Theor Appl Genet (1996) 93: 1090–1097

R. Pong-Wong · J. A. Woolliams

Estimating major gene effects with partial information using Gibbs sampling

Received: 3 April 1996 / Accepted: 12 April 1996

Abstract A method for estimating major gene effects using Gibbs sampling to infer genotype of individuals with unknown values, was compared with a standard mixed-model analysis. The purpose of this study was to evaluate the effect of including information of individuals with unknown genotypes on the estimates and their error variances (Ve) of the single-gene effects. When genotypes were known for all the individuals, results using the Gibbs method (GS) were similar to those obtained with the mixed model (MM). In the absence of selection, when information from individuals with unknown genotypes was included, GS yielded unbiased estimates of the major gene effects while reducing the Ve associated with them. This reduction in Ve depended on the gene frequency and mode of action of the major locus. For the additive effect, the reduction in Ve ranged from 29 to 69% of the total reduction which would have been obtained if all individuals had had a known genotype. Similarly the reduction in Ve found for the dominance effect ranged from 12 to 58%. Estimates using GS generally had small detectable biases when the polygenic heritability used in the analysis was inflated or estimated simultaneously. However, the benefit of using information from individuals with unknown genotypes was still maintained when comparing the mean square error of the estimates using either GS or MM when genotypes are only known for a subset of the population. When the population has been under selection, the use of Gibbs sampling to incorporate information of individuals without genotypes reduced substantially the bias and mean square error found for MM analysis on partial data. Nevertheless, there was some bias detected using Gibbs sampling. The gene frequency of the major gene in the base population was also well estimated despite its change over generations due to selection.

Key words Estimation · Gibbs-sampling · Major-gene · Selection · Dominance

Communicated by E. J. Eisen

R. Pong-Wong (✉) · J. A. Woolliams
Roslin Institute (Edinburgh), Roslin, Midlothian, EH25 9PS, UK

Introduction

Although quantitative traits are often considered to be influenced mainly by a large number of genes each having a small effect, single-genes with large effect affecting these traits have also been found. Examples of these are the κ-casein locus influencing milk protein content in dairy cattle (Bovenhuis 1992), the Booroola gene affecting reproduction in sheep (Piper and Bindon 1982), and the halothane locus which affects meat quality in pigs (Jensen and Barton-Gade 1985). Knowledge of the genotypes at these loci can be used to increase the accuracy of estimated breeding values of candidates for selection, thereby increasing the short-term genetic progress. However, it is important to establish reliable estimates of the single-gene effects or else genetic progress may be lost (Sales and Hill 1976).

When genotypes of individuals are known, the effect of the single locus upon a trait can be estimated without bias using standard mixed-model (MM) techniques (Kennedy et al. 1992). However, for reasons of practicality and economy, it is likely that most individuals, especially ancestors, will have an unknown genotype for the locus in question. Since mixed-model analysis requires a knowledge of the individuals' genotype, phenotypic information of individuals with unknown genotypes must be excluded from the analysis, thereby decreasing the accuracy of the estimates and introducing bias if the population has undergone selection.

Several techniques of estimating single-gene effects using information from animals with unknown genotypes have been reported (Hoeschele 1988; Hofer and Kennedy 1993; Kinghorn et al. 1993). They have been applied in segregation analyses where exact likelihood techniques cannot be used due to large and complex pedigrees, perhaps involving loops. These techniques use approximations to the likelihood in order to avoid the difficulty of computing all possible incidence matrices. Hofer and Kennedy (1993) have shown that using the approximations leads to bias and, therefore, alternative approaches avoiding them may prove superior.

Guo and Thompson (1992) showed that Gibbs sampling could be used to infer genotypes of individuals with unknown values. For a joint distribution, this method allows the estimation of the parameters for the marginal densities through sequentially sampling each variable from its conditional distribution, given the other variables (Casella and George 1992). The genotype of each animal can then be sampled conditional upon the genotypes of the other animals and, when a large number of samples are accumulated, their distribution will be proportional to the true probability distribution for the genotypes. Although computer intensive, this approach replaces difficult calculations with a series of random samples, allowing calculation of the genotype probability with great accuracy in large and complex pedigrees. Janss et al. (1995) extended this technique to the calculation of other hyperparameters, obtaining estimates of both the single-gene and the polygenic effects. This method has been used to detect major genes in a pig crossbred population (Janss et al. 1994).

The objectives of the present study were to evaluate, using simulations, the benefit of using a Gibbs sampling approach when genotypes of a given locus are known only on a subset of the population, and to extend findings previously reported by Pong-Wong and Woolliams (1994). The estimate and its error variance were compared with a standard mixed-model analysis carried out only with information from individuals with known genotypes. It examines some characteristics of Gibbs sampling when applied to populations under selection. The effect of gene frequency, the mode of action of the single-gene, errors in assumed polygenic parameters, and simultaneous estimation of the polygenic heritability were studied.

