likelihood after integrating out the general mean is the same irrespective of allele coding. The following derivation makes no assumptions on the prior densities of marker effects. Thus, the results apply to many models, including BLUP, BayesA and BayesB in [1].

The marginal likelihood for the mixed effects model is

$$p(\mathbf{y}|\mathbf{g},\boldsymbol{\theta}) = \int_{-\infty}^{\infty} p(\mathbf{y}|\mu,\mathbf{g},\boldsymbol{\theta})\,d\mu,$$

where $p(\mathbf{y} \mid \mu, \mathbf{g}, \boldsymbol{\theta})$ is the conditional density of $\mathbf{y}$, often a Gaussian density, and $\boldsymbol{\theta}$ contains all parameters in the distribution of $\mathbf{e}$, often only the residual variance parameter σ_e^2. Using the transformation result (7) in Appendix A and a change of integration variable $\mu_0 = \mu - \mathbf{v}'_m\mathbf{g}$, we can write

$$\begin{aligned} p(\mathbf{y}|\mathbf{g},\boldsymbol{\theta}) &= \int_{-\infty}^{\infty} p_0(\mathbf{y}|\mu - \mathbf{v}'_m\mathbf{g},\mathbf{g},\boldsymbol{\theta})\,d\mu \\ &= \int_{-\infty}^{\infty} p_0(\mathbf{y}|\mu_0,\mathbf{g},\boldsymbol{\theta})\,d\mu_0 \\ &= p_0(\mathbf{y}|\mathbf{g},\boldsymbol{\theta}), \end{aligned}$$

where p_0 denotes the 012 allele coding system. Hence, the marginal likelihood does not depend on allele coding, a property used in the following derivations.

The REML-likelihood is defined when **g** and **e** are multivariate Gaussian distributed, and equals

$$L(\boldsymbol{\theta},\boldsymbol{\eta}) = \int p(\mathbf{y}|\mathbf{g},\boldsymbol{\theta})p(\mathbf{g}|\boldsymbol{\eta})d\mathbf{g},$$

where $\boldsymbol{\eta}$ contains all parameters in the distribution of **g**, commonly only the genetic marker variance parameter σ_g^2. This likelihood is independent of allele coding, and, hence, REML parameter estimation is independent of allele coding. Note that maximum likelihood estimation is based on $L(\mu, \theta, \boldsymbol{\eta}) = \int p(\mathbf{y} \mid \mu, \mathbf{g}, \theta)p(\mathbf{g} \mid \boldsymbol{\eta})d\mathbf{g}$ and is affected by allele coding because, in this case, the general mean is not integrated out.

BLUP estimation of marker effects **g** assumes that the variance parameters $(\boldsymbol{\theta}, \boldsymbol{\eta})$ are known. The conditional distribution $p(\mathbf{g} \mid \mathbf{y}, \theta, \boldsymbol{\eta}) = p(\mathbf{y} \mid \mathbf{g}, \theta)p(\mathbf{g} \mid \boldsymbol{\eta})/L(\theta, \boldsymbol{\eta})$ is independent of allele coding. Hence, BLUP $\hat{\mathbf{g}}$ and associated uncertainties do not depend on the allele coding.

In Bayesian inference, the joint posterior after integrating out μ is

$$\begin{aligned} p(\mathbf{g},\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y}) &= \int p(\mu,\mathbf{g},\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})\,d\mu \\ &= p(\mathbf{y}|\mathbf{g},\boldsymbol{\theta})p(\mathbf{g}|\boldsymbol{\eta})p(\boldsymbol{\theta},\boldsymbol{\eta})/p(\mathbf{y}), \end{aligned}$$

where $p(\boldsymbol{\theta}, \boldsymbol{\eta}\ )$ is the joint prior for the parameters, and the denominator $p(\mathbf{y})$ is the integral of the numerator. All terms in the numerator are independent of allele coding, and by marginalization $p(\mathbf{y})$ satisfies the same. Hence, $p(\mathbf{g}, \boldsymbol{\theta}, \boldsymbol{\eta} \mid \mathbf{y})$ does not depend on allele coding.

The general intercept μ is, however, not independent of allele coding. For simplicity of the argument, we assume that parameters $(\boldsymbol{\theta}, \boldsymbol{\eta})$ are known, and omit showing these values. According to the transformation result (8) in Appendix A and a change of integration variable $\mu_0 = \mu - \mathbf{v}'_m\mathbf{g}$, the conditional expectation of the general mean is

$$\begin{aligned} \hat{\mu} &= \int \mu \int p(\mu,\mathbf{g}|\mathbf{y})d\mathbf{g}d\mu \\ &= \int\int \mu p_0(\mu - \mathbf{v}'_m\mathbf{g},\mathbf{g}|\mathbf{y})d\mu d\mathbf{g} \\ &= \int\int (\mu_0 + \mathbf{v}'_m\mathbf{g})p_0(\mu_0,\mathbf{g}|\mathbf{y})d\mu_0 d\mathbf{g} \qquad (2) \\ &= \int \mu_0 p_0(\mu_0|\mathbf{y})d\mu_0 + \int \mathbf{v}'_m\mathbf{g}p_0(\mathbf{g}|\mathbf{y})d\mathbf{g} \\ &= \hat{\mu}_0 + \mathbf{v}'_m\hat{\mathbf{g}}, \end{aligned}$$

where p_0 denotes density for the 012 allele coding, and $\hat{\mu}_0$ is the conditional expectation of the general mean when using the 012 allele coding. Thus, the general mean estimate is different by allele coding when $\mathbf{v}'_m\hat{\mathbf{g}}$ is not zero. When **g** and **e** are multivariate Gaussian distributed, the conditional expectations $\hat{\mathbf{g}}$ and $\hat{\mu}$ equal the BLUP and BLUE estimates, respectively.

Finally, the inference is indifferent to the allele being counted. This is demonstrated by studying the centered coding system and assuming that allele in the first marker is counted in the opposite way, i.e., the first column in **Z** is minus the first column in $\mathbf{Z}_c$, or $\mathbf{z}_1 = -\mathbf{z}_{c1}$. We see that $\mathbf{Zg} = \mathbf{Z}_c\tilde{\mathbf{g}}$ where the entries in $\tilde{\mathbf{g}}$ are equal to the entries in **g**, except for the first entry which equals minus the first entry in **g**. Since **g** and $\tilde{\mathbf{g}}$ have the same distribution, these two models are equivalent.