Methods

Model

A quantitative trait in a population was considered to be controlled by a polygenic effect together with a single locus with two alleles, a and A. The single-gene was assumed to have an additive effect (α) defined as half the difference between homozygotes [α=(AA – aa)/2] and a dominance effect (δ) defined as the deviation of the heterozygote from the average value of both homozygotes [δ=Aa–(AA+aa)/2]. In the unselected base population the favourable allele A had a frequency P, and the genotype frequencies were assumed to be in Hardy-Weinberg equilibrium. Polygenic and environmental variances were also assumed to be 50 units^2 each (i.e. h^2=0.5). In the genetic models considered α was either 0 or 10 units, while δ was 0, 10 or –10 units and P took values of 0.5 or 0.15.

Two population structures were simulated. The first was composed of 50 sires and 500 dams, randomly selected and mated hierarchically with one offspring per dam (ten per sire). Each animal had one phenotypic observation and the genotype of the single-gene was assumed to be known only for sires and offspring (i.e. 550 individuals with known genotypes, 500 with unknown).

The second population structure included two rounds of selection. From an unrelated base population of 1000 males and 1000 females, 50 sires and 500 dams were phenotypically selected to produce the next generation. Each female had four full-sib offspring (two males and two females) from which the next generation of parents was selected with the same criterion, to produce another generation. A total of 6000 individuals (2000/generation) was generated. All individuals had one phenotypic observation, but only 600 (10%) have a known genotype: all sires (100) and one individual per full-sib family (500) in the last generation.

Major-gene effect estimation

Methods used to estimate single-gene effects are the same as described by Kennedy et al. (1992) for the mixed-model approach (MM) and by Janss et al. (1995) for the Gibbs sampling scheme (GS). A full explanation of the methods has been reported previously by them.

Mixed Model

The analysis using MM was done using Henderson's mixed model equations as suggested by Kennedy et al. (1992). The analysis was carried out with the BLUP option of a DFREML programme (Meyer 1989) assuming a known polygenic heritability [$h^2=\sigma_a^2/(\sigma_a^2+\sigma_e^2)$ where σ_a^2 excludes the variance due to the major gene]. Only observations from animals with known genotypes were used, but all the available pedigree information was included to account for the covariance between observations. The genotypes of the animals was included in the model as a fixed-effect classification and the parameters α and δ were later calculated from the genotype estimates. In some cases polygenic genetic variance was also estimated (Meyer 1989) instead of assuming a known heritability.

Gibbs sampling

This analysis was done using the programme of Janss et al. (1995). The environmental variance, as well as the breeding values of all the animals and their genotypes for the locus in question were estimated together with the effect and frequency of the favourable allele. The samples accumulated (which are referred to as 'realisations' in this study) were later used to calculate the expectation and error variance (Ve) of the estimates.

In order to decrease computation within each replicate, only two realisations per replicate (i.e. the replicate of the data set for a given set of parameters) were used in the GS analysis, but Ves of the estimates were corrected for the sampling error associated with small samples. In principle when convergence to equilibrium distribution has been achieved, the chain of realisations of α forms a random sample from a distribution with expectation α^* and variance var(α^*), representing the estimate of α conditional on the data and its Ve about the true value α. If the number of realisations accumulated (n) is large, their average will be α^* and their variance will be Ve of α^* about α. Over all possible data sets the expected Ve (variance within replicate) would be equal to the variance of all replicates (variance between replicates). However, when few realisations are used, their expectation (α^{**}) is an estimate of α^*, but with a sampling-error variance which will be equal to var(α^*)/n. Therefore, about the true value, var(α^{**}) will be equal to var(α^*) + var(α^*)/n. For the case in which two independent realisations are used, var(α^{**}) about α will then be 3/2 var(α^*).

In order to test such an assumption, a preliminary study was done using GS analysis with either 2 or 500 realisations per replicate (data set). When n=500, realisations were taken at intervals of 20 samples between two consecutive realisations, with the first one obtained after 120 samples away from the arbitrary starting point (total length of the chain=10 100 samples). Realisations #100 and #500 (samples 2100 and 10 100 of the chain) were used for the analysis when n=2. Using an analysis of variance the Ves estimated within and between replicates were compared. Over 1000 replicates, it was found that the number of realisations used made no significant difference to the magnitude of these variances and, as expected, the variances components within and between replicates were of similar magnitude.