Genomic breeding values

Estimating breeding values

In breeding value evaluation, the main interest is in estimation of genomic breeding values for the genotyped animals. In other words, estimation of $\hat{\mathbf{a}} = \mathbf{Z}\hat{\mathbf{g}}$ where $\hat{\mathbf{g}}$ are solutions to the marker effects by a marker-based model like model (1). Because the marker effect solutions are the same for different allele coding systems, the estimated genomic breeding values are different due to differences in the coefficient matrix **Z**.

Allele coding does not, however, change relative differences between the estimated genomic breeding values, because $\mathbf{Z}\hat{\mathbf{g}} - \mathbf{Z}_0\hat{\mathbf{g}} = -\mathbf{1}_n(\mathbf{v}'_m\hat{\mathbf{g}})$ shows that they are just shifted by a constant. Let us define complete genomic breeding values as $\hat{\mathbf{a}}_d = \mathbf{1}_n\hat{\mu} + \mathbf{Z}\hat{\mathbf{g}}$. Substituting $\mathbf{Z} = \mathbf{Z}_0 - \mathbf{1}_n\mathbf{v}'_m$ and using equation (2) we obtain

$$\begin{aligned} \hat{\mathbf{a}}_d &= \mathbf{1}_n\hat{\mu} + (\mathbf{Z}_0 - \mathbf{1}_n\mathbf{v}'_m)\hat{\mathbf{g}} \\ &= \mathbf{1}_n(\hat{\mu} - \mathbf{v}'_m\hat{\mathbf{g}}) + \mathbf{Z}_0\hat{\mathbf{g}} \\ &= \mathbf{1}_n\hat{\mu}_0 + \mathbf{Z}_0\hat{\mathbf{g}}. \end{aligned}$$

Consequently, the estimated complete breeding values $\hat{\mathbf{a}}_d$ are the same irrespective of allele coding.

Equivalent model and allele coding

Assume that the marker effects have a Gaussian distribution $\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}_m\sigma_g^2)$ where $\mathbf{I}_m$ is an $m \times m$ identity matrix. The breeding values $\mathbf{a} = \mathbf{Zg}$ can be calculated directly without estimating $\hat{\mathbf{g}}$ by the model [2,3]

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{a} + \mathbf{e},$$

where the breeding values have prior density of $\mathbf{a} \sim N(\mathbf{0}, \mathbf{ZZ}'\sigma_g^2)$. Often the covariance matrix of the breeding values is scaled by a value such as $\nu_p = 2\sum_{i=1}^{m} p_i(1 - p_i)$ where p_i is the allele frequency of marker i. Then, the breeding values have a prior density of $\mathbf{a} \sim N(\mathbf{0}, \mathbf{G}\sigma_a^2)$ where the genomic relationship matrix is $\mathbf{G} = \mathbf{ZZ}'\frac{1}{v_p}$ and genetic variance is $\sigma_a^2 = \sigma_g^2 v_p$. Assuming that the residual distribution is $\mathbf{e} \sim N\,(\mathbf{0}, \mathbf{R})$, then the mixed model equations for the equivalent model are

$$\begin{bmatrix} \mathbf{1}_n'\mathbf{R}^{-1}\mathbf{1}_n & \mathbf{1}'_n\mathbf{R}^{-1} \\ \mathbf{R}^{-1}\mathbf{1}_n & \mathbf{R}^{-1} + \mathbf{G}^{-1}/\sigma_a^2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \tag{3}$$

The breeding value solutions $\hat{\mathbf{a}}$ from these mixed model equations (3) are the same as the genomic breeding values calculated by $\hat{\mathbf{a}} = \mathbf{Z}\hat{\mathbf{g}}$ where $\hat{\mathbf{g}}$ are marker effects estimated by the marker-based model (1).

Therefore, the conclusion for the marker-based models about relative differences between genomic breeding values being unaffected by allele coding is also true for the equivalent model, although different allele coding methods lead to different genomic relationship matrices. Similarly, variance component estimation by REML and Bayesian methods are unaffected by the allele coding due to equivalence of models with the marker-based models.

The mixed model equations in (3) are not well-defined when the genomic relationship matrix $\mathbf{G}$ is singular. However, mixed model equations not requiring an invertible $\mathbf{G}$ matrix do exist; see page 48 in [5]. The genomic relationship matrix $\mathbf{G}$ can be singular for several reasons. For example, there can be identical twins or clones that have the same genotypes. In addition, for the centered allele coding system the genomic relationship matrix is $\mathbf{G}_c = \mathbf{Z}_c\mathbf{Z}_c'$. The last row of $\mathbf{Z}_c = (\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\mathbf{Z}_0$ is equal to the sum of all the other rows. Hence, $\mathbf{Z}_c$ is not of full rank, and $\mathbf{G}_c$ is singular.

Prediction error variances and reliabilities

Gaussian models are often used in practical genomic evaluation of animals. In these models, reliabilities of estimated breeding values $\hat{\mathbf{a}}$ are calculated using elements of the inverse of the mixed model equations such as (3). Reliability of $\hat{\mathbf{a}}_i$ is

$$r_i^2 = 1 - \frac{\mathrm{PEV}_i}{\sigma_a^2 G_{ii}}, \tag{4}$$

where PEV_i is the prediction error variance, i.e., $\mathrm{Var}(a_i \mid \mathbf{y})$, of animal i, and G_{ii} is the diagonal element of animal i in the genomic relationship matrix $\mathbf{G}$; e.g. [6], p. 51 in [7]. The prediction error variance for animal i is the diagonal element of the inverse of the coefficient matrix of mixed model equations (3) for animal i. Alternatively,

$$\begin{aligned} \mathrm{PEV} &= \mathrm{Var}(\mathbf{a}|\mathbf{y}) \\ &= \mathrm{Var}(\mathbf{Zg}|\mathbf{y}) \\ &= \mathbf{Z}\mathrm{Var}(\mathbf{g}|\mathbf{y})\mathbf{Z}' \\ &= \mathbf{ZC}^g\mathbf{Z}', \end{aligned} \tag{5}$$

where $\mathbf{C}^g$ is the genomic marker effect submatrix in the inverse of the coefficient matrix of the mixed model equation for marker-based model (1) (see Appendix B). The submatrix $\mathbf{C}^g = \mathrm{Var}(\mathbf{g} \mid \mathbf{y})$ is the same irrespective of the allele coding method used as shown in the chapter on inference on marker-based models. Because the coefficient matrix $\mathbf{Z}$ is different depending on allele coding, PEV is also different depending on allele coding. Consequently, the reliability of $\hat{\mathbf{a}}$ depends on allele coding.

More generally, for any of the models considered in this paper, $\mathbf{a} = \mathbf{Zg}$ where $p(\mathbf{g} \mid \mathbf{y})$ is independent of allele coding and $\mathbf{Z}$ depends on allele coding. Therefore, the distribution $p(\mathbf{a} \mid \mathbf{y})$ and, in particular, the variance-covariance matrix $\mathrm{Var}(\mathbf{a} \mid \mathbf{y})$ and reliabilities of $\hat{\mathbf{a}}$ depend on allele coding.

The complete breeding value distribution $p(\mathbf{a}_d \mid \mathbf{y})$ does not depend on allele coding, unlike PEV associated with $\hat{\mathbf{a}}$. The proof is based on the demonstration that when applying any function f, the expectations are independent of the allele coding system,

$$\begin{aligned} &\mathrm{E}[f(\mathbf{a}_d)|\mathbf{y}] \\ &= \int f(\mathbf{1}_n\mu + \mathbf{Zg})p(\mu, \mathbf{g}|\mathbf{y})\, d\mu d\mathbf{g} \\ &= \int f(\mathbf{1}_n(\mu - \mathbf{v}_m'\mathbf{g}) + \mathbf{Z}_0\mathbf{g}) \\ &\qquad \times p_0(\mu - \mathbf{v}_m'\mathbf{g}, \mathbf{g}|\mathbf{y})\, d\mu d\mathbf{g} \\ &= \int f(\mathbf{1}_n\mu_0 + \mathbf{Z}_0\mathbf{g})p_0(\mu_0, \mathbf{g}|\mathbf{y})\, d\mu_0 d\mathbf{g} \\ &= \mathrm{E}_0[f(\mathbf{a}_d)|\mathbf{y}], \end{aligned}$$

where E_0 is the expectation when using the basic allele coding method. Therefore, the variance-covariance matrix $\mathrm{Var}(\mathbf{a}_d \mid \mathbf{y})$ and all higher order moments of the distribution are independent of allele coding. However, the result does not provide actual formulas for the moments.

A closed form formula of the variance-covariance matrix is derived for a Gaussian model. Assume $\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}_m\sigma_g^2)$ and $\mathbf{e} \sim N\,(\mathbf{0}, \mathbf{R})$. For this model, in Appendix B we obtain that

$$\begin{aligned} &\mathrm{Var}(\mathbf{a}_d|\mathbf{y}) = r\mathbf{1}_n\mathbf{1}_n' \\ &\quad + (\mathbf{I}_n - r\mathbf{1}_n\mathbf{1}_n'\mathbf{R}^{-1})\,\mathbf{ZC}^g\mathbf{Z}'(\mathbf{I}_n - r\mathbf{R}^{-1}\mathbf{1}_n\mathbf{1}_n'), \end{aligned}$$

where $r = 1/(\mathbf{1}_n'\mathbf{R}^{-1}\mathbf{1}_n)$ and $\mathbf{C}^g = [\mathbf{Z}'(\mathbf{R}^{-1} - r\mathbf{R}^{-1}\mathbf{1}_n\mathbf{1}_n'\mathbf{R}^{-1})\mathbf{Z} + \mathbf{I}_m/\sigma_g^2]^{-1}$. As demonstrated earlier in this section, this variance-covariance matrix is independent of allele coding. When $\mathbf{R} = \mathbf{I}_n\sigma_e^2$, the variance-covariance matrix simplifies to

$$\begin{aligned} &\mathrm{Var}(\mathbf{a}_d|\mathbf{y}) \\ &= \mathbf{1}_n\mathbf{1}_n'\sigma_e^2/n + \mathbf{Z}_c(\mathbf{Z}_c'\mathbf{Z}_c/\sigma_e^2 + \mathbf{I}_m/\sigma_g^2)^{-1}\mathbf{Z}_c', \end{aligned} \tag{6}$$

where $\mathbf{Z}_c$ is based on the centered allele coding method. The diagonal elements in (6) are different from the PEV$_i$s in (4) because they contain uncertainty about the unknown mean μ as well.

For the complete breeding values $\mathbf{a}_d$, we have shown above that prediction error variances are independent of allele coding and we have provided a formula for the Gaussian model. Reliabilities of $\hat{\mathbf{a}}_d$, however, can not be defined in a meaningful way. Substituting the diagonal elements from (6) for the PEV$_i$s in (4) is not appropriate since the denominator in (4) is Var($\mathbf{a}_i$) not Var($\mathbf{a}_d$)$_{ii}$. The denominator in the reliability formula should contain the marginal (unconditional) variance Var($\mathbf{a}_d$)$_{ii}$ = $\int$ Var(μ + $\mathbf{a}_i$ | μ)$d\mu$, but this variance is infinite.

McMC computations

Theoretical convergence rate

The convergence and mixing of an McMC algorithm depend on the parametrization of the model and on the algorithm used. Theoretical results about the geometric rate of convergence to the stationary distribution for Gibbs sampling algorithms are shown in [8], and as mentioned in [9] this rate also describes the mixing of the algorithm. Below we show specific results about the convergence rate ρ for Gibbs sampling algorithms for simulating from [μ, $\mathbf{g}$ | $\mathbf{y}$] in the marker-based model (1), where $\mathbf{g}$ is Gaussian $\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}_m\sigma_g^2)$ and $\mathbf{e}$ is Gaussian $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_n\sigma_e^2)$. These results provide some ideas about more general models and algorithms where theoretical results cannot be obtained.

Section 2.2 in [8] contains results about various Gibbs-sampling schemes for simulating from a multivariate distribution. Here we apply these results (see Appendix C for details) to two types of Gibbs updating schemes. The first scheme iterates between updating μ and a block of all components in $\mathbf{g}$, and will be called the block updating scheme hereinafter. The second scheme updates μ, g_1, ..., g_m sequentially one at a time, and will be called the single site updating scheme.

For the block scheme, the convergence rate is

$$\rho_1 = \frac{n}{\sigma_e^2}\bar{\mathbf{z}}' C_g^{-1}\bar{\mathbf{z}},$$

where $\bar{\mathbf{z}} = \mathbf{Z}'\mathbf{1}_n/n$ is a $m \times 1$ vector, and $\mathbf{C}_g = \mathbf{Z}'\mathbf{Z}/\sigma_e^2 + \mathbf{I}_m/\sigma_g^2$.

For the single site scheme, the convergence rate is

$$\rho_2 = \rho_{lv}(\mathbf{L}_g^{-1}(\mathbf{D}^{-1}\bar{\mathbf{z}}\bar{\mathbf{z}}'n/\sigma_e^2 - \mathbf{U}_g)),$$

where $\mathbf{L}_g$ is the matrix containing the lower triangle and the diagonal of $\mathbf{D}^{-1}$ $\mathbf{C}_g$, $\mathbf{D}$ is the diagonal of $\mathbf{C}_g$, and $\mathbf{U}_g$ is the matrix containing the upper triangle of $\mathbf{D}^{-1}$ $\mathbf{C}_g$. This single site Gibbs sampling algorithm is the stochastic counterpart of the Gauss-Seidel algorithm for solving the mixed model equations, and the convergence results are similar, see [10].