Because a small number of realisations per replicate were to be taken, several analyses were done to ensure that they were random

and independent samples. For the unselected population structure, it was found that sampling tended to converge to the true distribution and was independent of the initial point after approximately 100 samples from the starting point (Fig 1a). Two further tests were used to check the independence of the two realisations. Analysis of autocorrelations showed that correlation between samples was close to zero when the lag between them was around 50 samples (Fig 1b). The other test done was a cusum analysis. The cusum value at time t is the sum of deviations of each value from the overall mean cumulated until time t [i.e. $cusum(t, x)=\sum_{i=1}^{t}(x_i-\mu)$]. A cusum plot over time amplifies the trend within a given interval, allowing the detection of cyclicity in the chain. A change of trend in the chain would result in a change in the direction of the cusum curve and, therefore, the length of a cycle would be the lag between consecutive changes of direction of the cusum graph. In the case of the unselected population, the results suggested further long-term trends (to those observed with the autocorrelation study) in the realisations, with irregular cycles of the order of 100 samples (Fig 1c). Given these results, the two selected realisations were taken at samples 300 and 500 after the arbitrary starting point to ensure independence of the samples between themselves and between the starting point. A similar analysis was carried out for the population undergoing selection. In this case, the two realisations were taken at samples 1500 and 3000 after the arbitrary starting point, because of the more complex pedigree structure. However, it is important to point out that the protocol for obtaining realisations used here is specific to the situation of this study. The relatively simple pedigree structures considered made it unnecessary to use techniques such as "simulated tampering" (Geyer and Thompson 1995) and others (e.g. Lin et al. 1994) which have been found to be important for speeding up the mixing of chains when complex pedigree structures are involved.

The starting point for all cases assumed all polygenic breeding values and gene effects to be zero. Initial gene frequency was the gene frequency observed on those animals having known genotypes. All individuals with unknown genotype were first assigned with a heterozygous genotype.

Comparison between methods

The expectations of the gene effects and Ve obtained using both MM and GS methods on the same data were compared. A total of 1000 replicates per set of parameters was simulated. The same parameters were estimated with both methods. When unknown genotypes were present, GS analysis used the additional phenotypic information from such individuals. Values for Ve reported for GS are the mean of the variance components between and within replicates, estimated from an analysis of variance of the realisations.

All genotypes known

The purpose was to validate the equivalence between both methods in the cases considered here. The heritability of the polygenic effect was assumed to be known without error. Parameters used in this study assumed that the single locus had a totally additive effect (α=10; δ=0) and P=0.15.

Genotypes partially known

The effect of gene frequency and the mode of action of the single-gene were evaluated. Three variations were studied in this case: (1) the polygenic heritability was assumed to be known without error, (2) the heritability was assumed to be known but was biased upwards and (3) the heritability was unknown and was calculated from the data. The biased polygenic heritability was chosen to be that derived when the genetic variation associated with the single locus is included with the polygenic variance. Data sets were generated for different gene frequencies (P=0.15; 0.50) and effects of the locus on the trait (neutral: $\alpha=\delta=0$; additive: α=10, δ=0; dominant: α=10, δ=10; and recessive: α=10, δ=−10 when P=0.15). The prior distribution for