When the centered allele coding method $\mathbf{Z}_c = (\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'/n)\,\mathbf{Z}_0$ is used, $\bar{\mathbf{z}}$ is a vector of zeros and, hence, ρ_1 = 0. The centered allele coding method breaks dependency between the general mean and genetic marker effects, as seen from the variance-covariance matrix (derived in Appendix B for a more general situation)

$$\mathrm{Var}(\mu, \mathbf{g}|\mathbf{y}) = \begin{bmatrix} \sigma_e^2/n & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{Z}_c'\mathbf{Z}_c/\sigma_e^2 + \mathbf{I}_m\big/\sigma_g^2\right)^{-1} \end{bmatrix}.$$

Consequently, absorption of the general mean is done without needing to compute absorption explicitly. Note that, in general, this holds only when the residual variance-covariance matrix is $\mathbf{I}\sigma_e^2$. For the block McMC scheme, the convergence and mixing of the algorithms are of the same order as for non-McMC algorithms that simulate directly from the distribution of interest. For the single site McMC scheme, the centered allele coding method still breaks the dependence between the general mean and marker effects, but as ρ_2 > 0 illustrates, the individual marker effects g_1 , ..., g_m are not independent and the McMC samples are autocorrelated.

Data and methods

Data

Data for the XIIth QTLMAS workshop [11] were used to illustrate the theory. The simulated data had four generations. In each generation, 15 sires and 150 dams were selected randomly to produce the next generation. Each sire was mated to 10 dams and each mating produced 10 progeny. Thus, the base generation had 165 individuals, and the subsequent three generations, 1500 individuals each. In total, the analyzed data had 4665 animals with phenotypes. The simulated trait had a heritability of 0.30. The data had 6000 equally spaced SNP markers on six chromosomes. We deleted markers that had a minor allele frequency less than 1% among the phenotyped individuals and this reduced the number of markers to 5896.

The 012 allele coding method was used to make the base data set. So, the least frequent allele was counted. In addition, 210, 101, and centered allele coding data sets were analyzed.

Variance component analysis

The marker-based model (1) with common genetic variance was used to analyze the data: $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ and $\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}\sigma_g^2)$. McMC computations by a single site updating Gibbs sampler were used to calculate posterior mean estimates of the location (μ, $\mathbf{g}$) and dispersion (σ_g^2, σ_e^2) parameters. The length of the McMC chain was 100000 iterations, of which the burn-in period of 10000 was omitted. Every tenth sample was saved, giving 9000 saved samples. Effective sample sizes were calculated for all parameters using the initial monotone sequence approach [12]. The approach estimates the number of independent samples from the post burn-in samples.

Theoretical convergence rates were calculated for all the allele coding methods in the single site and block updating schemes. Here, the convergence rate also describes the mixing of the McMC chain and was measured by the correlation between successive McMC samples. Because of the Markov property, the k-lag correlation is ρ^k, where ρ is the convergence rate and k is the lag or distance between samples. In order to compare the theoretical convergence rate to the observed effective sample size, theoretical mixing was calculated relative to the 012 allele coding system as follows. Let ρ_{012} be the convergence rate from the 012 allele coding system, and ρ the convergence rates from another allele coding system. Mixing of the other allele coding system is equal to mixing of the 012 allele coding system when every k^{th} sample is taken from the McMC samples and $\rho_{012} = \rho^k$. Thus, the relative mixing is $k = \log(\rho_{012})/\log(\rho)$.

Parameter estimation by REML for both the marker-based model and the equivalent model and for all four allele coding methods was done using software DMU [13]. Both AI-REML and EM-REML were investigated. For the equivalent model and the centered allele coding method, the singular genomic relationship matrix was modified by multiplying the diagonals by 1.001 in order to be able invert the matrix. The effect of allele coding on convergence was investigated, and it was checked that the parameter estimates were the same.

Reliabilities of genomic breeding values

Reliabilities were calculated by different allele coding methods using (4) and elements of the inverse of the mixed model equations. The variance components were those estimated by the centered allele coding method using the single site McMC approach. Prediction error variances were calculated by both the marker-based model equations (5) and equivalent model equations (3). Mean, minimum, maximum and standard deviations of reliabilities were calculated for all allele coding methods. Also, correlations between reliabilities from all allele coding methods were calculated.

Results and Discussion

Posterior mean estimates of marker effects and variance components were almost equal between different allele coding methods (Table 1). Correlations of estimates of the marker effects between allele coding methods were higher than 99.98%. Only the general mean (μ) had a different estimate, as expected. The estimated variance components agreed well with those used to simulate the data. Additive genetic variance was $\hat{\sigma}_a^2 = \hat{\sigma}_g^2 v_p = 1.556$, where $v_p = 2\sum_{i=1}^{m} p_i(1-p_i) = 2323.00$, and p_i are the observed allele frequencies in the reference data. Thus, the heritability estimate was $\hat{h}^2 = \hat{\sigma}_a^2/(\hat{\sigma}_a^2 + \hat{\sigma}_e^2) = 0.34$, compared to the simulated value of 0.30.

Table 1 Posterior means of selected parameters by allele coding

	Allele coding			
Parameter	012	210	101	centered
μ	1.698	0.801	1.083	1.359
σ_g^2	6.700×10^{-4}	6.703×10^{-4}	6.691×10^{-4}	6.698×10^{-4}
σ_e^2	2.996	2.996	2.996	2.996

Effective sample sizes differed depending on allele coding method. The centered allele coding method had the best mixing, and the 210 allele coding method had the worst (Table 2). In particular, the increase in effective sample size was largest for the general mean. For the centered allele coding method, the general mean was independent from the marker effect **g**, which led to excellent mixing of this parameter. In general, the marker effects showed excellent mixing. With all allele coding methods, effective sample sizes were at least 5500 for all marker effects, and on average were equal to about 8800.

Theoretical convergence rates (Table 3) displayed the same results as the effective sample sizes discussed above. Note that our Gibbs sampler used single site updates for all parameters. For the single site updating algorithm, the 210 allele coding system was predicted to need 5.64 times more iterations than the 012 allele coding system. The number of effective samples for the general mean parameter (μ) was 5.11 times bigger for the 012 than for the 210 allele coding system. These figures were 0.48 and 0.48 for the 101 allele coding system, and 0.070 and 0.0051 for the centered allele coding system. Theoretical convergence rates for the block Gibbs sampler showed the same pattern as for the single site update (Table 3).

Surprisingly, the block Gibbs sampler was predicted to be worse than the single site Gibbs sampler for all allele coding systems except for the centered allele coding system. However, it is well known in the literature that block-updating schemes may sometimes be worse than

Table 2 Effective sample sizes in McMC computations by allele coding

	Allele coding			
Parameter	012	210	101	centered
μ	46	9	96	8961
σ_g^2	723	330	1001	1701
σ_e^2	7814	6861	7661	7720

Table 3 Predicted absolute and relative convergence rates in McMC computations by allele coding

		Allele coding			
		012	210	101	centered
single site	ρ_2	0.9974795	0.9995523	0.9947515	0.9647670
	relative	1.00	5.64	0.48	0.070
block	ρ_1	0.9995429	0.9998860	0.9989434	0.00
	relative	1.00	4.01	0.43	-

single site updating schemes, for examples see [8]. The excellent convergence rate of the block Gibbs sampler with centered allele coding was expected because, in this case, the Gibbs sampler is equal to Monte Carlo sampling.

Table 4 shows convergence of REML in parameter estimation. For the marker-based model, the convergence was independent of the allele coding system, whereas for the equivalent model, the convergence was fastest for the centered coding system and slowest for the 210 coding system, although the differences were small. The parameter estimates obtained (Table 5) were the same, with the exception of σ_e^2 for the centered coding system. The difference is due to the need to make the genomic relationship matrix $\mathbf{G}_c$ to be full rank by multiplication of the diagonals by 1.001. In summary, REML parameter estimation is only slightly affected by allele coding.

Reliabilities were affected depending on the allele coding method used. Differences were large (Table 6). Average reliabilities ranged from 0.37 with the 210 allele coding method to 0.80 with the centered allele coding method. The centered allele coding method gave higher reliabilities than achieved by any of the other allele coding methods. Reliabilities calculated by different allele coding methods were also different as judged by the correlation to each other (Table 7). Reliabilities calculated by the marker-based model and the equivalent model approaches were equal within the numerical rounding error.

The observed large differences in reliabilities using different allele coding methods can be explained by differences in estimation uncertainty. Different allele coding systems have different **Z** matrices. Consider first the 012 and 210 allele coding methods. The 012 allele coding system has a 0 coefficient when the individual is homozygous for the more frequent allele while the 210 allele coding system has a coefficient of 2 instead. In the marker-based model, uncertainty or the inverse of the coefficient matrix is the same irrespective of allele coding method. Reliability is calculated by multiplying the marker uncertainty by the **Z** matrix (5). Consequently, uncertainty is less in the 012 allele coding system than in the 210 allele coding system because the more frequent homozygous allele multiplies the marker solution and uncertainty by zero. Thus, homozygous genotypes for the more frequent allele do not increase uncertainty when estimating genomic breeding values. Thus, the 012 allele coding system will yield higher reliabilities than the 210 allele coding system, as was observed. This argument can be generalized as follows. In the genomic model considered, uncertainty of a genotype in estimating genomic breeding value is valued relative to a chosen base genotype. The further away an observed genotype is from the base genotype, the larger the coefficient in absolute value in the **Z** matrix and the higher the uncertainty in genomic breeding value. In the 012 allele coding system, the base genotype is homozygous for the more frequent allele, while in the 210 allele coding, it is homozygous for the less frequent allele. In the 101 allele coding system the base genotype is the heterozygote. The higher the number of heterozygous individuals is in the data the smaller will the uncertainty be, i.e., the higher the reliability will be for the 101 allele coding system. For the centered allele coding system the base genotype is the average genotype in the data. Thus, for this allele coding system the base population is roughly the population we work with [14], and it has the smallest average distance of observed genotypes from the base genotype. In practice, this can be expected to lead to the highest reliabilities.

Table 4 Number of iterations in REML by allele coding and model

Model	Allele coding	AI-REML	EM-REML
Marker	012	9	45
Marker	210	9	45
Marker	101	9	45
Marker	centered	9	45
G	012	7	34
G	210	9	44
G	101	7	31
G	centered	7	28

Different allele coding systems have different model design matrices **Z**, and, hence, imply different models. Thus, reliabilities from different allele coding systems are in fact from different statistical models. Comparison of reliabilities from different models is meaningless. However, the different allele coding systems lead to the same parameter estimates. If the correct allele coding method, i.e., statistical model, is known, it should be used. Because the true model is unknown and comparison of reliabilities by allele coding method is meaningless, some principles must be used to decide on which allele coding method should be used. These principles will not guarantee the use of a correct model or correct reliabilities. One such principle should be consistency of reliabilities between evaluations. The centered allele coding method changes model from one evaluation to the next because more marker data accumulate. Hence, according to the consistency principle, it cannot be recommended to computate

Table 5 REML estimates by allele coding

		Allele coding			
Model	**Parameter**	**012**	**210**	**101**	**centered**
Marker	σ_g^2	6.623 × 10^{-4}	6.623 × 10^{-4}	6.623 × 10^{-4}	6.623 × 10^{-4}
Marker	σ_e^2	2.993	2.993	2.993	2.993
G	σ_a^2	1.540	1.540	1.540	1.540
G	σ_e^2	2.993	2.993	2.993	2.992

reliabilities. Likewise, the base genotype in the 012 and 210 allele coding methods depend also on the observed allele frequencies, i.e., marker data.

The centered allele coding method is similar to that introduced in [4] where the allele frequencies were from an unselected base population. It was used in order to “give more credit to rare alleles than to common alleles when calculating genomic relationships”. As shown, inference is the same irrespective of the allele coding method when a fixed general mean is in the model. However, reliabilities are affected as shown. The use of base population allele frequencies in the centered allele coding method would remove the above mentioned problem of inconsistency between evaluations, but estimating these allele frequencies is elusive. Recently, [15] presented a method for adjusting the $\mathbf{G}_c$ relationship matrix to become a relationship matrix relative to the base population, thereby avoiding the estimation of base population allele frequencies.

The results in this paper are based on the assumption that phenotypes and genotypes are available for all animals in the analysis. This assumption may often not be satisfied. Models based on an extension of the genomic relationship matrix to include also non-genotyped animals have been presented by [16-18]. The results in the present paper about parameter estimates and estimated breeding values not depending on allele coding do not carry over to the models with an extended genomic relationship matrix.

Conclusions

We showed that, in theory, different allele coding methods led to the same inference in marker-based models when the model has a fixed general mean effect. Practical analyses led to the same conclusions. Also in theory, the centered allele coding method was expected to give better mixing properties when Markov chain Monte Carlo methods were used. This was also observed in practice. When an equivalent breeding value model was used, different allele coding methods proved to lead to the same inference as in the marker-based model. However, reliabilities of breeding values depend on the chosen allele coding system because different allele coding methods change the amount of uncertainty in the estimated breeding values.