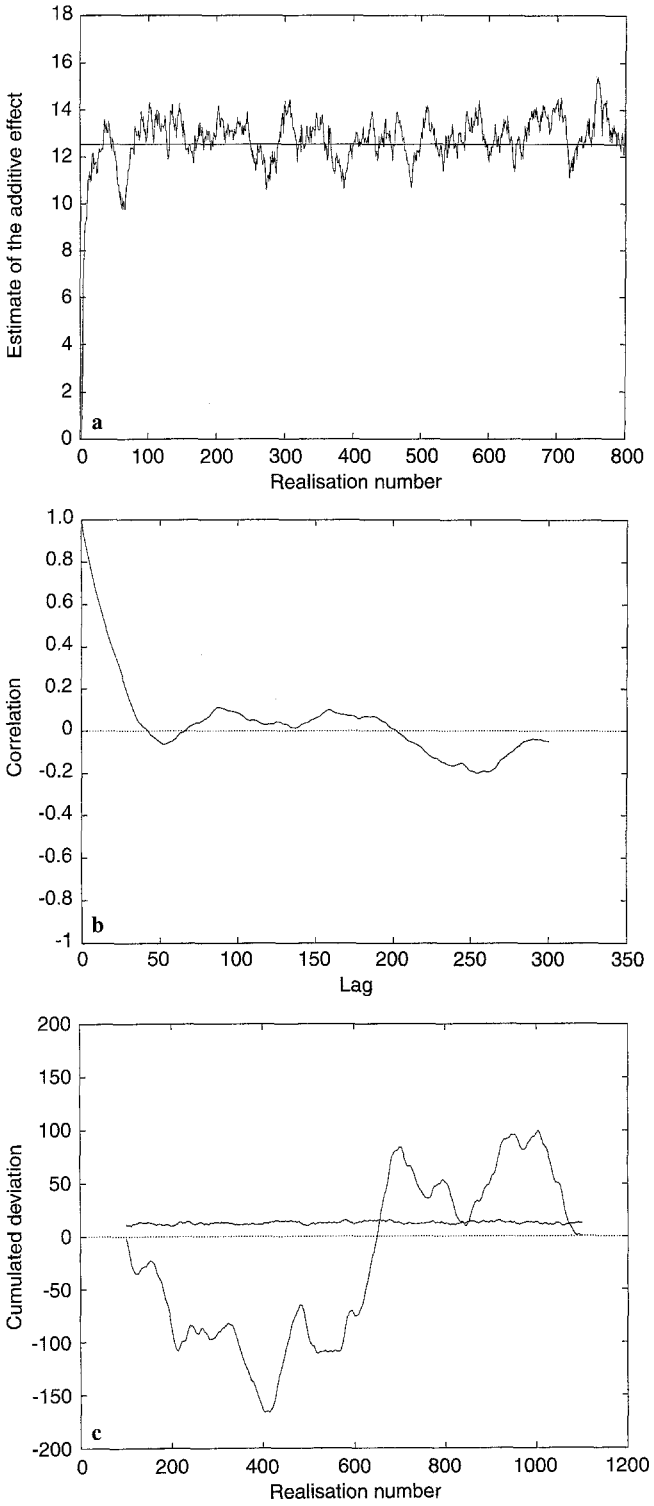

Fig. 1a–c Sampling properties of the additive effect (α) when using Gibbs sampling for the population structure without selection (α=10; δ=0; P=0.15). **a** Convergence to the true distribution over time from an arbitrary starting point. **b** Correlogram for 5000 realisations after removing a "burn-in" period of 100 realisations. **c** Cusum analysis for 1000 realisations after a "burn-in" period of 100 realisations (the corresponding values of the realisations are also shown for comparison)

gene frequency was uniform in the interval [0:1]; for gene effects a flat prior was used; and for the variance components a flat prior uniform in $\sigma^2>0$ was used.

Effect of selection

For this case only two sets of parameters were considered: when the single-gene was totally neutral (α=0, δ=0) and when it was recessive (α=10, δ=–10). In both cases it was assumed that P=0.15 and that the heritability was known without error.

Results

All genotypes known

Table 1 summarises the results obtained from both methods for 1000 replicates when all the genotypes were known. The expectation of α and δ and their respective Ves were similar for both MM and GS and were not significantly different from the true values simulated. Since GS does not infer genotypes in this case, estimation of the single-gene effects is reduced to the calculation of extra fixed effects. Small differences in results between both methods are due to sampling errors of GS, but if the number of realisations per replicate were to tend to infinity, GS results would converge to those obtained with MM.

Table 1 The comparison of the estimates of major gene effects and error variances (Ve) for mixed-model (MM) and Gibbs sampling (GS) approaches when genotypes for all animals are known and the major gene has an additive effect (α=10; δ=0). Ve for GS is the average of the components of variance between and within replicates (see Methods). Standard errors are given in parentheses

Effect	Estimate		Ve		
	MM	GS	MM	GS	Ratio
α	9.969 (0.033)	9.948 (0.041)	1.105	1.125	1.018
δ	0.036 (0.036)	0.067 (0.044)	1.339	1.349	1.007

Genotypes partially known

Using the true polygenic heritability

Results obtained from both methods when the polygenic heritability was known are shown in Table 2 for α and Table 3 for δ. The results of mixed models assuming all individuals had known genotypes (MM*) are also included in the tables as a comparison, since they are unbiased estimates of the gene effects and represent a lower bound to the error variance.

When only partial information existed, MM analysis yielded unbiased estimates of α and δ. Since the population was subjected only to random selection, no linkage disequilibrium accumulated between the polygenic effect and the genotypes of the single-gene and, therefore, exclusion of some information did not introduce bias in the estimation. Some estimates of δ using MM* were statistically different from the true value (Table 3). These are likely to be due to sampling error during the generation of data. Estimates using GS were not significantly different from the true value or from the MM* estimates. Inference of the dams' genotypes, and the use of the performance information on them, did not bias the estimates of the gene effects.

On the other hand, the use of extra information from animals with unknown genotypes (which can not be included in a true MM analysis) decreased the Ve of the estimates. This reduction varied with the true parameters of the single locus (α and δ) and its gene frequency. The smallest gain in accuracy (reduction of Ve) was when the gene was completely neutral for the trait (α=0, δ=0) and the greatest was for the case when the favourable allele was at a low frequency (P=0.15) and had a recessive effect (α=10, δ=–10).

Differences in the gain in accuracy obtained for the different set of parameters depended on how well the unknown genotypes were inferred (Table 4). The maximum gain would be achieved for the case when all unknown genotypes were sampled without error, yielding analogous results to MM* in Tables 2 and 3. The highest relative gain in accuracy was achieved when a rare recessive allele was segregating. Compared to the other modes of action with P=0.15, this case corresponded with a much greater con-

Table 2 The effect of the mode of action of the single gene on the estimates of its additive effect (α) and its error variance (Ve) when: (1) all individuals have known genotypes and a mixed-model is used (MM*); (2) only a subset have known genotypes and a mixed-model is used (MM); and (3) only a subset have known genotypes and Gibbs sampling is used (GS). Ve for GS is the average of the components of variance between and within replicates. Standard errors are given in parentheses

True parameters		Estimate of α			Ve		
α	δ	MM*	MM	GS	MM*	MM	GS
P = 0.50							
0	0	0.007 (0.015)	0.016 (0.021)	0.017 (0.026)	0.228	0.433	0.372
10	0	10.007 (0.015)	9.997 (0.021)	9.940 (0.022)	0.224	0.423	0.327
10	10	10.013 (0.015)	10.001 (0.020)	9.988 (0.022)	0.224	0.413	0.317
P = 0.15							
0	0	–0.057 (0.032)	–0.080 (0.047)	–0.050 (0.053)	1.058	2.242	1.864
10	0	10.007 (0.033)	10.039 (0.048)	9.924 (0.051)	1.096	2.293	1.748
10	10	9.953 (0.033)	9.970 (0.048)	9.928 (0.055)	1.110	2.303	1.959
10	–10	9.973 (0.034)	10.009 (0.048)	10.003 (0.048)	1.129	2.347	1.505

Table 3 The effect of the mode of action of the single-gene on the estimates of its dominance effect (δ) and its error variance (Ve) when: (1) all individuals have known genotypes and a mixed model is used (MM*); (2) only a subset have known genotypes and a mixed model is used (MM); and (3) only a subset have known genotypes and Gibbs sampling is used (GS). Ve for GS is the average of the components of variance between and within replicates. Standard errors are given in parentheses

True parameters		Estimate of δ					Ve		
α	δ	MM*		MM		GS	MM*	MM	GS
$P = 0.50$									
0	0	–0.036 (0.018)		–0.027	(0.025)	–0.048 (0.030)	0.333	0.634	0.596
10	0	–0.008 (0.018)		–0.031	(0.024)	–0.048 (0.026)	0.336	0.626	0.471
10	10	10.042 (0.018)		9.969	(0.025)	9.984 (0.028)	0.335	0.627	0.517
$P = 0.15$									
0	0	0.075 (0.036)		0.123	(0.052)	0.116 (0.061)	1.298	2.714	2.500
10	0	0.013 (0.036)		0.023	(0.050)	0.096 (0.056)	1.303	2.640	2.101
10	10	10.032 (0.035)		10.070	(0.052)	10.046 (0.061)	1.298	2.755	2.414
10	–10	–9.982 (0.038)		–9.971	(0.052)	–9.948 (0.052)	1.383	2.790	1.819

Table 4 The effect of the mode of action of the single-gene on the percentage of individuals with unknown genotype assigned to their correct genotype in each realisation and the subsequent gain in accuracy (for P=0.15; 'a' is the most common allele)

True parameters		Overall[a]	Within true genotype			Gain in accuracy[b]	
α	δ		aa	Aa	AA	α	δ
$P = 0.50$							
0	0	47.61	43.8	49.9	43.9	0.298	0.124
10	0	57.83	56.8	58.8	56.7	0.481	0.534
10	10	61.43	75.2	62.8	44.8	0.508	0.378
$P = 0.15$							
0	0	72.24	83.1	46.9	13.6	0.319	0.152
10	0	77.16	85.8	54.4	25.4	0.455	0.403
10	10	84.11	91.6	69.0	13.8	0.289	0.234
10	–10	74.13	83.3	50.2	51.5	0.692	0.585

[a] Weighted average over the three genotypes
[b] Gain in accuracy=$(Ve_{MM} - Ve_{GS})/(Ve_{MM} - Ve_{MM*})$ taken from Tables 2 and 3

fidence in correctly assigning individuals with the rarest genotype (see Table 4).

Using a biased heritability

When the analysis was done using a biased heritability, the reduction in Ve observed with GS was broadly comparable in magnitude with the results obtained when the true heritability had been used. Estimates for α were consistently biased downward for all the cases studied, but in all cases they were less than 2% of the true value. However, when the expected mean square errors (which include the bias) were calculated, the benefits of Gibbs sampling were only marginally reduced from the benefits realised for Ve (data not shown). The dominance effect appeared more robust to the effect of using the wrong polygenic heritability.

Simultaneously estimating variance components

When polygenic heritability was estimated simultaneously in the analysis, GS results showed bias from the true value for some cases in which the gene frequency (P) was 0.15 (Tables 5 and 6). However, the estimates in all these cases were not significantly different from the results obtained using MM* assuming the genotypes of all individuals to be known. Estimates obtained with MM* assuming all individuals with known genotypes, are considered to be the best linear unbiased estimate (BLUE), given the data. A considerable reduction in Ve was also observed for all the cases.

Effect of selection

A strong bias in the single-locus effects was observed using MM when the locus has an effect on the selected trait (Table 7). The results obtained using GS showed some small bias, but were consistent with the results using MM* (i.e. when all the individuals were assumed to have a known genotype). A small bias was also observed for GS when the locus is neutral on the selected trait. Estimates of Ve obtained with GS were half way between those obtained with MM and MM*.