Appendix A

In the following, we consider the effect of allele coding method on the densities $p(\mathbf{y} \mid \mu, \mathbf{g})$ and $p(\mu, \mathbf{g}, \mid \mathbf{y})$. For simplicity of presentation, the parameters in the distribution of **g** and **e** are omitted. Let p_0 denote density for 012 allele coding. Because the location parameters μ and marker effects **g** relate to the observations **y** only through $\mathbf{1}_n\mu + \mathbf{Zg}$, we first study this term. By substituting $\mathbf{Z} = \mathbf{Z}_0 - \mathbf{1}_n\mathbf{v}'_m$ into the term, we have

$$\begin{aligned}\mathbf{1}_n\mu + \mathbf{Zg} &= \mathbf{1}_n\mu + \mathbf{Z}_0\mathbf{g} - \mathbf{1}_n\mathbf{v}'_m\mathbf{g}\\ &= \mathbf{1}_n(\mu - \mathbf{v}'_m\mathbf{g}) + \mathbf{Z}_0\mathbf{g}.\end{aligned}$$

So, when different allele coding systems are used, the densities have equality by

$$p(\mathbf{y}|\mu, \mathbf{g}) = p_0(\mathbf{y}|\mu - \mathbf{v}'_m\mathbf{g}, \mathbf{g}). \tag{7}$$

By changing the integration variable $\mu_0 = \mu - \mathbf{v}'_m\mathbf{g}$, we obtain $\int p\ (\mathbf{y} \mid \mu, \mathbf{g})\ d\mu = \int p_0\ (\mathbf{y} \mid \mu_0, \mathbf{g})\ d\mu_0$ and, hence, $p(\mathbf{y}) = \int\int p(\mathbf{y} \mid \mu, \mathbf{g})p(\mathbf{g})d\mu d\mathbf{g} = p_0(\mathbf{y})$. From these results, we see that

$$\begin{aligned}&p(\mu, \mathbf{g}|\mathbf{y}) = p(\mathbf{y}|\mu, \mathbf{g})p(\mathbf{g})/p(\mathbf{y})\\ &= p_0(\mathbf{y}|\mu - \mathbf{v}'_m\mathbf{g}, \mathbf{g})p_0(\mathbf{g})/p_0(\mathbf{y})\\ &= p_0(\mu - \mathbf{v}'_m\mathbf{g}, \mathbf{g}|\mathbf{y}).\end{aligned} \tag{8}$$

Table 6 Summary statistics of genomic breeding value reliabilities by allele coding

Allele coding	min	mean	max	std
012	0.41	0.49	0.59	0.022
210	0.30	0.37	0.42	0.017
101	0.55	0.62	0.73	0.024
centered	0.72	0.80	0.95	0.026

Table 7 Correlations between genomic breeding value reliabilities by allele coding

Allele coding	210	101	centered
012	-0.22	0.32	0.43
210		0.82	0.59
101			0.91

The results (7) and (8) are fairly general in terms of distributional assumptions. The only requirements are that $p(\mathbf{y} \mid \mu, \mathbf{g})$ depends on μ and $\mathbf{g}$ only through $\mathbf{1}_n\mu + \mathbf{Zg}$, that an improper uniform prior is used for μ, and that $p(\mathbf{y})$ is finite. The later requirement is to assure the posterior distribution becomes a proper distribution, and this has to be proven for a model to be valid when an improper prior is used.

When $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, it is not difficult to show that

$$\int p(\mathbf{y}|\mu, \mathbf{g})\, d\mu \propto \exp(-\tfrac{1}{2}(\mathbf{y} - \mathbf{Zg})'\mathbf{M}(\mathbf{y} - \mathbf{Zg}))$$

with $\mathbf{M} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{1}_n\mathbf{1}_n'\mathbf{R}^{-1}/(\mathbf{1}_n'\mathbf{R}^{-1}\mathbf{1}_n)$. Therefore, $\int p\,(\mathbf{y} \mid \mu, \mathbf{g})\, d\mu < c_0$ where the constant c_0 is independent of $\mathbf{g}$. Thus, $p(\mathbf{y}) = \int\int p(\mathbf{y} \mid \mu, \mathbf{g})d\mu p(\mathbf{g})d\mathbf{g} < c_0 \int p(\mathbf{g})d\mathbf{g} = c_0$ is finite, irrespective of distribution of the marker effects $\mathbf{g}$.

Appendix B

We consider a Gaussian distribution model where $\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}_m\sigma_g^2)$ and $\mathbf{e} \sim N\,(\mathbf{0}, \mathbf{R})$. Consequently, the distribution $[\mu, \mathbf{g} \mid \mathbf{y}]$ is a multivariate Gaussian distribution. In the following, we derive the variance-covariance matrix for this distribution. The conditional density is

$$\begin{aligned} p(\mu, \mathbf{g}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{g}, \mu)p(\mathbf{g}) \\ &\propto \exp(-\tfrac{1}{2}(\mathbf{y} - \mathbf{1}_n\mu - \mathbf{Zg})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}_n\mu - \mathbf{Zg}) \\ &\qquad\qquad -\tfrac{1}{2}\mathbf{g}'\mathbf{g}/\sigma_g^2) \\ &\propto \exp\left(-\tfrac{1}{2}[\mu - \hat{\mu} \quad \mathbf{g}' - \hat{\mathbf{g}}']\mathbf{Q}\begin{bmatrix} \mu - \hat{\mu} \\ g - \hat{g} \end{bmatrix}\right), \end{aligned}$$

where

$$\mathbf{Q} = [\mathrm{Var}(\mu, \mathbf{g}|\mathbf{y})]^{-1} = \begin{bmatrix} r^{-1} & \bar{\mathbf{z}}_r' \\ \bar{\mathbf{z}}_r & \mathbf{C}_g \end{bmatrix}$$

With $r = 1/(\mathbf{1}_n'\mathbf{R}^{-1}\mathbf{1}_n)$, $\bar{\mathbf{z}}_r = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{1}_n$ and $\mathbf{C}_g = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{I}_m\sigma_g^{-2}$. The matrix $\mathbf{Q}$ is the coefficient matrix in the mixed model equations and the inverse of this matrix is

$$\mathrm{Var}(\mu, \mathbf{g}|\mathbf{y}) = \begin{bmatrix} r + r^2\bar{\mathbf{z}}_r'\mathbf{C}^g\bar{\mathbf{z}}_r & -r\bar{\mathbf{z}}_r'\mathbf{C}^g \\ \\ -r\mathbf{C}^g\bar{\mathbf{z}}_r & \mathbf{C}^g \end{bmatrix},$$

where $\mathbf{C}^g = (\mathbf{C}_g - r\bar{\mathbf{z}}_r\bar{\mathbf{z}}_r')^{-1}$. Note that the submatrix $\mathbf{C}^g = \mathrm{Var}(\mathbf{g} \mid \mathbf{y})$ is independent of allele coding, because, as shown in the main text, $p(\mathbf{g} \mid \mathbf{y})$ does not depend on allele coding.