The biases observed in Gibbs sampling increased the mean-square error only marginally and were considerably smaller than the mean-square errors obtained with MM.

Table 5 The effect of estimating the gene effects and the polygenic heritability simultaneously on the estimate of the additive effect (α) and its error variance (Ve) when: (1) all individuals have known genotypes using a mixed model (MM*); (2) only a subset have known genotypes using a mixed model (MM); and (3) only a subset have known genotype using Gibbs sampling (GS). Ve for GS is the average of the components of variance between and within replicates. Standard errors are given in parentheses

True parameters		Estimate of α			Ve		
α	δ	MM*	MM	GS	MM*	MM	GS
$P = 0.50$							
0	0	0.000 (0.015)	–0.007 (0.021)	–0.002 (0.023)	0.220	0.422	0.347
10	0	10.014 (0.015)	9.990 (0.020)	9.968 (0.021)	0.219	0.416	0.299
10	10	10.018 (0.015)	10.011 (0.020)	10.009 (0.021)	0.219	0.415	0.298
$P = 0.15$							
0	0	–0.063 (0.033)	–0.091 (0.047)	–0.053 (0.051)	1.092	2.235	1.784
10	0	9.957 (0.033)	9.936 (0.047)	9.876 (0.050)	1.096	2.260	1.642
10	10	9.931 (0.033)	9.955 (0.050)	9.868 (0.055)	1.130	2.445	2.033

Table 6 The effect of estimating the gene effects and the polygenic heritability simultaneously on the estimate of the dominance effect (δ) and its error variance (Ve) when: (1) all individuals have known genotypes using a mixed model (MM*); (2) only a subset have known genotypes using mixed model (MM); and (3) only a subset have known genotype using Gibbs sampling (GS). Ve for GS is the average of the components of variance between and within replicates. Standard errors are given in parentheses

True parameters		Estimate of δ			Ve		
α	δ	MM*	MM	GS	MM*	MM	GS
$P = 0.50$							
0	0	–0.011 (0.018)	–0.002 (0.025)	–0.016 (0.031)	0.332	0.649	0.620
10	0	–0.023 (0.018)	–0.010 (0.025)	–0.030 (0.027)	0.399	0.632	0.484
10	10	9.992 (0.018)	9.987 (0.025)	9.971 (0.027)	0.239	0.645	0.505
$P = 0.15$							
0	0	0.100 (0.036)	0.140 (0.052)	0.119 (0.060)	1.321	2.686	2.345
10	0	0.070 (0.036)	0.107 (0.053)	0.184 (0.057)	1.328	2.758	2.096
10	10	10.071 (0.037)	10.057 (0.054)	10.091 (0.061)	1.379	2.862	2.499

Table 7 The effect of selection on the estimates and their error variances (Ve) of major gene effects when: (1) all individuals have known genotypes using a mixed model (MM*); (2) only a subset have known genotype using a mixed model (MM); and (3) only a subset have known genotype using Gibbs sampling (GS). Ve for GS is the average of the components of variance between and within replicates. Standard errors are given in parentheses

True parameters		α			δ		
α	δ	MM*	MM	GS	MM*	MM	GS
Estimate							
0	0	0.012 (0.014)	0.019 (0.051)	–0.132 (0.042)	–0.007 (0.016)	0.019 (0.058)	–0.103 (0.042)
10	–10	9.980 (0.010)	7.738 (0.022)	9.960 (0.019)	–9.991 (0.009)	–11.886 (0.028)	–9.976 (0.015)
Ve							
0	0	0.204	2.650	1.212	0.238	3.203	1.176
10	–10	0.100	0.522	0.248	0.088	0.687	0.147

When the major-gene was recessive (α=10, δ=–10), the square roots of the mean square errors were 0.50 and 0.38 for α and δ when using GS, compared to 2.37 and 2.06 when using MM. This represent a reduction of approximately 80% using GS. For the case of the single-gene being neutral, the reduction was smaller, but still over 30%.

The gene frequency of the major gene in the base population was well estimated using GS. The accuracy of this estimation is better shown for the case when the major gene was recessive. For this case the average gene frequency observed in those animals with known genotypes was 0.46 (because of changes in the gene frequency over generations due to selection, and individuals with known genotypes were mainly from the last generation) compared with 0.149 obtained with GS (not significantly different from the simulated gene frequency in the base population, which was 0.15). When the single-gene was neutral the observed gene frequency and the estimate obtained with GS were 0.149 and 0.154 respectively.

Discussion

When genotypes were known for all individuals, analysis using MM or GS (using the priors defined here) can be considered equivalent. In both approaches, when all genotypes

are known without error the estimation of the single-locus effects is reduced to the estimation of an extra fixed effect. Wang et al. (1994) showed that, in a polygenic model with flat priors, the Gibbs sampling approach yields similar results, for both the random and fixed effects, as when solving the mixed-model equations directly.

Results obtained using GS showed a substantial improvement on the accuracy of the estimate when information on animals with an unknown genotype was available and included. This increase in the accuracy was dependent on the true parameters used for the single locus. For example, a high relative gain in accuracy was achieved when simulating a rare recessive allele. This can be explained by a better discrimination between two distinct major-locus genotype classes. The GS method employs samples from a posterior distribution which is a function of the probabilities conditional upon the current genotypes of ancestors and descendants, and these probabilities are calculated using transmission probabilities and penetrance values. The latter values are calculated for metric traits using a penetrance function which is conditional upon the data and the current values for the other parameters including the genotypic effects (Janss et al. 1995). With no effect of the locus, the posterior distribution will depend solely upon the transmission probabilities, with no influence from the penetrance function. The larger the separation of the genotypes (relative to the error variance, as in the recessive case) the more discriminatory the penetrance function becomes. The higher gain from the example with the rare recessive, compared to the rare dominant, may be ascribed to the additional benefit of more confidently identifying those individuals with the rarest genotype. Since estimation errors are $O(n^{-1})$, the relative gain from an additional genotype is greatest when n is smallest. In the dominance case the rarest genotype remains relatively poorly distinguished from the heterozygote.

Using a simple model assuming no polygenic effect, the expected proportion of individuals assigned to their true genotype was calculated analytically. For this case, the results obtained with simulation studies using GS were very similar to those obtained analytically (data not shown). The proportion assigned correctly were affected by the true parameters of the single-gene (α, δ and P) in a similar way to the studies in the presence of a polygenic effect. This change in the accuracy of sampling genotypes according to the mode of action and gene frequency was related to the expected reduction in Ve obtained when information from untyped individuals is included.

The reduction in Ve by including data from individuals with unknown genotypes can be compared to the reduction in Ve when all individuals are genotyped, the latter forming a lower bound to Ve. Without selection, in the cases studied the ratio of the reduction observed for using GS and the maximum possible reduction varied from 29% (α=0; δ=0; P=0.50) to 69% (α=10; δ=–10; P=0.15). A crude calculation, based on the rough approximation that the Ve for MM using n known genotypes is proportional to n^{-1}, shows that, using GS, ten individuals with data but unknown genotypes may be worth between three and seven individuals with both data and genotypes.

In practice, the estimation of the effects of the major gene will normally be carried out without full knowledge of the polygenic heritability. In these circumstances two approaches may be considered: (1) the estimate is made using the heritability obtained from analyses that ignore major gene effects (as is commonly the case now), which will consequently inflate the heritability of polygenic effects; and (2) the polygenic component will be estimated simultaneously with the gene effect. The use of GS when assuming a polygenic heritability that is biased upwards leads to gene effects that were biased downwards. This results from an over-optimistic view that genetic effects might be explained by the polygenes, consequently underestimating the other fixed and environmental effects. When using GS and simultaneously estimating both the gene effect and the polygenic variance, these biases were either small or absent. In the examples where biases were observed, the estimates were not significantly different from those obtained with MM*, assuming all known genotypes, which suggests that the biases were partly due to sampling errors. Nevertheless, in both approaches (i.e. using biased polygenic heritability or estimating it simultaneously), any biases observed were small and mean-square errors were smaller than those obtained when ignoring data from untyped individuals.

When a population has been under selection, MM analysis is not appropriate if genotypes are missing for some individuals. If the single locus has an effect on the selected trait, selection pressure would create and maintain a linkage disequilibrium between the polygenic effect and the genotypes of the locus in question. Kennedy et al. (1992) showed that this would lead to bias in the estimates if not all information is included in the analysis, unless the locus has no effect. This can be observed in our study where major biases were found using MM on the subset of typed individuals only.

The results with GS in the presence of selection are, therefore, of great importance. Our results showed that these major biases were largely removed and that gene effects and gene frequencies were estimated with considerably smaller mean-square errors. However, not all the bias was removed. These results are harder to explain: in some cases the MM* estimates were also significantly different from the true value and were similar to the GS estimates, so that some of these biases may be due to sampling error. One possible explanation of the bias observed with GS could be the lack of linearity when re-sampling genotypes. Then, if such assumption is violated, it would be anticipated that the unbiased property of the estimates may no longer hold. However, despite the small bias observed with GS, there is still a substantial gain when including information on individuals with unknown genotypes: the maximum bias observed with GS was 0.13 units, equivalent to less than 2% of the polygenic standard deviation; and the reduction in the square root of the mean square errors ranged from 31 to 80% of those obtained from using MM where information of individuals with unknown genotypes is ignored. The reduction of Ve achieved in these cases still implies that inclusion of around 3–4 individuals with an

unknown genotype would represent the use of an extra individual with known genotype.

The gene frequency of the single-gene in the base population was also estimated with great accuracy. This would suggest that accurate prior knowledge of such frequency would not be too essential to obtain accurate estimates of the single-gene effect, providing pedigree and selection information is available.