Variance-covariance matrix for the complete breeding value is

$$\begin{aligned} &\mathrm{Var}(\mathbf{a}_d|\mathbf{y}) \\ &= \begin{bmatrix} \mathbf{1}_n & \mathbf{Z} \end{bmatrix} \begin{bmatrix} r + r^2\bar{\mathbf{z}}_r'\mathbf{C}^g\bar{\mathbf{z}}_r & -r\bar{\mathbf{z}}_r'\mathbf{C}^g \\ -r\mathbf{C}^g\bar{\mathbf{z}}_r & \mathbf{C}^g \end{bmatrix} \begin{bmatrix} \mathbf{1}_n' \\ \mathbf{Z}' \end{bmatrix} \\ &= r\mathbf{1}_n\mathbf{1}_n' + r^2\mathbf{1}_n\bar{\mathbf{z}}_r'\mathbf{C}^g\bar{\mathbf{z}}_r\mathbf{1}_n' \\ &\quad -r\mathbf{ZC}^g\bar{\mathbf{z}}_r\mathbf{1}_n' - r\mathbf{1}_n\bar{\mathbf{z}}'\mathbf{C}^g\mathbf{Z}' + \mathbf{ZC}^g\mathbf{Z}' \\ &= r\mathbf{1}_n\mathbf{1}_n' \\ &\quad +(\mathbf{I}_n - r\mathbf{1}_n\mathbf{1}_n'\mathbf{R}^{-1})\,\mathbf{ZC}^g\mathbf{Z}'(\mathbf{I}_n - r\mathbf{R}^{-1}\mathbf{1}_n\mathbf{1}_n'). \end{aligned}$$

Appendix C

The results in [8] state that the convergence rate of a Gibbs sampler is equal to the largest modulus eigenvalue of a certain matrix $\mathbf{B}$ where the eigenvalues can be complex numbers. As mentioned in [9] this convergence rate is also a a measure of correlation between successive McMC samples, i.e., mixing of the algorithm. The closer the convergence rate is to zero the less correlated are the successive samples.

The $\mathbf{B}$ matrix is constructed as follows. Let $\mathbf{Q}$ be the inverse of the variance-covariance matrix of the target multivariate normal distribution, in our case the coefficient matrix in the mixed model equations. Assume that a Gibbs sampling scheme is used where the variables are grouped into s blocks and $\mathbf{Q}$ is split accordingly into s blocks. First, define

$$\mathbf{A} = \mathbf{I} - \mathrm{diag}(\mathbf{Q}_{11}^{-1}, \ldots, \mathbf{Q}_{ss}^{-1})\mathbf{Q}.$$

Let $\mathbf{L}$ be the block lower triangular matrix with blocks in the lower diagonal being those of $\mathbf{A}$, and let $\mathbf{U} = \mathbf{A} - \mathbf{L}$. Thus, $\mathbf{U}$ is a strictly upper triangle matrix with zeros in the diagonal. Then the matrix of interest is

$$\mathbf{B} = (\mathbf{I} - \mathbf{L})^{-1}\mathbf{U}.$$

We consider the genomic marker model (1) where $\mathbf{g}$ is Gaussian $\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}_m\sigma_g^2)$ and $\mathbf{e}$ is Gaussian $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_n\sigma_e^2)$. The conditional distribution $[\mu, \mathbf{g} \mid \mathbf{y}]$ is a multivariate normal distribution with some mean vector and a variance-covariance matrix with inverse

$$\mathbf{Q} = [\mathrm{Var}(\mu, \mathbf{g}|\mathbf{y})]^{-1} = \begin{bmatrix} n/\sigma_e^2 & \mathbf{1}_n'\mathbf{Z}/\sigma_e^2 \\ \mathbf{Z}'\mathbf{1}_n/\sigma_e^2 & \mathbf{C}_g \end{bmatrix},$$

where $\mathbf{C}_g = \mathbf{Z}'\mathbf{Z}/\sigma_e^2 + \mathbf{I}_m/\sigma_g^2$; see Appendix B. We consider two McMC updating schemes. The first scheme iterates two blocks: μ and $\mathbf{g}$. The second scheme updates successively all parameters: μ, g_1, ..., g_m. For the first McMC updating scheme we have

$$\mathbf{A} = \mathbf{I} - \begin{bmatrix} (n/\sigma_e^2)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_g^{-1} \end{bmatrix} \mathbf{Q} = \begin{bmatrix} 0 & -\bar{\mathbf{z}}' \\ -\mathbf{C}_g^{-1}(n\bar{\mathbf{z}}/\sigma_e^2) & \mathbf{0} \end{bmatrix},$$

where $\bar{\mathbf{z}} = \mathbf{Z}'\mathbf{1}_n/n$ is a $m \times 1$ vector. Hence,

$$\mathbf{B} = \left(\mathbf{I} - \begin{bmatrix} 0 & \mathbf{0} \\ -\mathrm{C}_g^{-1}(n\bar{\mathbf{z}}/\sigma_e^2) & \mathbf{0} \end{bmatrix}\right)^{-1} \begin{bmatrix} 0 & -\bar{\mathbf{z}}' \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0} \\ -\mathrm{C}_g^{-1}\bar{\mathbf{z}}n/\sigma_e^2 & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} 0 & -\bar{\mathbf{z}}' \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} 0 & -\bar{\mathbf{z}}' \\ \mathbf{0} & \mathrm{C}_g^{-1}\bar{\mathbf{z}}\bar{\mathbf{z}}'n/\sigma_e^2 \end{bmatrix}.$$

The convergence rate is

$$\rho_1 = \rho_{lv}(\mathbf{B}) = \rho_{lv}(\mathbf{C}_g^{-1}\bar{\mathbf{z}}\bar{\mathbf{z}}'n/\sigma_e^2) = \frac{n}{\sigma_e^2}\,\bar{\mathbf{z}}'\mathbf{C}_g^{-1}\bar{\mathbf{z}},$$

where $\rho_{lv}(\mathbf{B})$ of a matrix $\mathbf{B}$ is a notation for the maximum modulus eigenvalue of $\mathbf{B}$. The final equality follows from a general property for a square matrix form **Cvv'** where **v** is a vector, saying that it only has one eigenvalue different from zero which is equal to **v'Cv**.

For the second McMC update scheme,

$$\mathbf{A} = \mathbf{I} - \begin{bmatrix} (n/\sigma_e^2)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{-1} \end{bmatrix} \mathbf{Q} = \mathbf{I} - \begin{bmatrix} 1 & \bar{\mathbf{z}}' \\ \mathbf{D}^{-1}(n\bar{\mathbf{z}}/\sigma_e^2) & \mathbf{D}^{-1}\mathbf{C}_g \end{bmatrix} = \begin{bmatrix} 0 & -\bar{\mathbf{z}}' \\ -\mathbf{D}^{-1}(n\bar{\mathbf{z}}/\sigma_e^2) & \mathbf{I}_m - \mathbf{D}^{-1}\mathbf{C}_g \end{bmatrix},$$

where **D** is the diagonal of $\mathbf{C}_g$. Hence,

$$\mathbf{B} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{D}^{-1}\bar{\mathbf{z}}n/\sigma_e^2 & \mathbf{L}_g \end{bmatrix}^{-1} \begin{bmatrix} 0 & -\bar{\mathbf{z}}' \\ \mathbf{0} & -\mathbf{U}_g \end{bmatrix},$$

where $\mathbf{U}_g$ is an upper triangular matrix containing the upper triangle of $\mathbf{D}^{-1}\ \mathbf{C}_g$ and $\mathbf{L}_g$ is a matrix containing the diagonal and lower triangle of $\mathbf{D}^{-1}\ \mathbf{C}_g$. Therefore,

$$\mathbf{B} = \begin{bmatrix} 1 & \mathbf{0} \\ -\mathbf{L}_g^{-1}\mathrm{D}^{-1}\bar{\mathbf{z}}n/\sigma_e^2 & \mathbf{L}_g^{-1} \end{bmatrix} \begin{bmatrix} 0 & -\bar{\mathbf{z}}' \\ \mathbf{0} & -\mathbf{U}_g \end{bmatrix} = \begin{bmatrix} 0 & \bar{\mathbf{z}}' \\ \mathbf{0} & \mathbf{L}_g^{-1}(\mathbf{D}^{-1}\bar{\mathbf{z}}\bar{\mathbf{z}}'n/\sigma_e^2 - \mathbf{U}_g) \end{bmatrix}.$$

The convergence rate is

$$\rho_2 = \rho_{lv}(\mathbf{B}) = \rho_{lv}(\mathbf{L}_g^{-1}(\mathbf{D}^{-1}\bar{\mathbf{z}}\bar{\mathbf{z}}'n/\sigma_e^2 - \mathbf{U}_g)).$$

Acknowledgements
Two anonymous reviewers are thanked for their corrections and valuable comments on the first version of the paper. IS acknowledges support from Finnish FABA funded from the TEKES (the Finnish Funding Agency for Technology and Innovation) project "Genomic information in genetic evaluations and breeding programs". OFC acknowledges support from grant 3405-10-0137 funded under the GUDP program by the Danish Ministry of Food, Agriculture and Fisheries, the Milk Levy Fund, Viking Genetics and Nordic Cattle Genetic Evaluation.

Author details
[1]Biotechnology and Food Research, MTT Agrifood Research Finland, FI-31600 Jokioinen, Finland. [2]Aarhus University, Faculty of Agricultural Sciences, Dept of Genetics and Biotechnology, Blichers Allé 20, P.O. BOX 50, DK-8830, Tjele, Denmark.

Authors' contributions
IS wrote the first drafts of the manuscript and OFC helped to revise and finalize it. IS and OFC derived the formulae together. IS did all the data analysis except the REML computations which were done by OFC.

Competing interests
The authors declare that they have no competing interests.

Received: 14 February 2011 Accepted: 26 June 2011
Published: 26 June 2011

References

1. Meuwissen THE, Hayes BJ, Goddard ME: Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 2001, **157**:1819-1829.
2. Goddard M: Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 2009, **136**:245-257.
3. Strandén I, Garrick DJ: Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J Dairy Sci* 2009, **92**:2971-2975.
4. VanRaden PM: Efficient methods to compute genomic predictions. *J Dairy Sci* 2008, **91**:4414-4423.
5. Henderson CR: *Applications of linear models in animal breeding* Guelph, Ontario, Canada: University of Guelph; 1984.
6. Henderson CR: Best linear unbiased estimation and prediction under a selection model. *Biometrics* 1975, **31**:423-447.
7. Mrode RA: *Linear models for the prediction of animal breeding values* Wallingford, UK: CABI Publishing; 2005.
8. Roberts GO, Sahu SK: Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J Roy Statist Soc Ser B* 1997, **59**:291-317.
9. Papaspiliopoulos O, Roberts GO, Sköld M: A general framework for the parametrization of hierarchical models. *Statist Sci* 2007, **22**:59-73.
10. Barrett R, Berry M, Chan TF, Demmel J, Donato J, Dongarra J, Eijkhout V, Pozo R, Romine C, Van der Vorst H: *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods.* 2 edition. Philadelphia, PA: SIAM; 1994.

11. Crooks L, Sahana G, De Koning DJ, Lund MS, Carlborg Ö: **Comparison of analyses of the QTLMAS XII common dataset. II: genome-wide association and fine mapping.** *BMC proceedings, BioMed Central Ltd* 2009, **3**:S2.
12. Geyer CJ: **Practical Markov chain Monte Carlo.** *Statist Sci* 1992, **7**:473-483.
13. Madsen P, Jensen J: *A users guide to DMU, version 6, Release 5.0* Aarhus University; 2011.
14. Hayes BJ, Visscher PM, Goddard ME: **Increased accuracy of artificial selection by using the realized relationship matrix.** *Genet Res* 2009, **91**:47-60.
15. Powell JE, Visscher PM, Goddard ME: **Reconciling the analysis of IBD and IBS in complex trait studies.** *Nat Rev Genet* 2010, **11**:800-805.
16. Legarra A, Aguilar I, Misztal I: **A relationship matrix including full pedigree and genomic information.** *J Dairy Sci* 2009, **92**:4656-4663.
17. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ: **Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluations of Holstein final score.** *J Dairy Sci* 2010, **93**:743-752.
18. Christensen O, Lund M: **Genomic prediction when some animals are not genotyped.** *Genet Sel Evol* 2010, **42**:2.

doi:10.1186/1297-9686-43-25
Cite this article as: Strandén and Christensen: **Allele coding in genomic evaluation.** *Genetics Selection Evolution* 2011 **43**:25.