Furthermore, the biases observed using GS are small compared with those found with other methods. Hofer and Kennedy (1993) used three different methods for estimating the single-gene effect when genotype information is missing. When genotypes were known for 10% of the population (all sires and half the dams), and assuming the polygenic to be heritability known, all methods showed bias in the estimate, ranging from 1 to 43%. The population structure they used was similar to the unselected population employed in the present study but with larger full-sib families.

Additional to the mode of action of the single-gene, its allele frequency, the effect of selection and the uncertainty of the polygenic variance, there are other factors which may affect the reduction in Ve for estimates obtained when information of untyped animals is included. Increasing the number of offspring with known genotypes (in this study one offspring/dam has a known genotype) will increase the accuracy of sampling genotypes and a higher reduction in Ve would be obtained. In very large full-sib families, knowledge of the genotype of one parent and all the offspring would determine the genotype of the other parent with only negligible error. However, in practice the maximum number of individuals to be genotyped is usually limited. Thus, increasing the number of individuals genotyped per family generally represents fewer families with no individual genotyped and, therefore, less untyped ancestors with information included in the analysis. Further studies are required to assess the ideal selection of individuals to be genotyped in order to maximize the gain in accuracy from including information of untyped relatives. The case studied here, with few offspring per dam, is common in cattle data.

The effective inclusion of information from individuals with known performance records but unknown genotypes is one example of the benefits of using Gibbs sampling in the analysis of field data. The results show that the technique may be of great importance in enabling breeders to combine information of individual loci with the prediction of residual polygenic breeding values. In practice most of the ancestral animals would have performance records, but it is unlikely that they would have a known genotype. Since many populations would be under selection, the inclusion of these ancestors would decrease the bias due to linkage disequilibrium between genotypes and the polygenic effects. For the dairy cattle situation, where large half sib-families are common, genotyping of a few sires would allow the inclusion of performance records of the dams, so increasing the power of the analysis. The need for additional computing power will be a small cost compared to the gain in accuracy of the estimate.

Acknowledgements The authors gratefully acknowledge financial support from The Milk Marketing Board of England and Wales, the Ministry of Agriculture, Fisheries and Food, and to Prof. R. Thompson for helpful discussions. We also thank Dr. L. Janss for providing the basic programme for the Gibbs sampling, and to the two referees who provided useful comments which led to a significant improvement to this paper.

References

Bovenhuis H (1992) The relevance of milk protein polymorphisms in dairy cattle breeding. PhD Thesis, Wageningen Agricultural University, The Netherlands

Casella G, George EI (1992) Explaining the Gibbs sampler. Am Stat 46, 167–174

Guo SW, Thompson EA (1992) A Monte Carlo method for combined segregation and linkage analysis. Am J Hum Gen 51:1111–1126

Geyer CJ, Thompson EA (1995) Annealing Markov Chain Monte Carlo with application to ancestral inference. J Amer Stat Assoc 90:909–920

Hoeschele I (1988) Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. Theor Appl Gen 76:81–92

Hofer A, Kennedy BW (1993) Genetic evaluation for a quantitative trait controlled by polygene and a major locus with genotypes not or only partly known. Genet Sel Evol 25:537–555

Janss LLG, Thompson R, Van Arendonk JAM (1995) Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. Theor Appl Gen 91:1137–1147

Janss LLG, Van Arendonk JAM, Brascamp EW (1994) Identification of a single-gene affecting intramuscular fat in Meishan crossbreds using Gibbs sampling. Proc. 5th World Cong Genet Appl Livest Prod 18:361–364

Jensen P, Barton-Gade P (1985) Performance and carcass characteristics of pigs with known genotypes for Halothane susceptibility. In: Stress susceptibility and meat quality in pigs. EAAP Publications, pp 33–80

Kennedy BW, Quinton M, Van Arendonk JAM (1992) Estimation of effects of single genes on quantitative traits. J Anim Sci 70:2000–2012

Kinghorn BP, Kennedy BW, Smith C (1993) A method of screening for genes of major effect. Genetics 134:351–360

Lin S, Thomson EA, Wijsman E (1994) An algorithm for Monte Carlo estimation of genotype probabilities on complex pedigrees. Ann Hum Genet 58:343–357

Meyer K (1989) Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. Genet Sel Evol 21:317–340

Piper LR, Bindon BM (1982) Genetic segregation for fecundity in Booroola merino sheep. Proc. 1st Cong Sheep and Beef Cattle Breed. Vol. 1, Technical, pp 395–400

Pong-Wong R, Woolliams JA (1994) Recovery of information on major gene effects using Gibbs sampling when genotypes are known for a subset of the population. Proc. 5th World Cong Genet Appl Livest Prod 21:256–259

Sales J, Hill WG (1976) Effect of sampling errors on efficiency of selection indices. 2. Use of information on associated traits for improvement of a single important trait. Anim Prod 23:1–14

Wang CS, Rutledge JJ, Gianola D (1994) Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. Genet Sel Evol 26:91–116